2025年北京市高等教育教学成果奖

支撑材料

成果名称：见微知著，追求卓越——面向本-研贯通培养的《细胞生物学》课程体系建设

成果完成人：卫涛涛　丁文军　胡俊杰　陆忠兵

成果完成单位：中国科学院大学

目　录

1. 国家级一流本科课程证书（《细胞生物学》课程）
2. 国家级一流本科课程公示截图（《细胞生物学实验》课程）
3. 北京市课程思政示范课程证书（《细胞生物学》课程）
4. 北京高校优质本科教材课件（重点项目）证书（《细胞生物学》课程）
5. 国家科学技术进步奖证书
6. 湖北省科技进步奖证书
7. 上海市科学技术奖证书
8. 江西省自然科学奖证书
9. 教育部“101”教材《细胞与分子生物学》封面
10. 科普图书《细胞拼图》封面
11. 中国科学院大学生命科学学院本科毕业生作为第一作者在Nature（2篇）、Science、Cell（4篇）上发表的代表性论文

国家级一流本科课程

证

书

★ ★ ★ ★ ★

课 程 类 别：线下一流课程

课 程 名 称：细胞生物学

课 程 负 责 人：卫涛涛

主要建设单位：中国科学院大学

教 育 部

2020 年 11 月

证 书 编 号：2020130218

附件

第三批国家级一流本科课程公示名单

四、线下一流课程（1842 门）

序号	课程名称	课程负责人	课程团队其他主要成员	主要建设单位
264	细胞生物学实验	胡俊杰	李栋、王磊、蔡华清、田烨	中国科学院大学

荣誉证书

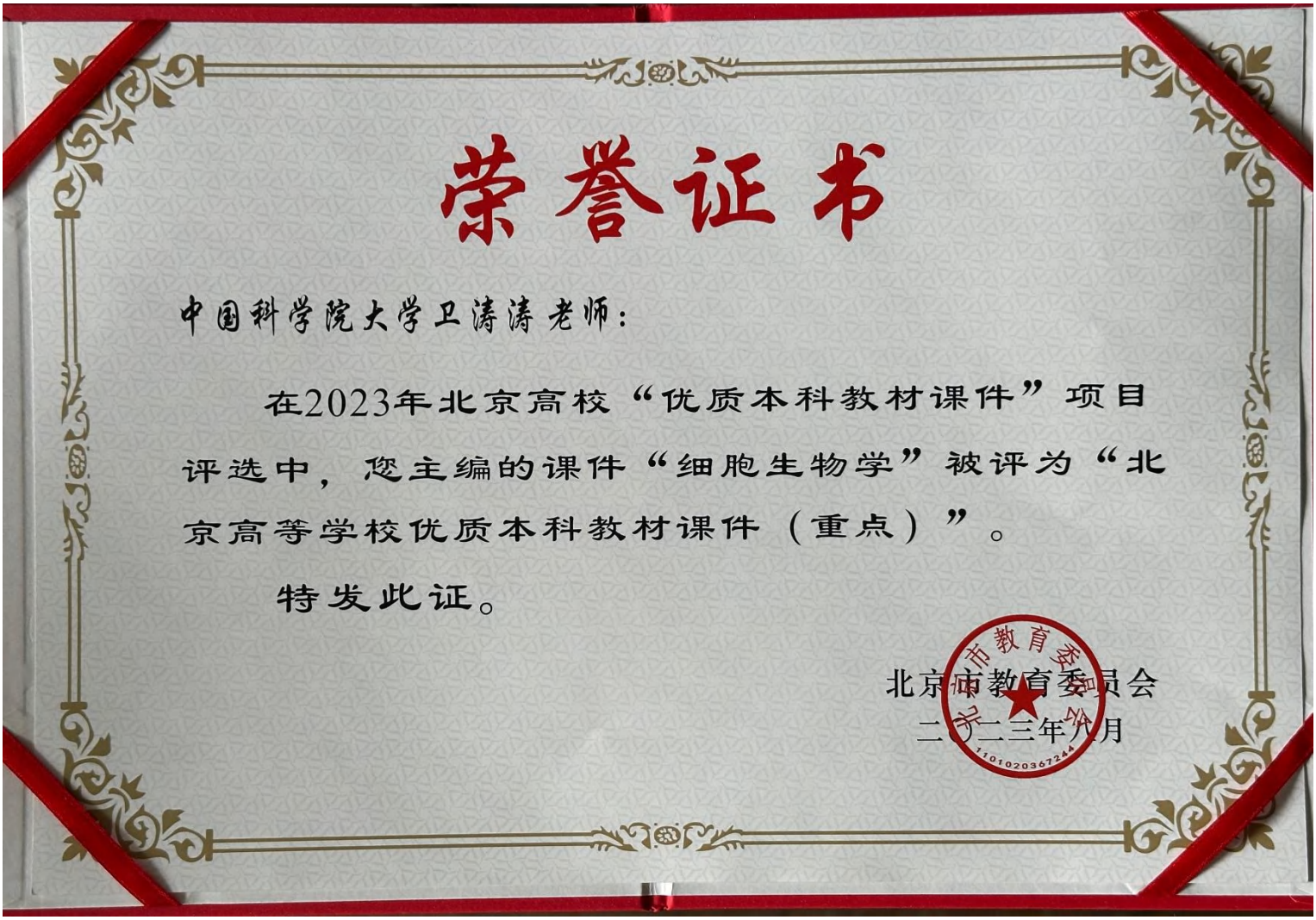
中国科学院大学：

你校卫涛涛负责的“细胞生物学”入选北京市课程思政示范课程，授课教师入选课程思政教学名师和教学团队。

特发此证！

中共北京市委教育工作委员会　　北京市教育委员会

二〇二二年一月



荣誉证书

中国科学院大学卫涛涛老师：

在2023年北京高校“优质本科教材课件”项目评选中，您主编的课件“细胞生物学”被评为“北京高等学校优质本科教材课件（重点）”。

特发此证。

北京市教育委员会

二〇二三年八月

国家科学技术进步奖

证 书

为表彰国家科学技术进步奖获得者，特颁发此证书。

项目名称：区域环境污染人群暴露风险防控技术及其应用

奖励等级：二等

获 奖 者：丁文军

中华人民共和国国务院

2018 年 12 月 12 日

证书号：2018-J-231-2-02-R07

科学技术奖励证书

项目名称：卵巢衰老评估、预警和干预的创新性研究与应用

奖励类别：科技进步奖

获奖等级：壹等奖

获 奖 人：丁文军

证书编号：2021J-190-1-035-028-R02

湖北省人民政府

2022年1月

上海市科学技术奖

证　书

为表彰上海市自然科学奖获得者，特颁发此证书。

项目名称：心力衰竭的分子细胞学机制研究

获 奖 者：陆忠兵

奖励等级：一等奖

上海市人民政府

2024年07月18日

证书号：20232015-1-R03

江西省自然科学奖

证　书

为表彰江西省自然科学奖获得者，特颁发此证书。

项目名称：贵金属、碳纳米材料微观结构及模拟生物酶分子机制研究

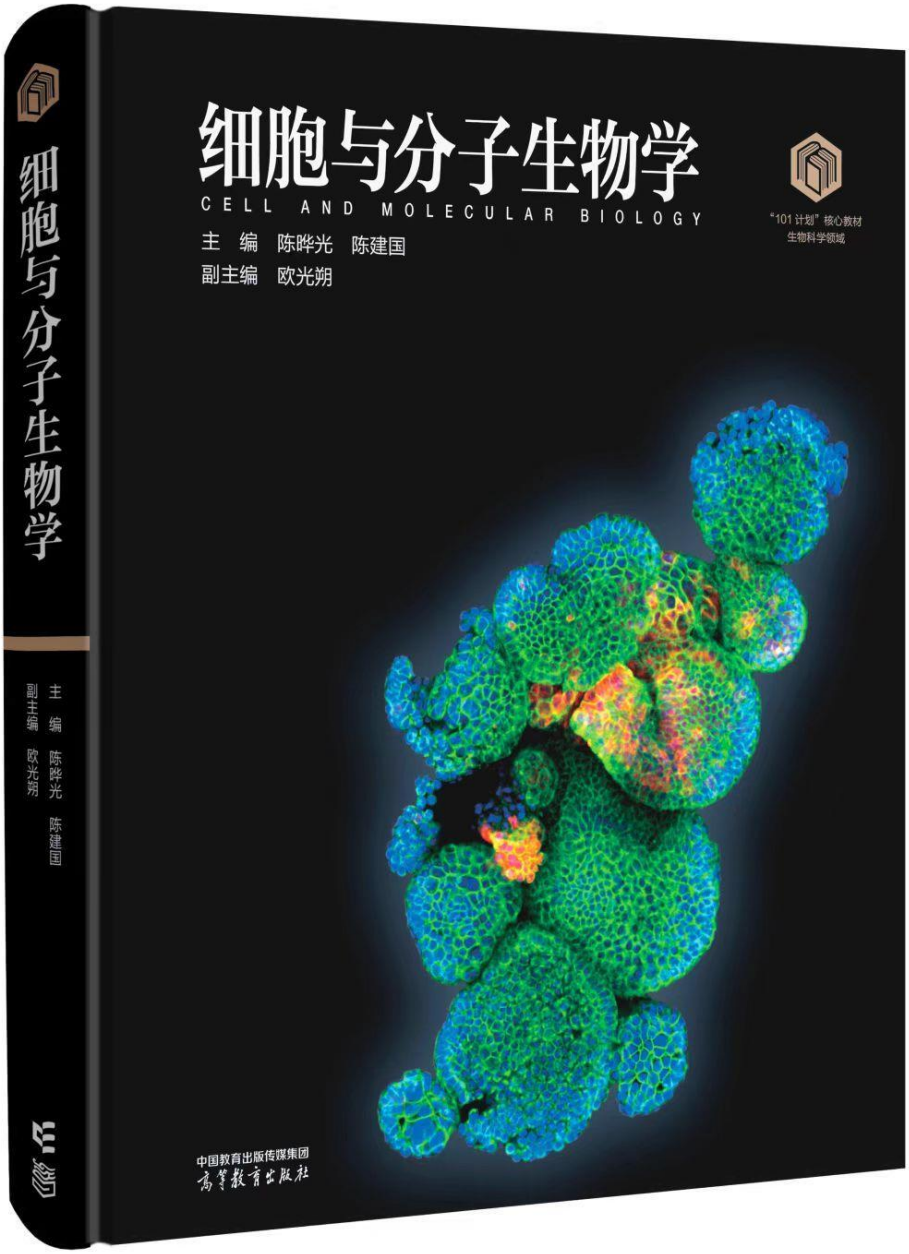
奖励等级：二等奖

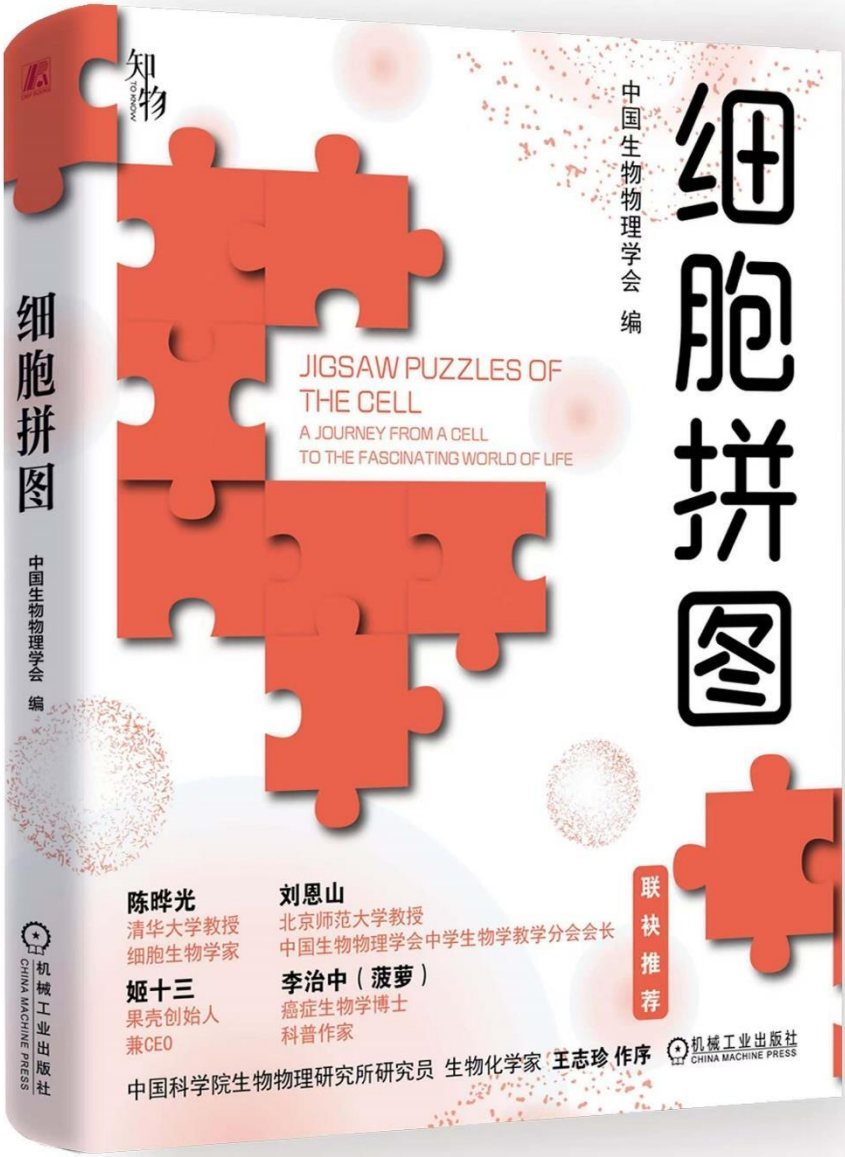
获 奖 者：卫涛涛

江西省人民政府

2021年 6月15日

证书号：Z-20-2-03-R05


细胞与分子生物学
CELL AND MOLECULAR BIOLOGY
主 编 陈晔光 陈建国
副主编 欧光朔
"101 计划"核心教材
生物科学领域
中国教育出版传媒集团
高等教育出版社
细胞与分子生物学
主 编 陈晔光 陈建国
副主编 欧光朔


知物
TO KNOW
细胞拼图
中国生物物理学会 编
JIGSAW PUZZLES OF THE CELL
A JOURNEY FROM A CELL TO THE FASCINATING WORLD OF LIFE
联袂推荐
陈晔光
清华大学教授
细胞生物学家
刘恩山
北京师范大学教授
中国生物物理学会中学生物学教学分会会长
姬十三
果壳创始人
兼CEO
李治中（菠萝）
癌症生物学博士
科普作家
中国科学院生物物理研究所研究员 生物化学家 王志珍 作序
机械工业出版社
CHINA MACHINE PRESS

Basis of the H2AK119 specificity of the Polycomb repressive deubiquitinase

https://doi.org/10.1038/s41586-023-05841-y

Received: 26 February 2022

Accepted: 14 February 2023

Published online: 29 March 2023

Check for updates

Weiran Ge[1,2,5], Cong Yu[1,2,5], Jingjing Li[1,2,5], Zhenyu Yu[1], Xiaorong Li[1], Yan Zhang[1], Chao-Pei Liu[1], Yingfeng Li[1], Changlin Tian[3], Xinzheng Zhang[1,2], Guohong Li[1,2], Bing Zhu[1,2]✉ & Rui-Ming Xu[1,2,4]✉

Repression of gene expression by protein complexes of the Polycomb group is a fundamental mechanism that governs embryonic development and cell-type specification[1–3]. The Polycomb repressive deubiquitinase (PR-DUB) complex removes the ubiquitin moiety from monoubiquitinated histone H2A K119 (H2AK119ub1) on the nucleosome[4], counteracting the ubiquitin E3 ligase activity of Polycomb repressive complex 1 (PRC1)[5] to facilitate the correct silencing of genes by Polycomb proteins and safeguard active genes from inadvertent silencing by PRC1 (refs. 6–9). The intricate biological function of PR-DUB requires accurate targeting of H2AK119ub1, but PR-DUB can deubiquitinate monoubiquitinated free histones and peptide substrates indiscriminately; the basis for its exquisite nucleosome-dependent substrate specificity therefore remains unclear. Here we report the cryo-electron microscopy structure of human PR-DUB, composed of BAP1 and ASXL1, in complex with the chromatosome. We find that ASXL1 directs the binding of the positively charged C-terminal extension of BAP1 to nucleosomal DNA and histones H3–H4 near the dyad, an addition to its role in forming the ubiquitin-binding cleft. Furthermore, a conserved loop segment of the catalytic domain of BAP1 is situated near the H2A–H2B acidic patch. This distinct nucleosome-binding mode displaces the C-terminal tail of H2A from the nucleosome surface, and endows PR-DUB with the specificity for H2AK119ub1.

Polycomb-group (PcG) proteins repress gene expression through post-translationally modifying histones and promoting the formation of facultative heterochromatin[1,2,10]. The archetypal PcG complexes PRC1 and PRC2 affect developmental programs and human diseases through H2AK119 ubiquitination and H3K27 methylation, respectively[11–13]. The distinct PR-DUB complex fulfils its PcG function through a mechanism antagonistic to that of PRC1; namely, by deubiquitination of H2AK119ub1 on the nucleosome[4,14]. This property of PR-DUB ensures the correct levels of H2AK119ub1 for compaction of the Polycomb chromatin domain, and protects active genes from silencing by PRC1 (refs. 6–9). Human PR-DUB is composed of BAP1—a ubiquitin C-terminal hydrolase (UCH)—and one of the ASXL (additional sex combs-like) proteins (ASXL1–ASXL3). Both BAP1 and the ASXL proteins are evolutionarily conserved, multidomain proteins that are frequently mutated in cancer[15–19] (Fig. 1a). Notably, some cancer-associated truncations of ASXL1–ASXL3 behave as gain-of-function mutations of PR-DUB, making BAP1 a potential therapeutic target in leukaemia[20–22].

BAP1 is inactive on its own, but binding to the deubiquitinase adaptor (DEUBAD) domain of ASXL1–ASXL3 activates BAP1's deubiquitinase activity[4,23]. The activated BAP1 complex specifically deubiquitinates H2AK119ub1 in the context of the nucleosome, despite being capable of promiscuously deubiquitinating free histones and peptide substrates[4,23] (Fig. 1b). The crystal structure of *Drosophila* PR-DUB, consisting of Calypso (Caly) and the DEUBAD domain of Asx (Asx^{DEU}), shows that Asx^{DEU} interacts with the UCH37-like (ULD) domain of Caly to form a complete ubiquitin-binding pocket[24,25]. The Caly–Asx^{DEU} structure lacks a bound substrate, although we can learn a great deal about ubiquitin binding from the highly homologous structure of the UCH37–$RPN13^{DEU}$ complex (UCH37 is also known as UCH-L5) bound to ubiquitin[26,27]. However, what determines the nucleosome-dependent H2AK119ub1 specificity remains unclear. To solve this problem, we determined the structure of PR-DUB in complex with the H2AK119ub1 chromosome, which is formed by the binding of linker histone H1 (H1.4) to the 187-bp DNA nucleosome core particle (NCP) with H2AK119ub1, termed NCP(H2AK119ub1).

Overall structure

The complexes of a human BAP1–ASXL1 heterodimer, with a catalytically inactive C91S mutant version of full-length BAP1 and a 378-residue N-terminal fragment of ASXL1, bound to NCP(H2AK119ub1) with and without linker histone H1 (H1.4), were prepared for structural determination by single-particle cryo-electron microscopy (cryo-EM) (Fig. 1a and Extended Data Fig. 1). The structure of the H1-bound complex was determined at an overall resolution of 3.0 Å, and mask-based local refinement improved the resolution of the relatively flexible region

[1]National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. [2]School of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]Division of Life Sciences and Anhui Provisional Engineering Laboratory of Peptide Drugs, University of Science and Technology of China, Hefei, China. [4]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China. [5]These authors contributed equally: Weiran Ge, Cong Yu, Jingjing Li. ✉e-mail: zhubing@ibp.ac.cn; rmxu@ibp.ac.cn

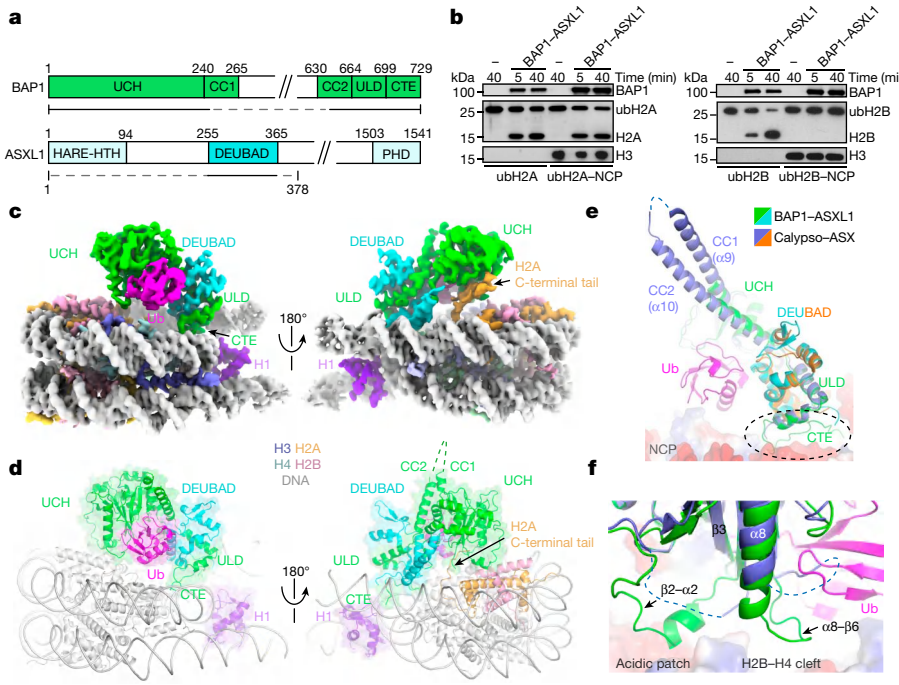


Fig. 1 | Structure of PR-DUB bound to the H2AK119ub1 chromatosome. **a**, Domain organization of human BAP1 and ASXL1. Domain names and residue numbers marking their boundaries are labelled. The underlines denote the fragments used in this study, and the dashed segments indicate disordered regions. CC, coiled-coil domain. **b**, Western blot detection of deubiquitination by the BAP1–ASXL1(1–378) complex using free histone and nucleosome substrates, monoubiquitinated at H2AK119 (left) or H2BK120 (right). Lanes with a '–' sign are mock reactions without enzyme added. The top, middle and bottom blots used anti-BAP1, anti-H2A (or anti-H2B) and anti-H3 antibodies, respectively. H3 lanes were run on a separate gel as sample-processing controls. **c**, Two views of the composed density map of the BAP1–ASXL1–H1–NCP complex. The colour code is used throughout. Ub, ubiquitin. **d**, Atomic model of the complex structure, with the cartoon diagrams of BAP1, ASXL1, H1 and ubiquitin superimposed with a transparent surface rendition, shown in the same views as above. The dashed lines denote disordered regions. **e**,**f**, Superposition with the *Drosophila* Caly–ASX^{DEU} structure (PDB_id, 6HGC) reveals structural differences in the coiled-coil domains (CC1 and CC2) and the CTE (**e**), and in the β2–α2 and α8–β6 loops (**f**). NCP is shown in an electrostatic-potential surface representation.

surrounding BAP1, ASXL1 and the ubiquitin moiety to 3.9 Å (Fig. 1c, Extended Data Fig. 2 and Extended Data Table 1). The high-quality density maps allowed reliable building of the atomic model (Extended Data Fig. 3). In the absence of H1, a relatively lower resolution (3.8 Å overall) complex is visualized (Extended Data Fig. 4). The two structures are highly similar (Extended Data Fig. 4e), and we use the higher-resolution one for most of the analysis here, unless explicitly specified otherwise.

A heterodimer of BAP1 and the DEUBAD domain of ASXL1 (ASXL1^{DEU}) is bound on one side of the nucleosome (Fig. 1d). The catalytic UCH domain, the ASXL1-binding ULD domain, an approximately 15-residue positively charged segment of BAP1's C-terminal extension (CTE) and a nearly intact ASXL1^{DEU} (amino acids (aa) 249–341) are well resolved, whereas a large non-conserved internal region (aa 245–645) and several exposed loops of BAP1, and around 250 N-terminal residues of ASXL1, are disordered (Fig. 1a). The overall structure of the BAP1–ASXL1^{DEU} heterodimer is similar to that of Caly–Asx^{DEU}. Major differences include the longer antiparallel helices, α9 and α10, in the coiled-coil domain of Caly; and the ordered CTE and β2–α2 and α8–β6 loops in BAP1 (Fig. 1e,f and Extended Data Fig. 5a–c). Despite the shorter ordered α9–α10 helices in BAP1, their role in bringing the UCH and ULD domains spatially close—thus allowing the formation of a complete ubiquitin cleft by the BAP1 UCH and the ASXL1^{DEU} domains—is maintained (Fig. 1d,e). It has been suggested that α9–α10 of Caly might facilitate the homodimerization and deubiquitination of nucleosomal H2AK119ub1 (refs. 24,25), but our multi-angle light-scattering (MALS) analysis does not provide evidence of dimerization of the human BAP1–ASXL1(1–378) complex (Extended Data Fig. 1d). In the middle of the ordered β2–α2 loop of BAP1, a conserved positively charged sequence motif spanning residues 56–60 adopts a helical form and is located in the acidic patch of histones H2A and H2B (Fig. 1f and Extended Data Fig. 5b). In addition, the tip of the α8–β6 loop (aa 220–223) is located above the cleft formed between histones H2B and H4, and the C-terminal portion of this loop interacts with ubiquitin (Fig.1f and Extended Data Fig. 5c).

Nucleosome anchoring through the BAP1 CTE

A section of the positively charged BAP1 CTE has previously been shown to be important for nucleosome binding and deubiquitination of H2AK119ub1 (ref. 23). In the structure, residues 699–715 of the BAP1 CTE interact with both histones and DNA near the nucleosomal dyad (superhelical location 0; SHL 0) (Figs. 1e and 2a). Residues 699–706 form a finger-like structure that pokes into a nucleoprotein pocket composed of the N-terminal ends of α3 of H3 and α2 of H4 of the same heterodimer, and two DNA strands spanning the minor groove at SHL 0 (Fig. 2b). Although Arg699 and Arg700 only contact the DNA backbone, Arg701 is inserted into the DNA minor groove and makes contacts with the bases and the sugar group (Fig. 2b,c). The side chain of Gln702 interacts with the DNA backbone and Ser47 of H4, and its main chain, together with that of Gly703, contacts Met120 and Lys122 of H3 (Fig. 2b). Outside the finger region, the remaining ordered CTE portion stretches along the DNA backbone, with Arg708 and Arg713 making direct contacts. The CTE sequence is strictly conserved among vertebrate BAP1 proteins, but a 35-residue insertion is found in *Drosophila* Caly (Extended Data Fig. 5a).

In agreement with previous findings, a partial truncation of CTE from residue 711 (ref. 23), or removing the entire CTE from residue 697,

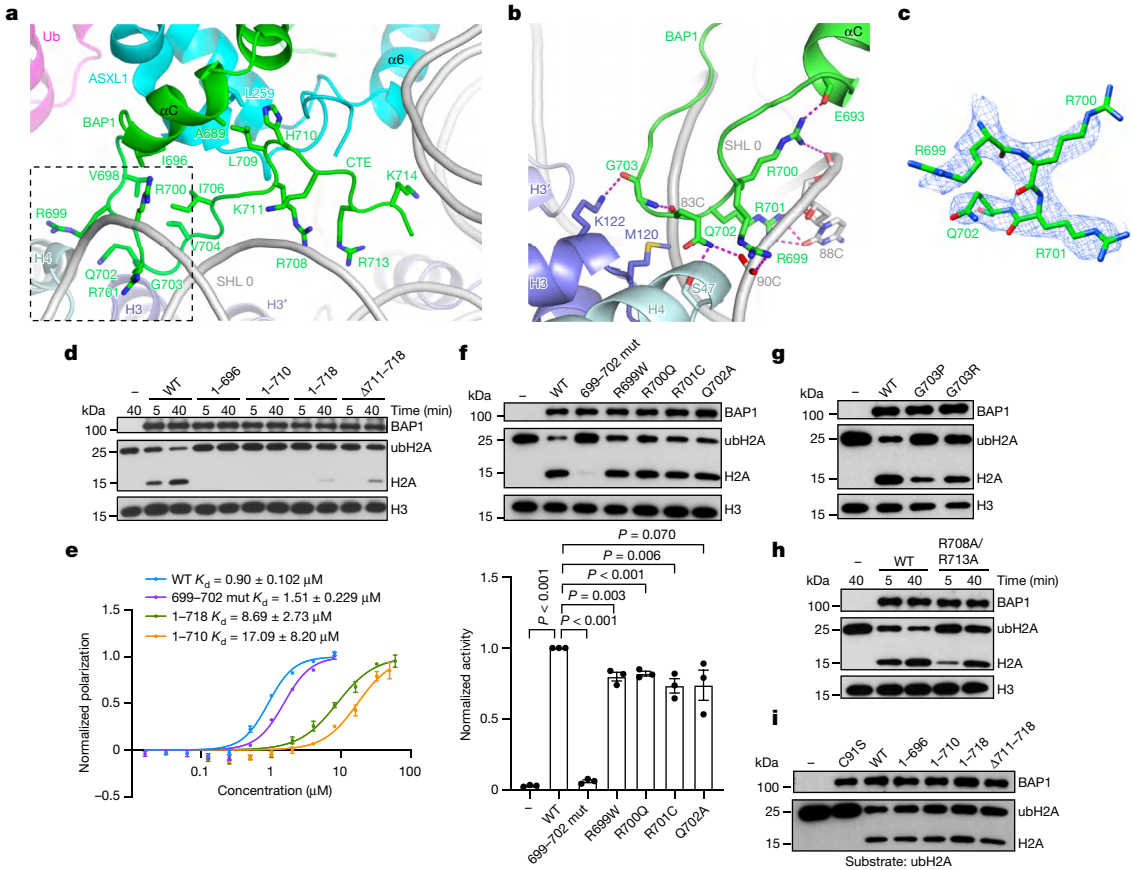


Fig. 2 | Structure and function of the BAP1 CTE. a, The BAP1 CTE is ordered from I696 to P715, and side chains of key residues are depicted in a stick model. **b**, Enlarged view of the interface between BAP1's CTE finger motif and NCP. Residues and nucleotides that are involved are indicated. Magenta dashed lines represent hydrogen bonds. **c**, EM density map surrounding residues 699–702 of the BAP1 CTE finger motif, contoured at 7σ. **d**, Western blot detections of nucleosomal H2AK119 deubiquitination of PR-DUB assembled with wild-type (WT) BAP1 or the indicated BAP1 mutants and ASXL1(1–378). Deubiquitination reactions were monitored at 5 min and 40 min; '–' signs indicates lanes without enzyme added. The top, middle and bottom blots used anti-His tag, anti-H2A and anti-H3 antibodies. Quantification shown in Extended Data Fig. 5d. **e**, Fluorescence polarization measurements of the binding affinities of the indicated BAP1–ASXL1(1–378) complexes to NCP(H2AK119ub1). K_d values were obtained by fitting the data to a model of specific binding with Hill slope in GraphPad (n = 3 biologically independent experiments, mean ± s.e.m.). **f**–**h**, Western blot analyses of the indicated BAP1 mutants, assayed as described in **d**. Blottings were performed at 10 min (**f**,**g**), or at 5 min and 40 min (**h**). H3 lanes are sample-processing controls. The bottom panel in **f** is the quantification of the top blot (n = 3 biological independent experiments, mean ± s.e.m.), normalized against the value of the wild-type complex. P values were derived from an unpaired two-tailed t-test, with no adjustment made for multiple comparisons. Quantifications of **g**,**h** are shown in Extended Data Fig. 5e,f. **i**, Deubiquitination of K119ub1 of free histone H2A by the indicated wild-type or mutant BAP1 complexes with ASXL1(1–378), detected by western blot.

abolished the nucleosomal H2AK119 deubiquitination activity of PR-DUB in vitro (Fig. 2d and Extended Data Fig. 5d). A shorter deletion from residue 719, or removal of an internal segment spanning residues 711–718, yielded substrantially reduced levels of enzymatic activity. BAP1 CTE truncation mutants also exhibited reduced nucleosome-binding abilities, as judged by fluorescent polarization experiments (Fig. 2e). BAP1–ASXL1(1–378) complexes assembled with BAP1(1–710) and BAP1(1–718) bound to NCP approximately 18- and 10-fold more weakly, respectively, than the wild-type complex, which has a dissociation constant (K_d) of 0.9 μM (Fig. 2e).

In comparison, BAP1 CTE point mutations had less effect (Fig. 2f). Changing all four residues of BAP1 699–702 to alanines (699–702 mut) seriously compromised DUB activity, although nucleosome binding was only mildly affected (Fig. 2e). Three cancer missense mutations, R699W, R700Q and R701C (ref. 28), as well as a Q702A mutation, led to modest reductions of activity compared to that of the wild-type complex (Fig. 2f). Gly703 at the tip of the CTE finger is situated in a shallow depression on the nucleosome surface near the interface between two H3–H4 heterodimers (Fig. 2a,b). Introducing a sharp bend with a G703P mutation or a G703R bulky side-chain substitution reduced the nucleosomal H2AK119ub1 DUB activity by about half (Fig. 2g and Extended Data Fig. 5e). Furthermore, alanine substitutions of the DNA-backbone-interacting CTE residues Arg708 and Arg713 also resulted in a marked reduction of BAP1's nucleosomal DUB activity (Fig. 2h and Extended Data Fig. 5f).

Because the BAP1 CTE is located distant from the ubiquitin-binding pocket and the enzymatic active site, it is possible that it influences H2AK119ub1 DUB activity by positioning the BAP1–ASXL1(1–378) complex correctly with respect to the nucleosome, rather than by directly affecting the enzymatic ability of PR-DUB. Indeed, various BAP1 CTE truncation complexes showed similar levels of enzymatic activity to that of the wild-type BAP1 complex toward free histone H2AK119ub1 (Fig. 2i), unlike their activities toward the nucleosomal substrate (Fig. 2d).

ASXL1 directs the nucleosome binding of the CTE

The structure shows that stabilization of BAP1's αC, to which the CTE is immediately joined, through interaction with ASXL1^{DEU}, is crucial for the spatial positioning of the CTE (Fig. 3a). In particular, Thr262 and Leu267 of ASXL1 contact αC residues through hydrophobic and van der

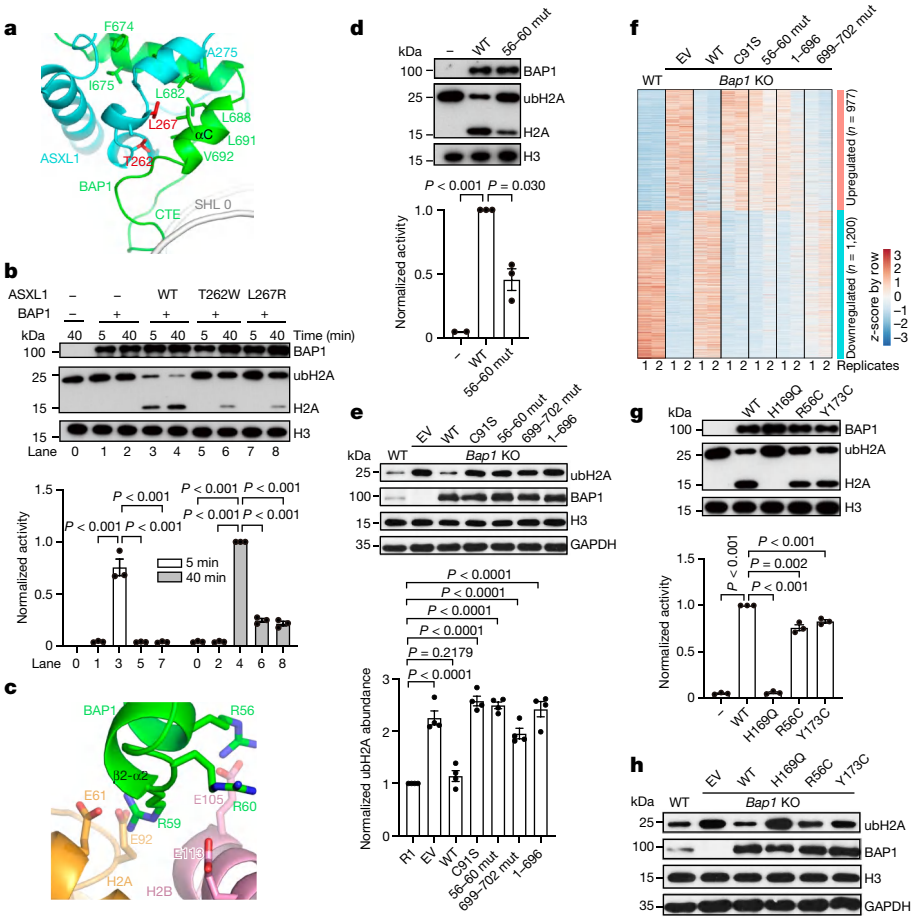


Fig. 3 | Nucleosome binding is key to H2AK119ub1 deubiquitination. **a**, Interaction between αC of BAP1^{ULD} (green) and ASXL1^{DEU} (cyan), involving Thr262 and Leu267 (red), directs the proper nucleosome binding of the BAP1 CTE. **b**, Top, western blot detection of H2AK119 deubiquitination by BAP1 or its complexes with wild-type or mutant forms of ASXL1(1–378), detected with anti-His tag (top), anti-H2A (middle) and anti-H3 (bottom) antibodies at 5 min and 40 min. H3 lanes here and in **d**,**g** were run on separate gels as sample-processing controls. Bottom, quantification of the western blot, normalized against the wild-type value at 40 min. Unshaded and grey-shaded bars indicate values at 5 min and 40 min, respectively. $n = 3$ biologically independent replicates; data are mean ± s.e.m. *P* values were derived from an unpaired two-tailed *t*-test, with no adjustments made for multiple comparisons. **c**, Magnified view of the interface between the RRSRR motif (aa 56–60) of BAP1 and the H2A–H2B acidic patch. **d**, Deubiquitination activity with the five RRSRR residues mutated to alanine (56–60 mut), measured as in **b** at 10 min. **e**, Top, western blot detection of H2AK119ub1 and BAP1 in wild-type R1 mouse ES cells and in *Bap1* KO cells transfected with an empty vector (EV) or with plasmids re-expressing the indicated BAP1 constructs, using anti-H2AK119ub1, anti-BAP1, anti-H3 and anti-GAPDH antibodies. Bottom, quantification of the levels of ubH2A, normalized against wild-type R1. Data are mean ± s.e.m. ($n = 4$ biologically independent experiments). **f**, Heat maps of transcriptional changes in BAP1-regulated genes in wild-type cells and the indicated rescued cell lines (two replicates each). **g**, In vitro deubiquitination activity of the indicated BAP1 cancer mutations, detected as in **b** at 10 min. **h**, Western blot detection of the levels of ubH2A in *Bap1* KO cells rescued with the indicated BAP1 constructs. See Extended Data Fig. 9 for quantification and transcriptional changes.

Waals interactions. Replacing them with bulky residues should interfere with the packing of BAP1 αC and alter CTE's mode of NCP binding. Not unexpectedly, T262W or L267R ASXL1 mutant complexes showed clear reductions in H2AK119ub1 DUB activity (Fig. 3b), but no appreciable difference in nucleosome-binding affinity was exhibited by the T262W mutant complex (Extended Data Fig. 5g). This result indicates that the stable positioning of BAP1 αC against ASXL1 directs the CTE to enable correct nucleosome binding. On another note, ASXL1^{DEU} does not seem to be involved in nucleosome binding, as a shorter ASXL1 fragment, ASXL1(1–330)—which lacks the entire C-terminal helix that is closest to NCP within ASXL1—did not show any appreciable difference in nucleosome binding (Extended Data Fig. 5g).

Targeting the H2A–H2B interface

Almost diametrically opposite to the CTE-finger-binding site on the nucleosomal surface, the β2–α2 loop in the BAP1 UCH domain is situated next to the H2A–H2B acidic patch (Fig. 1f). The acidic patch is a common protein–nucleosome interaction site[29], but a previous report found that it had no obvious involvement in binding the BAP1–ASXL1 complex[23]. The density for the loop segment encompassing residues 55–70 of BAP1 is relatively weak, but sufficient to locate a highly conserved positively charged sequence motif, RRSRR (spanning residues 56–60), in the vicinity of the H2A–H2B acidic patch (Fig. 3c and Extended Data Fig. 5b). Owing to the ambiguity of side-chain positions, we mutated all five residues to alanine (56–60 mut) and found that this change significantly compromised the H2AK119ub1 DUB activity (Fig. 3d), although no notable difference in nucleosome binding was detected (Extended Data Fig. 5g). As with the CTE mutations, the β2–α2 loop mutation did not affect the deubiquitination of the free histone H2AK119ub1 substrate (Extended Data Fig. 5h), providing evidence of the importance of the RRSRR motif for nucleosomal H2AK119ub1 deubiquitination that involves the H2A–H2B acidic patch.

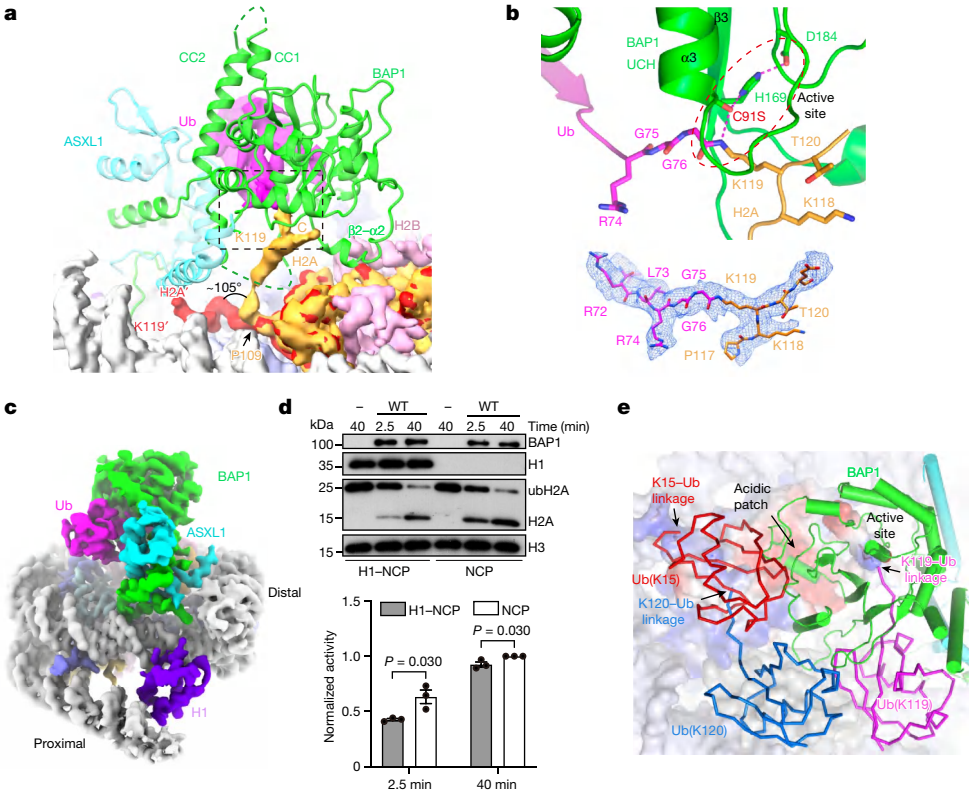


Fig. 4 | Conformational change of the H2A C-terminal tail and the role of H1 in PR-DUB activity. a, Comparison of the two H2A C-terminal tails from the same NCP reveals a large conformational difference. The density map of H2A in the BAP1–ASXL1-bound side is shown in yellow-orange and that in the opposite side (H2A′) is shown in red. **b**, Top, magnified view of the H2AK119–Ub isopeptide bond next to the BAP1 catalytic triad (C91, H169 and D184). Bottom, the composite density map, contoured at 4σ, showing the covalent linkage of H2AK119 and ubiquitin. **c**, Head-on view of H1 binding near the nucleosome dyad. Density regions for H1, NCP, BAP1, ASXL1 and ubiquitin are coloured in purple, grey, green, cyan and magenta, respectively. Linker-DNA arms near (proximal) and far from (distal) H1 are labelled. **d**, Western blot detection (top) and quantification (bottom) comparing the deubiquitinase activity with NCP and chromatosome substrates ($n = 3$ biologically independent replicates, mean ± s.e.m.). The H3 lanes are sample-processing controls. Statistical testing was performed in the same way as in Fig. 2f. **e**, Distinct spatial locations of H2AK15ub (red; Protein Data Bank (PDB) ID: 7E8I), H2BK120ub (blue; PDB ID: 4ZUX) and H2AK119ub (magenta), and the histone–ubiquitin isopeptide bonds relative to the NCP-bound BAP1–ASXL1. The NCP is shown in a surface representation, with H2A and H2B coloured according to electrostatic potential.

BAP1 mutations alter gene regulation

Consistent with the in vitro results, knockout (KO) of *Bap1* in mouse embryonic stem (ES) cells yielded an increased level of H2AK119ub1. Western blot analyses showed that transfecting the *Bap1* KO cells with mammalian expression vectors carrying coding sequences for the BAP1(C91S), BAP1(56–60 mut), BAP1(699–702 mut) and BAP1(1–696) mutants all resulted in higher levels of H2AK119ub1 as compared to transfection of the wild-type BAP1 plasmid (Fig. 3e). RNA-sequencing (RNA-seq) analysis identified several hundred up- and downregulated genes in *Bap1* KO cells compared to wild-type ES cells, in agreement with previous data[7,8] (Extended Data Fig. 6). Gene expression analyses of PR-DUB-regulated genes by RNA-seq revealed similar effects to those that were observed in the western blot analyses, with the CTE-finger alanine-substitution mutant showing partial rescues, consistent with the cumulative effect of the CTE constituents (Figs. 2d and 3e,f and Extended Data Fig. 7a,b).

Again consistent with the in vitro results (Fig. 2f), the R699W and R701C cancer-associated mutations in the BAP1 CTE led to mild reductions of DUB activity in cells (Extended Data Fig. 8a), and RNA-seq analysis revealed corresponding changes of BAP-regulated genes (Extended Data Fig. 8b–d). Three non-CTE cancer-related mutations—H169Q in the catalytic active site, R56C in the β2–α2 helix (Fig. 3c) and Y173C on β3, which is involved in stabilizing α8 (Fig. 1f)—also affected H2AK119 DUB activity to varying degrees both in vitro and in cells (Fig. 3g,h and Extended Data Fig. 9a), with the catalytic H169Q mutation showing the most significant defect, and resulted in corresponding gene expression changes (Extended Data Fig. 9b–d).

In addition, frameshift and premature truncation mutations of ASXL1 are frequently associated with myeloid leukaemias and Bohring–Opitz syndrome. Ectopic expression of these mutants resulted in reduced levels of H2AK119ub1, leading to the hypothesis that ASXL1's C-terminal region, which is absent in these mutants, might have an autoinhibitory activity[22]. We directly tested this hypothesis by fusing two C-terminal segments, one spanning residues 1401–1541 and the other residues 1306–1541, to the N-terminal 378 residues of ASXL1. Both fusion proteins contain N- and C-terminal domains that are conserved in ASXL1–ASXL3, and their complexes with BAP1 showed in vitro DUB activities that were indistinguishable from that attained with the ASXL1 N-terminal fragment only (Extended Data Fig. 10a,b), suggesting that there are alternative mechanisms for the elevated DUB activity that is associated with these ASXL1 cancer mutations.

The H2A tail in the active site of BAP1

The covalent linkage between the lysine side chain and ubiquitin is flexible and has not (to our knowledge) previously been observed in protein–nucleosome complexes with ubiquitinated histones[30,31]. Our 3.7-Å-resolution unsharpened global refinement map shows that the C-terminal tail of H2A, starting from Pro109, turned by around 105°

toward the active site of BAP1 (Fig. 4a and Extended Data Fig. 2c), with the Cα of Lys119 rising by around 26 Å above the nucleosomal surface from the Cα of Pro109. By contrast, the H2A tail on the opposite side of the nucleosome, shown as H2A′ in the superimposed figure, remains in the same conformation as that in the free NCP (Fig. 4a). H2A C-terminal residues beyond Lys118, as observed in the high-resolution NCP structure[32], will clash with the BAP1 CTE, suggesting that nucleosome binding of BAP1–ASXL1 displaces the C-terminal tail of H2A in the NCP structure, and that this enables the active deubiquitination of H2AK119ub1 (Fig. 4a). It is worth noting that H2AK118 may be monoubiquitinated inside the cell, although it appears that only one of the two consecutive lysines—predominantly Lys119 (K119)—is modified at a time[33,34]. The structure shows that the C-terminal tail of H2A may be further extended to place an ubiquitinated H2AK118 at the active site for deubiquitination (Fig. 4b). Finally, the 3.9-Å-resolution locally refined map clearly shows the placement of the isopeptide bond between the C terminus of ubiquitin and the ε amino group of H2AK119, next to the catalytically inactivated C91S residue of BAP1 (Fig. 4b), thus providing a rare glimpse into the geometric arrangement of the scissile isopeptide bond and the active-site residues of hydrolases of the UCH family. No interaction is observed between BAP1 and the side chains of other H2A residues that neighbour Lys119, supporting the notion that the H2AK119 specificity of PR-DUB results from the particular spatial configuration of nucleosome binding, and not from its intrinsic preference for the amino acid sequence of H2A.

Functional roles of the linker histone

H1 is present in both euchromatic and heterochromatic regions, and is generally associated with gene repression[35]. It can coexist with H2AK119ub1 in the same nucleosome[36], and we showed that PR-DUB is active toward a chromatosome substrate (Extended Data Fig. 1c). In our structure, the globular domain of H1 is visualized bound to nucleosomal DNA in an on-dyad mode[37] (Fig. 4c). It is situated closer to the proximal arm of the linker DNA than to the distal arm, and there is no direct contact with the PR-DUB complex. However, an H2AK119ub1 chromatosomal substrate exhibited a modestly reduced DUB activity (Fig. 4d). This difference might be attributable to the asymmetrical contact between the linker H1 histone and the DNA, which appears to limit the binding of PR-DUB to one side of the nucleosomal surface owing to steric constraint (Extended Data Fig. 10c). Without the linker histone, independent binding of PR-DUB to either side of the nucleosomal surface seems to be possible, although only at a low percentage (Extended Data Fig. 4c,f). Nevertheless, our demonstration that PR-DUB can deubiquitinate nucleosomal and chromatosomal substrates alike suggests that H2AK119ub1 and H1 can separately modulate gene expression. They might also work synergistically, as indicated by the observation that deubiquitination of H2AK119 promotes the dissociation of H1 from the nucleosome[38].

Conclusions

Our study reveals that the distinct nucleosome-binding mode of PR-DUB displaces the H2A C-terminal tail on the nucleosome surface and allows H2AK119ub1 access to the enzymatic active site of BAP1. In this mode of binding, the isopeptide bond between ubiquitin and other nucleosomal ubiquitination sites—such as H2BK120 and H2AK13 or H2AK15—cannot reach the active site of PR-DUB owing to distance and geometric constraints (Fig. 4e), thus specifying the physiologically relevant H2AK119ub1 specificity of PR-DUB. Consistent with the structural findings, mutations of BAP1 residues that are involved in catalysis or NCP binding, including cancer-associated mutations, all resulted in the reduction of DUB activity and the alteration of gene expression, to varying degrees. The DEUBAD domain of ASXL1 has at least two roles in H2AK119 deubiquitination: formation of the ubiquitin-binding cleft together with the UCH domain of BAP1; and securing the CTE finger for interaction with the nucleosome, through contact with the αC of BAP1. Frequent cancer-associated truncations of ASXL1–ASXL3 retain an intact DEUBAD domain, and ectopic expression of ASXL1 truncation mutants resulted in substantially reduced levels of cellular H2AK119ub1, indicative of an autoinhibitory effect of ASXL1's C-terminal domain. Our biochemical analysis suggests that further effects of the C-terminal domains of ASXL proteins might be important for tumorigenesis.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05841-y.

1. Simon, J. A. & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.* **10**, 697–708 (2009).
2. Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome regulation by Polycomb and Trithorax: 70 years and counting. *Cell* **171**, 34–57 (2017).
3. Blackledge, N. P. & Klose, R. J. The molecular principles of gene regulation by Polycomb repressive complexes. *Nat. Rev. Mol. Cell Biol.* **22**, 815–833 (2021).
4. Scheuermann, J. C. et al. Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* **465**, 243–247 (2010).
5. Wang, H. et al. Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431**, 873–878 (2004).
6. Campagne, A. et al. BAP1 complex promotes transcription by opposing PRC1-mediated H2A ubiquitylation. *Nat. Commun.* **10**, 348 (2019).
7. Conway, E. et al. BAP1 enhances Polycomb repression by counteracting widespread H2AK119ub1 deposition and chromatin condensation. *Mol. Cell* **81**, 3526–3541 (2021).
8. Fursova, N. A. et al. BAP1 constrains pervasive H2AK119ub1 to control the transcriptional potential of the genome. *Genes Dev.* **35**, 749–770 (2021).
9. Bonnet, J. et al. PR-DUB preserves Polycomb repression by preventing excessive accumulation of H2Aub1, an antagonist of chromatin compaction. *Genes Dev.* **36**, 1046–1061 (2022).
10. Beck, D. B. et al. Chromatin in the nuclear landscape. *Cold Spring Harb. Symp. Quant. Biol.* **75**, 11–22 (2010).
11. Aloia, L., Di Stefano, B. & Di Croce, L. Polycomb complexes in stem cells and embryonic development. *Development* **140**, 2525–2534 (2013).
12. Chen, Z., Djekidel, M. N. & Zhang, Y. Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. *Nat. Genet.* **53**, 551–563 (2021).
13. Parreno, V., Martinez, A. M. & Cavalli, G. Mechanisms of Polycomb group protein function in cancer. *Cell Res.* **32**, 231–253 (2022).
14. Gaytán de Ayala Alonso, A. et al. A genetic screen identifies novel polycomb group genes in *Drosophila*. *Genetics* **176**, 2099–2108 (2007).
15. Harbour, J. W. et al. Frequent mutation of *BAP1* in metastasizing uveal melanomas. *Science* **330**, 1410–1413 (2010).
16. Pena-Llopis, S. et al. BAP1 loss defines a new class of renal cell carcinoma. *Nat. Genet.* **44**, 751–579 (2012).
17. Dey, A. et al. Loss of the tumor suppressor BAP1 causes myeloid transformation. *Science* **337**, 1541–1546 (2012).
18. Carbone, M. et al. BAP1 and cancer. *Nat. Rev. Cancer* **13**, 153–159 (2013).
19. Katoh, M. Functional and cancer genomics of ASXL family members. *Br. J. Cancer* **109**, 299–306 (2013).
20. Wang, L., Birch, N. W., Zhao, Z., Nestler, C. M. & Shilatifard, A. Epigenetic targeted therapy of stabilized BAP1 in *ASXL1* gain-of-function mutated leukemia. *Nat. Cancer* **2**, 515–526 (2021).
21. Abdel-Wahab, O. et al. *ASXL1* mutations promote myeloid transformation through loss of PRC2-mediated gene repression. *Cancer Cell* **22**, 180–193 (2012).
22. Balasubramani, A. et al. Cancer-associated *ASXL1* mutations may act as gain-of-function mutations of the ASXL1–BAP1 complex. *Nat. Commun.* **6**, 7307 (2015).
23. Sahtoe, D. D., van Dijk, W. J., Ekkebus, R., Ovaa, H. & Sixma, T. K. BAP1/ASXL1 recruitment and activation for H2A deubiquitination. *Nat. Commun.* **7**, 10292 (2016).
24. De, I. et al. Structural basis for the activation of the deubiquitinase Calypso by the Polycomb protein ASX. *Structure* **27**, 528–536 (2019).
25. Foglizzo, M. et al. A bidentate Polycomb repressive-deubiquitinase complex is required for efficient activity on nucleosomes. *Nat. Commun.* **9**, 3932 (2018).
26. Sahtoe, D. D. et al. Mechanism of UCH-L5 activation and inhibition by DEUBAD domains in RPN13 and INO80G. *Mol. Cell* **57**, 887–900 (2015).
27. Vander Linden, R. T. et al. Structural basis for the activation and inhibition of the UCH37 deubiquitylase. *Mol. Cell* **57**, 901–911 (2015).
28. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
29. McGinty, R. K. & Tan, S. Principles of nucleosome recognition by chromatin factors and enzymes. *Curr. Opin. Struct. Biol.* **71**, 16–26 (2021).
30. Morgan, M. T. et al. Structural basis for histone H2B deubiquitination by the SAGA DUB module. *Science* **351**, 725–728 (2016).
31. Wang, H. et al. Structure of the transcription coactivator SAGA. *Nature* **577**, 717–720 (2020).

32. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J. Mol. Biol.* **319**, 1097–1113 (2002).
33. Lan, L. et al. Monoubiquitinated histone H2A destabilizes photolesion-containing nucleosomes with concomitant release of UV-damaged DNA-binding protein E3 ligase. *J. Biol. Chem.* **287**, 12036–12049 (2012).
34. Nakata, S. et al. The dynamics of histone H2A ubiquitination in HeLa cells exposed to rapamycin, ethanol, hydroxyurea, ER stress, heat shock and DNA damage. *Biochem. Biophys. Res. Commun.* **472**, 46–52 (2016).
35. Zlatanova, J. & Vanholde, K. Histone H1 and transcription—still an enigma. *J. Cell Sci.* **103**, 889–895 (1992).
36. Jason, L. J. M., Finn, R. M., Lindsey, G. & Ausio, J. Histone H2A ubiquitination does not preclude histone H1 binding, but it facilitates its association with the nucleosome. *J. Biol. Chem.* **280**, 4975–4982 (2005).
37. Zhou, B. R. et al. Distinct structures and dynamics of chromatosomes with different human linker histone isoforms. *Mol. Cell* **81**, 166–182 (2021).
38. Zhu, P. et al. A histone H2A deubiquitinase complex coordinating histone acetylation and H1 dissociation in transcriptional regulation. *Mol. Cell* **27**, 609–621 (2007).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Plasmid construction

Full-length cDNA of human BAP1 was amplified by PCR, and bacterial expression plasmids for full-length BAP1 and BAP1 fragments encompassing residues 1–696, 1–710, 1–718 and an internal deletion of residues 711–718 (Δ711–718) were cloned into a modified pRSFDuet-1 vector (Novagen), termed pRSFDuet-Smt, in which a 6×His-SUMO cassette was inserted between the NcoI and BamHI sites. A plasmid with the coding sequence for ASXL1 residues 1–480 was purchased from Sangon Biotech, and bacterial expressing plasmids for ASXL1 fragments spanning residues 1–330 or 1–378 were cloned into the pGEX-6P-1 vector (GE Healthcare). Plasmids for the expression of C-terminal glutathione *S*-transferase (GST)-fused ASXL1(1–378)-$(GGGGS)_2$-(1401–1541) and ASXL1(1–378)-$(GGGGS)_2$-(1306–1541) fragments, termed ASXL1-C1 and ASXL1-C2, respectively, were constructed by linking the C-terminal pieces, which were amplified from human cDNA, with the N-terminal region by PCR. Expression plasmids for BAP1 and ASXL1 mutants were generated by site-directed mutagenesis (Supplementary Table 1).

A bacterial expression plasmid for human H2B was generated by cloning the cDNA into pET30a (Novagen), and a co-expression plasmid for human H3 and H4 was obtained by inserting their cDNAs into the same pCDFDuet-1 vector (Novagen). A plasmid expressing C-terminal 6×His-tagged full-length human H1.4 was constructed using a pET-28a vector (Novagen).

DNA fragments of 187 bp for nucleosome assembly were generated from a plasmid with four tandem copies of the 187-bp fragment cloned into a pBluescript vector (Addgene). Each DNA fragment contains a 147-bp Widom 601 positioning sequence flanked by a 17-bp and a 23-bp linker-DNA sequence[39] (Supplementary Table 1). An EcoRV restriction site is engineered at both ends of the DNA for isolation of the 187-bp fragment.

Expression and purification of BAP1–ASXL1 complexes

Wild-type or mutant BAP1 and ASXL1(1–378) proteins were co-expressed in *Escherichia coli* BL21(DE3) RIL cells. The plasmids were co-transformed into the *E. coli* cells, then a single colony was picked and cultured overnight in 150 ml LB medium containing 50 μg ml^{-1} kanamycin, 100 μg ml^{-1} ampicillin and 34 μg ml^{-1} chloramphenicol at 37 °C. The start-up culture was then used to inoculate 5-l or 2.4-l cultures, and protein production was induced with the addition of 0.25 mM isopropylthio-β-galactoside (IPTG) when the cell density reached optical density at 600 nm ($OD_{600\,nm}$) values between 0.6 and 0.8, followed by shifting the temperature to 16 °C and continuing shaking the culture for 17 h. Cells were then collected by centrifugation, resuspended in 20 mM Tris-HCl pH 8.0, 500 mM NaCl, 5% (v/v) glycerol, 10 mM imidazole and 1 mM phenylmethylsulfonyl fluoride (PMSF), and lysed by pressure rupture followed by ultrasonication. Cell debris was then cleared by centrifugation, and the supernatant was incubated with Ni-excel affinity resins (GE Healthcare), followed by washing the resins first with the lysis buffer, and subsequently with 20 mM Tris-HCl pH 7.5, 500 mM NaCl, 5% glycerol and 25 mM imidazole, and the bound proteins were eluted in this buffer by adjusting the imidazole concentration to 200 mM.

For purification of the BAP1(C91S)–ASXL1(1–378) complex, GST and His-SUMO tags were cleaved after the Ni-excel step by the addition of PreScission and SUMO proteases during overnight dialysis in buffer A-100 (20 mM Tris-HCl pH 7.5, 100 mM NaCl, 5% glycerol, 1 mM ethylenediaminetetraacetic acid (EDTA) and 1 mM dithiothreitol (DTT)). The dialysed sample was then loaded onto a HiTrap SP cation exchange column (GE Healthcare) pre-equilibrated with buffer A-100, and a linear gradient of 0–70% buffer A-1000, which differs from A-100 in NaCl concentration (100 mM to 1 M), was developed to elute the bound proteins. The eluted fractions were analysed by SDS–PAGE, and the ones enriched with the BAP1(C91S)–ASXL1(1–378) complex were pooled and concentrated, before loading onto a Superdex 200 HiLoad 16/60 sizing column in the A-500 buffer. The second peak from the sizing column contains the protein complex of interest, and the fractions were pooled and concentrated for complex formation with the nucleosome.

Wild-type and mutant BAP1–ASXL1(1–378) complexes used in fluorescence polarization experiments and H2A and H2B deubiquitinase assays were purified following similar procedures to that outlined for the BAP1(C91S)–ASXL1(1–378) complex above, except that the size-exclusion step was performed with a Superdex 200 10/300 GL in the A-100 buffer.

For the nucleosomal H2AK119ub1 deubiquitination assay, mutant BAP1–ASXL1 complexes were first purified through Ni-excel resins, followed by cleavage of the GST tag using PreScission protease, leaving the His-SUMO on BAP1. The GST tag was removed by applying the sample through an SP column. The target protein complex was eluted with a linear gradient of 0–70% buffer A-1000. The BAP1 complexes with ASXL1-C1 and ASXL1-C2 fragments were successively purified through Ni-excel and glutathione (GSH) sepharose (GE Healthcare) resins. The target protein was finally eluted by 20 mM Tris-HCl pH 7.5, 500 mM NaCl, 5% glycerol and 30 mM GSH. These proteins were then flash-frozen in liquid nitrogen and stored at −80 °C before use. All of the purification steps were performed at 4 °C.

Preparation of histones, nucleosomal DNA and nucleosomes

Chemically synthesized full-length human H2A monoubiquitinated at K119 (H2AK119ub1) and H2B monoubiquitinated at K120 (H2BK120ub1) were purchased from Hefei KS-V Peptide Biotechnology.

Human histones H2B and H2A were expressed in an insoluble form in *E. coli*, and purified following a method described before[40]. Human H3 and H4 were co-expressed in soluble forms in the BL21 (DE3) RIL strain of *E. coli*[41], and purified through successive HiTrap Heparin and Superdex 200 HiLoad 16/60 (GE Healthcare) column chromatography steps. Purified H3–H4 tetramers in 20 mM Tris-HCl pH 7.5, 2 M NaCl, 1 mM EDTA and 0.1% β-mercaptoethanol were concentrated and flash-frozen in liquid nitrogen and stored at −80 °C.

Linker histone H1 was expressed in the BL21 (DE3) Rosetta strain of *E. coli*. Collected cells were resuspended in 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 6 M guanidine-HCl and 1 mM PMSF, followed by cell rupture by pressure and ultrasonication. Cellular debris was cleared by centrifugation, and the supernatant was loaded onto a Ni-excel column. Then, the loaded column was washed with buffer containing 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 5% glycerol and 7 M urea, first without imidazole, then with the addition of 15 mM imidazole, and finally with 200 mM imidazole to elute the bound H1. The eluted sample was dialysed against 20 mM Tris-HCl pH 7.5, 500 mM NaCl and 1 mM EDTA three times to remove urea. Then, the sample was purified through a HiTrap Heparin HP cation exchange column (GE Healthcare). Highly purified H1 fractions were pooled and flash-frozen in liquid nitrogen and stored at −80 °C.

The plasmid carrying the tandem 187-bp nucleosomal DNA fragment was produced using the DH5α strain of *E. coli* in cell culture with TB medium. Large-scale plasmid purification was performed following a published procedure[42]. The 187-bp fragment was excised from the plasmid by EcoRV digestion, isolated by polyethylene glycol (PEG) precipitation and further purified through ethanol precipitation.

The 5′-FAM-labelled 167-bp DNA fragment used for the assembly of nucleosomes in fluorescent polarization experiments was generated by PCR using a 5′-FAM-labelled primer. The sequence of the 167-bp fragment, which contains a centrally located 147-bp Widom sequence, is shown in Supplementary Table 1. The PCR product was purified using the Omega E.Z.N.A. gel extraction kit in the dark.

The assembly of histone octamers and reconstitution of NCPs by salt dialysis were performed according to a published protocol[40], and the assembly of the H1-containing chromatosome was performed according to a previously described procedure[43]. In brief, H1 was added to the NCP dialysis mixture at a molar ratio of 0.8:1 when the salt concentration

reached 600 mM, followed by further dialysis at 600 mM NaCl for 3 h, and then against the HE buffer (10 mM HEPES pH 8.0 and 1 mM EDTA) for 4 h.

Assembly of the BAP1–AXSL1–NCP complexes

For assembly of the BAP1(C91S)–ASXL1(1–378)–H1–NCP complex, purified BAP1–ASXL1 and H1–NCP complexes were separately dialysed into 20 mM HEPES pH 8.0, 50 mM NaCl, 5% glycerol, 1 mM EDTA and 1 mM DTT, and precipitants were removed by centrifugation. Then, BAP1–ASXL1 and H1–NCP complexes were mixed at a 3:1 molar ratio and incubated on ice for 30 min.

For the BAP1(C91S)–ASXL1(1–378)–NCP complex, purified BAP1–ASXL1 and NCP were first dialysed into 20 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA and 1 mM DTT separately. The BAP1–ASXL1 complex was then added in a threefold molar excess to NCP, and the mixture was dialysed into 20 mM HEPES pH 8.0, 50 mM NaCl, 5% glycerol, 1 mM EDTA and 1 mM DTT.

The above samples were then subjected to GraFix (ref. 44). One hundred microlitres of sample was loaded to the 5-ml tubes with a 10–40% linear glycerol gradient and a 0–0.15% linear glutaraldehyde gradient in 20 mM HEPES pH 8.0 and 50 mM NaCl, followed by centrifugation at 30,000 rpm for 17 h using a Hitachi SW50 rotor at 4 °C. After centrifugation, the sample was separated into 25 fractions, 200 µl each, from top to bottom, and the fractions were examined by 4% native PAGE and negative-staining EM. Fractions with homogeneous, properly sized particles were picked and subjected to glycerol removal by dialysis and concentration by centrifugation.

Cryo-EM sample preparation and data collection

Quantifoil R2/1 300 mesh grids were glow-discharged for 1 min using a Gatan Plasma System. Three microlitres of BAP1–ASXL1–NCP sample, with or without H1, was placed onto the grid, which was plunge-frozen in liquid ethane by using a Vitrobot (Thermo Fisher Scientific) with settings at 10 °C, 100% humidity, 10 s incubation time and 3.5 s blot time.

Automated data acquisition was performed using a modified SerialEM (ref. 45) script to perform beam-image shift data collection[46]. Images for the histone-H1-containing complex were collected on a Talos Arctica 200 kV machine (Thermo Fisher Scientific) equipped with a K2 Summit camera (Gatan) in super-resolution mode at a nominal magnification of 130,000 and a pixel size of 0.5 Å, with a defocus range from −1.5 µm to −2.0 µm. The exposure time was 5.2 s, with a total exposure dose of 50 electrons per $Å^2$ over 32 frames. Images for the nucleosome complex were collected using a Titan Krios 300 kV machine (FEI) equipped with the same type of camera, and a dosage of 60 electrons per $Å^2$ was used for data collection. A total of 21,466 micrographs for the BAP1(C91S)–ASXL1(1–378)–H1–NCP sample and 6,269 micrographs for the BAP1(C91S)–ASXL1(1–378)–NCP sample were collected (Extended Data Table 1).

Image processing and 3D reconstruction

For the BAP1(C91S)–ASXL1(1–378)–H1–NCP complex, all micrographs were motion-corrected and dose-weighted with patch-based schemes in cryoSPARC (ref. 47), and contrast transfer function (CTF) parameters were estimated with the patch-based method in cryoSPARC. Micrographs with poor Thon rings were discarded, and around 100 micrographs were randomly selected for reference-free particle picking using cryoSPARC blob picker. The picked particles were used for an initial round of two-dimensional (2D) classification, and 2D class averages with clear features were selected as templates for particle auto-picking of the whole dataset. A total of 12,435,535 particles were auto-picked and extracted with twofold binning using a box size of 256 pixels. After several rounds of reference-free 2D classification, a total of 4,818,641 particles were selected for cryoSPARC ab initio reconstruction. After converting the cryoSPARC particle location file to a star file using the csparc2star.py script in pyem (ref. 48), several rounds of global three-dimensional (3D) classification and one round of no-alignment 3D classification in RELION 3.0 were performed[49]. A total of 443,790 particles were selected and re-extracted without binning using a box size of 352 pixels and subjected to non-uniform refinement, yielding a cryo-EM density map at an overall resolution of 2.75 Å. Multibody refinement analysis revealed that the BAP1–ASXL1 and ubiquitin module is placed, with respect to NCP, relatively dynamically (data not shown). To improve the quality of the density for this module, the density belonging to the NCP and H1 was subtracted from the 2D images and these images were subjected to two rounds of no-alignment 3D classification with a soft-edged mask covering the BAP1–ASXL1–Ub region. A total of 95,225 particles in classes with obviously improved densities for the BAP1–ASXL1–Ub module were selected and re-extracted using a box size of 352 pixels for local refinement with the soft-edged mask, which resulted in a much-improved density at a resolution of 3.9 Å. Meanwhile, global refinement using these particles without density subtraction yielded a 3.0-Å-resolution density map for the entire complex, as evidenced in a region near the catalytic active site of BAP1. The map was post-processed in deepEMhancer to improve interpretability[50]. All resolutions were estimated according to the gold-standard Fourier shell correlation criterion value of 0.143. The 2.75-Å and 3.0-Å global refinement maps and the 3.9-Å local refinement map were used to generate a composite map in PHENIX (ref. 51). A flow chart schematically summarizing the procedures of image processing and 3D reconstruction is shown in Extended Data Fig. 2.

For the BAP1(C91S)–ASXL1(1–378)–NCP complex without H1, image pre-processing steps similar to that described above were used. A total of 5,181,645 particles were picked, and 1,006,755 of these were kept after 2D classification. Then, global 3D classification, particle subtraction, 3D classification without alignment and particle re-extraction were performed and 56,370 particles were kept for the final refinements. Global refinement using these particles yielded a 4.2-Å-resolution density map, and the density was further improved by density modification in PHENIX (ref. 52). Meanwhile, local refinement using these 56,370 particles yielded a 4.4-Å-resolution density map in the region of the BAP1–ASXL1–Ub module. A flow chart schematically summarizing the procedures of image processing and 3D reconstruction is shown in Extended Data Fig. 4.

Model building, refinement and validation

An initial model of the BAP1(C91S)–ASXL1(1–378)–H1–NCP complex was built with the help of homologous substructures, including the cryo-EM structure of the H1-containing chromatosome (PDB ID: 7K5Y; ref. 37), and crystal structures of Caly–ASX (PDB ID, 6HGC) and UCH-L5–RPN13–Ub (PDB ID, 4UEL) complexes[24,26]. UCSF Chimera was used for docking the homologous structures into the density, and model adjustment and editing were performed with Coot (ref. 53). De novo building of various regions without previous structural knowledge, including the BAP1 CTE region and the H2A tail surrounding the K119–Ub isopeptide bond, and multiple rounds of manual model adjustment and rebuilding, were performed in Coot. Subsequently, the model was subjected to real-space refinement using PHENIX and density-constrained stereochemistry refinement using Rosetta (ref. 54), and, finally, the model was validated with MolProbity (ref. 55). Statistics for model refinement and validation are shown in Extended Data Table 1. Cryo-EM density map and structural model figures were prepared using Chimera (ref .[56]), ChimeraX (ref. 57), Coot and PyMOL (Schrödinger).

In vitro deubiquitination assay

Deubiquitination reactions with free histone H2BK120ub1 and H2AK119ub1 substrates were performed with 10 pmol of histone mixed with 1 pmol of the BAP1–ASXL1(1–378) complex in 20 µl buffer at 25 °C for defined durations; reactions with nucleosomal substrates carrying corresponding ubiquitinated histones were performed using 3.3 pmol

nucleosome and 3.3 pmol of the BAP1–ASXL1(1–378) complex in 20 µl buffer. Reactions were stopped by the addition of 10 µl SDS loading buffer.

Separate deubiquitination assays of nucleosomal H2AK119ub1 were performed with 3.2–3.5 pmol of NCP mixed with wild-type or mutant His-SUMO-tagged BAP1–ASXL1(1–378) complexes at enzyme-to-substrate molar ratios of 0.3:1 or 0.5:1 in 20-µl reactions. These reactions were allowed to proceed at 25 °C for a defined duration before being stopped by the addition of 20 µl SDS loading buffer. When comparing the activities of BAP1(WT)–ASXL1(1–378) and BAP1(C91S)–ASXL1(1–378) for chromatosomal H2AK119ub1, 4.4 pmol NCP was deubiquitinated by the enzyme complex in a 40-µl system at 25 °C for 20 min. The enzyme was twofold serially diluted from the double molar amount of substrate, and the reaction was terminated by 20 µl SDS loading buffer. The results were analysed by western blot with anti-His tag (Invitrogen, MA1-21315, 1:1,000 dilution), anti-GST (Beyotime, AF0174, 1:4,000 dilution), anti-H2A (Millipore, 07-146; Cell Signaling Technology, 12349s, 1:5,000 dilution), anti-H2B (Abcam, ab1790, 1:5,000 dilution), anti-H3 (Abcam, ab1791, 1:10,000 dilution) and anti-BAP1 (Cell Signaling Technology, 13271s, 1:1,000 dilution) antibodies.

All of the reactions above were performed in a buffer with 50 mM Tris-HCl pH 7.5, 100 mM NaCl, 8 mM $MgCl_2$ and 1 mM DTT, except for the reactions in Extended Data Fig. 8b, in which 150 mM NaCl was used and 0.1% NP40 was added. The uncropped and unprocessed scans are supplied in Supplementary Fig. 1.

Quantification of deubiquitination activity was based on scanned intensities of western blots from three biologically independent experiments: catalytic activity = [H2A]/([H2Aub] + [H2A]), and the value was then normalized with that of the wild type in the same blot. Bar charts represent plots of mean ± s.e.m., and *P* values denote the results of an unpaired two-tailed *t*-test.

Fluorescence polarization assay

The of 5′-FAM-labelled 167-bp NCP (100 nM) was mixed with twofold serially diluted samples of BAP1–ASXL1 in buffer containing 20 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM EDTA and 1 mM DTT. The mixtures were incubated at 4 °C for 30–60 min before being measured with an Envision multimode plate reader (Perkin Elmer). The mP values after subtracting background were fitted with GraphPad Prism v.9 using the specific binding model with Hill slope: normalized polarization value = (measured mP value − mP value of background)/mP_{max}. The calculated K_d shows symmetrical confidence intervals with a 95% confidence level. Measurements were taken from three distinct samples and each sample was measured three times.

MALS

One hundred microlitres of purified BAP1(C91S)–ASXL1(1–378) complex at a concentration of around 2 mg ml^{-1} was injected into the Superdex 200 Increase 10/300 GL column (GE Healthcare) at a flow rate of 0.4 ml per min in a buffer of 20 mM Tris-HCl, pH 7.5, 100 mM NaCl and 1 mM EDTA. The chromatography system was coupled with the 18-angle light-scattering detector (DAWN HELEOS II, Wyatt Technology) and differential refractive index detector (Optilab rEX, Wyatt Technology), and the data were analysed by the ASTRA software.

Cell lines and cell culture

R1 mouse ES cells were grown on 0.1% gelatin-coated dishes in 2i/LIF-containing DMEM medium (Gibco) supplemented with 15% fetal bovine serum (Hyclone), 2 mM glutamine (Sangon Biotech), 100 U ml^{-1} penicillin, 0.1 mg ml^{-1} streptomycin (Sangon Biotech), 0.1 mM non-essential amino acids (Sigma), 1 mM sodium pyruvate (Gibco), 50 mM β-mercaptoethanol and 1,000 U ml^{-1} leukaemia inhibitory factor (Novoprotein), and GSK3b and MEK 1/2 inhibitors (Selleckchem) to final concentrations of 3 µM and 1 µM, respectively.

The stable *Bap1* KO cell line was generated by CRISPR–Cas9. sgRNA sequences were as follows: sgRNA1: ACCGAAGGTCGTGGGTAACG; sgRNA2: AGGGGCGAGAGCGTTTCCGC.

BAP1 rescue cells were generated by the PiggyBac system. *Bap1* KO cells were mock-transfected or transfected with an EF1α-driven plasmid encoding wild-type BFP-P2A-HA-BAP1 or BAP1 truncations or mutations, and another vector carrying the transposase using Lipofectamine 3000 (Thermo Fisher Scientific), according to the manufacturer's instructions. BFP-positive cells were selected by fluorescence-activated cell sorting (FACS) 24 h after transfection. After further culturing for about seven days, BFP-positive cells within a specific BFP signal range were collected as stable BAP1 rescue cells. A representative FACS gating strategy is shown in Supplementary Fig. 2.

Cellular western blotting

For the analysis of BAP1 with total protein lysates, mouse ES cells were lysed in SDS lysis buffer (50 mM Tris-HCl pH 6.8, 2% SDS and 5 mM DTT) and boiled at 99 °C for 20 min. Lysates were centrifuged at 16,000*g* for 20 min, and the supernatant was collected for western blot analysis. For analysis of histone modifications, cells were washed with 1×PBS and resuspended in 0.4 M HCl overnight at 4 °C. Cells were then pelleted by centrifugation at 16,000*g* for 30 min at 4 °C, and the supernatant containing histones was precipitated on ice with 33% trichloroacetic acid for at least 30 min, followed by centrifugation at 16,000*g* for 20 min at 4 °C. The pellets were washed twice with ice-cold acetone and then dissolved in double-distilled water. The Bradford protein assay was used to determine protein concentration, and primary antibodies included: anti-BAP1 (Proteintech, 10398-1-AP; dilution, 1:500), anti-H2AK119ub1 (Cell Signaling Technology, 8240; dilution: 1:2,000), anti-H3 (Abcam, ab1791; dilution 1:80,000) and anti-GAPDH (ABclonal, AC001; dilution: 1:100,000).

The quantification of the levels of ubH2A in wild-type R1 and BAP1 mutant rescue cell lines was derived from scanned intensities of western blots. The grey values of ubH2A and H3 were collected by ImageJ v.2.1.0. H3 is a loading control. [ubH2A]/[H3] is used to represent the ubH2A level, and the value is normalized relative to that of the wild-type R1. The values were analysed using Microsoft Excel and plotted using GraphPad Prism v.9.0.0. Bar charts represent plots of mean ± s.e.m. ($n = 4$ biologically independent experiments). *P* values denote the results of an unpaired two-tailed *t*-test.

RNA-seq and data analysis

Total RNA was extracted from cultured cells using TRIzol reagent (Thermo Fisher Scientific), according to the manufacturer's instructions. The generation and sequencing of RNA-seq libraries of wild-type and *Bap1* KO mouse ES cells, and *Bap1* KO rescue cell lines (*Bap1* KO + BAP1(R699W); *Bap1* KO + BAP1(R701C); and *Bap1* KO + BAP1(699–702 mut)) were performed by Gene+; and those of the *Bap1* KO rescue cell lines (*Bap1* KO + empty vector; *Bap1* KO + wild-type BAP1; *Bap1* KO + BAP1(C91S); *Bap1* KO + BAP1(56–60 mut); *Bap1* KO + BAP1(699–702 mut); *Bap1* KO + BAP1(1–696); *Bap1* KO + BAP1(H169Q); *Bap1* KO + BAP1(R56C); and *Bap1* KO + BAP1(Y173C)) were performed by ANOROAD. All samples were collected in two replicates.

RNA-seq analysis was performed as previously described[58]. After removing adapters and low-quality bases with a Phred quality score of less than 20 using Trimmomatic (v.0.36)[59], paired-end reads were aligned to the annotated mouse transcripts (mm10 Gencode vM15 release) using Hisat2 (v.2.1.0)[60]. The expression levels of genes were calculated using StringTie (v.1.3.3b)[61]. Read counts were calculated using featureCounts (ref. 62). The R package edgeR (ref. 63) was used for identifying differentially expressed genes, and only those with a read count higher than 50 in at least one sample of a comparison were kept. A given gene was considered differentially expressed if the false discovery rate (FDR) was < 0.05, $P < 0.01$ and fold change ≥ 2.

Statistics and reproducibility

Deubiquitination assays in Figs. 1b and 2i and Extended Data Fig. 5h were repeated three times independently with similar results. Consistent results were obtained for that of Extended Data Fig. 1c from three independent experiments under different conditions. Similar chromatography and electrophoresis results to those shown in Extended Data Fig. 1a,b were obtained in six separate protein purifications. The reconstitution of nucleosome or chromatosome experiments shown in Extended Data Fig. 1e were consistently repeated seven times. The GraFix results shown in Extended Data Fig. 1f were repeated three times independently, all having similar electrophoresis patterns.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The cryo-EM density maps have been deposited in the Electron Microscopy Data Bank (EMDB) with accession numbers EMD-34431 and EMD-34432 for the chromatosomal and nucleosomal complexes, respectively. Constituent and consensus maps for the chromatosomal complex have been deposited with accession numbers EMD-35179, EMD-35180, EMD-35181 and EMD-35182. The atomic coordinates for the structure of PR-DUB bound to the chromatosome have been deposited in the PDB with the accession code 8H1T. The coordinates for the structures of *Drosophila* Caly in complex with DEUBAD of ASX (PDB ID: 6HGC), H2AK15-ubiquitnated NCP in complex with the BRCA1–BARD complex (PDB ID: 7E8I), the H2BK120-ubiqutinated NCP in complex with the SAGA DUB module (PDB ID: 4ZUX), the UCH-L5 complex with DEUBAD of RPN13 and ubiquitin (PDB ID: 4UEL) and the chromatosome with linker histone H1.4 (PDB ID: 7K5Y) were downloaded from the RCSB Protein Data Bank (https://www.rcsb.org). The RNA-seq data have been deposited in the Genome Sequence Archive at the National Genomics Data Center, China National Center for Bioinformation–Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA006013) and are publicly accessible at https://ngdc.cncb.ac.cn/gsa. Two public RNA-seq datasets with the accession numbers GSE162739 and GSE161995 were used in this study. The list of differentially expressed genes (DEGs) defined in a previous study[7] was downloaded from ref. 7 (https://www.sciencedirect.com/science/article/pii/S1097276521005001?via%3Dihub#app2), with the original file name of 1-s2.0-S1097276521005001-mmc3.xlsx. The list of DEGs defined in a previous study[8]. was downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161995), with the original file name of GSE161995_BAP1ff_cnRNA_seq_DESeq2_Table_GEO.txt. A threshold of $P_{adj} < 0.05$ and fold change > 1.5 was used to extract the DEGs. Source data are provided with this paper.

39. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276**, 19–42 (1998).
40. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol.* **304**, 3–19 (1999).
41. Fang, Q. et al. Human cytomegalovirus IE1 protein alters the higher-order chromatin structure by targeting the acidic patch of the nucleosome. *eLife* **5**, e11911 (2016).
42. Dyer, P. N. et al. in *Chromatin and Chromatin Remodeling Enzymes* Part A (eds Allis, C.D. & Wu, C.) 23-44 (2004).
43. Song, F. et al. Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science* **344**, 376–380 (2014).
44. Kastner, B. et al. GraFix: sample preparation for single-particle electron cryomicroscopy. *Nat. Methods* **5**, 53–55 (2008).
45. Schorb, M., Haberbosch, I., Hagen, W. J. H., Schwab, Y. & Mastronarde, D. N. Software tools for automated transmission electron microscopy. *Nat. Methods* **16**, 471–477 (2019).
46. Wu, C., Huang, X., Cheng, J., Zhu, D. & Zhang, X. High-quality, high-throughput cryo-electron microscopy data collection via beam tilt and astigmatism-free beam-image shift. *J. Struct. Biol.* **208**, 107396 (2019).
47. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
48. Asarnow, D., Palovcak, E. & Cheng, Y. UCSF pyem v0.5. *Zenodo* https://doi.org/10.5281/zenodo.3576630 (2019).
49. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
50. Sanchez-Garcia, R. et al. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun. Biol.* **4**, 874 (2021).
51. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
52. Terwilliger, T. C., Ludtke, S. J., Read, R. J., Adams, P. D. & Afonine, P. V. Improvement of cryo-EM maps by density modification. *Nat. Methods* **17**, 923–927 (2020).
53. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
54. Wang, R. Y. et al. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife* **5**, e17219 (2016).
55. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
56. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
57. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
58. Zhang, J. et al.Highly enriched BEND3 prevents the premature activation of bivalent genes during differentiation. *Science* **375**, 1053–1058 (2022).
59. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
60. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
61. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
62. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
63. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

Acknowledgements We thank B. Zhu, X. Huang, L. Chen, X. Li and other staff members at the Center for Biological Imaging of the Institute of Biophysics (IBP), Chinese Academy of Sciences (CAS), for support in cryo-EM data collection; X. Yu at the Protein Science Core Facility of IBP for help with MALS analyses; Y. Gu for participation at an early stage of the study; and J. Chen for logistic support. The study was supported in part by grants from the Chinese Ministry of Science and Technology (2019YFA0508900), the Natural Science Foundation of China (31991162, 92153302 and 91853204), the K. C. Wong Education Foundation (GJTD-2020-06), the Youth Innovation Promotion Association (2018125) and the Strategic Priority Research Program (XDB37010100) of CAS.

Author contributions R.-M.X. & B.Z. designed and supervised the project. G.L. contributed experimental designs and advised on enzymatic activity assays. X.Z. advised on cryo-EM analysis. C.T. advised on the synthesis of ubiquitinated histones. W.G. performed protein expression and purification and cryo-EM sample preparation, analysed the structure and performed biochemical experiments. C.Y. prepared cryo-EM samples, collected data and solved and analysed the structures. J.L. analysed the cellular effects of BAP1 mutations. Z.Y. supervised the cryo-EM sample preparation and fluorescent polarization experiments. X.L. advised on enzymatic activity assays. Y.Z. performed bioinformatic analysis. C.-P.L. helped with structural model building. Y.L. contributed to the analysis of cellular effects of BAP1 mutations. W.G., C.Y. and R.-M.X. wrote the manuscript, and all authors read and edited the manuscript.

Competing interests The authors declare no competing interests.

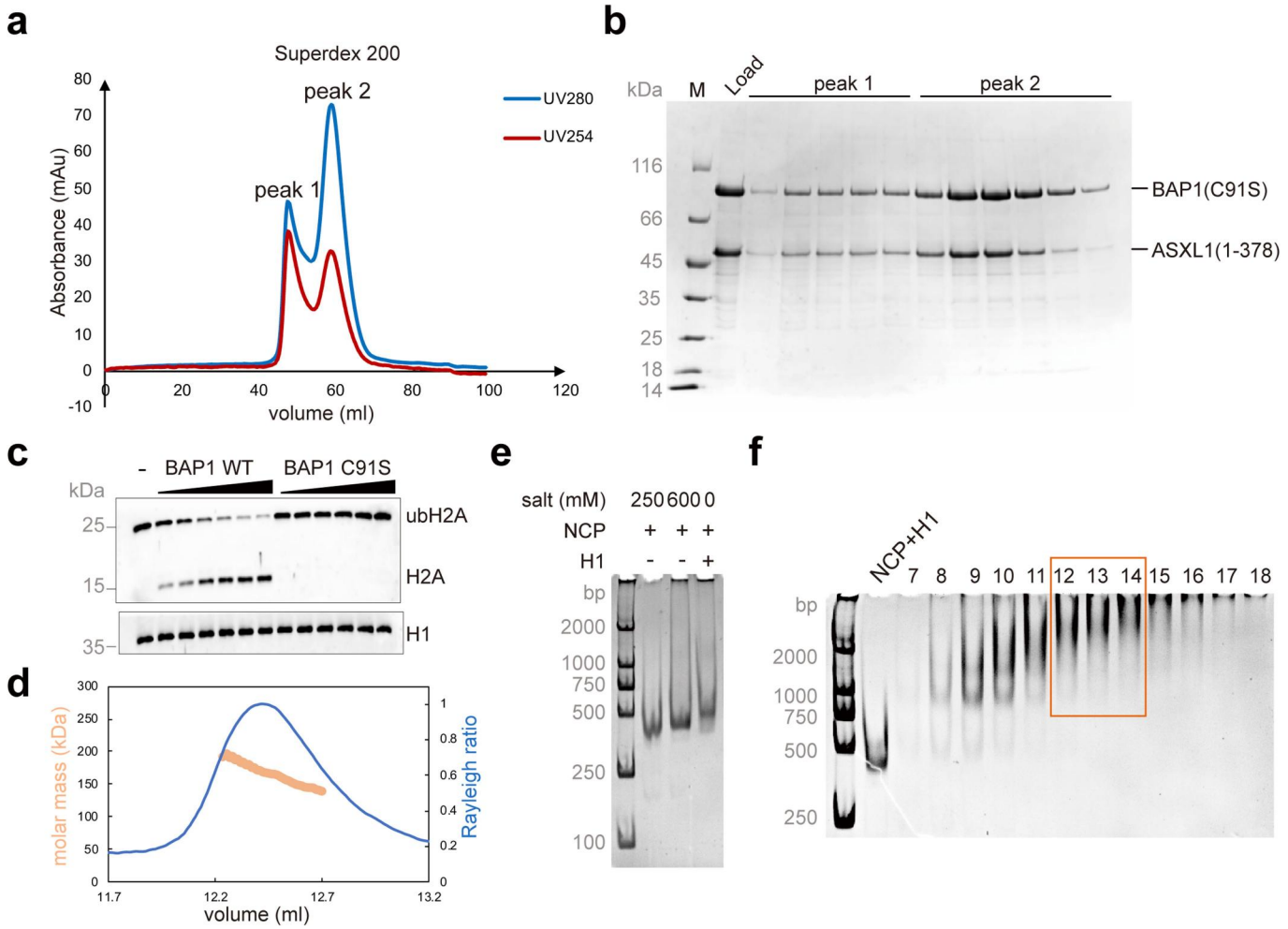
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05841-y.

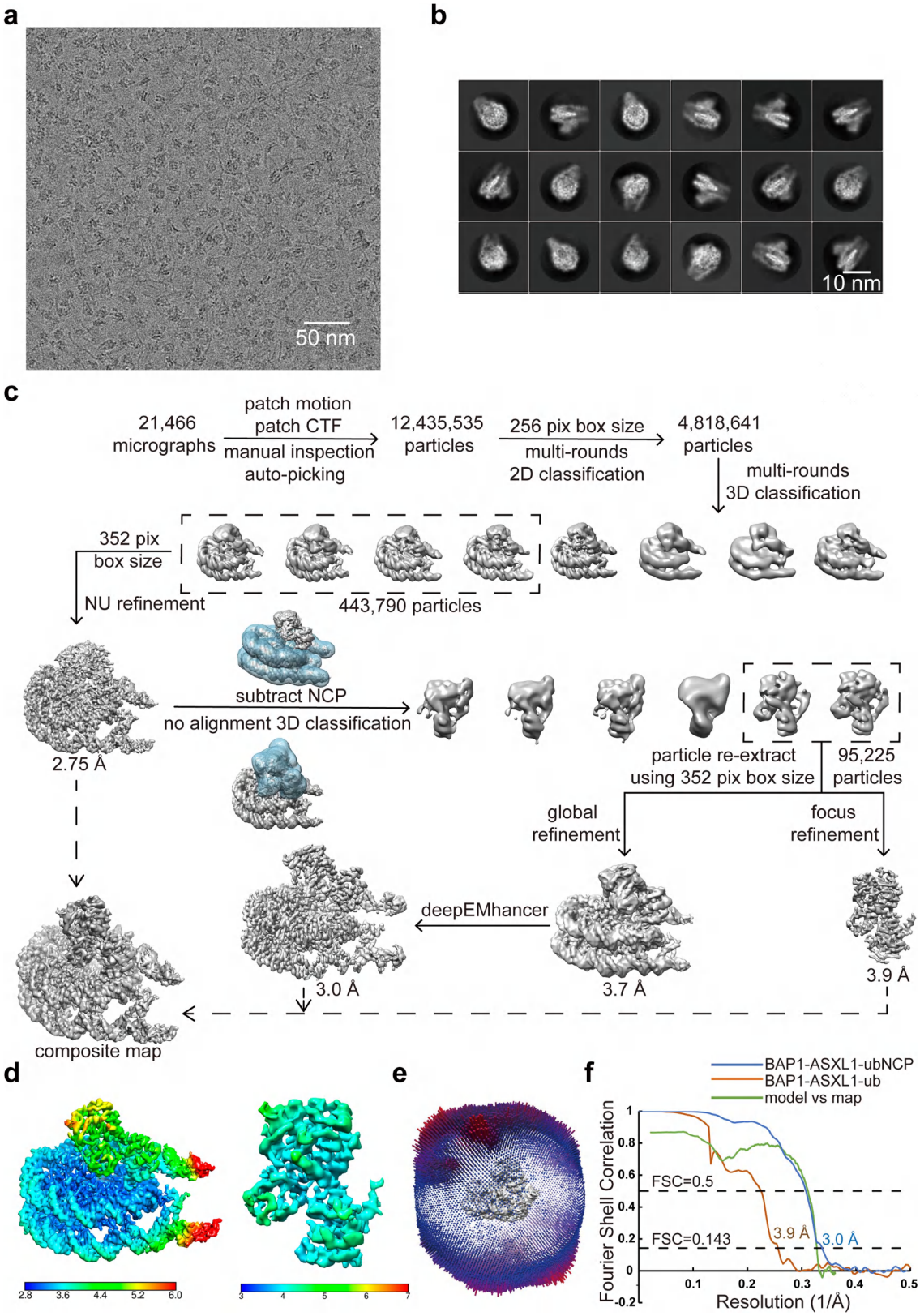
Correspondence and requests for materials should be addressed to Bing Zhu or Rui-Ming Xu.

Peer review information *Nature* thanks Robert Klose and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

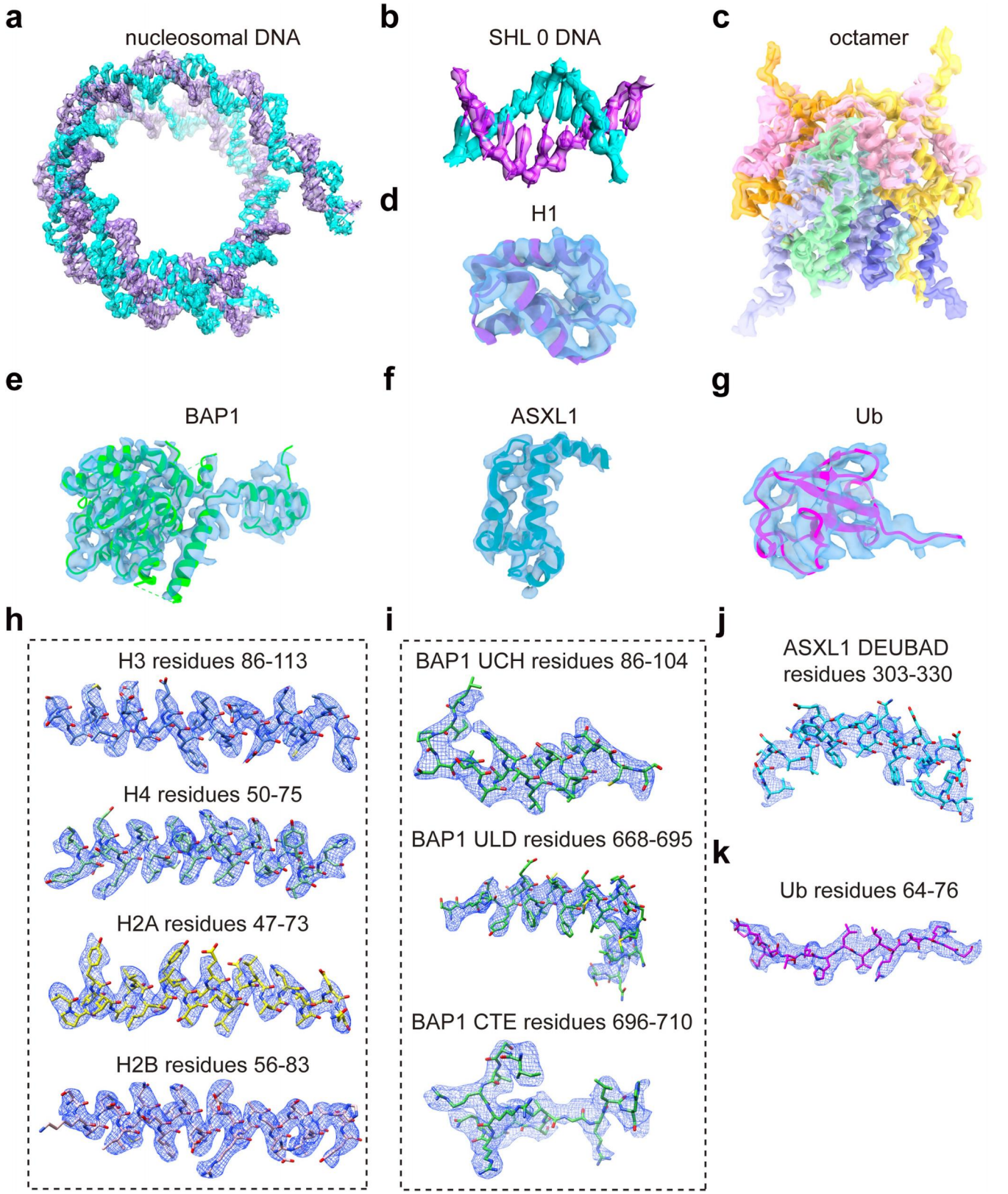
Reprints and permissions information is available at http://www.nature.com/reprints.



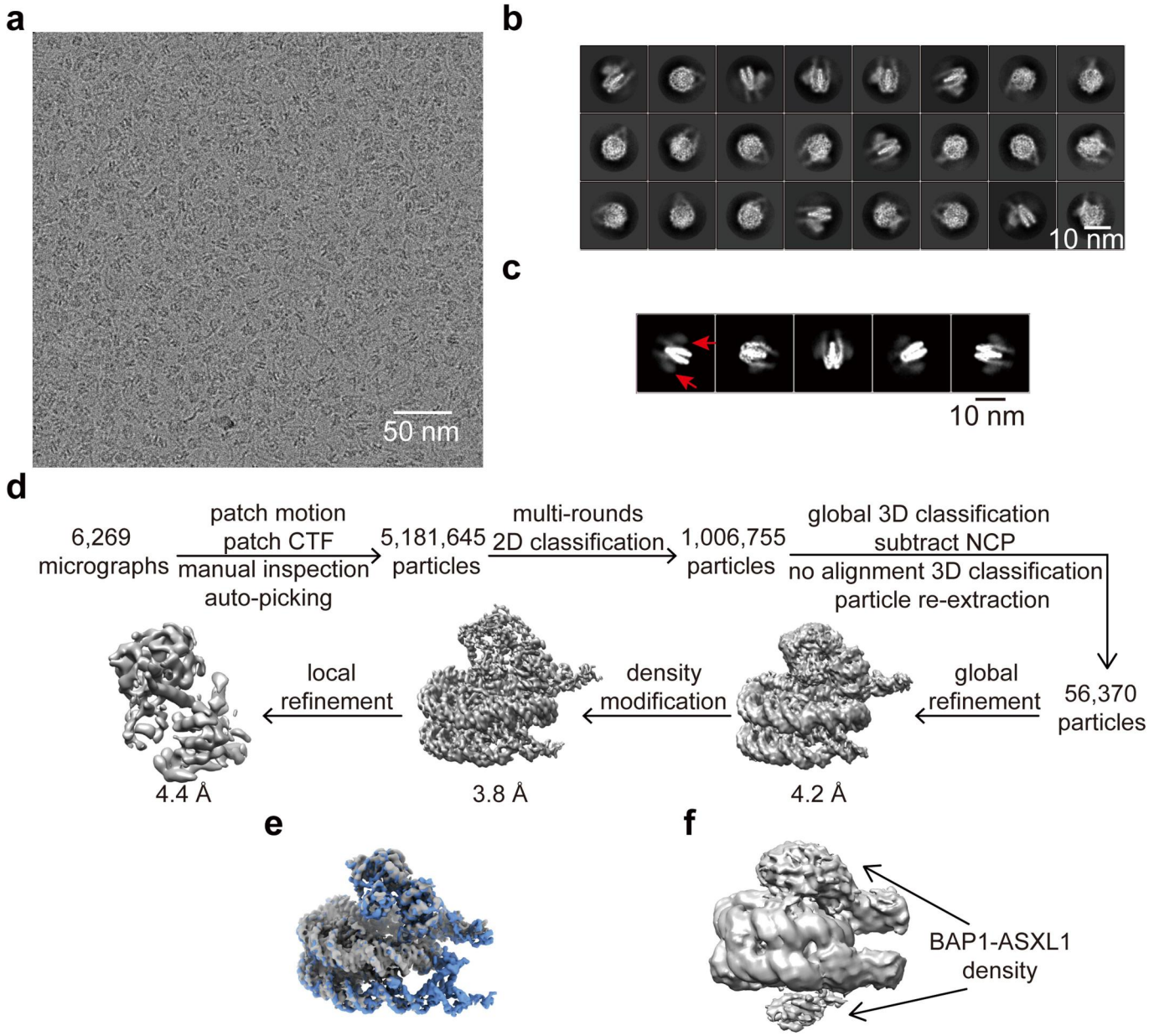
Extended Data Fig. 1 | Cryo-EM sample preparation. a, Elution profile of size-exclusion column chromatography (SEC) of the BAP1–ASXL1(1–378) complex. Peak 1 was eluted close to the void volume of the column. **b**, Coomassie-stained SDS–PAGE analysis of fractions from SEC shown in (a). M, molecular weight marker; values are labelled on the left. Peak 2 fractions were collected for assembling complex with NCP. **c**, Western blot analyses of DUB activities of the wild-type and the C91S mutant BAP1 complexes with ASXL1(1–378). Same amount of the 187-bp DNA H2AK119ub1 chromatosome substrate was used in each reaction. From left to right, lane 1, no enzyme added; lanes 2–7 and lanes 8–13, successive twofold dilutions of the wild-type and C91S BAP1 enzyme complexes, respectively, detected with anti-H2A (top panel) and anti-His tag (bottom panel) antibodies. **d**, SEC-MALS analysis of BAP1 C91S–ASXL1(1–378) complex at the concentration of 2.2 mg/ml. Calculated molar mass = 164∓8.2 kDa, which is far below the 244.2 kDa calculated molecular mass of a dimer of the BAP1–ASXL1(1–378) complex. The data were analysed by the ASTRA software. **e**, GoldView-stained native PAGE shows the reconstituted NCP, NCP in a 600 mM NaCl buffer, at which point H1 was added and dialysed to zero salt to assemble the chromatosome. **f**, GoldView-stained native PAGE of fractions after GraFix. The chromatosome alone was used as a control. Fractions 12–14, highlighted in the red box, were used for cryo-EM sample preparation.



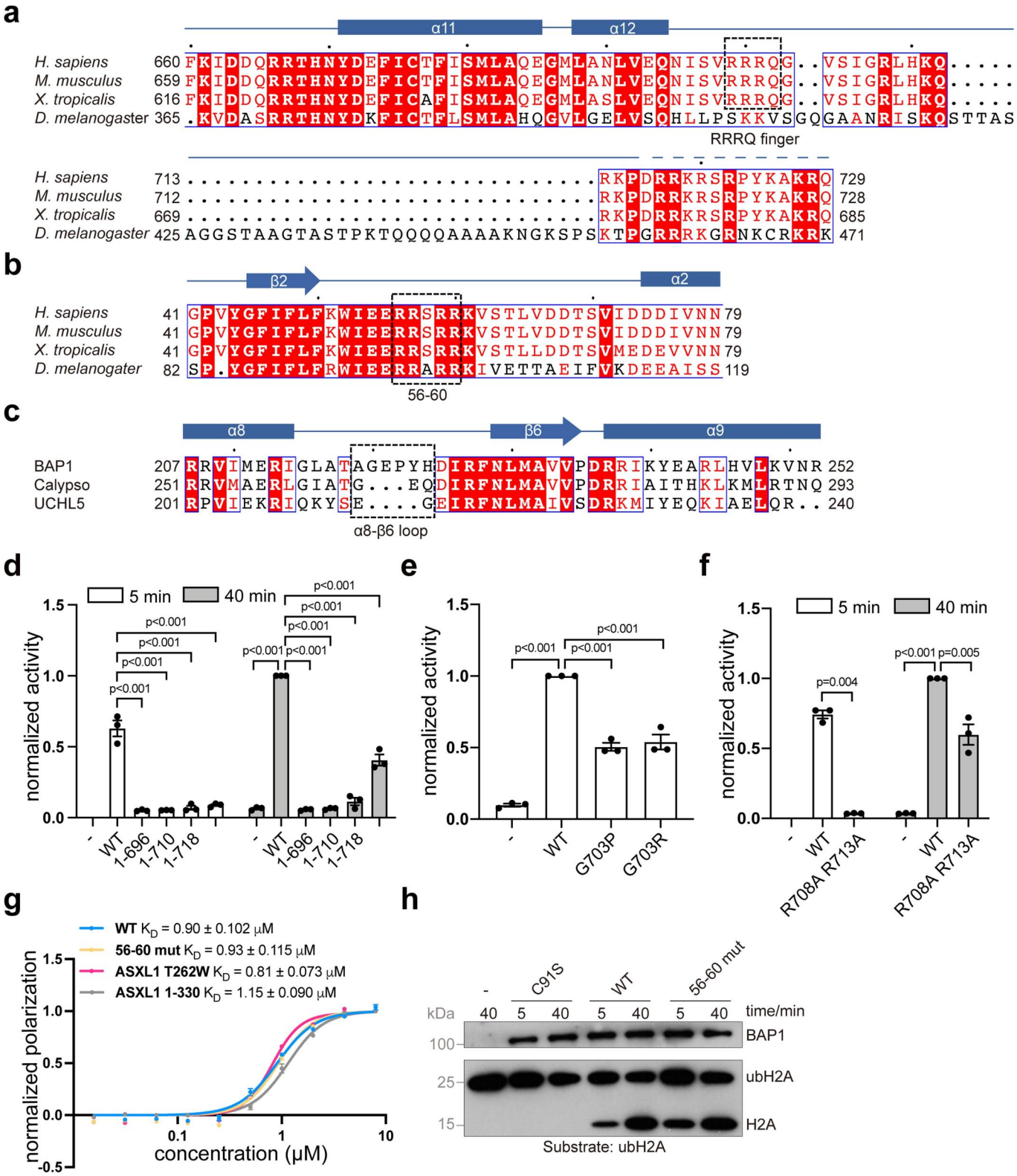
Extended Data Fig. 2 | Structure determination of the BAP1–ASXL1(1–378)–H1–NCP(H2AK119ub1) complex. a, A representative electron micrograph low-pass-filtered to 15 Å from 21,466 micrographs. **b**, Selected reference-free 2D class average. **c**, Workflow of cryo-EM data processing. **d**, The 3.0-Å global refinement map of the full complex (left), and the 3.9-Å local refinement BAP1–ASXL1(1–378)–ubiquitin subcomplex (right). For convenience of presentation, a composite map was constructed by combining indicated global and focused refinement maps. **e**, Angular distribution of the full complex particles in the final reconstructions. **f**, Gold-standard Fourier correlation of the BAP1–ASXL1(1–378)–H1–NCP(H2AK119ub1) complex.



Extended Data Fig. 3 | EM density maps of various components or regions of the PR-DUB–substrate complex. a, 601 nucleosomal DNA. **b**, The SHL 0 region of nucleosomal DNA. **c**, Human histone octamer. Two H3s are coloured in blue and light blue, H4s in pale cyan and pale green, H2As in yellow-orange and bright orange, and H2Bs in pink and light pink. **d**, The globular domain of linker histone H1.4. **e**, Human BAP1. **f**, The DEUBAD domain of ASXL1. **g**, The ubiquitin moiety of H2AK119ub1. **h**, Selected regions of core histones. **i**, BAP1 regions. **j**, The DEUBAD domain of ASXL1. **k**, The C-terminal region of the ubiquitin moiety of H2AK119ub1. Atomic models are shown in a stick representation, and the maps are contoured at 5σ-7σ.

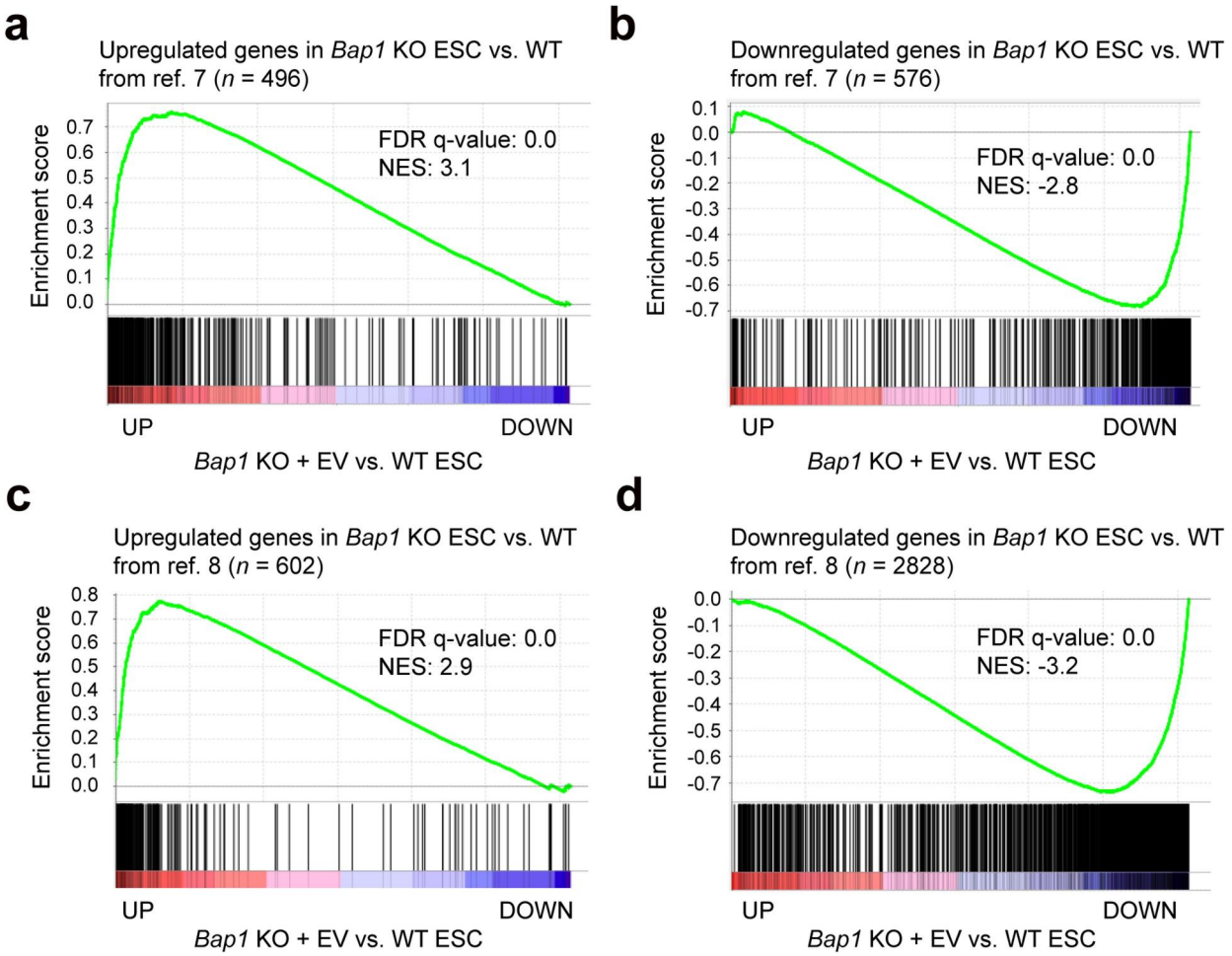


Extended Data Fig. 4 | Cryo-EM analysis of the BAP1–ASXL1(1–378)–NCP(H2AK119ub1) complex without H1. a, A representative electron micrograph low-pass-filtered to 15 Å from 6,269 micrographs. **b**, Selected reference-free 2D class average. **c**, 2D class averages of two BAP1–ASXL1 complexes bound to NCP in higher contrast. The red arrows indicate the density for the bound BAP1–ASXL1 complexes. **d**, Workflow of cryo-EM data processing. **e**, Alignment of the EM densities for the BAP1–ASXL1(1–378)–NCP(H2AK119ub1) complexes with (blue) and without (grey) H1. **f**, 3D classification map of two BAP1–ASXL1(1–378) complexes bound to the nucleosome.

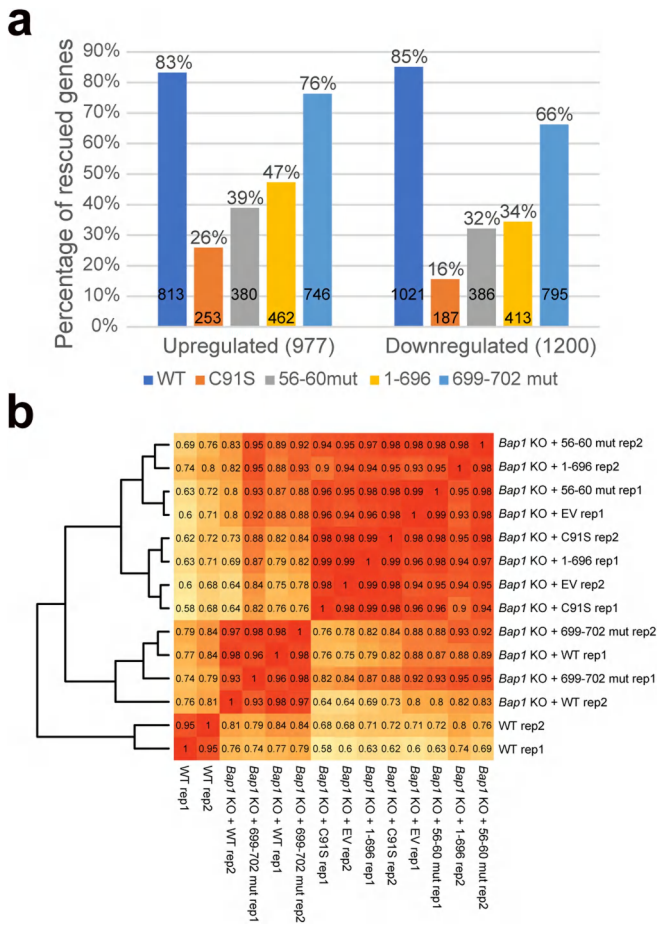


Extended Data Fig. 5 | See next page for caption.

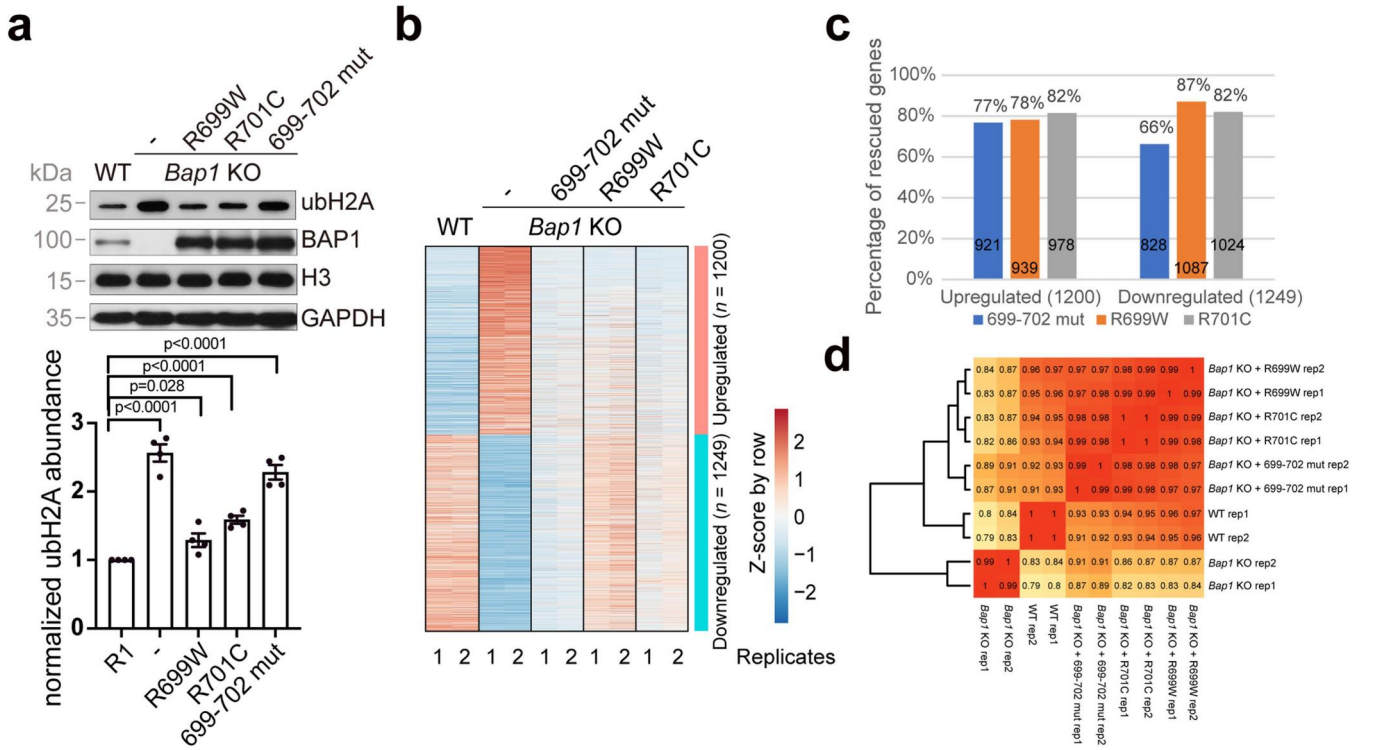
Extended Data Fig. 5 | BAP1 residues in distinct regions involved in H2AK119 deubiquitination and nucleosome binding. a, Sequence alignment of the CTE region of BAP1 orthologues. The RRRQ (aa 699–702) CTE finger motif is boxed in dashed lines. **b**, Alignment of the RRSRR motif (aa 56–60) in the β2–α2 loop of BAP1. **c**, Alignment of the sequences encompassing the α8–β6 loop of BAP1, Caly and UCH-L5. Above the sequence, BAP1 secondary structure elements are depicted, and period signs denote whole 10th residues of BAP1. **d**–**f**, Quantification of catalytic activities of western blot measurements in Figs. 2d, g and h, respectively. The values for indicated wild-type and BAP1 mutant complexes measured at 5 and 40 min in Fig. 2d,h are shown with clear and grey-shaded bars, respectively. The wild-type values (at 40 min) were set to unity, and all values were derived from three replicates, and represented as mean ± s.e.m. *P* values, displayed at the top of the bar charts, denote the results of an unpaired two-tailed *t*-test. **g**, Fluorescence polarization measurements of binding affinities of the BAP1–ASXL1 complexes carrying indicated BAP1 (56–60 mut) or ASXL1 (T262W, 1–330) mutations to the NCP(H2AK119ub1) assembled with 167-bp 601 DNA containing a 5′-FAM fluorophore. The data were analysed in the same way as in Fig. 2e (n = 3 biologically independent experiments; mean ± s.e.m.). **h**, Western blot analysis of free histone H2AK119 deubiquitination by wild-type and indicated BAP1 mutants in complex with ASXL1(1–378). A '–' sign indicates a mock reaction without enzyme added. Reactions were monitored at the indicated time points, and detected with anti-His tag (top) or anti-H2A (bottom) antibodies.



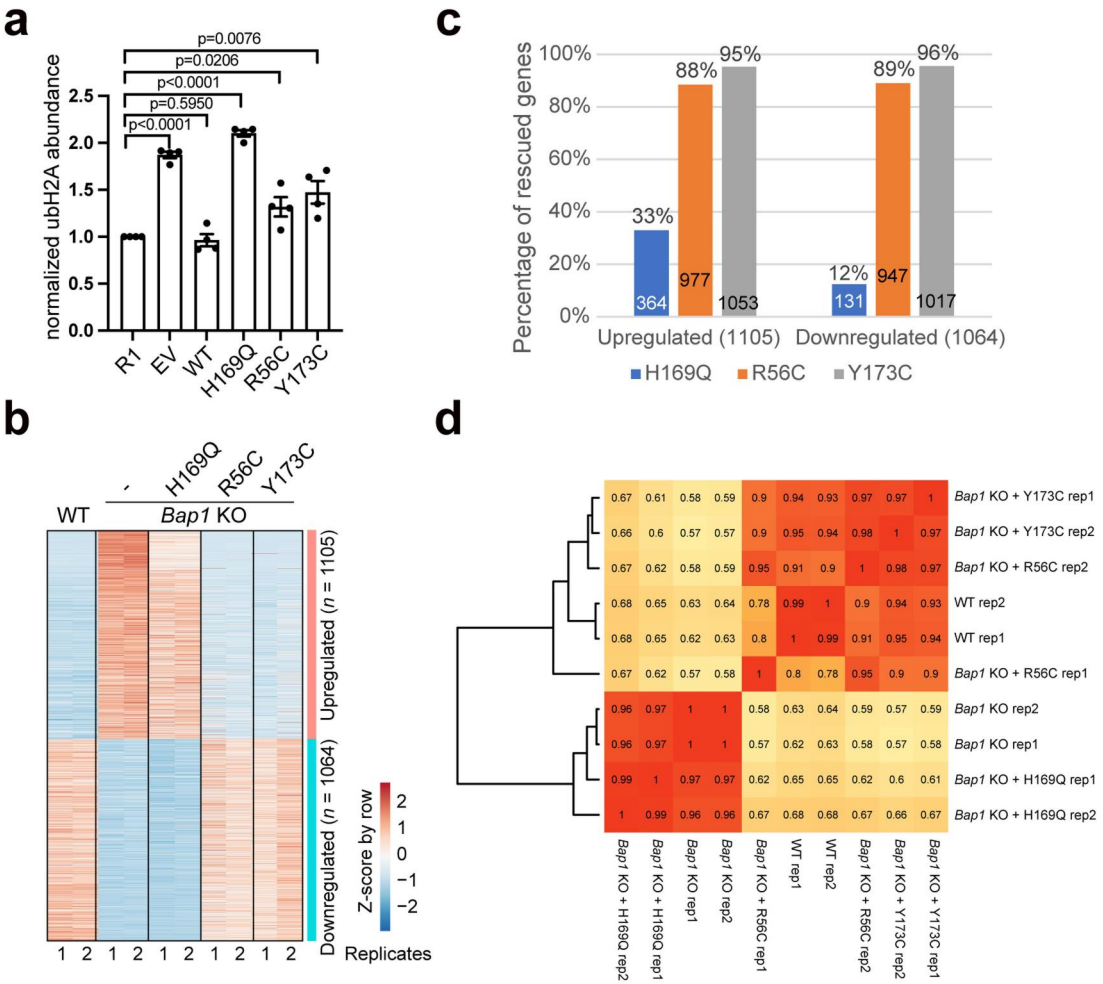
Extended Data Fig. 6 | Gene set enrichment analysis. a,**b**, Upregulated (**a**) and downregulated (**b**) genes in *Bap1* KO ES cells versus wild-type ES cells from a previous study[7] with respect to the global transcriptional changes observed in our *Bap1* KO + EV ES cells versus wild-type ES cells. **c**,**d**, Upregulated (**c**) and downregulated (**d**) genes analysed in a previous study[8] relative to the global transcriptional changes we observed.



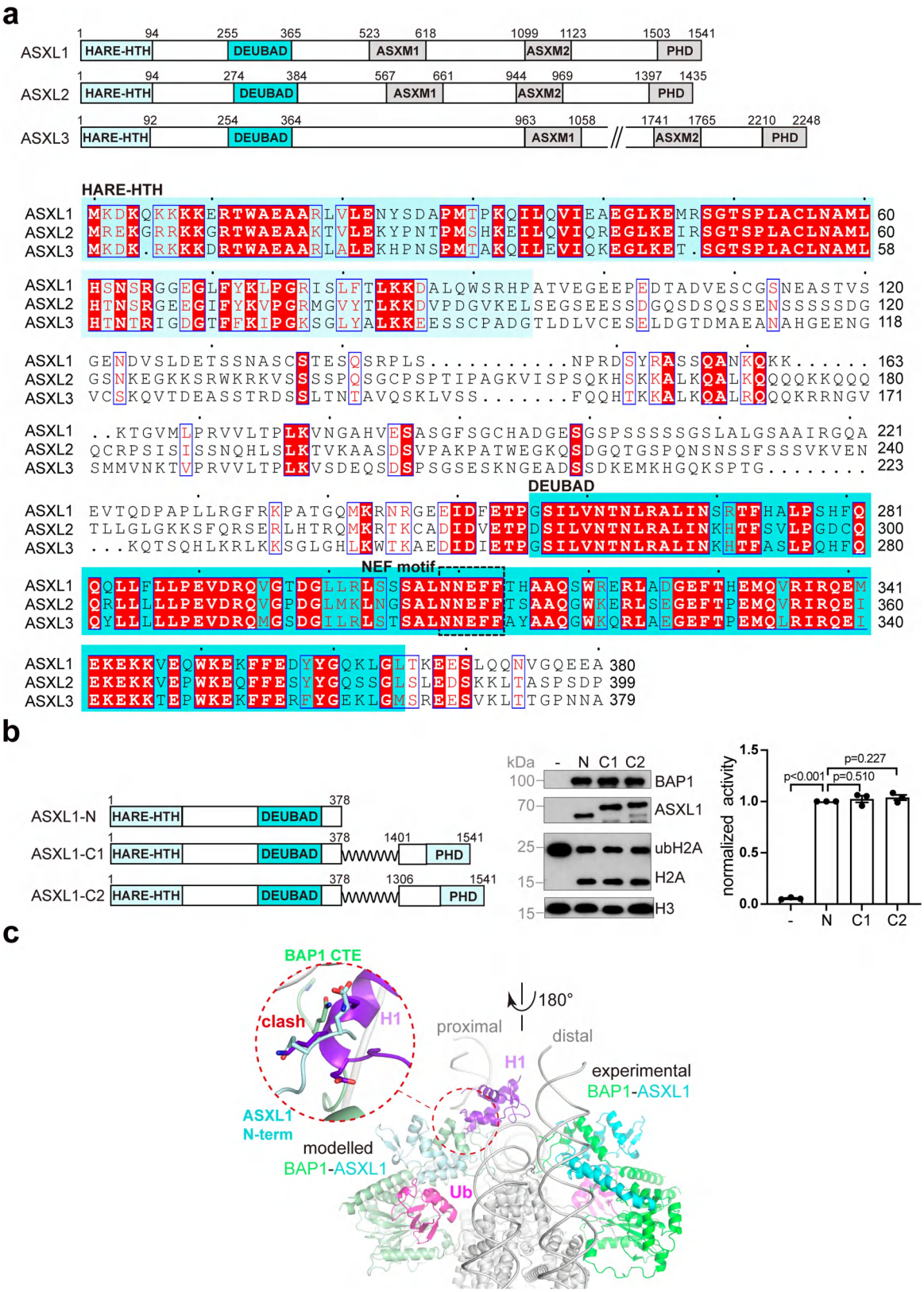
Extended Data Fig. 7 | Gene expression analysis of *Bap1* mutants. **a**, Percentage of rescued BAP1-regulated genes (differentially expressed genes between *Bap1* KO + EV ES cells and wild-type ES cells) in the indicated wild-type and mutant BAP1 rescue cells. **b**, Expression correlation among the 14 RNA-seq samples of mouse ES cells on BAP1-regulated genes (differentially expressed genes between *Bap1* KO + EV and wild-type R1). There are two replicates for each cell line. Pearson correlation coefficients are indicated.



Extended Data Fig. 8 | Cellular DUB activities and changes in gene expression for BAP1 CTE cancer-associated mutations. a, Western blot detection of the levels of ubH2A and BAP1 in *Bap1* KO cells rescued with the indicated BAP1 mutant plasmids. Quantification of the western blot results is represented in a bar chart of mean ± s.e.m. ($n = 4$ biologically independent experiments). *P* values, displayed at the top of the bar chart, denote the results of an unpaired two-tailed *t*-test. **b**, Heat maps of transcriptional levels of BAP1-regulated genes in wild-type and indicated rescue cell lines. **c**, Percentage of rescued BAP1-regulated genes (differentially expressed genes between *Bap1* KO and wild-type ES cells) in the indicated mutant BAP1 rescue cells. **d**, Expression correlation among the 10 RNA-seq samples.



Extended Data Fig. 9 | Quantification of cellular DUB activities and RNA-seq analysis for non-CTE BAP1 cancer-associated mutations. a, Quantification of western blot detection of H2AK119ub and BAP1, as shown in Fig. 3h, in *Bap1* KO cells transfected with an empty vector (EV), or plasmids re-expressing wild-type or the indicated BAP1 mutants. Bar charts represent plots of mean ± s.e.m. ($n = 4$ biologically independent experiments). *P* values, displayed at the top of the bar chart, denote the results of an unpaired two-tailed *t*-test. **b**, Heat maps of transcriptional levels of BAP1-regulated genes in wild-type and indicated rescue cell lines. There are two replicates for each cell line. **c**, Percentage of rescued BAP1-regulated genes (differentially expressed genes between *Bap1* KO and wild-type ES cells) in each indicated mutant BAP1 rescue cells. **d**, Expression correlation among the 10 RNA-seq samples.



Extended Data Fig. 10 | Structure and function of ASXL1 domains and the effect of H1 on H2AK119 deubiquitination. a, Domain structures of ASXL1–ASXL3 and alignment of their N-terminal sequences encompassing HARE-HTH and DEUBAD domains. **b**, Left, schematic representation of ASXL1 constructs used in the deubiquitinase assay. Wavy lines represent a 2×(GGGGS) linker. Middle, western blot analysis of nucleosomal H2AK119 deubiquitination of BAP1 complexes with the indicated ASXL1 fragments. H3 lanes are sample-processing controls. Right, quantification of catalytic activities based on western blot measurements ($n = 3$ biologically independent experiments, mean ± s.e.m.). *P* values, displayed at the top of the bar chart, denote the results of an unpaired two-tailed *t*-test. No adjustment was made for multiple comparisons. **c**, A BAP1–ASXL1(1–378) complex modelled to the proximal DNA side (left) of the chromatosome based on the observed binding mode in the distal DNA side (right) sterically clashes with H1 (see inset).

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

Samples	BAP1-ASXL1-H1-NCP^{K119ub}	BAP1-ASXL1-NCP^{K119ub}
Data collection and processing		
Microscope	Talos Arctica	Titan Krios
Detector	Gatan Bio-Quantum K2	Gatan Bio-Quantum K2
Magnification	130,000	130,000
Voltage (kV)	200	300
Electron exposure (e^-/Å^2)	50	60
Defocus range (μm)	-1.5 to -2.0	-1.5 to -2.0
Pixel size (Å)	0.5	0.52
Symmetry imposed	C1	C1
Initial particle images (no.)	12,435,535	5,181,645
Final particle images (no.)	95,225	56,370
Map resolution (Å)	3.0 (global) 3.9 (local)	3.8 (global) 4.4 (local)
FSC threshold	0.143	0.143
Refinement		
Initial model used (PDB code)	7k5y/6hgc/4uel	
Model resolution range (Å)	2.8 - 6	
Map sharpening B factor (Å^2)	-65	
Model composition		
Nonhydrogen atoms	17400	
Protein residues	1323	
Nucleotides	334	
B factors (Å^2)		
Protein	110.95	
Nucleotides	149.03	
R.m.s. deviations		
Bond lengths (Å)	0.003	
Bond angles (°)	0.512	
Validation		
MolProbity score	1.53	
Clashscore	11.0	
Poor rotamers (%)	0	
Ramachandran plot		
Favored (%)	97.76	
Allowed (%)	2.24	
Disallowed (%)	0	

nature portfolio

Corresponding author(s): Rui-Ming Xu, Bing Zhu

Last updated by author(s): Jan 27, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SerialEM 3.8 Grey values of bands in western blots were collected by ImageJ 2.1.0. Data of FP experiments were collected by EnVision Manager 1.13.3009.1409. Data of SEC-MALS was collected by ASTRA 5.3.4.20. RNA-seq data was collected by Illumina NovaSeq 6000 and DNBSEQ-T7 Flow cytometry data was collected by BD FACS Aria III
Data analysis	CryoSPARC v3.2 RELION-3.0.8 UCSF pyem v0.5 DeepEMhancer Rosetta 3.13 Coot 0.8.9 PHENIX 1.18.2 Pymol 1.7 UCSF Chimera 1.13 ChimeraX 1.4 MolProbity 4.3 Quantifications of western blots were analyzed by Excel and plotted by GraphPad Prism 9.0.0. Data of FP was firstly calculated by Microsoft Excel, then analyzed and fitted to specific binding model with Hill slope by GraphPad Prism 9.0.0. Data of SEC-MALS was analyzed by ASTRA 5.3.4.20 and plotted by Microsoft Excel. Trimmomatic v0.36 Hisat2 v2.1.0 StringTie v1.3.3.b

featureCounts v1.6.3
edgeR v3.38.4
GSEA v4.0.3
R v3.4.0
Flow cytometry data was analyzed using FlowJo v10.8.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The cryo-EM density maps have been deposited in the Electron Microscopy Data Bank (EMDB) with accession numbers 34431, 34432, 35179, 35180, 35181 and 35182. The atomic coordinates have been deposited in the Protein Data Bank (PDB) under the accession code 8H1T.

The coordinates for the structures of Drosophila CALY in complex with DEUBAD of ASX (PDB id: 6HGC), H2AK15-ubiquitnated NCP in complex with BRCA1-BARD complex (PDB id: 7E8I), the H2BK120-ubiqutinated NCP in complex with the SAGA DUB module (PDB id: 4ZUX), the UCH-L5 complex with DEUBAD of RPN13 and ubiquitin (PDB id: 4UEL), and the chromatosome with linker histone H1.4 (PDB id: 7K5Y) were downloaded from the RCSB Protein Data Bank (https://www.rcsb.org).

The RNA-Seq data have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA006013) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa.

Two public RNA-seq data sets under the accession numbers GSE162739 and GSE161995 were used in this study. The list of differentially expressed genes (DEGs) defined in ref. 7 was downloaded from https://www.sciencedirect.com/science/article/pii/S1097276521005001?via%3Dihub#app2 and the original file name is 1-s2.0-S1097276521005001-mmc3.xlsx. The list of DEGs defined in ref. 8 was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161995 and the original file name is GSE161995_BAP1ff_cnRNA_seq_DESeq2_Table_GEO.txt. A threshold of P-adj < 0.05 and fold change > 1.5 was used to extract the DEGs.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes are reported in the figure legends or the Methods section. For western blots and fluorescence polarization, each assay was conducted at least 3 times independently. No sample-size calculation was performed.
Data exclusions	In Extended Data Fig. 5g, two points of ASXL1 T262W at the concentration of 8 uM were excluded because at this concentration only 2 samples were measured, which is not sufficient for statistical analysis. However, the exclusion does not affect the calculation of the Kd value appreciably, as these data points are located in the already plateaued region.
Replication	All replications were successful, and replication numbers are either described in the figure legends or the Methods section. Protein purifications were conducted at least 6 times and showed similar chromatography and electrophoresis patterns. Reconstruction of nucleosome and chromatosome were repeated at least for 6 times and gave similar results. GraFix was repeated 3 times independently and showed similar electrophoresis patterns. For RNA-seq, two replications showed similar results.
Randomization	For cryo-EM structure determination, all particles were randomly split into two half sets during 3D refinement. For all other experiments, all data were used for analysis, therefore no randomization was needed.
Blinding	RNA-seq data were generated by commercial vendors who did not have the knowledge about the cell lines being sequenced. Blinding does not apply to other experiments, as no groups were assigned.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	anti-BAP1, Proteintech, 10398-1-AP, western blot (1:500) anti-BAP1, Cell Signaling Technology, 13271s, western blot (1:1000) anti-H2AK119ub1, Cell Signaling Technology, 8240, western blot (1:2000) anti-H3, Abcam, ab1791, western blot (1:80k) anti-GAPDH, ABclonal, AC001, western blot (1:100k) anti-H2A, Millipore, 07-146, western blot (1:5000) anti-H2A, Cell Signaling Technology, 12349s, western blot (1:2000) anti-H2B, Abcam, ab1970, western blot (1:5000) anti-His tag, Invitrogen, MA1-21315, western blot (1:1000) anti-GST, Beyotime, AF0174, western blot (1:4000)
Validation	Commercially available products, validations are available on manufacturers' websites. anti-BAP1, Proteintech, 10398-1-AP: https://www.ptgcn.com/products/BAP1-Antibody-10398-1-AP.htm anti-BAP1, Cell Signaling Technology, 13271s: https://www.cellsignal.com/products/primary-antibodies/bap1-d7w7o-rabbit-mab/13271 anti-H2AK119ub1, Cell Signaling Technology, 8240: https://www.cellsignal.cn/products/primary-antibodies/ubiquityl-histone-h2a-lys119-d27c4-xp-rabbit-mab/8240?site-search-type=Products&N=4294956287&Ntt=8240&fromPage=plp&_requestid=4288435 anti-H3, Abcam, ab1791: https://www.abcam.cn/histone-h3-antibody-nuclear-marker-and-chip-grade-ab1791.html anti-GAPDH, ABclonal, AC001: https://abclonal.com/catalog-antibodies/GAPDHRabbitpAb/AC001 anti-H2A, Millipore, 07-146: https://www.sigmaaldrich.cn/CN/zh/product/mm/07146 anti-H2A, Cell Signaling Technology, 12349s: https://www.cellsignal.cn/products/primary-antibodies/histone-h2a-d6o3a-rabbit-mab/12349?site-search-type=Products&N=4294956287&Ntt=12349s&fromPage=plp&_requestid=4289536 anti-H2B, Abcam, ab1970: https://www.abcam.cn/prolactinprl-antibody-me-122-ab1970.html anti-His tag, Invitrogen, MA1-21315: https://www.thermofisher.cn/cn/zh/antibody/product/6x-His-Tag-Antibody-clone-HIS-H8-Monoclonal/MA1-21315 anti-GST, Beyotime, AF0174: https://beyotime.com/product/af0174.htm

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Mouse ES cell line R1 (ATCC) was a gift from Guohong Li, Institute of Biophysics. Stable Bap1 KO cell line was generated by CRISPR/Cas9. BAP1 WT or mutation rescue cell lines (Bap1 KO + empty vector, Bap1 KO + WT BAP1, Bap1 KO + C91S BAP1, Bap1 KO + 56-60mutant BAP1, Bap1 KO + 699-702mutant BAP1, Bap1 KO + 1-696 BAP1, Bap1 KO + R56C BAP1, Bap1 KO + H169Q BAP1, Bap1 KO + Y173C BAP1, Bap1 KO + R699W BAP1, Bap1 KO + R701C BAP1) were generated by PiggyBac system, Bap1 KO cells were mock transfected or transfected with EF1α-driven vectors encoding BFP-P2A-HA-BAP1 wild type or BAP1 truncation/mutations and another vector encoding transposase using Lipofectamine 3000 (Thermo Fisher Scientific),

	according to manufacturer's instructions. BFP-positive cells were selected by FACS 24 h post-transfection. After further cultured for about seven days, BFP-positive cells within specific BFP signal range were collected as stable BAP1 rescue cells.
Authentication	Mouse ES cell line R1 was frequently checked by the morphological features, but not authenticated. Bap1 KO cell line and BAP1 WT or mutation rescue cell lines were verified by western blot and RNA-seq
Mycoplasma contamination	The cell lines were not tested for mycoplasma contamination but no indication of contamination was observed.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	BAP1 rescue cells were generated by PiggyBac system. Bap1 KO cells were mock transfected or transfected with EF1α-driven vectors encoding BFP-P2A-HA-BAP1 wild type or BAP1 truncation/mutations and another vector encoding transposase using Lipofectamine 3000 (Thermo Fisher Scientific). BFP-positive cells were sorted by FACS 24h post-transfection. After further cultured for about seven days, the cells were collected and sorted by FACS the second time and BFP-positive cells within specific BFP signal range were collected as stable BAP1 rescue cells
Instrument	Cell sorting was done using BD FACS Aria III or BD Influx
Software	BD FACS Diva v8.0.1 software was used for data collection and FlowJo v10.8.1 was used for data analysis
Cell population abundance	For the first FACS sorting, there were about 4-10% BFP-positive cells; For the second FACS sorting, there were about 50% BFP-positive cells. In order to make all BAP1 rescue cells has similar BAP1 expression level, the cells in the same BFP signal range were collected, which were about 25%
Gating strategy	Untransfected cells as negative control to set gating strategy. BAP1 WT or mutations rescue cells were gated using BFP signal

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Structures of the human CST-Polα–primase complex bound to telomere templates

https://doi.org/10.1038/s41586-022-05040-1

Received: 16 December 2021

Accepted: 29 June 2022

Published online: 13 July 2022

Check for updates

Qixiang He[1], Xiuhua Lin[1], Bianca L. Chavez[1], Sourav Agrawal[1], Benjamin L. Lusk[1] & Ci Ji Lim[1]✉

The mammalian DNA polymerase-α–primase (Polα–primase) complex is essential for DNA metabolism, providing the de novo RNA–DNA primer for several DNA replication pathways[1–4] such as lagging-strand synthesis and telomere C-strand fill-in. The physical mechanism underlying how Polα–primase, alone or in partnership with accessory proteins, performs its complicated multistep primer synthesis function is unknown. Here we show that CST, a single-stranded DNA-binding accessory protein complex for Polα–primase, physically organizes the enzyme for efficient primer synthesis. Cryogenic electron microscopy structures of the CST-Polα–primase preinitiation complex (PIC) bound to various types of telomere overhang reveal that template-bound CST partitions the DNA and RNA catalytic centres of Polα–primase into two separate domains and effectively arranges them in RNA–DNA synthesis order. The architecture of the PIC provides a single solution for the multiple structural requirements for the synthesis of RNA–DNA primers by Polα–primase. Several insights into the template-binding specificity of CST, template requirement for assembly of the CST-Polα–primase PIC and activation are also revealed in this study.

Polα–primase initiates DNA synthesis during genome replication[2,5]. The four-subunit Polα–primase (PRIM1–PRIM2–POLA1–POLA2, also known as p49–p58–p180–p70; Fig. 1a) complex does this by using its primase domain (PRIM1–PRIM2) to de novo synthesize a short 7–11-nucleotide (nt) RNA primer, which is then extended by its DNA polymerase domain (POLA1)[2,5]. The enzyme requires modulation by highly conserved single-stranded DNA (ssDNA)-binding proteins (accessory proteins) such as the replication protein A (RPA)[6] and CTC1–STN1–TEN1 (CST; Fig. 1a; ref. [7]) complexes. These two protein complexes share functions[8,9] and structural homology[10].

At mammalian telomeres, CST inhibits telomerase extension of telomere overhangs[11,12] and coordinates the subsequent complementary strand synthesis with Polα–primase. This latter step is known as the telomere C-strand fill-in process and is essential for telomere stability[3,13]. Beyond telomere maintenance, the CST-Polα–primase machinery is also critical for restarting stalled DNA replication forks genome-wide and regulating DNA end resectioning at double-stranded DNA (dsDNA) breaks[11,14–19]. Despite a central and essential role in eukaryotic DNA metabolism, how Polα–primase performs its complicated enzymatic actions and how its accessory proteins modulate each step are unknown[2,5]. In this work we report the cryogenic electron microscopy (cryo-EM) structures of the human CST-Polα–primase preinitiation complex (PIC) bound to three distinct types of telomere overhang, which reveal several key structure–function insights into how the enzyme synthesizes de novo RNA–DNA primer with the aid of CST.

Template length requirements for PIC

To reconstitute the human CST-Pol–α–primase co-complex for structural determination, we purified recombinant CST and Polα–primase complexes from insect cells and combined them stoichiometrically (Fig. 1a,b and Extended Data Fig. 1a). We found that CST and Polα–primase directly interact, regardless of nucleic acids or the intrinsically disordered $POLA1_{1–337}$ domain of Polα–primase that interacts with other accessory proteins[6,20] (Extended Data Fig. 1b). In agreement with our finding, a recent structural study showed that CST-Polα–primase can form a recruitment co-complex in the absence of a template DNA or the disordered amino-terminal region of POLA1 (ref. [21]). Our recombinant Polα–primase is enzymatically active, as tested using telomeric ssDNA templates, and the enzyme can be stimulated by CST (Fig. 1c). However, the stimulation effect by CST was dependent on the length of the telomeric template (Fig. 1d and Extended Data Fig. 1c,d); a six-repeat telomeric template (6×TEL, $(TTAGGG)_6$), but not a three-repeat telomeric template (3×TEL, $(TTAGGG)_3$), resulted in stimulation of Polα–primase by CST. This result indicates that CST stimulation of Polα–primase activity requires a telomeric ssDNA template longer than 18 nt, a dimension that CST, alone, can bind tightly[11,22,23].

Next we investigated whether CST and Polα–primase form a new co-complex structure, with or without the presence of telomeric templates. Negative-stain EM analysis showed that CST-Polα–primase formed a new co-complex structure upon adding telomeric templates (Extended Data Fig. 2a), one that is different from that of CST or Polα–primase alone (Extended Data Fig. 2b). This new structure (hereafter termed PIC, as no nucleotides were added) was formed using telomeric ssDNA templates as short as 15 nt ($(TTAGGG)_2TTA$; Extended Data Fig. 2a). Without template added, the CST-Polα–primase co-complex failed to yield this new PIC structure (Extended Data Fig. 2b).

All of the above findings show that a short telomeric template (≥15 nt) is sufficient to trigger CST and Polα–primase to assemble into a preinitiation complex, but a longer template (>18 nt) is required to stimulate enzyme activity.

[1]Department of Biochemistry, University of Wisconsin–Madison, Madison, WI, USA. ✉e-mail: ciji.lim@wisc.edu

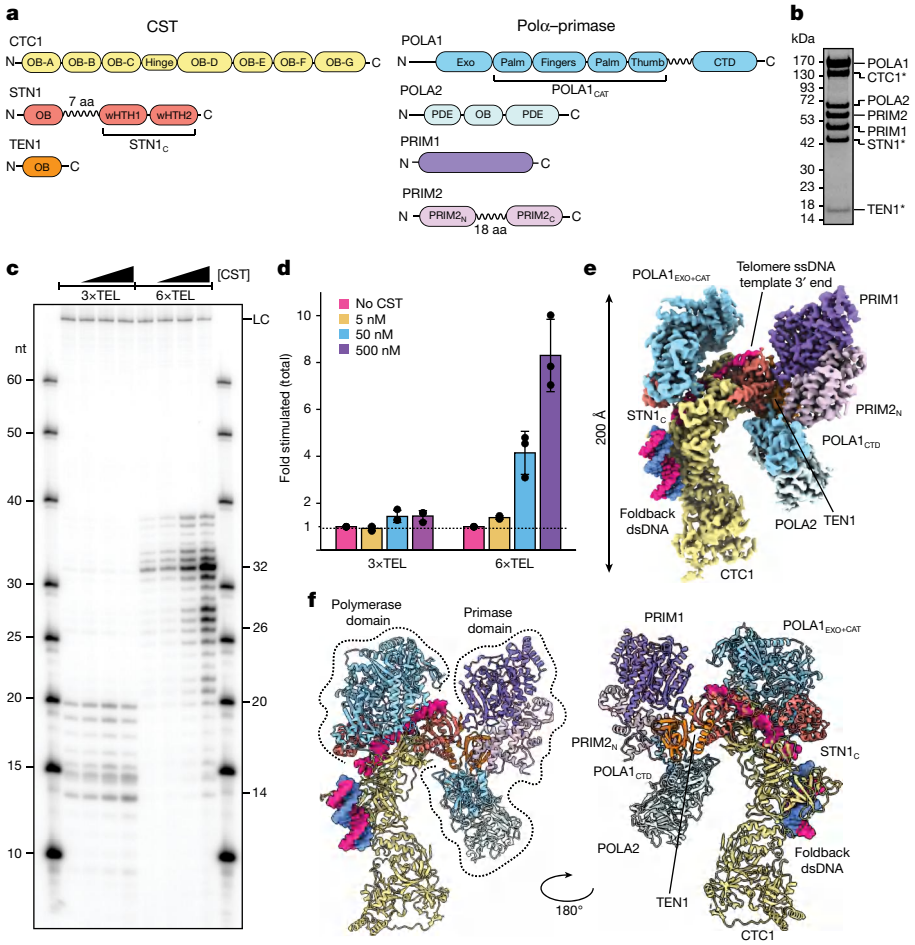


Fig. 1 | The architecture of the human CST-Polα–primase PIC bound to a telomere overhang. a, The structures of the CST and Polα–primase domains. The zigzag lines indicate flexible linkers connecting domains. aa, amino acids. **b**, Coomassie-stained denaturing protein gel of 20 µg of the purified reconstituted CST-Polα–primase co-complex. The labels marked with asterisks are the CST subunits; all other labels are the Polα–primase subunits. Molecular mass markers are shown on the left. **c**, Direct assay of Polα–primase using telomeric templates with the titration of CST (0, 5, 50 and 500 nM) for enzyme stimulation. The first and last lanes are labelled oligonucleotide ladders. The numbers on the left are oligonucleotide ladder sizes; the numbers on the right are predicted sizes of RNA–DNA primers made by the enzyme using the repetitive telomeric sequence templates (see Extended Data Fig. 1). **d**, Total lane activity quantified from **c**. The bars represent the mean values and the error bars show the standard deviation as calculated from several independent experiments ($n = 3$). Data points from each experiment are represented by circles in each corresponding bar. The dotted line marks the fold stimulation of 1 (no stimulation), and the values were calculated by dividing the raw counts from each relevant lane by the counts without CST added. **e**, The front view of the cryo-EM map of the CST-Polα–primase PIC bound to the 4×TEL-FB template. **f**, A ribbon representation of the atomic model built from the cryo-EM map in **e**. For **e**,**f**, each domain is coloured as in **a**. The foldback dsDNA in **e**,**f** is modelled on the basis of the cryo-EM map in Fig. 4a.

Structure of the human CST-Polα–primase PIC

To investigate the mechanism underlying the assembly of the CST-Polα–primase PIC and the stimulation mechanism, we used cryo-EM single-particle analysis to solve the cryo-EM structures of CST-Polα–primase bound to three different telomere templates: 3×TEL, 4×TEL-FB (a 15-base-pair foldback stem-loop preceding a $(TTAGGG)_4$ 3′ overhang) and 6×TEL (Supplementary Table 1). The three structures were solved at a global resolution (reported at a Fourier shell correlation (FSC) level of 0.143) of 3.6 Å, 3.4 Å and 3.7 Å for the 3×TEL, 4×TEL-FB and 6×TEL co-complexes, respectively (Supplementary Figs. 1–6 and Supplementary Table 2).

The three cryo-EM maps shared the same PIC structure, with map-to-map FSC values of about 3.7–3.8 Å (Extended Data Fig. 3), consistent with our negative-stain EM analysis (Extended Data Fig. 2a). The overall structural similarity of the cryo-EM maps shows that CST-Polα–primase forms the same PIC structure regardless of the telomeric template length or structure (single-stranded or overhang). As the three co-complexes share the same consensus structure, hereafter, unless otherwise stated, we used the 4×TEL-FB co-complex as the representative model to describe our findings (Fig. 1e).

The cryo-EM map resolution allowed us to dock most, if not all, of the seven subunits of CST and Polα–primase (from AlphaFold models[24] and past structural studies[2,10,25,26]), and de novo build the protein-bound DNA template, thus revealing the overall architecture of the human CST-Polα–primase PIC structure (Fig. 1e,f).

Polymerase and primase separated by CST

From the cryo-EM structure, we gained several key insights into the architecture of the PIC. The PIC has a 1:1:1 molar stoichiometry of CST/Polα–primase/template (Fig. 1e,f). Interestingly, the template-bound CST caused Polα–primase to undergo a large conformational change in the PIC structure, changing the enzyme from a compact shape to a segregated form (Fig. 1f and Extended Data Fig. 4). This conformational change led to a separation of the RNA[26] and DNA[27] catalytic centres of Polα–primase, exposing the DNA catalytic centre of POLA1 that was buried in the apo state of the enzyme[2,25] (Extended Data Fig. 4).

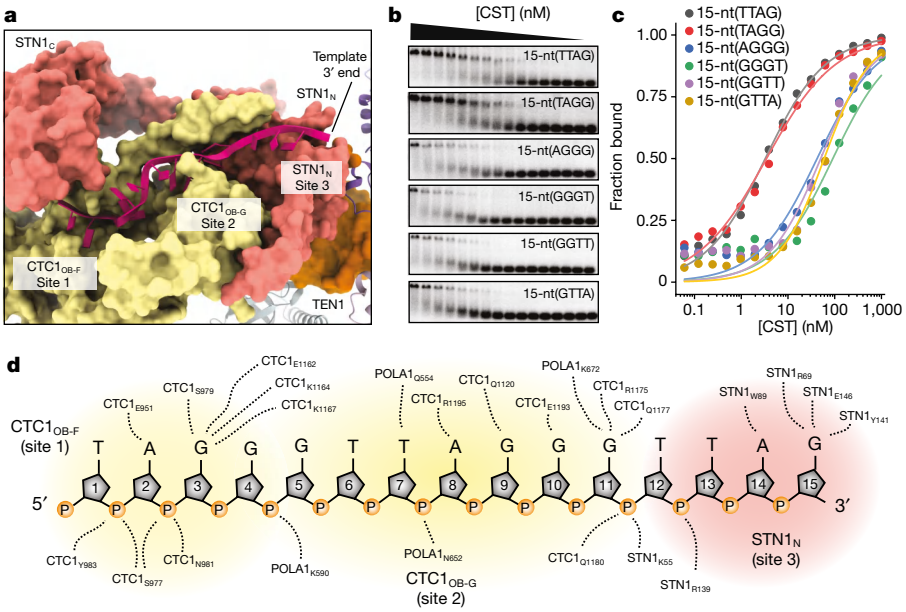


Fig. 2 | The molecular and structural basis for the binding preference of CST to telomeric sequences. a, Template bound by the CTC1 and STN1 subunits of CST in the PIC structure. CTC1, STN1 and TEN1 are depicted as surfaces; the rest of the PIC structure is represented as ribbon diagrams. The 15-nt modelled template, coloured deep pink and illustrated as filled nucleotides, is bound by three binding sites spanning the $CTC1_{OB\text{-}F}$, $CTC1_{OB\text{-}G}$ and $STN1_N$ domains. These binding sites are labelled sites 1, 2 and 3, respectively. **b**, Gel electrophoresis mobility shift assays to determine the binding affinity of CST to six different permutations of a 15-nt telomeric sequence. The highest CST concentration used was 1 μM and a twofold serial dilution was used for the protein titration. **c**, Binding analysis and curve fitting of the electrophoresis mobility shift assay data in **b**. The circles are the data points for the fraction bound; the lines are fitted curves. Two other independent experiments are shown in Extended Data Fig. 7c. **d**, Protein–DNA interactions as shown in **a**.

The template-bound CST serves as a scaffold to mediate the conformational change of Polα–primase, and this is achieved through two spatially separated interaction sites between CST and Polα–primase; $STN1_N$–TEN1 binds the primase domain (PRIM1–PRIM2–$POLA1_{CTD}$–POLA2), and CTC1 and $STN1_C$ support the polymerase domain ($POLA1_{EXO}$–$POLA1_{CAT}$; Fig. 1f). TEN1 seemingly makes contacts with three of the four (PRIM1, $PRIM2_N$ and POLA2) subunits of the primase domain, with the $PRIM1_{84–101}$ loop sitting in the interface between $STN1_N$ and TEN1 (Fig. 1f and Extended Data Fig. 5a–d).

The template-bound CST interacts with the polymerase domain through several contact points (Fig. 1f): the antiparallel β-strands of $CTC1_{OB\text{-}G}$ interact with the flexible loop between residues 547 and 556 of $POLA1_{EXO}$ (Extended Data Fig. 5e); the template interacts with $POLA1_{EXO}$, whereby $POLA1_{EXO}$ uses an α-helical region and other nearby charged residues to contact the ssDNA template, as seen in yeast Pola1 (ref. [28]; Extended Data Fig. 5f); and $STN1_C$ interacts with $POLA1_{EXO}$ (Extended Data Fig. 5g)—coulombic surface analysis reveals a positively charged surface that $POLA1_{EXO}$ potentially uses to interact with a negatively charged surface on $STN1_C$ (Extended Data Fig. 5h,i).

We found that the $PRIM2_N$ conformation is slightly changed in the PIC structure (as compared to its apo-state structure[25]) to accommodate the primase domain conformation, with the large subdomain of $PRIM2_N$ (ref. [29]) rotated about 40° from the small subdomain that interacts with PRIM1 (Extended Data Fig. 6a).

The cryo-EM map densities of the thumb domain of $POLA1_{CAT}$ and $PRIM2_C$ are poorly resolved (Supplementary Fig. 7), indicating substantial flexibility in these domains. This flexibility of the thumb domain of $POLA1_{CAT}$ in the PIC allows the thumb domain to curl up when $POLA1_{CAT}$ binds to a primer–template molecule[27,28,30,31] (Extended Data Fig. 6b). The flexibility of the $PRIM2_C$ domain is consistent with it being flexibly tethered to the rest of the primase domain by an 18-amino-acid linker (Fig. 1a). Below we will further analyse and discuss the PIC structure flexibility and functions (see the section entitled $PRIM2_C$ is poised for primer synthesis).

Molecular basis for telomere binding by CST

Fifteen nucleotides of the 4×TEL-FB template were resolved and modelled in the cryo-EM structure of our PIC (Fig. 2a). The telomeric sequence, 5′-TAGGGTTAGGGTTAG-3′, assigned to the 15-nt template model, was based on the best model-to-map cross-correlation value (Extended Data Fig. 7a,b). This sequence assignment is further supported by DNA-binding experiments (Fig. 2b,c and Extended Data Fig. 7c). The template strand polarity is further confirmed by the fold-back dsDNA density at the template 5′ end in the cryo-EM map (Fig. 1e,f).

Given that the telomeric ssDNA portion of the 4×TEL-FB template is 24 nt, there remains 8 nt of unresolved template at the 3′ end (based on the assigned sequence). This unresolved portion of the template, which extends beyond $STN1_N$ into the primase region, is presumably flexible, either bound or unbound by a protein. In support of this view, $PRIM2_C$, the domain of primase that engages this part of the template during primer synthesis, binds the template poorly[32] and is flexibly linked to the rest of the primase[29]. The $PRIM2_C$ domain is also poorly resolved in the cryo-EM map of our PIC structure (Supplementary Fig. 7).

Our template model extends from the previously determined ssDNA-binding anchor site of CST[10] and reveals how CST uses other binding sites to engage a longer telomeric ssDNA. The bound template adopts an extended conformation that spans 5′ to 3′ from $CTC1_{OB\text{-}F}$ to $STN1_N$. These protein–DNA interactions are further supported by interactions of the $POLA1_{EXO}$ domain (Fig. 2d). The modelled template conformational homogeneity and its contact with the CTC1, $STN1_N$ and $POLA1_{EXO}$ domains (Fig. 2a,d) suggest that this portion of the template is stably bound by CST-Polα–primase.

These protein–ssDNA interactions are clustered across three separate ssDNA-binding sites in $CTC1_{OB\text{-}F/G}$, $CTC1_{OB\text{-}G}$ and $STN1_N$ (Extended Data Fig. 7d). The first binding site in $CTC1_{OB\text{-}F/G}$ shares similar interactions with the previously determined ssDNA-binding anchor site of CST[10]. This binding site utilizes a network of hydrogen bonds to contact the phosphate groups and bases of A2 and G3 of the template (Fig. 2d

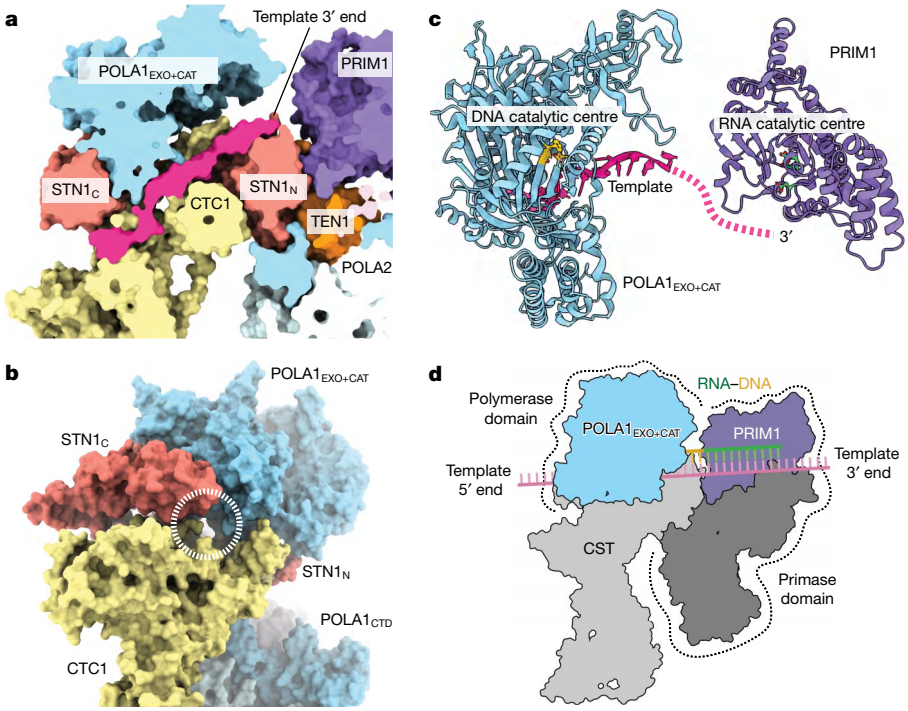


Fig. 3 | CST arranges the catalytic centres of Polα–primase in RNA–DNA synthesis order along the template. a, A cross-sectional view of the molecular tunnel formed by CST and the polymerase domain, with the ssDNA (deep pink) template threading through. **b**, The molecular tunnel viewed from the 5′ end of the template but with the template model hidden to reveal the tight-fitting tunnel. The entrance of the tunnel is marked by a white dashed circle. **c**, CST separation of the polymerase and primase domains of Polα–primase puts the RNA and DNA catalytic centres in the correct RNA–DNA primer synthesis order. Catalytic side chains are shown as green and yellow for the RNA and DNA catalytic centres, respectively. Only the $POLA1_{EXO+CAT}$ domain, PRIM1 subunit and template are shown. The rest of the PIC structure is hidden for clarity. A dashed deep pink line is drawn to illustrate a longer template reaching across PRIM1. **d**, A model illustrating how CST segregates Polα–primase into the polymerase and primase domains (dashed lines) and arranges them along an extended DNA template in the RNA-to-DNA synthesis order.

and Extended Data Fig. 7e,f): Y983, S977 and N981 of CTC1 hydrogen bond with A2-p and G3-p; E951, S979, E1162 and K1167 of CTC1 hydrogen bond with the A2 and G3 bases; and K1164 of CTC1 seemingly forms a cation–π interaction with the G3 base (Extended Data Fig. 7f). Extensive hydrogen bonds are also observed at the second binding site at $CTC1_{OB\text{-}G}$ (Fig. 2d and Extended Data Fig. 7g): E1120, R1175, Q1177, R1193 and R1195 of CTC1 hydrogen bond with the bases of A8, G9, G10 and G11. This binding site is further supported by the POLA1(EXO) domain binding to several template residues in its proximity: N652 of POLA1 hydrogen bonds with A8-p, and Q554 and K672 of POLA1 hydrogen bond with T7 and G11 bases, respectively (Fig. 2d and Extended Data Fig. 7h).

The third binding site in STN1(N) uses a wider variety of interactions to engage the template (Fig. 2d and Extended Data Fig. 7i,j): K55 and R139 of STN1 form salt bridges with T12-p and T13-p, respectively; R69 of STN1 forms a hydrogen bond with the G15 base; and W89 and Y141 form a π–π interaction with the A14 and G15 bases, respectively. Coulombic surface analysis confirmed that the interactions utilized by the first two binding sites of CTC1 are electrostatic (highly positively charged surface), whereas the third binding site in $STN1_N$ is neutral (Extended Data Fig. 7k). Amino acid sequence conservation analysis[33] suggests that most of the PIC residues identified for ssDNA binding are well conserved (Supplementary Fig. 8). As similarly shown for ssDNA molecules bound by POT1 (ref. [34]) or RPA[35], the extended conformation of the template is supported by several DNA base-stacking interactions: G3 stacks with G4, G5 stacks with T6, T7 stacks with A8 and G9, G10 stacks with G11, and T13 stacks with A14 (Fig. 2d).

The function of several of these identified DNA-binding residues has been demonstrated: most of the site 1 residues have been rigorously validated as the ssDNA-binding anchor site of CST[10], the alterations to R1193 and R1195 of CTC1 in site 2 were shown to abolish ssDNA binding by CST[10], and the alterations to W89, R139 and Y141 of STN1 in site 3 reduced the affinity of CST for short telomeric DNA[36].

Collectively, the above protein–DNA analysis provides a structural basis for CST binding to telomeric ssDNA: the modelled 15-nt telomeric template (5′-TAGGGTTAGGGTTAG-3′) explains why CST prefers G-rich sequences when binding short ssDNA molecules[11,22]. The preference of CST for G-rich sequences is defined by the specific and nonspecific interactions between CST and the template, which explain why CST can or cannot tolerate base changes at specific template positions[22]. For example, the T12 and T13 residues of the template do not participate in base stacking or protein–DNA interactions; thus, their base changes can be tolerated[22].

PIC directs the template 3′ end to PRIM1

The polymerase domain sits on top of the template-bound CST (Fig. 1f), and this interaction is mediated by the $POLA1_{EXO}$ binding to the template (see the second binding site in Fig. 2d), CTC1 and $STN1_C$ (Fig. 1f and Extended Data Fig. 5g). Surface representation of the PIC structure revealed that this assembly formed a tight-fit tunnel for the CST-bound template (Fig. 3a), directing the 3′ end of the template towards the RNA catalytic centre of PRIM1 (Fig. 3b, c). This tunnel was also recently reported for the *Tetrahymena* telomere replication holoenzyme[37], suggesting that this molecular tunnel is likely to be a conserved structural feature for CST-Polα–primase in the eukaryotic kingdom.

The function of this molecular tunnel regarding RNA–DNA primer synthesis is unclear, but we can draw some insights from our PIC structure. On the basis of our protein–ssDNA interaction analysis and PIC architecture (Figs. 1f and 2d), the positioning of the polymerase domain on top of the template-bound CST may help stabilize the template binding of CST or direct the template 3′ end towards the RNA catalytic centre of PRIM1 (Fig. 3b, c) or both. Further analysis of the PIC structure flexibility yielded more insights into other possible functions of the tunnel (see the section below entitled POLA1 swivels on top of CST–template).

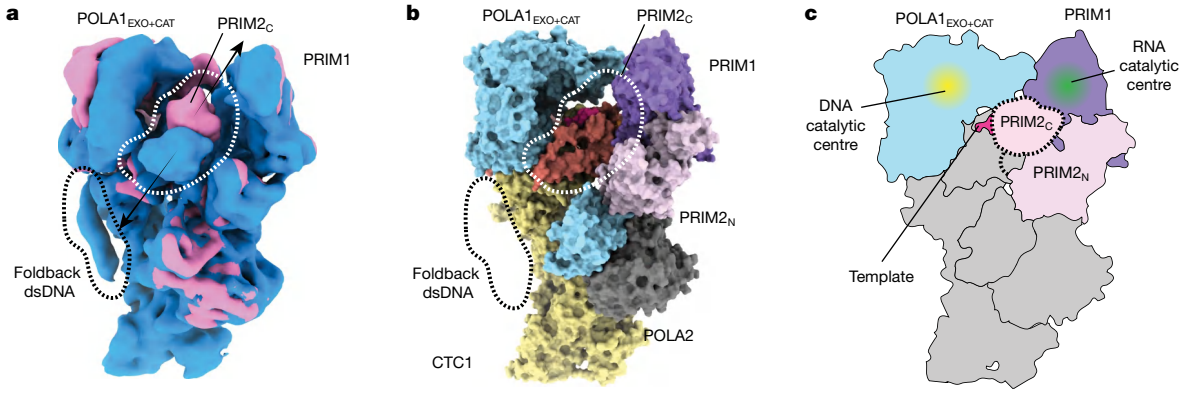


Fig. 4 | PRIM2$_C$ is flexibly poised to perform de novo RNA primer synthesis and facilitate primer handover. a, A 3DVA end-to-end model of the PIC structure revealing flexibility in the EM density of PRIM2$_C$ and the foldback dsDNA. The first and last frames (EM maps) of the principal component analysis mode are shown as blue and pink surfaces. The black dashed line encircles the dsDNA region of motion; the white dashed line represents the boundary of the PRIM2$_C$ domain motion. The black arrows represent the direction of motion of the PRIM2$_C$ domain. These maps are filtered at a nominal resolution of 8 Å. **b**, A direct comparison of the consensus PIC model, depicted as a surface representation, with the flexibility range of the foldback dsDNA and PRIM2$_C$ domain overlaid. **c**, A model of the PIC structure illustrating the proximity of the CST-bound template and the RNA and DNA catalytic centres of Polα–primase to the flexible PRIM2$_C$ domain (dashed outline). The template is coloured deep pink; the RNA and DNA catalytic centres are highlighted with green and yellow spots, respectively. The rest of the PIC is coloured grey.

The PIC structure is not compatible with the ability of CST to form a decameric nucleoprotein complex[10]: the polymerase domain on top of CST would be a steric hindrance for CST to form an inverted dimer, which is the first step of CST oligomerization (Extended Data Fig. 8).

Catalytic centres are in synthesis order

RNA–DNA primers made by Polα–primase are in the order of RNA to DNA synthesis[2,5], and CST stimulates this complicated process at several steps[7,38,39]. Our PIC structure reveals how CST, as an ssDNA-binding accessory protein of Polα–primase, arranges the template and Polα–primase into a conformation that stimulates enzyme activity (Fig. 3c,d). Specifically, CST positions the DNA and RNA catalytic centres of the enzyme in the order of 5′ to 3′ along an extended template.

We propose that this unique arrangement can stimulate Polα–primase activity at several enzymatic steps by increasing the template binding affinity, removing potential template secondary structures and facilitating efficient switching of RNA to DNA synthesis. First, the PIC structure shows that the template is bound mainly by CST (Fig. 2d), which provides the molecular reasoning for: how CST modulates Polα–primase template affinity[39], how CST recruits Polα–primase to telomeres[13], and why DNA binding is necessary for stimulation of Polα–primase activity[40]. Second, the structure also reveals that the ssDNA template ahead of Polα–primase is unfolded (inferred by the extended conformation of the template) by the CST-polymerase domain (Fig. 3a), which would explain how CST-Polα–primase resolves ssDNA secondary structures at stalled DNA replication forks[15,41]. Third, positioning the RNA catalytic centre of PRIM1 downstream of the template after the DNA catalytic centre of POLA1 (Fig. 3d) provides the correct sequence of action for RNA and DNA synthesis. Presumably, this order facilitates an efficient handover of RNA primer–template from PRIM1 to POLA1 for DNA elongation (for further analysis and discussion, see the section below entitled PRIM2$_C$ is poised for primer synthesis).

POLA1 swivels on top of CST–template

Three-dimensional (3D) variability analysis (3DVA)[42] of the cryo-EM structure of our PIC revealed that the polymerase domain swivels on top of the template-bound CST (Extended Data Fig. 9 and Supplementary Video 1), with the pivot point being the interaction between POLA1$_{EXO}$ and CTC1 (Extended Data Fig. 5e). The swivel range is seemingly restricted and controlled by the STN1$_C$ domain; a flexible 7-amino-acid linker connects this domain to the STN1$_N$ domain (Extended Data Fig. 9). This large swivelling motion is consistent with the modest interactions between POLA1$_{EXO}$ of the polymerase domain and its three contact points: CTC1, STN1$_C$ and template (Extended Data Fig. 5e–g). The function of this elaborate mechanism is unclear but may be related to template movement during RNA or DNA synthesis. A speculative but plausible explanation is that this swivelling motion comes from a Brownian ratchet process used by CST-Polα–primase to facilitate template translocation during RNA primer synthesis.

PRIM2$_C$ is poised for primer synthesis

The carboxy-terminal domain of PRIM2, PRIM2$_C$, facilitates primer initiation by binding the 5′-triphosphate of the primer and coordinating the handover of the RNA primer–template to POLA1 (refs. [25,26,32,43]). These functions depend on the domain flexibility of PRIM2$_C$; PRIM2$_C$ is linked to PRIM2$_N$ through an 18-amino-acid linker[2,32]. How the flexibility of the PRIM2$_C$ domain mediates the multiple functions of PRIM2$_C$ is unclear.

3DVA analysis revealed that the PRIM2$_C$ domain is flexible compared to the rest of the rigid PIC structure (Fig. 4a,b and Supplementary Video 2). From the range of flexibility, we found that PRIM2$_C$ can position itself in front of the template, the RNA catalytic centre of PRIM1 or DNA catalytic centre of POLA1. These are key positions where PRIM2$_C$ has to be to perform its multiple functions (Fig. 4b). We thus propose that the PIC structure represents a single architectural platform that CST-Polα–primase uses to meet the multiple conformational requirements for RNA–DNA primer synthesis (Fig. 4c). The PRIM2$_C$ position and flexibility would enable it to find and bind the template that protrudes from the CST-polymerase domain tunnel and initiate RNA primer synthesis with the nearby PRIM1 (as shown in the next section). After maturation of the RNA primer, the PRIM2$_C$-bound primer–template requires only a simple tilt for the conveniently positioned POLA1$_{CAT}$ to engage the 3′ end of the primer.

Template-bound primase structure in PIC

One of the PRIM2$_C$ positions in our 3DVA analysis showed PRIM2$_C$ in contact with the RNA catalytic centre of PRIM1 (Fig. 4a, pink-coloured

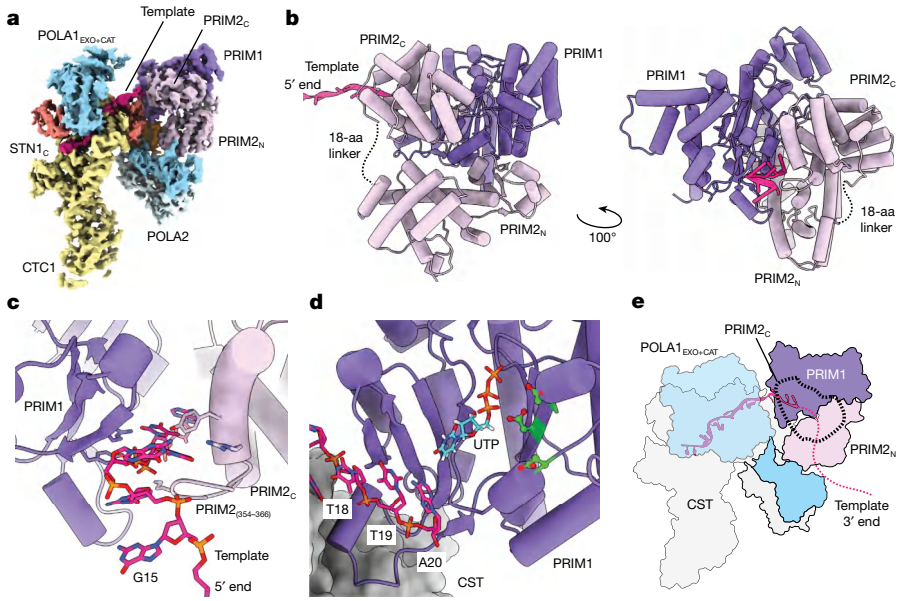


Fig. 5 | The human primase structure that is poised for RNA primer synthesis. a, The cryo-EM structure of a PIC structure trapped in an advanced preinitiation state, as compared to the consensus PIC structure in Fig. 1e. The PRIM2$_C$ domain density is sufficiently well resolved to be docked with its atomic model. **b**, A model of the primase structure poised to initiate RNA primer synthesis. The template (deep pink) is shown as a ladder of nucleotides; the PRIM1 and PRIM2 subunits are depicted as ribbon–tube helix representations. The rest of the PIC structure is hidden for visual clarity. **c**, Template entry into the PRIM1–PRIM2$_C$ crevice. The side chains of protein residues involved in template interactions are shown. The template and side chains are coloured by heteroatom. **d**, The template reaches in front of RNA catalytic residues of PRIM1 (side chains shown in green and coloured by heteroatom). The 3′-end residue, A20, of the ssDNA template is in position for base pairing with a UTP RNA molecule bound by the RNA catalytic centre of PRIM1. The UTP molecule is from the 4BPW model in the Protein Data Bank[26] and is superimposed on our structure to illustrate its proximity to our template model. **e**, A model illustrating how the template is bound by CST-Polα–primase in an advanced preinitiation state poised for NTP incorporation. The 5′-end region of the template is stably bound by the CST-polymerase domain (the polymerase domain is on top of the template and thus is shown semi-transparent for visual clarity) and its 3′ end is engaged by the primase (PRIM1–PRIM2$_C$) for primer initiation. The template makes a sharp turn on contacting the primase. The black dashed circle depicts the PRIM2$_C$ position. The deep pink dotted line represents the predicted extended pathway of a longer template. The POLA1$_{CTD}$ domain is drawn as a solid sky-blue shape.

density, and Supplementary Video 1), suggesting a state in which PRIM2$_C$ could bind both the unresolved 3′ end of the template and PRIM1—an advanced PIC state in which the primase is poised to incorporate the first set of NTPs.

Thus, using further 3D heterogeneity analysis[44,45], we attempted to filter out a conformational state in which PRIM1 and PRIM2$_C$ are in contact. The analysis yielded a subset of particles, leading to a 4.3-Å (FSC = 01.43) cryo-EM map (Fig. 5a and Extended Data Fig. 10a). This map has the PRIM2$_C$ density sufficiently well resolved for us to dock in a PRIM2$_C$ atomic model. Structural studies have shown that the PRIM2$_C$ domain exists as two structural variants: an all-helical variant form[46] and a β-sheet-containing version[47]. The all-helical variant is proposed to be the variant that is more capable of binding ssDNA[46]. We found that the all-helical variant docks well into our PRIM2$_C$ density (Extended Data Fig. 10b), revealing the human primase structure that is poised to initiate RNA primer synthesis (Fig. 5b).

In this advanced preinitiation state, PRIM2$_C$ binds both PRIM1 and template (Fig. 5b,c and Extended Data Fig. 10c,d). The binding of PRIM2$_C$ stabilizes the previously unresolved 3′ portion of the template (Fig. 2) and allows us to extend the template by a further 5 nt (5′-GGTTA-3′; Extended Data Fig. 10e). We now see that the template 3′ end reaches into the RNA catalytic crevice that is formed by PRIM1 and PRIM2$_C$ (Extended Data Fig. 10f). The template entry into the crevice is seemingly guided by several protein–DNA interactions that are contributed by both proteins (Fig. 5c and Extended Data Fig. 10d), with many of these participating residues well conserved (Supplementary Fig. 9). In particular, the identified R56 residue of PRIM1 has been shown to be critical for primase RNA primer synthesis[26], but its underlying mechanism is unknown. Our structure reveals that the PRIM1$_{R56}$ plays an essential role in template binding and positioning the template for RNA primer initiation.

The template extension into the PRIM1–PRIM2$_C$ crevice revealed that the A20 residue of the template (the last modelled nucleotide) is positioned in front of the RNA catalytic centre of PRIM1 (Fig. 5d) and is thus poised for PRIM1 to catalyse the formation of the first NTP of de novo RNA primer synthesis. This extended template path connects seamlessly with the template path seen in a previously solved structure of a template-bound PRIM2$_C$ (ref. [25]; Extended Data Fig. 10g). By combining our PIC structure and that structure, we can visualize how a more prolonged template traces along the molecular surface of a CST-Polα–primase preinitiation complex (Fig. 5e); the template makes a sharp turn on entering the PRIM1–PRIM2$_C$ crevice and is guided to exit the PRIM1 catalytic centre through its interaction with PRIM2$_C$.

The flexibility of the PRIM2$_C$ domain (Fig. 4a and Supplementary Video 2) suggests that this primase preinitiation structure is a transient one, which is consistent with the rate-limiting steps of primase RNA primer synthesis[32,48]. As the template threads through the crevice formed between PRIM1 and PRIM2$_C$ (Fig. 5e and Extended Data Fig. 10f), it is likely that PRIM1 or PRIM2$_C$ first binds the template before interacting with the other protein to form the observed structure.

The 4×TEL-FB template is nearly sufficient in length (about 24 nt of telomere overhang) to reach into the RNA catalytic centre of PRIM1–PRIM2$_C$ (Fig. 5d). In other words, for CST-Polα–primase PIC to be productive, it requires a template of at least this length. A shorter template, such as 3×TEL, would not be productive as it will be too short to reach the primase after CST binding. The 3×TEL template can still trigger CST-Polα–primase to form the PIC (because 3×TEL is still long enough

for CST to bind; Fig. 2), just not a productive one (Fig. 1c,d). Longer templates, such as 6×TEL, will result in the formation of a PIC that is productive (Fig. 1d).

Discussion

Our work provides the structural basis to explain how the ssDNA-binding accessory protein complex CST dictates the specificity of Polα–primase for the template and organizes the enzyme to perform its complicated multistep RNA–DNA primer synthesis during telomere replication. CST achieves this by driving Polα–primase to form a preinitiation complex structurally poised to stimulate the enzyme activity at several levels. This co-complex structure also provides an elegantly simple solution to the complex changing needs of the biphasic Polα–primase functions. Given the structural similarity of CST to RPA[6,10], it will be interesting to determine whether RPA shares the same stimulatory mechanism.

The CST-Polα–primase PIC structure ties in several previous studies and places them in a coherent framework that can be used to generate new structure-guided hypotheses and questions about the downstream catalytic mechanisms of the enzyme. How the swivelling action of the polymerase domain participates in template translocation during RNA primer synthesis and how the primase determines the length of the RNA primer and performs primer handover are directions for future research. Notably, our PIC structure is compatible with the archaeal primase models[49,50]. Future structure–function studies involving further enzymatically advanced co-complexes will yield answers to this next set of questions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05040-1.

1. Attali, I., Botchan, M. R. & Berger, J. M. Structural mechanisms for replicating DNA in eukaryotes. *Annu. Rev. Biochem.* **90**, 77–106 (2021).
2. Baranovskiy, A. G. & Tahirov, T. H. Elaborated action of the human primosome. *Genes* **8**, 62 (2017).
3. Lim, C. & Cech, T. R. Shaping human telomeres: from shelterin and CST complexes to telomeric chromatin organization. *Nat. Rev. Mol. Cell Biol.* **22**, 283–298 (2021).
4. Bonnell, E., Pasquier, E. & Wellinger, R. J. Telomere replication: solving multiple end replication problems. *Front. Cell Dev. Biol.* **9**, 668171 (2021).
5. Pellegrini, L. The Pol alpha-primase complex. *Subcell. Biochem.* **62**, 157–169 (2012).
6. Braun, K. A., Lao, Y., He, Z., Ingles, C. J. & Wold, M. S. Role of protein-protein interactions in the function of replication protein A (RPA): RPA modulates the activity of DNA polymerase alpha by multiple mechanisms. *Biochemistry* **36**, 8443–8454 (1997).
7. Casteel, D. E. et al. A DNA polymerase-α·primase cofactor with homology to replication protein A-32 regulates DNA replication in mammalian cells. *J. Biol. Chem.* **284**, 5807–5818 (2009).
8. Miyake, Y. et al. RPA-like mammalian Ctc1-Stn1-Ten1 complex binds to single-stranded DNA and protects telomeres independently of the Pot1 pathway. *Mol. Cell* **36**, 193–206 (2009).
9. Surovtseva, Y. V. et al. Conserved telomere maintenance component 1 interacts with STN1 and maintains chromosome ends in higher eukaryotes. *Mol. Cell* **36**, 207–218 (2009).
10. Lim, C. et al. The structure of human CST reveals a decameric assembly bound to telomeric DNA. *Science* **368**, 1081–1085 (2020).
11. Chen, L. Y., Redon, S. & Lingner, J. The human CST complex is a terminator of telomerase activity. *Nature* **488**, 540–544 (2012).
12. Zaug, A. J. et al. CST does not evict elongating telomerase but prevents initiation by ssDNA binding. *Nucleic Acids Res.* **49**, 11653–11665 (2021).
13. Kelich, J. M., Papaioannou, H. & Skordalakes, E. Pol alpha-primase dependent nuclear localization of the mammalian CST complex. *Commun. Biol.* **4**, 349 (2021).
14. Wang, F. et al. Human CST has independent functions during telomere duplex replication and C-strand fill-in. *Cell Rep.* **2**, 1096–1103 (2012).
15. Stewart, J. A. et al. Human CST promotes telomere duplex replication and general replication restart after fork stalling. *EMBO J.* **31**, 3537–3549 (2012).
16. Mirman, Z. et al. 53BP1–RIF1–shieldin counteracts DSB resection through CST- and Polα-dependent fill-in. *Nature* **560**, 112–116 (2018).
17. Gu, P. et al. CTC1-STN1 coordinates G- and C-strand synthesis to regulate telomere length. *Aging Cell* **17**, e12783 (2018).
18. Gu, P. & Chang, S. Functional characterization of human CTC1 mutations reveals novel mechanisms responsible for the pathogenesis of the telomere disease Coats plus. *Aging Cell* **12**, 1100–1109 (2013).
19. Chen, L. Y., Majerska, J. & Lingner, J. Molecular basis of telomere syndrome caused by CTC1 mutations. *Genes Dev.* **27**, 2099–2108 (2013).
20. Zhu, W. et al. Mcm10 and And-1/CTF4 recruit DNA polymerase alpha to chromatin for initiation of DNA replication. *Genes Dev.* **21**, 2288–2299 (2007).
21. Cai, S. W. et al. Cryo-EM structure of the human CST-Polalpha/primase complex in a recruitment state. *Nat. Struct. Mol. Biol.* https://doi.org/10.1038/s41594-022-00766-y (2022).
22. Hom, R. A. & Wuttke, D. S. Human CST prefers G-rich but not necessarily telomeric sequences. *Biochemistry* **56**, 4210–4218 (2017).
23. Bhattacharjee, A., Wang, Y., Diao, J. & Price, C. M. Dynamic DNA binding, junction recognition and G4 melting activity underlie the telomeric and genome-wide roles of human CST. *Nucleic Acids Res.* **45**, 12311–12324 (2017).
24. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
25. Baranovskiy, A. G. et al. Mechanism of concerted RNA-DNA primer synthesis by the human primosome. *J. Biol. Chem.* **291**, 10006–10020 (2016).
26. Kilkenny, M. L., Longo, M. A., Perera, R. L. & Pellegrini, L. Structures of human primase reveal design of nucleotide elongation site and mode of Pol alpha tethering. *Proc. Natl Acad. Sci. USA* **110**, 15961–15966 (2013).
27. Coloma, J., Johnson, R. E., Prakash, L., Prakash, S. & Aggarwal, A. K. Human DNA polymerase alpha in binary complex with a DNA:DNA template-primer. *Sci. Rep.* **6**, 23784 (2016).
28. Perera, R. L. et al. Mechanism for priming DNA synthesis by yeast DNA polymerase alpha. *Elife* **2**, e00482 (2013).
29. Baranovskiy, A. G. et al. Crystal structure of the human primase. *J. Biol. Chem.* **290**, 5635–5646 (2015).
30. Baranovskiy, A. G. et al. Activity and fidelity of human DNA polymerase alpha depend on primer structure. *J. Biol. Chem.* **293**, 6824–6843 (2018).
31. Baranovskiy, A. G. et al. Structural basis for inhibition of DNA replication by aphidicolin. *Nucleic Acids Res.* **42**, 14013–14021 (2014).
32. Baranovskiy, A. G. et al. Insight into the human DNA primase interaction with template-primer. *J. Biol. Chem.* **291**, 4793–4802 (2016).
33. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).
34. Lei, M., Podell, E. R. & Cech, T. R. Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. *Nat. Struct. Mol. Biol.* **11**, 1223–1229 (2004).
35. Fan, J. & Pavletich, N. P. Structure and conformational change of a replication protein A heterotrimer bound to ssDNA. *Genes Dev.* **26**, 2337–2347 (2012).
36. Bhattacharjee, A., Stewart, J., Chaiken, M. & Price, C. M. STN1 OB fold mutation alters DNA binding and affects selective aspects of CST function. *PLoS Genet.* **12**, e1006342 (2016).
37. He, Y., Song, H., Chan, H. et al. Structure of *Tetrahymena* telomerase-bound CST with polymerase α-primase. *Nature* https://doi.org/10.1038/s41586-022-04931-7 (2022).
38. Goulian, M., Heard, C. J. & Grimm, S. L. Purification and properties of an accessory protein for DNA polymerase alpha/primase. *J. Biol. Chem.* **265**, 13221–13230 (1990).
39. Goulian, M. & Heard, C. J. The mechanism of action of an accessory protein for DNA polymerase alpha/primase. *J. Biol. Chem.* **265**, 13231–13239 (1990).
40. Zaug, A.J., Goodrich, K.J., Song, J.J. et al. Reconstitution of a telomeric replicon organized by CST. *Nature* https://doi.org/10.1038/s41586-022-04930-8 (2022).
41. Zhang, M. et al. Mammalian CST averts replication failure by preventing G-quadruplex accumulation. *Nucleic Acids Res.* **47**, 5243–5259 (2019).
42. Punjani, A. & Fleet, D. J. 3D variability analysis: resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702 (2021).
43. Zerbe, L. K. & Kuchta, R. D. The p58 subunit of human DNA primase is important for primer initiation, elongation, and counting. *Biochemistry* **41**, 4891–4900 (2002).
44. Scheres, S. H. et al. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nat. Methods* **4**, 27–29 (2007).
45. Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* **17**, 1214–1221 (2020).
46. Agarkar, V. B., Babayeva, N. D., Pavlov, Y. I. & Tahirov, T. H. Crystal structure of the C-terminal domain of human DNA primase large subunit: implications for the mechanism of the primase-polymerase alpha switch. *Cell Cycle* **10**, 926–931 (2011).
47. Vaithiyalingam, S., Warren, E. M., Eichman, B. F. & Chazin, W. J. Insights into eukaryotic DNA priming from the structure and functional interactions of the 4Fe-4S cluster domain of human DNA primase. *Proc. Natl Acad. Sci. USA* **107**, 13684–13689 (2010).
48. Sheaff, R. J. & Kuchta, R. D. Mechanism of calf thymus DNA primase: slow initiation, rapid polymerization, and intelligent termination. *Biochemistry* **32**, 3027–3037 (1993).
49. Yan, J., Holzer, S., Pellegrini, L. & Bell, S. D. An archaeal primase functions as a nanoscale caliper to define primer length. *Proc. Natl Acad. Sci. USA* **115**, 6697–6702 (2018).
50. Greci, M. D., Dooher, J. D. & Bell, S. D. The combined DNA and RNA synthetic capabilities of archaeal DNA primase facilitate primer hand-off to the replicative DNA polymerase. *Nat. Commun.* **13**, 433 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

Methods

Cloning

Human *POLA1* (NP_058633.2) and *POLA1*$_{\Delta 1-337}$ were cloned into the pOET1 transfer vector (Mirus Bio). Human *PRIM1* (NP_000937.1), *PRIM2* (NP_000938.2) and *POLA2* (NP_002680.2) cDNAs were individually cloned into the pFastBac1 expression vector (Invitrogen). The POLA1 and POLA1$_{\Delta 1-337}$ expression vectors have N-terminal twin-Strep-tags whereas those of the POLA2, PRIM1 and PRIM2 constructs have N-terminal hexa-histidine (6×His) tags. The cDNAs of the human CST subunits, *CTC1* (AAI11784), *STN1* (NP_079204) and *TEN1* (NP_001106795) were cloned into a single pBAC4x-1 transfer vector (Novagen). The CTC1 open reading frame has an N-terminal 3×Flag whereas both STN1 and TEN1 have 6×His tags.

Insect cell culture and baculovirus generation

Recombinant baculovirus encoding CST or POLA1 (or its truncation mutant) was produced using the flashBAC ultra system (Mirus Bio). The baculoviruses encoding PRIM1, PRIM2 and POLA2 were made using the Bac-to-Bac system (Invitrogen). All viruses were made using *Spodoptera frugiperda* (SF9) cells (Invitrogen). The viruses were amplified to a titre of >1.0×10^8 pfu ml^{-1} before using them for insect cell infection. The virus titres were measured by a flow cytometry assay (Expression Systems).

Expression and purification of CST and Polα–primase

One or two litres of *Trichoplusia* (Tni) cells (Expression System) were infected with recombinant baculoviruses at a cell density of $1.5–2.0 \times 10^6$ cells per millilitre. A single baculovirus stock was used for CST expression, and four baculoviruses (POLA1, POLA2, PRIM1 and PRIM2) were used to co-express Polα–primase. Each infection was carried out at a multiplicity of infection of 1. The infected Tni cells were further incubated in an orbital shaker (27 °C, 130 r.p.m.) for 66–68 h before collection.

The CST protein complex was purified according to an established protocol[10]. The infected Tni cells were collected by centrifugation at 1,500*g* for 30 min at 4 °C. The cell pellets were resuspended in lysis buffer (25 mM HEPES–NaOH pH 7.5, 300 mM NaCl, 15 mM imidazole, 1 mM dithiothreitol (DTT), 1× Xpert Protease Inhibitor Cocktail Solution (GenDEPOT)) at 50 ml lysis buffer per litre of cells and subjected to ultrasonication for cellular lysis. The cell lysate was clarified by high-speed centrifugation, and the clarified lysate was incubated with Ni-NTA agarose resin (Qiagen) for 2 h at 4 °C using a rotator. The resin was washed three times with wash buffer (same as the lysis buffer but without protease inhibitors) and eluted with Ni-NTA elution buffer (25 mM HEPES–NaOH pH 7.5, 300 mM NaCl, 300 mM imidazole, 1 mM DTT and protease inhibitors). The eluate was then incubated with anti-Flag resin (GenScript) overnight at 4 °C. The resin was washed three times with wash buffer, and the CST complex was eluted with Flag elution buffer (25 mM HEPES–NaOH pH 7.5, 300 mM NaCl, 1 mM DTT, 0.4 mg ml^{-1} 3×Flag peptide (APExBIO)).

The protocol for purification of Polα–primase is the same as that for purification of CST except that the buffers contain 150 mM NaCl instead of 300 mM NaCl and a twin-Strep-tag pulldown was performed instead of a Flag-tag pulldown. The Ni-NTA eluate was incubated with Strep-Tactin XT 4Flow-resin (IBA LifeScience) overnight at 4 °C. The resin was washed three times with wash buffer (25 mM HEPES–NaOH pH 7.5, 150 mM NaCl, 1 mM DTT) before eluting the Polα–primase complex with 1× BXT elution buffer (IBA LifeScience). The purity and integrity of both protein complexes were checked with SDS–polyacrylamide gel electrophoresis (PAGE) analysis before they are concentrated to ≥5.0 mg ml^{-1}. The complexes were used in experiments or snap-frozen in small aliquots for long-term storage at −80 °C.

Co-complex formation

Purified CST and Polα–primase were mixed together at a 2:1 molar ratio and incubated at 4 °C overnight. Strep-Tactin XT 4Flow-resin was added to the mixture and incubated for 1 h at 4 °C. The resin was washed four times with wash buffer (25 mM HEPES–NaOH pH 7.5, 150 mM NaCl, 1 mM DTT) before the co-complex was eluted with 1× BXT elution buffer supplemented with 1 mM $MgCl_2$ (0.5 h incubation). The integrity and purity of the co-complex were confirmed with SDS–PAGE before experimental use or snap-freezing for long-term storage at −80 °C.

Pulldown assay

The protein bait and prey samples were incubated for 1 h at 4 °C before Strep-Tactin XT 4Flow-resin was added to capture the twin-Strep-tagged bait proteins. The resin was washed three times with wash buffer before the captured samples were eluted with 1× BXT elution buffer (incubated for 0.5 h). The pulled-down protein samples were analysed with Coomassie-stained SDS–PAGE.

Agarose gel binding assay

Radioactive oligonucleotides, 5′-labelled with ^{32}P, were used in the binding assays. The estimated specific activity was at least 200,000 c.p.m. pmol^{-1}, and 500 c.p.m. was used for each reaction. For binding, purified protein complexes and DNA were mixed and incubated as a 10 µl reaction volume in binding buffer (25 mM HEPES–Na pH 7.5, 150 mM NaCl, 2 mM $MgCl_2$, 0.2 mM EGTA, 0.1% NP-40, 10% glycerol, 1 mM TCEP) for 2 h on ice. The samples were then loaded into a pre-chilled 1× TBE 0.7% agarose gel and the gel was run at 7 V cm^{-1} for 1.5 h in a cold room (4 °C). The gels were vacuum-dried at 80 °C for 1.5 h before exposure to a storage phosphor screen overnight. The screen was then imaged with a Typhoon FLA 9000 scanner (GE Lifesciences). Binding analysis was carried out using GelAnalyzer software (GelAnalyzer 19.1; www.gelanalyzer.com), and curve fitting was performed with OriginPro software (OriginLab, USA). The Hill binding equation was used for the curve fitting.

Enzyme direct assay

A standard reaction (20 µl) consisted of NTPs, dNTPs, 5 µM DNA template and 50 nM Polα–primase in direct assay buffer (of 50 mM of HEPES–NaOH pH 7.5, 50 mM NaCl, 5 mM $MgCl_2$ and 2 mM DTT). Addition of CST concentrations was as indicated in the figure legends or annotations. The NTP mixture used was ATP rich; 2 mM ATP, and 50 µM UTP and 50 µM CTP. The dNTP mixture used was: 0.13 µM [^{32}P]dCTP, 2.37 µM dCTP, 50 µM dATP and 50 µM dTTP. DNA templates of various telomeric repeats $(TTAGGG)_n$ were ordered from a vendor (IDT). The enzymatic reactions were initiated by adding the DNA template. The reactions were incubated at 37 °C for 1 h before they were quenched by DNA precipitation. A random-sequence 80-nt oligonucleotide was added as a loading control in this step. After DNA precipitation, each sample was dissolved in 20 µl 1× formamide loading dye and 10 µl of it was loaded into each well of a 10% 1× TBE 7 M urea PAGE gel. Electrophoresis was performed at a constant 45 W until the bromophenol blue dye reached a third of the way from the gel bottom. The gels were vacuum-dried at 80 °C for 1.5 h before exposure to a storage phosphor screen. After adequate exposure, the gels were imaged on a Typhoon FLA9000 gel imager. Enzyme activity analysis was carried out using the GelAnalyzer software.

Negative-stain EM sample preparation, data collection and image analysis

Samples were immediately diluted to 75–100 nM before depositing onto a glow-discharged EM grid (CF400-Cu-UL, EMS). The grids were then subjected to 2% (w/v) uranyl formate (Structure Probe) staining using an established protocol[51]. The negative-stain EM sample grids were imaged on a 120 kV transmission electron microscope (Talos L120C, Thermo Fisher). Image datasets were collected with SerialEM[52] using low-dose mode. 2D classification analysis was performed using cryoSPARC2 (ref. [53]).

Cryo-EM sample preparation and data collection

Cryo-EM sample grids for the co-complex were prepared as described for CST[10]. Briefly, 1.2/1.3 µm 300-mesh holey grids (C-Flat or Quantifoil)

were glow-discharged for 20 s at 15 mA (PELCO easiGlow) before use. Protein samples were supplemented with 4 mM CHAPSO detergent[54] before being added to freshly glow-discharged EM grids and then vitrified using Vitrobot (Thermo Fisher) with a 4.5 s blot time at 4 °C and 95% relative humidity.

Suitable cryo-EM sample grids for data collection were first screened on an Arctica 200 kV transmission electron microscope (Thermo Fisher) equipped with a Bioquantum K3 energy filter-detector (Gatan). The CST-Polα–primase–3×TEL sample dataset (6,252 videos) was collected at the National Cryo-EM Facility at the Frederick National Laboratory for Cancer Research. The videos were collected using Latitude software at a calibrated pixel size of 1.08 Å per pixel, a defocus range of −1.0 to −2.5 µm, a total dose of 50.0 $e^- Å^{-2}$ (40 frames) and the K3 detector (20-eV energy slit) in counting mode. The CST-Polα–primase–4×TEL-FB sample dataset (7,794 videos) was collected at the University of Wisconsin–Madison Cryo-EM Research Center. The videos were collected using SerialEM software[52] (version 4.0) at a calibrated pixel size of 1.1 Å per pixel, a defocus range of −0.7 to −2.8 µm, a total dose of 50.0 $e^- Å^{-2}$ (40 frames) and the K3 detector (20 eV energy slit) in counting mode. The CST-Polα–primase–4×TEL-FB sample dataset (8,658 videos) was collected at the University of Wisconsin–Madison Cryo-EM Research Center. The videos were collected using SerialEM software at a calibrated pixel size of 1.1 Å per pixel, a defocus range of −0.7 to −2.8 µm, a total dose of 51.1 $e^- Å^{-2}$ (40 frames) and the K3 detector (20 eV energy slit) in counting mode.

Cryo-EM single-particle analysis

In general, all three datasets were processed using similar image processing pipelines (Supplementary Figs. 5–7). All video frames were motion-corrected (5 × 5 patch)[55] and their contrast transfer function (CTF)[56] was estimated in Relion. The motion-corrected micrographs were curated, on the basis of their CTF parameters, and subjected to particle picking using the Topaz software[57]. The picked particles were extracted with a binning factor of 5, and particles corresponding to intact samples were selected through two rounds of hetero refinement (six classes) in the cryoSPARC software[53]. The selected particles were re-extracted and centred at the original pixel size before 3D reconstruction using non-uniform refinement[45] in cryoSPARC. The refined particles were exported back to Relion for several rounds of CTF refinement (global and per-particle) and particle polishing[55,58]. These particles were then imported back to cryoSPARC for final reconstruction to obtain the cryo-EM maps of the consensus PIC structure. The final reconstruction was performed using the non-uniform refinement algorithm in CryoSPARC.

To improve the resolution of the consensus maps, masked 3D classification and local refinement were performed individually for the CTC1–$STN1_N$–ssDNA, polymerase–$STN1_C$ and primase domains (Supplementary Figs. 4–6). The masks were generated in ChimeraX software[59,60] using the docked atomic models as segmentation guides. Each domain was subjected to particle subtraction and 3D classification (eight classes) in Relion to identify a subset of particles that yielded an intact domain conformation. During particle subtraction, the particles were recentred and re-extracted at a smaller box size (256 pixels) for 3D classification. Correspondingly, a matching 3D model was reconstructed at a similar dimension to that of the subtracted particle images and filtered to 12 Å. This 3D model was then used as an initial model for the eight-class 3D classification step. The 3D classification was performed with either no alignment or with a small-angle search with local searches enabled. Appropriate 3D class(es) were selected and reverted to their original images for a final 3D reconstruction. The three local refined domain EM densities were combined using the consensus map as a guide. The conversion of files from cryoSPARC to Relion was performed using UCSF PyEM scripts (https://github.com/asarnow/pyem).

Model building, refinement and validation

All seven protein subunits of the PIC were built using computed models from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/). The models were first manually moved into their respective positions in the 4×TEL-FB cryo-EM map and then flexibly fitted using the Jiggle fit function in Coot[61]. The ssDNA molecule was built starting from the 4-nt model shown in the previous CST structure[10]. The assembled PIC model was visually checked in Coot before subjecting it to real-space refinement in PHENIX[62] and visually checked again using Coot. Model-to-map validations were performed using the PHENIX cryo-EM comprehensive validation program.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The described cryo-EM maps and coordinate files have been deposited in the Electron Microscopy Data Bank and the Protein Data Bank (PDB) under the accession codes EMD-27097 (co-complex–3×TEL-consensus); EMD-27099 (co-complex–3×TEL-local-refined-merged); EMD-27104, PDB ID 8D0B (co-complex–4×TEL-local-refined-merged and consensus maps and model); EMD-27107, PDB ID 8D0K (co-complex–4×TEL–$PRIM2_C$-advanced PIC-local-refined-merged and consensus maps and model); and EMD-27109 (co-complex–6×TEL-local-refined-merged and consensus). The coordinates were assembled and built from individual predicted models from the AlphaFold Protein Structure Database https://alphafold.ebi.ac.uk/.

51. Rames, M., Yu, Y. & Ren, G. Optimized negative staining: a high-throughput protocol for examining small and asymmetric protein structure by electron microscopy. *J. Vis. Exp.* https://doi.org/10.3791/51087 (2014).
52. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
53. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
54. Chen, J., Noble, A. J., Kang, J. Y. & Darst, S. A. Eliminating effects of particle adsorption to the air/water interface in single-particle cryo-electron microscopy: bacterial RNA polymerase and CHAPSO. *J. Struct. Biol. X* **1**, 100005 (2019).
55. Zivanov, J., Nakane, T. & Scheres, S. H. W. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCrJ* **6**, 5–17 (2019).
56. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
57. Bepler, T. et al. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat. Methods* **16**, 1153–1160 (2019).
58. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**, e42166 (2018).
59. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
60. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
61. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
62. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).

Acknowledgements We thank T. Tahirov and A. Baranovskiy from the University of Nebraska Medical Center for initial discussion and suggestions in setting up our experiments. We thank T. Cech and A. Zaug from the University of Colorado Boulder for help, discussion and suggestions. We also thank our departmental colleagues S. Butcher and J. Kimble for helpful input, discussion and critical reading of the manuscript. We also thank the members of the laboratory of C.L. for helpful discussions and suggestions. This research was, in part, supported by the Cryo-EM Research Center in the Department of Biochemistry at the University of Wisconsin–Madison and the National Cancer Institute's National Cryo-EM Facility at the Frederick National Laboratory for Cancer Research under contract 75N91019D00024. Support for this research was provided to C.L. by the National Institutes of Health, the National Institute of General Medical Sciences (R00GM131023) and the University of Wisconsin–Madison, Office of the Vice-Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation and the Department of Biochemistry.

Author contributions Q.H. and C.L. designed the expression constructs for recombinant protein production, with support from B.L.L., and Q.H. and X.L. performed insect cell culture

and baculovirus production for insect cell infection. Q.H. expressed and purified the recombinant human protein complexes. Q.H., B.L.C. and C.L. conducted the negative-stain EM data collection and sample screening. Q.H. and C.L. prepared the cryo-EM samples for data collection. C.L. processed and analysed the cryo-EM datasets with support from Q.H. C.L. and Q.H. performed model building and refinement. C.L. and X.L. performed the enzyme assays, and C.L., X.L. and Q.H. analysed the data. Q.H., X.L., S.A. and B.L.C. performed the electrophoresis mobility shift assays, and Q.H., S.A., X.L. and B.L.C. analysed the data. C.L. directed the project and designed the experiments with input from Q.H., X.L., B.L.C. and S.A. C.L. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

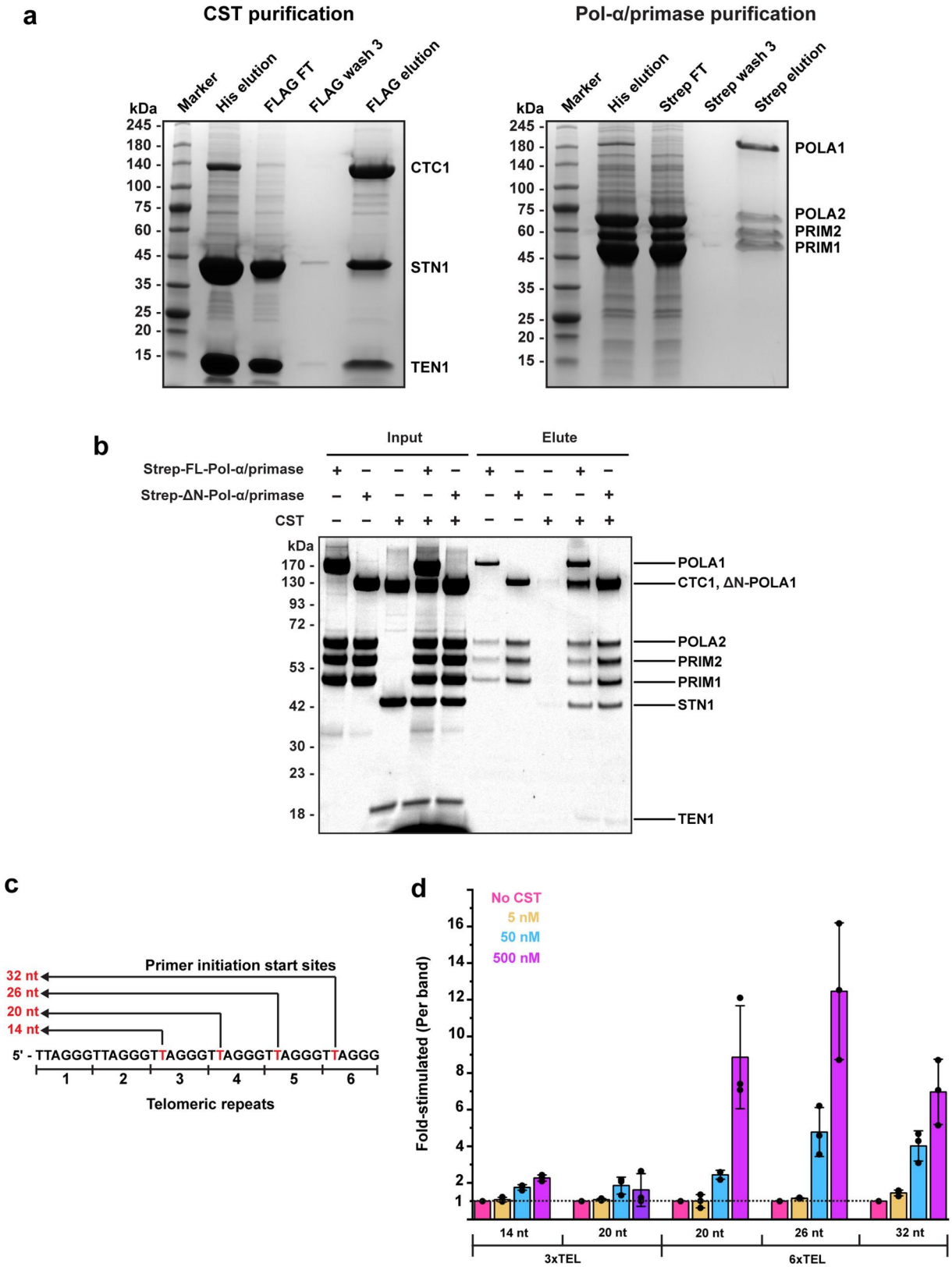
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05040-1.

Correspondence and requests for materials should be addressed to Ci Ji Lim.

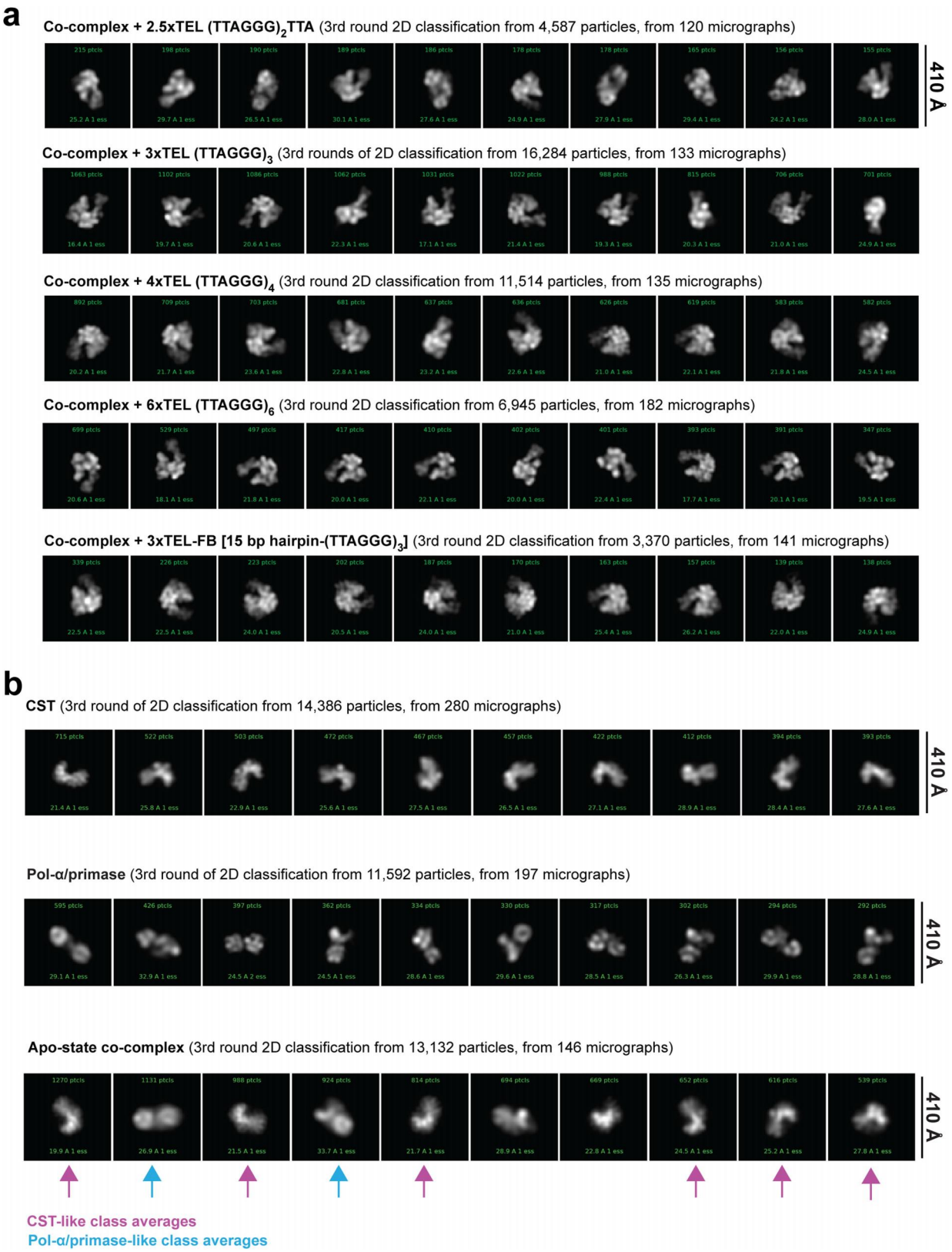
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

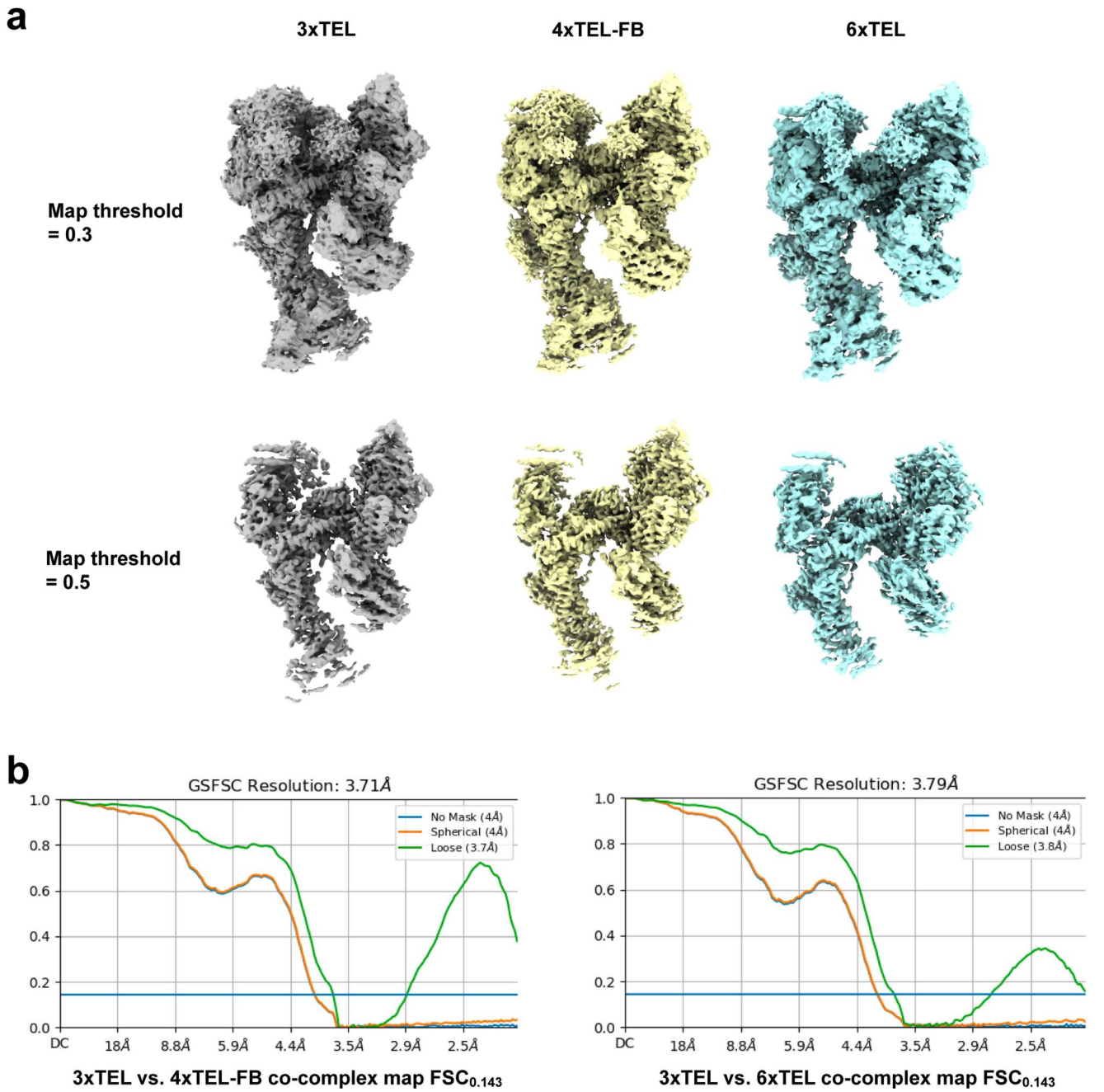


Extended Data Fig. 1 | See next page for caption.

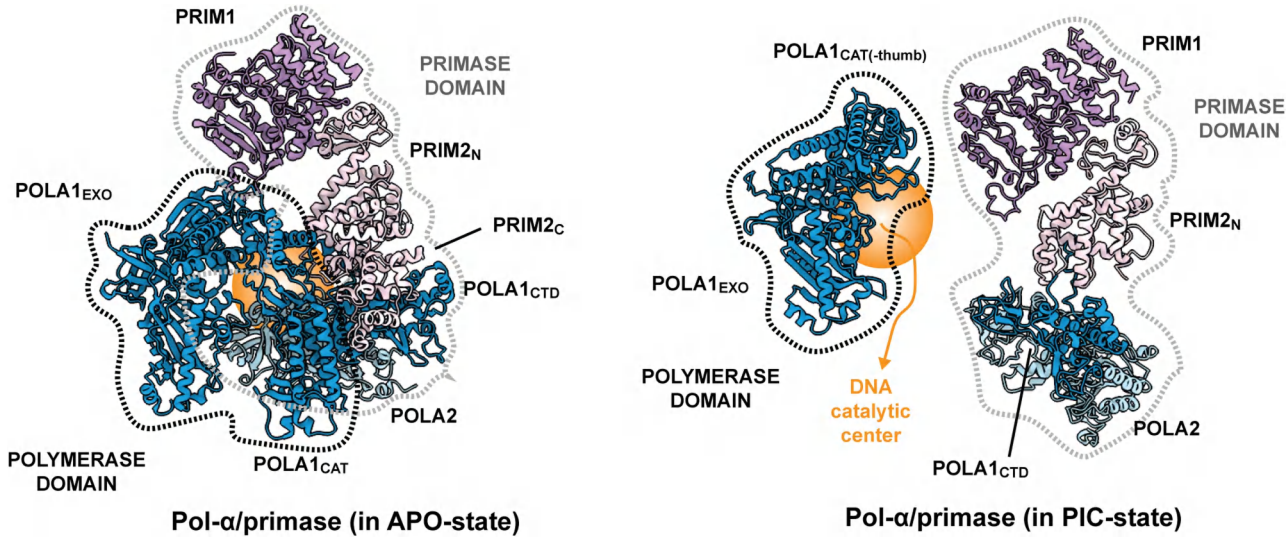
Extended Data Fig. 1 | Recombinant human CST and pol-α/primase purification and their biochemical characterization assays. (**a**) Tandem affinity tags purification of CST and pol-α/primase complexes from baculovirus-infected insect cells. CST heterotrimeric complex was purified by first using Ni-NTA beads to pull down 6xHis-STN1 and 6xHis-TEN1 and then using anti-FLAG beads to pull-down 3xFLAG-CTC1. Pol-α/primase 4-subunit complex was purified by first using Ni-NTA beads to pull down 6xHis-PRIM1, 6xHis-PRIM2, and 6xHis-POLA2 and then using Strep beads to pull-down Strep-POLA1. FT: flowthrough. Trident (GeneTex) high-range prestained protein marker was used in both coomassie-stained denaturing protein gels. The results are reproducible across multiple independent experiments ($n > 5$). (**b**) Pull-down assay to characterize direct protein-protein interactions between CST and full-length or truncated ($\Delta POLA1_{1\text{-}337}$) pol-α/primase (ΔN-Pol-α/primase). Strep beads were used with pol-α/primase as bait and CST as prey. All input lanes are loaded as 10% of the total sample. Molecular weight markers are annotated for size reference. The pull-down results are reproducible across multiple independent experiments ($n = 3$). (**c**) Predicted RNA-DNA primers that are made by pol-α/primase using the repetitive telomeric DNA template in ATP-rich NTP reaction conditions. Primer product sizes are calculated assuming the polymerase reaches the end of the template. (**d**) Fold-stimulation analysis of direct enzyme assays done at predicted RNA-DNA primer product level (see Fig. 1c). CST does not stimulate pol-α/primase with the 3xTEL template, whether measured at the individual primer product level or total activity (see Fig. 1d). CST can stimulate pol-α/primase using the 6xTEL template. Data are plotted using bars for mean values and error bars for standard deviation as calculated from three independent experiments ($n = 3$). Data points from each experiment are represented by circle-filled markers in each corresponding bar. The dashed line marked fold-stimulation of one (no stimulation) and values were calculated by dividing the band intensity counts over the corresponding counts from the band and lane without CST added.



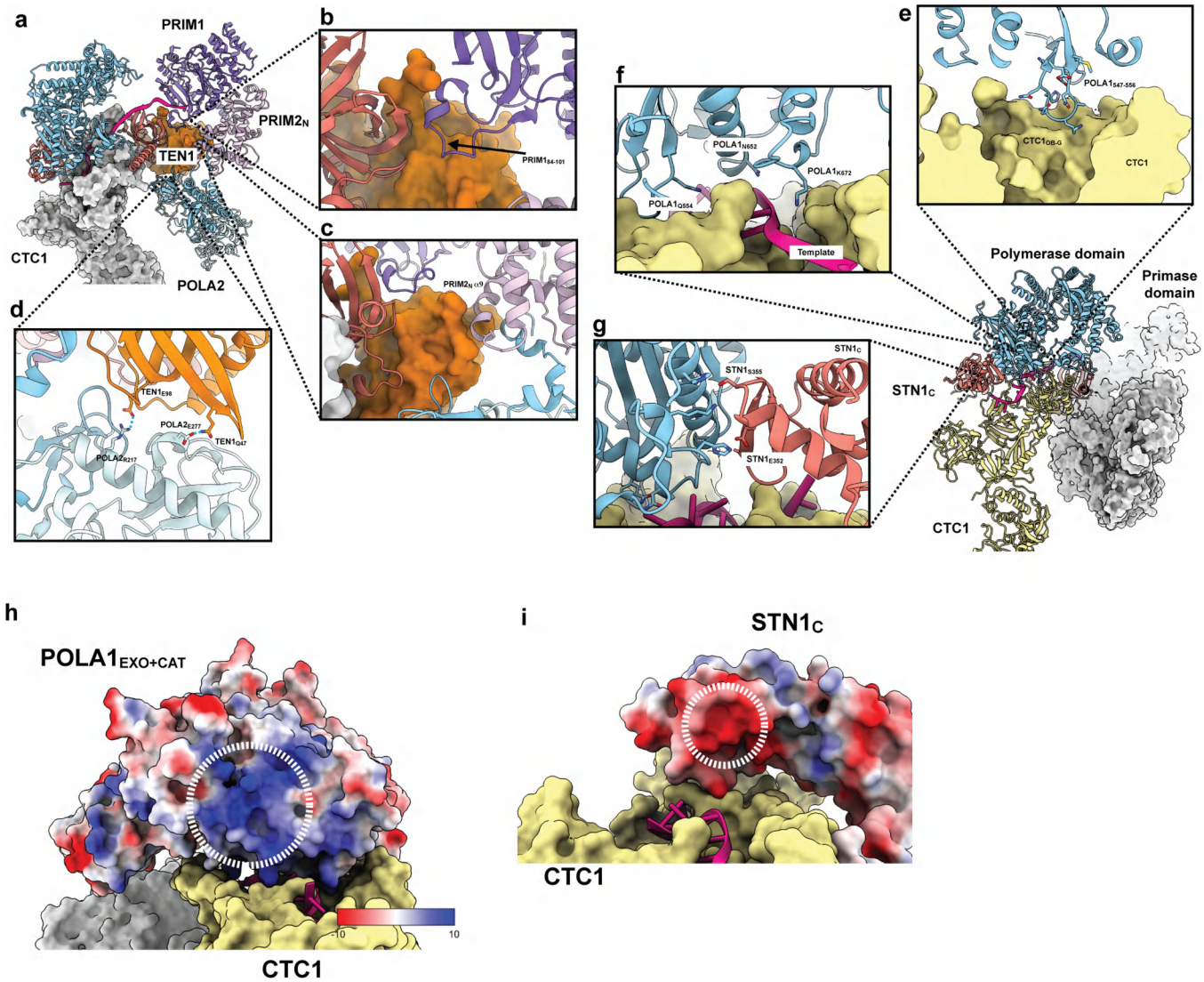
Extended Data Fig. 2 | Negative-stain electron microscopy 2D classification analysis of CST-pol-α/primase co-complex bound to different telomeric ssDNA template lengths and structures. (**a**) Top ten (by distribution) class averages of CST-pol-α/primase preincubated with various lengths of telomeric single-stranded DNA, from 2.5 to 6 repeats of TTAGGG, and structures (ssDNA versus ss/dsDNA). 3xTEL-FB is a DNA template that has a 15 bp hairpin structure at the 5′ end of the GG(TTAGGG)$_2$TTAG overhang. (**b**) Top ten (by distribution) 2D class averages of each apo-state CST, pol-α/primase, and CST-pol-α/primase co-complex. No new PIC structure was observed in the absence of a template. Instead, individual CST or pol-α/primase 2D class averages were seen in the co-complex analysis. CST-like and pol-α/primase class averages are pointed out by color-coded arrows for visual comparison.



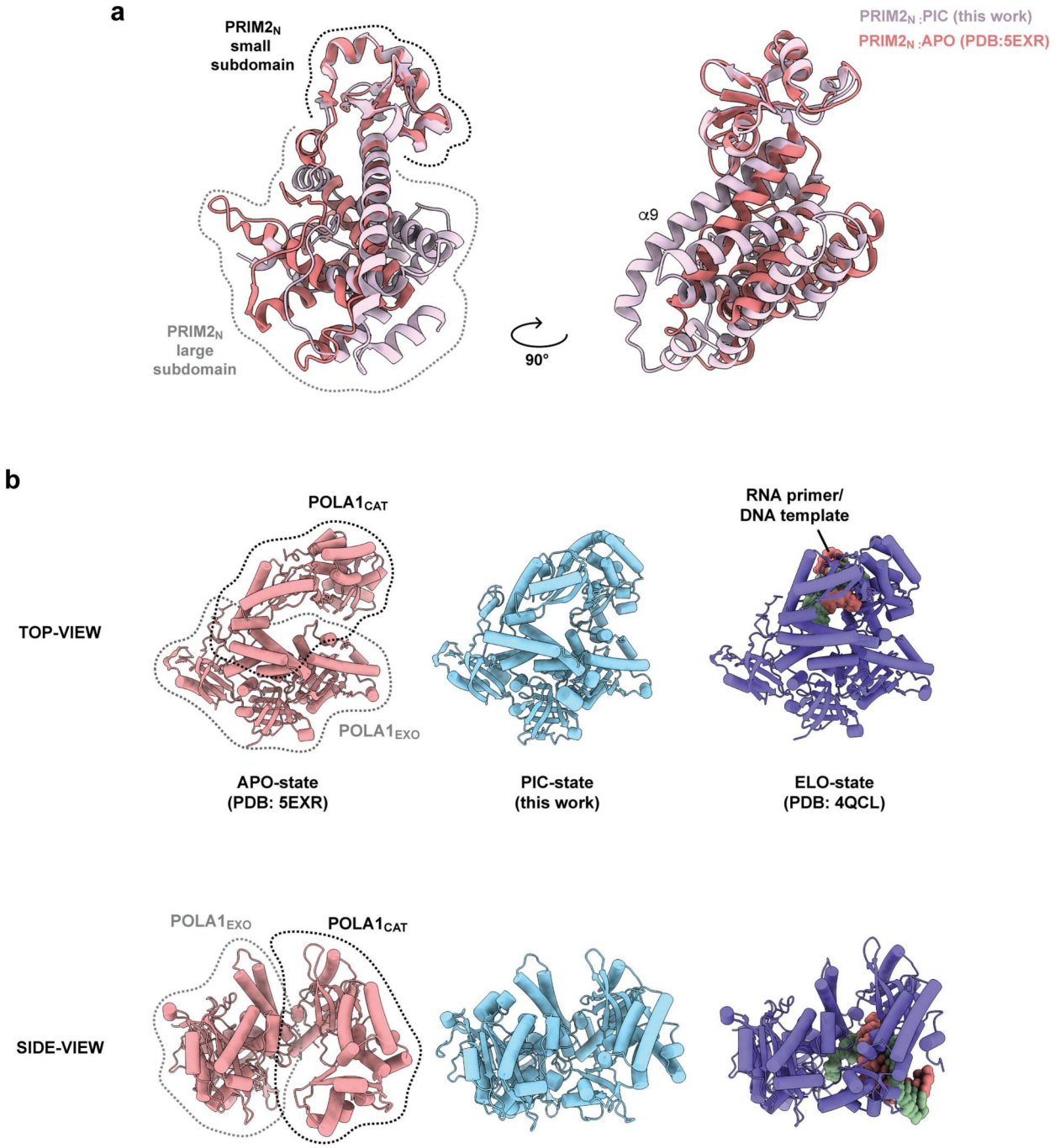
Extended Data Fig. 3 | 3xTEL, 4xTEL-FB, and 6xTEL co-complexes share the same overall PIC structure. (**a**) Comparison of the three co-complexes consensus cryo-EM maps at two different map thresholds. (**b**) Fourier-shell correlation (FSC) values of the 3xTEL PIC cryo-EM map aligned against the 4xTEL-FB map or 6xTEL map. The FSC values are reported at the level of 0.143.



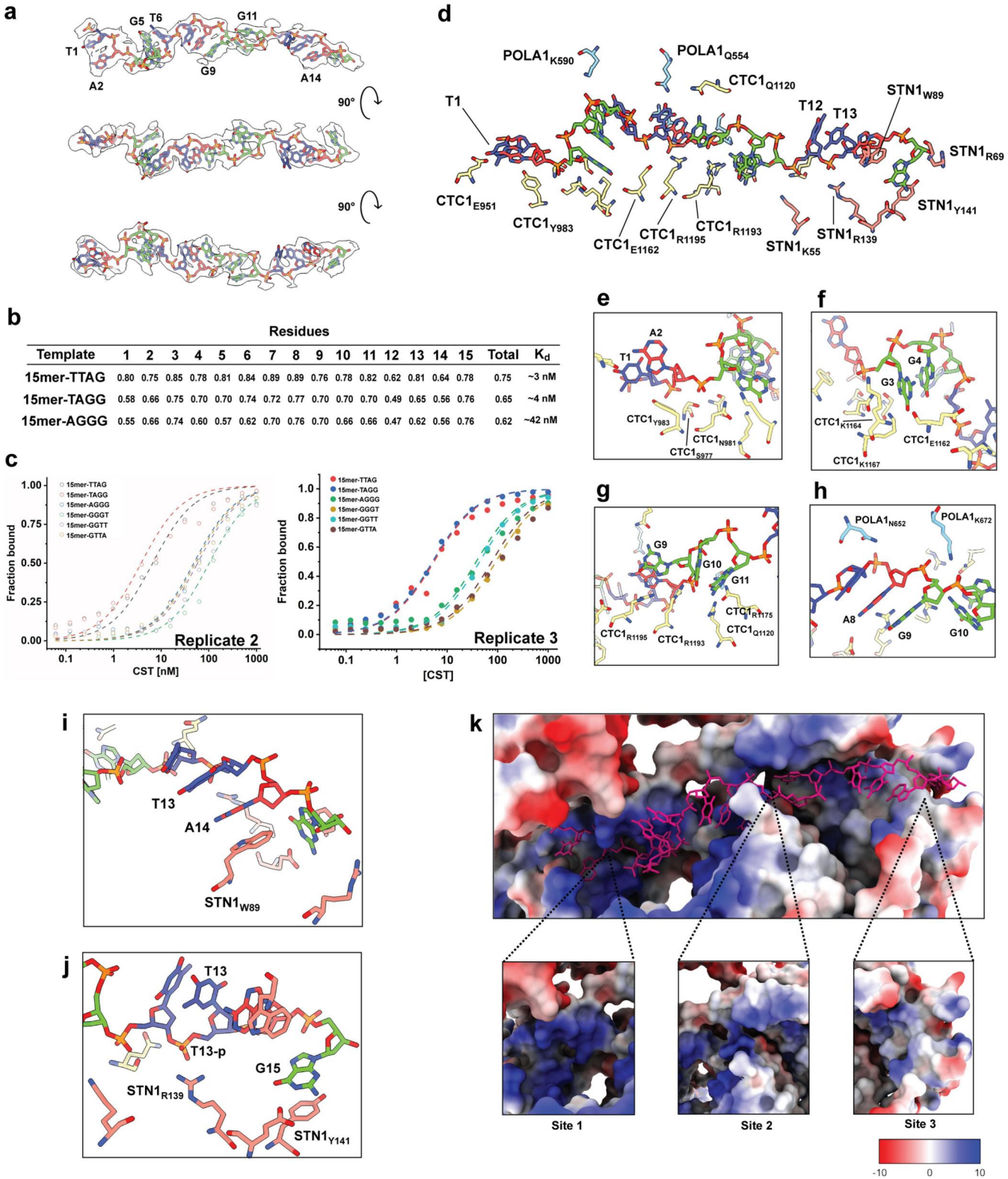
Extended Data Fig. 4 | Pol-α/primase conformation comparison in its apo-state vs. its preinitiation-state. Pol-α/primase architecture changes from a compact shape to a segregated form upon forming the preinitiation complex (PIC) with a template-bound CST. The APO and PIC structures are depicted as ribbon cartoons. The CST complex and template are hidden in the PIC state to better demonstrate the segregated architecture of the PIC structure – separated into polymerase and primase domains. The lobe boundaries are loosely defined by the dashed lines – black for the polymerase domain and grey for the primase domain. The POLA1 DNA catalytic center is highlighted in orange to illustrate its accessibility upon forming the PIC. $POLA1_{CAT}$ thumb subdomain and $PRIM2_C$ domain are not illustrated in the PIC structure because of their flexibility (see main text).



Extended Data Fig. 5 | CST and Pol-α/primase interactions. (**a**) Overview of TEN1 interactions with primase domain. TEN1 is depicted as surface representation while the primase subunits are drawn as ribbons. (**b**) TEN1 and $STN1_N$ interface sits the $PRIM1_{84-101}$ loop. (**c**) TEN1 also made contacts with the alpha-helix nine (α9) of the $PRIM2_N$ domain. (**d**) Two hydrogen bonds were identified between TEN1 and POLA2 subunit; $TEN1_{E98}$ to $POLA2_{R217}$ and $TEN1_{Q47}$ to $POLA2_{E277}$. (**e**) The loop between residues 547-556 of POLA1 sits on the $CTC1_{OB-G}$. The POLA1 loop is depicted with heteroatom-coloured sidechains. (**f**) The template bound by CTC1 forms hydrogen bonds with POLA1 charged residues Q554, N652, and K672. (**g**) $STN1_C$ interacts with POLA1 using negatively charged residues E352 and S355. On the opposite surface, POLA1 uses the positively charged residues H627, H658, and K661. (**h**, **i**) The surfaces where POLA1 and $STN1_C$ contact has complementary charged residues. In panel **h**, the POLA1 $STN1_C$-interacting surface, encircled by a white dashed circle, is positively charged. The colour key for the coulombic electrostatic potential is shown at the bottom right of the panel. In panel **i**, for $STN1_C$, its POLA1-interacting surface, circle by a white dashed circle, is negatively charged. The white dashed circles represent the area of contact for the two protein subunits.

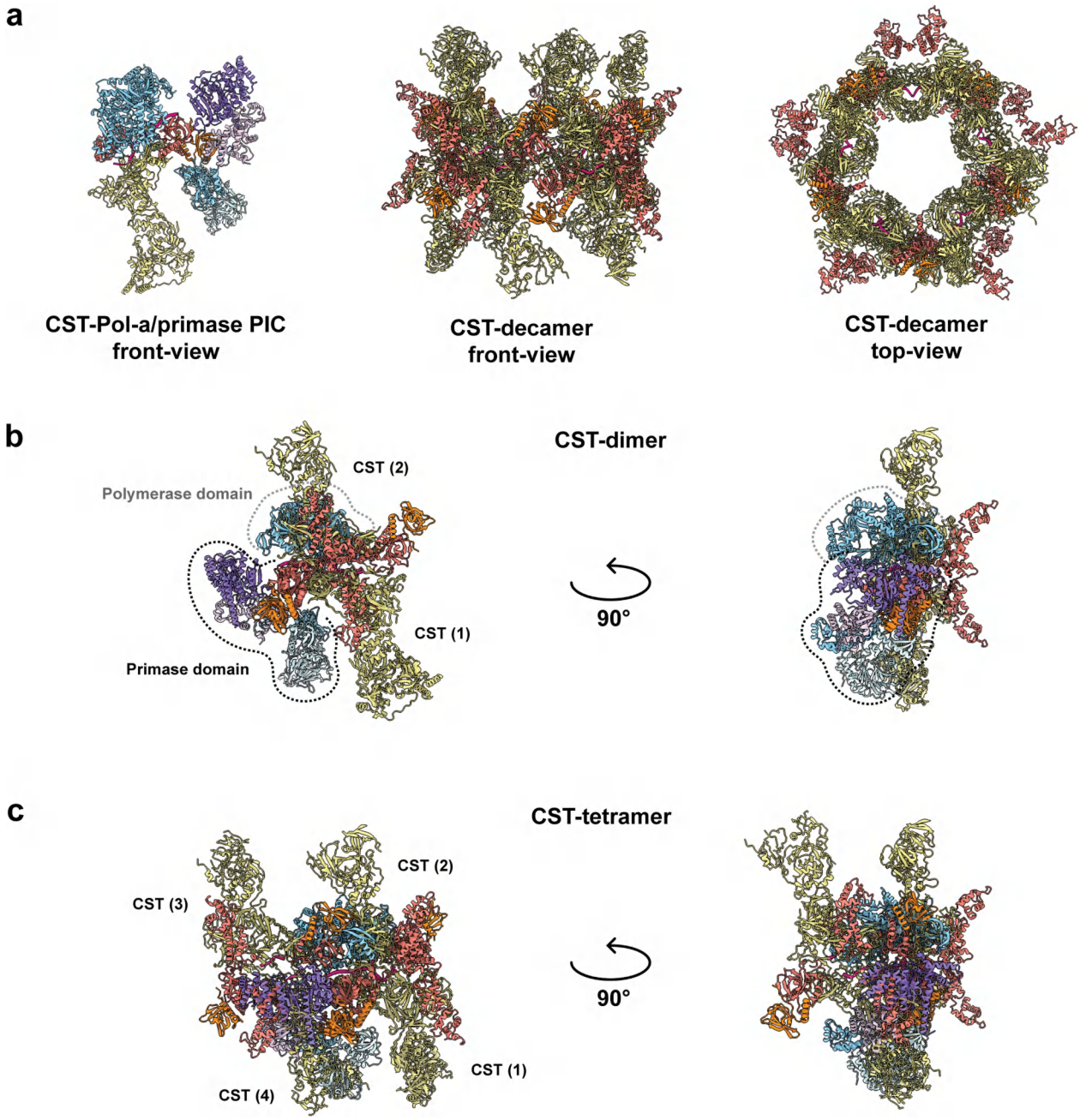


Extended Data Fig. 6 | Comparisons of PRIM2$_N$ conformation in enzyme apo and preinitiation states and POLA1$_{EXO-CAT}$ domain conformation in the enzyme apo, preinitiation and elongation state. (**a**) The PRIM2$_N$ structure from our PIC model was compared against the pol-α/primase model in the APO-state (PDB: 5EXR). The PRIM2$_N$ small and large subdomains were the same for both enzymatic states but the domain relative position had rotated by about 40°. The structures were aligned using the small subdomain to illustrate the large domain shift. (**b**) Top-down view of the POLA1$_{EXO-CAT}$ domain at various enzymatic states; apo (APO), preinitiation (PIC), and DNA elongation (ELO). To note, the APO- and ELO-state structures were obtained from either pol-α/primase (PDB: 5EXR) or POLA1 catalytic domain (PDB: 4QCL) alone. The PIC-state was derived from the CST-pol-α/primase co-complex structure. The RNA-DNA primer-template structure of the ELO-state is depicted as a space-filled model. The side-view perspective of each enzyme state is also shown for comparison.

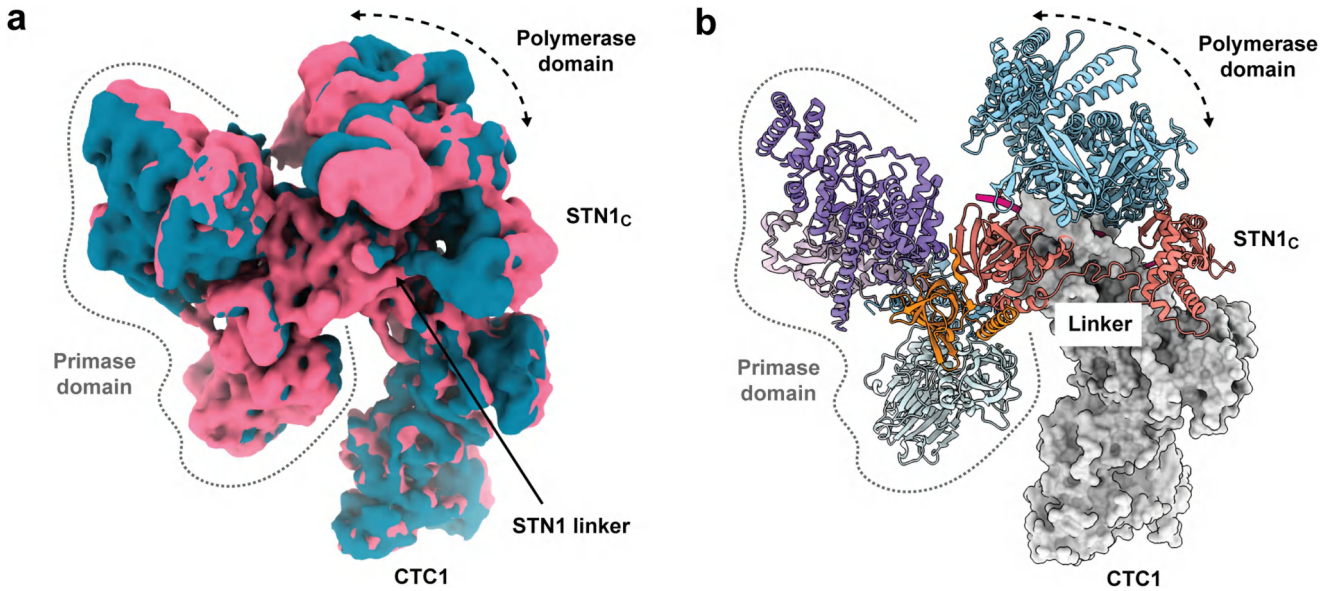


Extended Data Fig. 7 | See next page for caption.

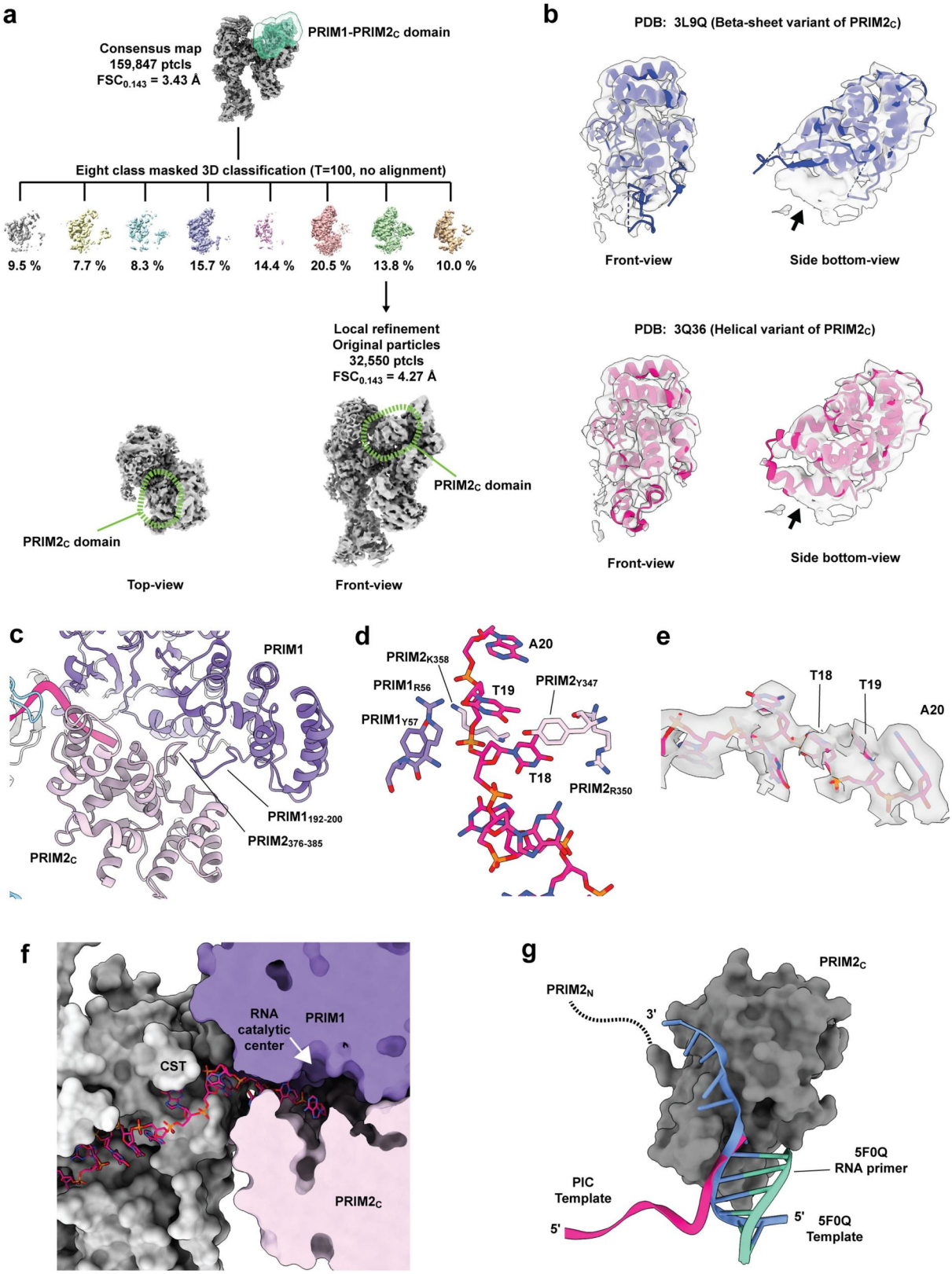
Extended Data Fig. 7 | Analysis of PIC-template interactions. (**a**) Model of the 15 nt telomeric sequence built into template cryo-EM density (shown as semi-transparent grey coloured volume). The model is depicted as an atomic model that is color-coded T=blue, A=red, and G=green. The modelled sequence is 5′- TAGGGTTAGGGTTAG – 3′ (15mer-TTAG). (**b**) Model-to-map cross-correlation (CC) analysis reported at per residue level and chain average. The experimental binding affinities of CST to these templates are provided. The template 15mer-TAGG and 15mer-AGGG sequences are 5′- AGGGTTAGGGTTAGG – 3′ and 5′- GGGTTAGGGTTAGGG – 3′ respectively. (**c**) Independent experiments and quantification for Fig. 2b. See Supplementary Table 1 for template sequences. (**d**) Overview of protein-DNA interactions involved in template binding. (**e**, **f**) $CTC1_{OB\text{-}F}$ interactions with the template at the binding site 1. (**g**) CTC1OB-G residues binding to the template at binding site two are shown. (**h**) POLA1 residues binding to the template at binding site 2 are shown. (**i**) STN1 W89 base stacks with A14 of the template. This is supported by T13-T14 base stacking. (**j**) Other protein-DNA interactions between $STN1_N$ and template at binding site 3. STN1 R139 hydrogen bonds with T13 phosphate backbone and STN1 Y141 base stacks with G15 aromatic base. **(k)** The 15 nt DNA template model is depicted as a deep pink atomic model. Sites 1 and 2 are positively charged while site 3 is more neutral. Site 1 DNA-binding residues are mostly made from $CTC1_{OB\text{-}F}$, site 2 residues are from $CTC1_{OB\text{-}G}$ and POLA1, and site 3 residues are from the $STN1_N$ domain. The electrostatic potential colour key is in units of kcal/(mol·*e*) at 298 K.



Extended Data Fig. 8 | CST-pol-α/primase PIC structure is incompatible with CST decamer formation at early oligomerization states. (**a**) Dimension comparison of the PIC model to the CST decamer model (ten copies of CST; PDB: 6W6W). (**b**) The PIC primase domain is not sterically hindered by the CST inverted dimer and thus could exist. However, its polymerase domain directly clashes with CST dimer formation. (**c**) The PIC primase and polymerase domains clash with CST tetramer formation. As in panel **b**, the polymerase domain clashes with the dimer formation, while the primase domain is now in the way of the fourth CST in the CST tetramer model. Subunits are coloured as in Fig. 1 for all panels.



Extended Data Fig. 9 | Polymerase domain swivels on top of the template-bound CST. (**a**) Three-dimensional (3D) variable analysis (3DVA) principal component analysis revealed a mode that showed the polymerase domain rotating about on the top of the template-bound CST. The first and last frames of the PCA mode are shown here as pink and navy blue cryo-EM maps. The cryo-EM maps are filtered to a resolution of 8 Å. The polymerase domain swivelling motion is seemly controlled by the $STN1_C$ domain and rotation angle restrained by the STN1 linker. (**b**) The corresponding components in A depicted on the PIC atomic model. CTC1 is represented as surface and coloured light grey. The rest of the PIC are drawn as ribbons. The color scheme follows Fig. 1.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Cryo-EM heterogeneity analysis reveals template-bound human primase structure at an advanced preinitiation state. (**a**) The 4xTEL-FB PIC structure was masked with a PRIM1-PRIM2$_C$ for particle subtraction and 3D classification to isolate a subset of particles that allowed higher-resolution 3D reconstruction of the PRIM2$_C$ domain. An eight classes 3D classification step without image alignment was used to sieve out a conformation (~13.8%) that has an extra EM density in front of the PRIM1 subunit. This subtracted particles subset (32,550 particles) was reverted to their original images before subjecting to 3D refinement to obtain a 4.3 Å cryo-EM map. The resolved PRIM2$_C$ domain density is circled in green dashed lines. (**b**) Two crystal structure variants of the PRIM2$_C$ domain are docked into the cryo-EM density that was resolved in Fig. 5 and panel **a**. The all-helical variant (PDB: 3Q36) is a better fit than the β-sheet-containing version (PDB: 3L9Q), as can be seen in the region pointed out by the black arrow. (**c**) PRIM1 and PRIM2$_C$ have minimal contacts except for two loop regions; PRIM1$_{192-200}$ and PRIM2$_{376-385}$. (**d**) PRIM1 and PRIM2 interactions with the template region into the primase crevice. (**e**) The five additional template residues were built based on the resolved primase density. The last three nucleotides are annotated. (**f**) The PRIM1 and PRIM2$_C$ domains form a crevice that the template sits in for RNA primer synthesis. The crevice has the PRIM1 RNA catalytic center facing the template. (**g**) The 3′ end of the extended 20 nt template model in the PIC structure connects seamlessly with the template model in the previously solved RNA-DNA primer-template-bound PRIM2$_C$ crystal structure (PDB: 5F0Q).

nature portfolio

Corresponding author(s): Qixiang He, Xiuhua Lin, Bianca L. Chavez, Sourav Agrawal, Benjamin L. Lusk, Ci Ji Lim

Last updated by author(s): Jun 12, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Cryo-EM datasets were collected using Latitude (at NCEF) or SerialEM (version 3.8.7, at CEMRC) software package.
Data analysis	Cryo-EM data was processed by Relion (3), CTTFind4 (4.1), Topaz (0.2.5), CryoSPARC (3.3), Daniel Asarnow's pyEM scripts (open source - https://github.com/asarnow/pyem). Modeling was done using UCSF Chimera X (1.2), Coot (0.9.2), and PHENIX (1.19.2). Binding and enzyme direct assays were quantified using GelAnalyzer (19.1) and curved fitted and plotted using Origin software (2022).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The described cryo-EM maps and coordinates files are deposited in the Electron Microscopy Data Bank (EMDB) and PDB under accession codes: EMD-27097 (co-complex-3xTEL-consensus), EMD-27099 (co-complex-3xTEL-local-refined-merged), EMD-27104; PDB ID 8D0B (co-complex-4xTEL local-refined-merged and consensus maps and model), EMD-27107; PDB ID 8D0K (co-complex-4xTEL-PRIM2C-advanced PIC-local-refined-merged and consensus maps and model), and EMD-27109 (co-complex-6xTEL-local-refined-merged and consensus). The coordinates are assembled and built from individual predicted models from the AlphaFold Protein Structure Database - https://alphafold.ebi.ac.uk/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size for cryo-EM data collection was determined by the availability of electron microscope time. This is further defined by the consideration for the number of particles that are needed to reconstruct a high-resolution cryo-EM map for structure-function analysis. The number of movies collected, and the final particles used for 3D reconstructions are comparable to other high-quality cryo-EM studies. Biochemical assays that require quantification were conducted across at least three independent experiments, using reagents from at least three different batches, and thus are comparable to other published studies.
Data exclusions	Poor resolution data was excluded during cryo-EM data processing, a standard step in cryo-EM single-particle analysis to obtain the highest-possible structure resolution. Only outlier data points (such as misloading of samples in gels; only 1 or 2 lanes in entirety) were excluded in our DNA-binding assays.
Replication	All biochemical experiments were repeated at least in three separate experiments using at least two different batches of protein samples. Thus, they are all reproducible.
Randomization	Randomization is not relevant to this study because the studies did not involve human subjects or live animals
Blinding	Blinding is not relevant to this study because the studies did not involve human subjects or live animals

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	SF9 and Tni insect cell lines for baculovirus generation and proteins production respectively. The SF9 cells were obtained from Invitrogen, ThermoFisher while the Tni cells were from Expression systems.
Authentication	No authentication of cell lines were performed as they were purchased from vendors and their cell morphology and doubling time are consistent with the vendors specifications and comparable to similar studies.
Mycoplasma contamination	Cells were not tested for mycoplasma since they are doubling at a healthy rate and morphology looked fine.
Commonly misidentified lines (See ICLAC register)	No misidentified cell lines were used in this work.

RESEARCH ARTICLE SUMMARY

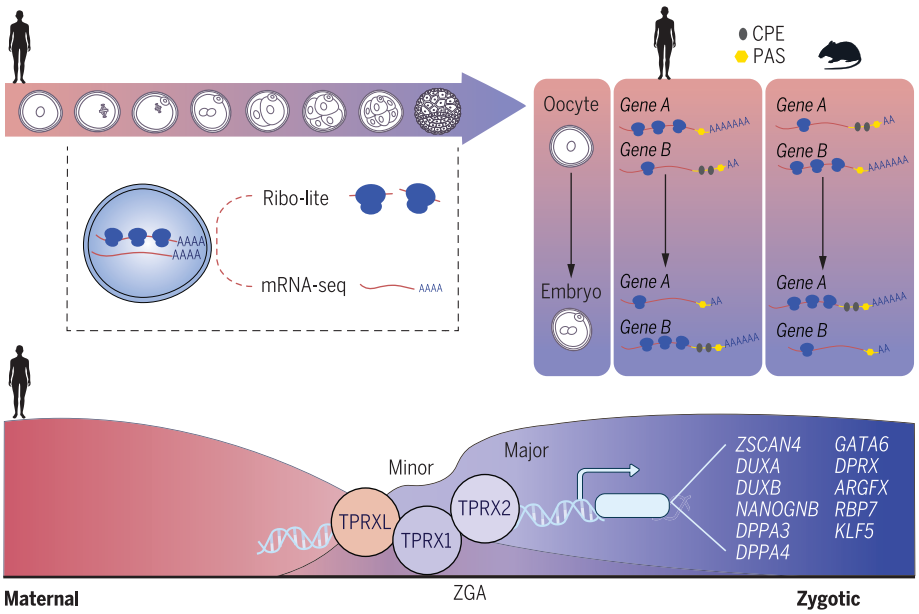
DEVELOPMENT

Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation

Zhuoning Zou†, Chuanxin Zhang†, Qiuyan Wang†, Zhenzhen Hou†, Zhuqing Xiong, Feng Kong, Qiujun Wang, Jinzhu Song, Boyang Liu, Bofeng Liu, Lijuan Wang, Fangnong Lai, Qiang Fan, Wenrong Tao, Shuai Zhao, Xiaonan Ma, Miao Li, Keliang Wu, Han Zhao*, Zi-Jiang Chen*, Wei Xie*

INTRODUCTION: During the mammalian oocyte-to-embryo transition (OET), translation plays a critical role in regulating meiotic resumption, zygotic genome activation (ZGA), and early embryonic development. ZGA marks the first transcription event in a new life and the onset of the embryonic program. However, how mammalian ZGA is initiated remains poorly understood. For example, although key ZGA transcription factors (TFs) have been well characterized in other species such as zebrafish and fly, which TFs control human ZGA remains elusive.

RATIONALE: Studying the translatomes in human oocytes and early embryos is critical for understanding their posttranscriptional regulation during human OET and identifying candidate ZGA regulators. In particular, the TFs regulating ZGA are expected to arise either from translation of oocyte-inherited transcripts or from transcription during the early phase of ZGA, and their motifs should be enriched at enhancers and promoters of ZGA genes. However, translatome profiling in human oocytes and early embryos is severely hindered by the scarcity of research materials. Therefore, we sought to first profile the translatomes and transcriptomes from the same low-input samples of human oocytes and early embryos using an ultrasensitive method. Combined with analyses of the assay for transposase-accessible chromatin sequencing (ATAC-seq) datasets in human early embryos that we reported previously, candidate TFs for human ZGA were then identified. Their potential roles in genome activation and early development were assessed by gene knockdown in human embryos and overexpression analyses in human embryonic stem cells (hESCs).



Translatome-transcriptome co-profiling in human oocytes and early embryos reveals key human ZGA factors. The translatome and transcriptome were co-profiled in human oocytes and early embryos. Comparison with mouse reveals widespread species-specific translation from oocytes (full-grown oocytes depicted) to embryos, in part driven by distinct configurations of regulatory elements in 3′UTRs, including CPEs. The TPRX family TFs TPRXL/1/2 are highly translated around ZGA and are identified as critical regulators of human ZGA.

RESULTS: By combining the ultrasensitive Ribo-lite (ligation-free, ultra-low-input, and enhanced Ribo-seq) with Smart-seq2 in a method we call Ribo-RNA-lite (R2-lite), we jointly profiled the translatome and transcriptome across eight stages of human oocytes and early embryos. Through comparison with their counterparts in mouse, we found not only genes with conserved translational activities but also widespread, differentially translated genes functioning in epigenetic reprogramming, transposon defense, and small RNA biogenesis. Species-specific translation is in part driven by different configurations of regulatory elements such as cytoplasmic polyadenylation element (CPE) and polyadenylation signal site (PAS) in the 3′ untranslated regions (3′UTRs).

Using the R2-lite data, we found that a group of PRD-like homeobox TFs became highly translated before or during ZGA, with their motifs enriched in distal open chromatin regions (putative enhancers) near activated genes upon ZGA. These TFs include TPRXL, which is encoded by a CPE-containing maternal transcript subjected to translation up-regulation upon meiotic resumption, and TPRX1/2, which are expressed during the early phase of ZGA (minor ZGA). The joint knockdown of *TPRX1/2/L* [*TPRX* triple KD (TKD)] led to severe defects in development and ZGA. About 31% of ZGA genes that preferentially contain PRD-like TF-binding motifs at promoters and nearby putative enhancers were down-regulated in *TPRX* TKD embryos. These TPRX target genes include *ZSCAN4*, *DUXB*, *DUXA*, *NANOGNB*, *DPPA4*, *GATA6*, *DPRX*, *ARGFX*, *RBP7*, and *KLF5*, many of which encode key transcription regulators. Finally, ectopically expressed TPRXs could bind and activate a subset of ZGA genes in hESCs.

CONCLUSION: Here, we charted the translational landscapes during the human OET. Comparison of data in mouse identified widespread differentially translated genes, in part driven by species-specific configuration of key regulatory elements in the 3′UTRs. This dataset further identified a group of PRD-like homeobox TFs, including TPRXL, TPRX1, and TPRX2, that are highly translated around ZGA. *TPRXs* are required for proper ZGA and preimplantation development and are also sufficient to activate key ZGA genes when ectopically expressed in hESCs. Therefore, these data not only reveal the conservation and divergence of translational regulation during OET but also identify critical TF regulators of human ZGA. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: hanzh80@yahoo.com (H.Z.); chenzijiang@hotmail.com (Z.-J.C.); xiewei121@tsinghua.edu.cn (W.X.)
†These authors contributed equally to this work.
Cite this article as Z. Zou *et al.*, *Science* **378**, eabo7923 (2022). DOI: 10.1126/science.abo7923

S READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abo7923

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

RESEARCH ARTICLE

DEVELOPMENT

Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation

Zhuoning Zou[1,2,3]†, Chuanxin Zhang[4,5,6,7,8]†, Qiuyan Wang[1,2]†, Zhenzhen Hou[4,5,6,7,8]†, Zhuqing Xiong[1,9], Feng Kong[1,2], Qiujun Wang[1,2], Jinzhu Song[4,5,6,7,8], Boyang Liu[4,5,6,7,8], Bofeng Liu[1,2], Lijuan Wang[1,2], Fangnong Lai[1,2], Qiang Fan[1,2], Wenrong Tao[4,5,6,7,8], Shuai Zhao[4,5,6,7,8], Xiaonan Ma[4,5,6,7,8], Miao Li[4,5,6,7,8], Keliang Wu[4,5,6,7,8], Han Zhao[4,5,6,7,8,10]*, Zi-Jiang Chen[4,5,6,7,8,10,11,12]*, Wei Xie[1,2]*

Translational regulation plays a critical role during the oocyte-to-embryo transition (OET) and zygotic genome activation (ZGA). Here, we integrated ultra-low-input ribosome profiling (Ribo-lite) with messenger RNA sequencing to co-profile the translatome and transcriptome in human oocytes and early embryos. Comparison with mouse counterparts identified widespread differentially translated gene functioning in epigenetic reprogramming, transposon defense, and small RNA biogenesis, in part driven by species-specific regulatory elements in 3′ untranslated regions. Moreover, PRD-like homeobox transcription factors, including *TPRXL*, *TPRX1*, and *TPRX2*, are highly translated around ZGA. *TPRX1/2/L* knockdown leads to defective ZGA and preimplantation development. Ectopically expressed TPRXs bind and activate key ZGA genes in human embryonic stem cells. These data reveal the conservation and divergence of translation landscapes during OET and identify critical regulators of human ZGA.

Human infertility is a rising global issue, and understanding human oocyte maturation and early embryonic development is critical in deciphering its etiology. After entering meiosis, mammalian oocytes are arrested at prophase I and only resume meiosis upon hormone stimulation. The genome stays transcriptionally silenced from late-stage full-grown oocytes (FGOs) to fertilized embryos before zygotic genome activation (ZGA) (*1*). As a result, the oocyte-to-embryo transition (OET) is often driven by posttranscriptional regulation (*2*). This involves stage-specific translational regulation of pretranscribed mRNAs in oocytes, including translational activation of "dormant" maternal mRNAs, which initially stay untranslated after being transcribed (*1*, *3*).

In *Xenopus* oocytes, translation is mediated by *cis*-acting sequences in the 3′ untranslated regions (3′UTRs), including cytoplasmic polyadenylation element (CPE) bound by cytoplasmic polyadenylation element binding protein 1 (CPEB1) and the polyadenylation hexanucleotide signal (PAS) recognized by the cleavage and polyadenylation specificity factor (CPSF) (*3*–*6*). CPEB1 suppresses polyadenylation in FGOs but promotes polyadenylation during meiotic resumption upon phosphorylation. In the mouse, CPEB1 is essential for fertility and oocyte maturation, and its phosphorylation similarly results in translational activation of several dormant RNAs (*7*, *8*). However, how translation is regulated in human OET is poorly studied.

ZGA marks the first transcription event in a new life (*9*), but how it is initiated in mammals remains elusive. DUX in the mouse and its counterpart, DUX4, in humans activate a subset of genes preferentially expressed during minor ZGA, the early wave of ZGA (e.g., *Zscan4/ZSCAN4* and *MERVL/HERVL*) (*10*–*12*). However, *Dux* deficiency in mice causes only minor defects in ZGA and is compatible with development (*13*–*16*). Similarly, knockdown (KD) of *DUX4* in human embryos disturbs gene expression without affecting preimplantation development during the studied period (*17*). Moreover, *Dux* and *DUX4* themselves are activated during minor ZGA (*10*–*12*). Thus, additional key transcription factors (TFs) for mammalian ZGA remain to be discovered. ZGA regulators may arise from the translation of maternally deposited RNAs, as exemplified by Nanog, SoxB1, and Pou5f1 in zebrafish and Zelda in fly (*9*). Recently, we reported an ultrasensitive ribosome profiling method [ligation-free, ultra-low-input and enhanced Ribo-seq (Ribo-lite)] that can accommodate as few as 50 murine embryonic stem cells (mESCs) or single mouse oocytes (*18*). Here, we applied Ribo-lite coupled with RNA sequencing (RNA-seq) to human oocytes and early embryos to investigate the translational regulation during human OET and to identify possible regulators of human ZGA.

Results

Co-profiling of translatome and transcriptome landscapes of human oocytes and preimplantation embryos by Ribo-RNA-lite

We combined Ribo-lite (*18*) with Smart-seq2 (a low-input RNA-seq method) (*19*) by splitting the lysed samples for transcriptome and translatome co-profiling in a method we call Ribo-RNA-lite (R2-lite) (Fig. 1A and fig. S1A). The global patterns of ribosome-protected fragments (RPFs) and mRNAs from R2-lite in 1000 human embryonic kidney 293 (HEK293) cells correlated with those of the conventional Ribo-seq data (*20*) and the reference Smart-seq2 data (*18*), respectively (fig. S1B). RPFs, but not mRNAs, showed typical Ribo-seq features, including the expected mapping rates (fig. S1C), the depletion of signals from 3′UTRs (fig. S1, D and E), and the characteristic 3-nucleotide (nt) periodicity (fig. S1F). Exogenous External RNA Controls Consortium (ERCC) RNA added to the samples was well detected by mRNA-seq but not by Ribo-lite (fig. S1G). These data demonstrate that R2-lite could efficiently profile the transcriptome and translatome from the same low-input sample.

Next, we performed R2-lite in human oocytes and early embryos (using four to 25 oocytes or embryos for each stage and two replicates; table S1), including FGOs, metaphase I (MI) oocytes, metaphase II (MII) oocytes, one-cell (1C), two-cell (2C), four-cell (4C), and eight-cell (8C) embryos, as well as inner cell mass (ICM) from the blastocysts (fig. S1A). We also included primed human embryonic stem cells (hESCs) as a control. Ribo-lite captured 7790 to 10,854 expressed genes [fragments per kilobase per million (FPKM) > 1] and mRNA-seq detected 7689 to 9056 genes (FPKM > 1) across stages (fig. S2A and FPKM in table S2) with high reproducibility ($R > 0.75$) (fig. S2, B and C). The variations between replicates were comparable to the Ribo-seq data in mouse oocytes and early embryos (*18*, *21*), as well as mESCs (*22*) (fig. S2D). Ribo-lite data also showed the expected low 3′UTR read rates (fig. S3A) and prominent 3-nt periodicity (fig. S3, B and C)

[1]Center for Stem Cell Biology and Regenerative Medicine, MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China. [2]Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China. [3]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China. [4]Center for Reproductive Medicine, Shandong University, Jinan, Shandong 250012, China. [5]Key Laboratory of Reproductive Endocrinology of Ministry of Education, Shandong University, Jinan, Shandong 250012, China. [6]Shandong Key Laboratory of Reproductive Medicine, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong 250012, China. [7]Shandong Provincial Clinical Research Center for Reproductive Health, Jinan, Shandong 250012, China. [8]National Research Center for Assisted Reproductive Technology and Reproductive Genetics, Shandong University, Jinan, Shandong 250012, China. [9]Peking University-Tsinghua University-National Institute of Biological Sciences Joint Graduate Program, School of Life Sciences, Tsinghua University, Beijing 100084, China. [10]Research Unit of Gametogenesis and Health of ART-Offspring, Chinese Academy of Medical Sciences, China. [11]Shanghai Key Laboratory for Assisted Reproduction and Reproductive Genetics, Shanghai 200135, China. [12]Center for Reproductive Medicine, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200135, China.
***Corresponding author. Email: hanzh80@yahoo.com (H.Z.); chenzijiang@hotmail.com (Z.-J.C.); xiewei121@tsinghua.edu.cn (W.X.)**
†These authors contributed equally to this work.

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

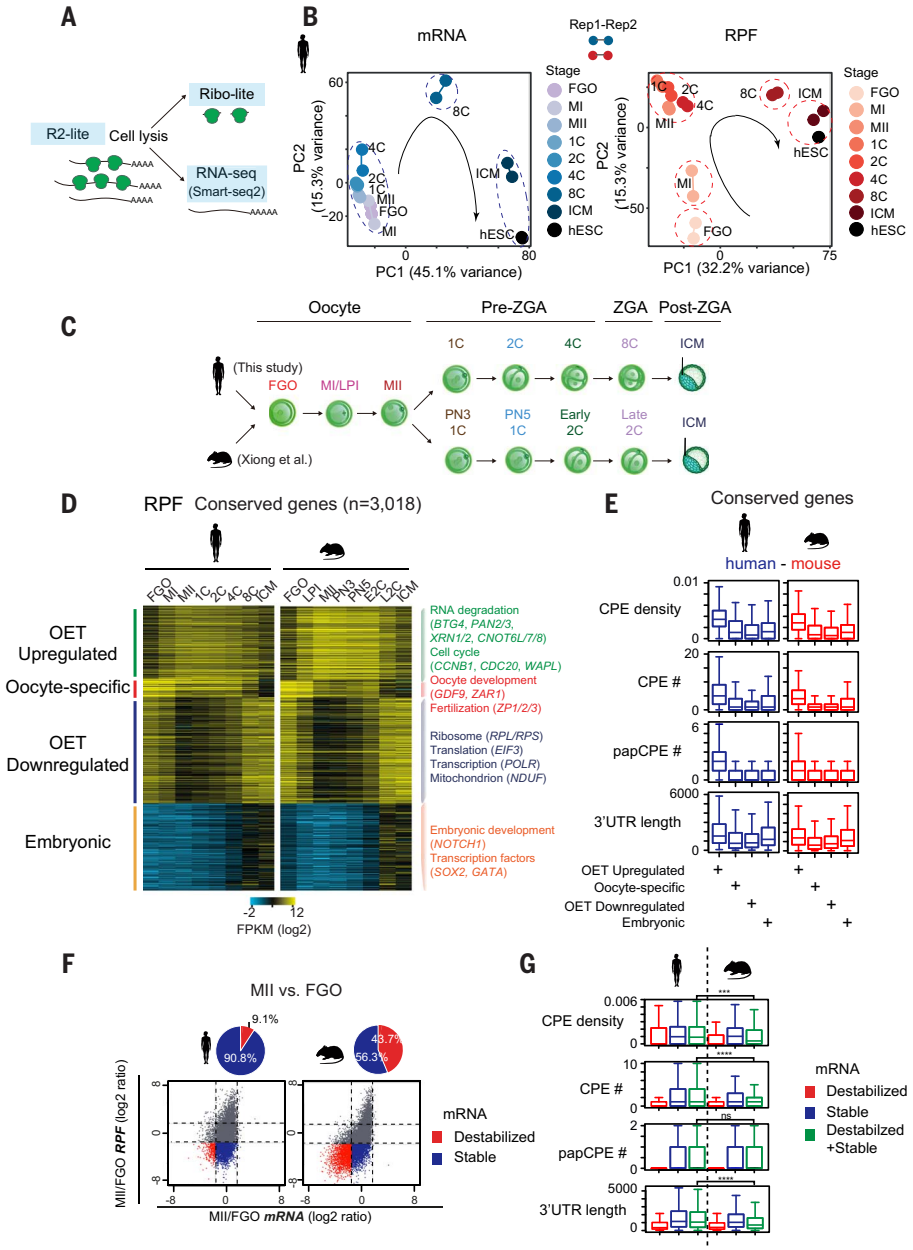


Fig. 1. Genome-wide mapping of translatome and transcriptome in human oocytes and mouse early embryos. (**A**) Schematic of R2-lite. (**B**) Principal component analysis (PCA) of transcriptomes and translatomes for human oocytes and early embryos. (**C**) Schematic of human and mouse oocytes and embryos used for comparison. (**D**) Heatmaps showing the RPF of human and mouse homologous genes with conserved translation patterns. The enriched Gene Ontology (GO) terms and genes are also listed. (**E**) Boxplots of CPE density (count per nucleotide), CPE number, papCPE number, and 3′UTR length (in nucleotides) of human-mouse conserved genes of each class in (D). (**F**) Top, pie charts showing the percentages of destabilized and stable maternal mRNAs among FGO-MII translationally down-regulated genes in human and mouse. Bottom, scatterplots comparing RPF fold change (log2 ratio) with mRNA fold change (log2 ratio) from FGO to MII oocytes in human and mouse. Dashed lines indicate a threefold change. (**G**) Boxplots of CPE density (count per nucleotide), CPE number, papCPE number, and 3′UTR length (in nucleotides) of all FGO-MII translationally down-regulated genes (green, sum of stable and destabilized maternal mRNA), destabilized (red), and stable (blue) mRNAs in human and mouse. *P* values (unpaired two-sample *t* test) are as follows: ***$P < 0.001$; ****$P < 0.0001$; ns, not significant.

comparable to the bulk HEK293 reference (*20*) and low-input hESCs. One of the "gold standards" of translation profiling in oocytes is "dormant RNAs," which show low translation efficiency (TE) in FGOs but become actively translated upon meiotic resumption (*1*, *3*). We collected a number of well-known dormant mRNAs reported in mouse, as well as potential dormant mRNAs from the proteome of human oocytes (*23*), including key regulators for meiotic maturation (*MOS*, *WEE2*, and *FBXO43*), maternal RNA clearance (*BTG4*, *CNOT6L/7/8*, and *DCP1A*), cell-cell adhesion (*LIMA1* and *CTNNB1*), mRNA splicing regulators (*WTAP* and *MAGOH*), a DNA methylation regulator (*TET3*), TF (*NFYB*), pyridoxal phosphate-binding protein (*ACCSL*), and epoxide hydrolase (*LTA4H*) (*18*, *23*). Most of these mRNAs exhibited translational repression in FGOs (low TE or RPF/mRNA) but became activated for translation in MII oocytes (fig. S4A), as supported by TE analysis (fig. S4, B and C), a result reproduced in replicates (fig. S4, A and B). This contrasted with their mRNA levels, which were much less dynamic (fig. S4D, "*BTG4*" and "*CNOT6L*"). The moderate increases of mRNAs from FGOs to MII oocytes were likely caused by increased capture efficiency of lengthened poly(A) tails upon meiotic resumption by mRNA-seq (*18*, *19*, *24*, *25*). Finally, we also identified differentially translated genes (DEGs) or RPF DEGs between every two consecutive stages, which revealed marker genes and gene functional enrichment that generally match the expected human developmental processes (fig. S5). For example, translation up-regulation was observed for the marker genes *ZSCAN4*, *ZSCAN5A/B/C*, and *DUXA/B* from 2C to 4C and *GDF3/PDGFRA/DNMT3L/POU5F1* from 8C to ICM. Therefore, R2-lite successfully captures transcriptional and translational activities in human oocytes and early embryos.

Dynamic human translatome in oocytes and early embryos

Human ZGA occurs in two waves: minor ZGA at 1C to 4C and major ZGA around 8C (*26*). For convenience, we use "ZGA" to refer to major ZGA unless otherwise noted. We found that the transcriptomes were similar from FGO to 4C, followed by a major switch around 8C (Fig. 1B). By contrast, translatomes were more dynamic, showing multiple transitions upon meiotic resumption (FGO to MI and MI to MII) and the onset of major ZGA (from 4C to 8C) (Fig. 1B and fig. S6A). At each stage, the translatome and transcriptome exhibited distinct landscapes in human oocyte and pre-ZGA embryos, but they tended to show more concordance after ZGA (fig. S6B). On the basis of their translation dynamics, we classified genes into four major classes (fig. S6C, left, and table S3). The first class, "OET up-regulated" genes, were translated at low levels in FGOs but at high levels in MII oocytes. Examples included those involved in mRNA destabilization (*BTG4*, *PAN2/3*, and *CNOT6/6L/7/8*) and protein ubiquitination (*FBX*s and *RNF*s). This was consistent with the notion that maternal mRNA and protein clearance are required for proper OET and ZGA (*2*). This class also included key cell-cycle genes (i.e., *CCNB1* and *CDC20*) that regulate meiotic resumption and progression. The second class were "oocyte-specific" genes. These

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

were highly translated from the FGO stage, but their translation was gradually down-regulated during OET, and they were barely translated after ZGA (e.g., *ZP1/2/3/4* and *GDF9*). The third class, "OET down-regulated" genes, were highly translated in FGOs, but their translation ceased in MII oocytes, before it was reactivated in early embryos. This class included genes that were involved in ribosome biogenesis (*RPS/RPL*), the mitochondrial respiratory chain (*NDUF* dehydrogenase), and transcription (*POLR*). The fourth class, "embryonic" genes, were transcribed and translated around or after ZGA and preferentially functioned in transcription and embryonic development. The corresponding mRNA levels for these four classes were generally correlated with the RPF levels, but there were differences (fig. S6C, arrows). Thus, translation undergoes dynamic regulation in human oocytes and early embryos.

Conservation of human and mouse translational landscapes during OET

We investigated to what extent gene expression during OET is conserved between human and mouse by comparing their translatomes and transcriptomes from the equivalent stages (*18*), first focusing on the homologous genes (Fig. 1C). The mouse data that we used to compare with the data from human MI oocytes were from a slightly earlier stage of late prometaphase I (LPI). We also grouped mouse genes into the same four classes on the basis of their translation dynamics (table S3). Overall, a substantial percentage of human genes in the four classes (51.9% of OET up-regulated, 29.7% of oocyte-specific, 61.1% of OET down-regulated, and 45.7% of embryonic) fell in the same classes in mouse (Fig. 1D and fig. S6C; see fig. S6D for examples). Together, they constituted 50.5% (3018 of 6106) of human genes from the four classes combined. The similar translation dynamics of these genes across species suggest strong evolutionary constraints because of their essential roles during OET (e.g., RNA degradation, oocyte development, translation, and embryonic development) (Fig. 1D).

Translation repression in FGOs and subsequent activation upon meiotic resumption are critically controlled by CPE through dynamic regulation of deadenylation and polyadenylation (*3*, *4*). CPEs are most effective when they are near PASs [PAS proximal CPEs (papCPEs)] within ~100 nt in mouse (*18*), consistent with previous studies (*6*, *27*, *28*). A similar trend was observed for humans: The translation of genes with papCPEs tended to be repressed in FGOs and reactivated in MII oocytes (fig. S7, A and B). CPEs within 100 to 150 nt of PASs were also effective but only when more than one was present, and the number of such genes was limited (fig. S7A). By contrast, the translation of genes with no CPEs was preferentially down-regulated during the same period; the translation of those with non-papCPEs remained largely constant (fig. S7, A and B). Globally, CPE densities, CPE counts, and papCPE numbers were highest for OET up-regulated transcripts and lowest for OET down-regulated transcripts both in human and mouse (Fig. 1E). It has been shown that short 3′UTRs promote deadenylation in zebrafish and mammalian embryos (*29*–*31*). Consistently, OET up-regulated and OET down-regulated transcripts showed longer and shorter 3′UTRs, respectively, in human and mouse (Fig. 1E). Therefore, the relationship between CPE and translation regulation is likely conserved between human and mouse.

Widespread maternal RNA destabilization occurs in mouse MII oocytes

Translation of the OET down-regulated genes was down-regulated from FGO to MII oocytes. This was frequently accompanied by decreases in mRNAs in mouse (1445 of 3305 genes, or 43.7%), but not in human (only 260 of 2829, or 9.1%) (Fig. 1F, also see fig. S6C, mRNA-OET down-regulated, compare mouse and human). We considered the transcripts with mRNA decrease to be "destabilized," which may reflect either the deadenylation or decay of RNA, and those that did not show evident mRNA decrease to be "stable." Indeed, two recent studies showed that although widespread deadenylation occurs from FGOs to MII oocytes in both species, human retains much longer poly(A) tails in MII oocytes compared with mouse in MII oocytes (*32*, *33*). Stable and destabilized genes were generally enriched for ribosomal protein and ATP-binding factors, respectively, and mitochondrion genes were enriched in both groups (fig. S8A).

We then investigated what factors might underlie such a differential deadenylation. Several key factors involved in decapping and deadenylation, such as *BTG4*, *CNOT6L/7/8*, and *PAN2/3*, but not *CNOT6*, were expressed more highly in mouse than in human (fig. S8B), although their true translation levels remained unknown because of the lack of spike-in calibration in the Ribo-lite data. The lower numbers of CPEs, especially papCPEs, in the 3′UTRs are associated with decreased translation and increased RNA instability and RNA deadenylation in mouse oocytes upon meiotic resumption (*18*). Consistently, genes belonging to the OET down-regulated class generally showed overall low CPE counts and short 3′UTRs both in mouse and human (Fig. 1E), but mouse genes showed even lower CPE counts, CPE densities, and shorter 3′UTRs (Fig. 1G). The destabilized genes in human (9.1%) also showed low CPE and papCPE counts and short 3′UTRs compared with stable genes, similar to those in mouse (Fig. 1G). These data suggest that for OET down-regulated genes, mouse tends to have fewer CPEs and shorter 3′UTRs than human, accompanied by more-deadenylated and less-stable RNAs in MII oocytes. A similar trend was observed for oocyte-specific genes (fig. S6C).

Human OET down-regulated genes were reactivated for translation before or during ZGA (Fig. 1D). For some genes, the resumed translation occurred before major ZGA, suggesting translation reactivation of existing mRNAs in anticipation of ZGA (fig. S8C, "pre-ZGA up"). We found that these genes were enriched for transcription regulators. Such pre-ZGA translation up-regulation did not simply arise from the minor ZGA, because the majority of the corresponding transcripts (81.7% for the stable group and 53.5% for the destabilized group) were insensitive to treatment with the transcription inhibitor α-amanitin (*24*) (fig. S8D). Furthermore, stable mRNAs were more likely to be up-regulated in pre-ZGA embryos than destabilized mRNAs in both human (54.4 versus 28.5%, respectively) and mouse (48.8 versus 37.5%, respectively) (fig. S8E), suggesting that their stability in oocytes may benefit their potential functions in early development and ZGA. Such translation up-regulation is relative to other genes in the genome because of the lack of the spike-in calibrated absolute measurement of translation by Ribo-seq for low-input samples (*18*).

CPE-mediated regulatory codes underline the divergence of human and mouse translatome during OET

Of human-mouse homologous genes, 49.5% showed species-specific translation dynamics (Fig. 2A). For example, the component for Polycomb repressive complex 2, *EED*, belonged to the OET up-regulated class in mouse but to the embryonic class in human (fig. S9A). Its presence in mouse but not in human pre-ZGA embryos was consistent with the propagation of oocyte H3K27me3 to early embryos as non-canonical imprints in mouse, but not human, embryos (*34*–*36*). Other divergent gene pairs included those involved in protein degradation, histone methyltransferases, and demethylases (fig. S9A). We also confirmed such species-specific translation dynamics using two additional mouse translatome profiling datasets (fig. S9, B and C).

We then investigated whether CPEs may underpin species-specific translation regulation. We identified 366 genes belonging to the human OET up-regulated class and the mouse OET down-regulated class. These genes tended to have higher CPE density and more papCPEs in human than in mouse (Fig. 2B). Conversely, genes that fell into human OET down-regulated and mouse OET up-regulated classes (N = 271) had lower CPE density and fewer papCPEs in human than in mouse (Fig. 2B). We then chose individual genes to assess the translational activities of 3′UTRs using green fluorescent protein (GFP) reporters. h*DROSHA* and m*Drosha* are human OET up-regulated

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

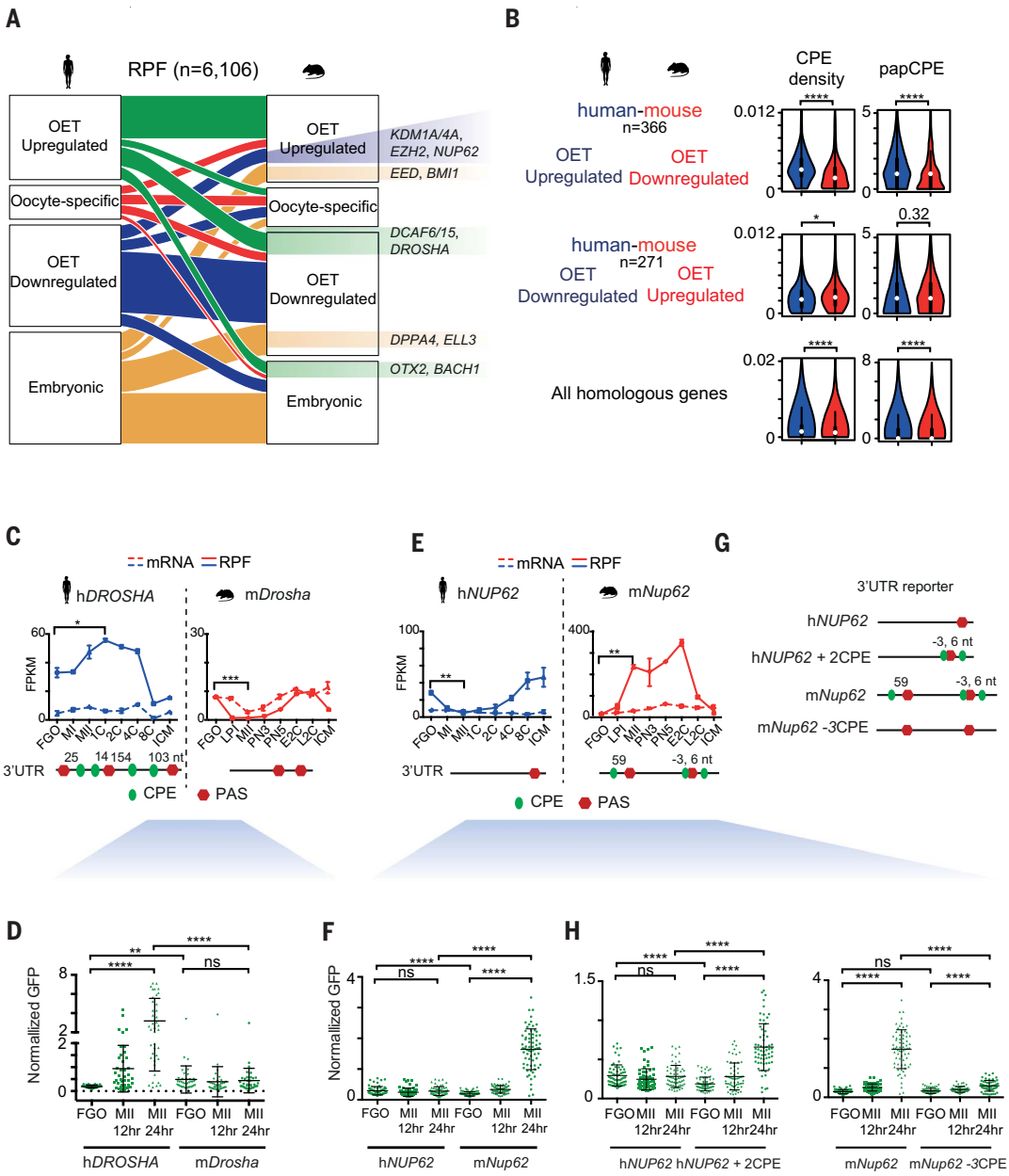


Fig. 2. Divergence of translatomes between human and mouse oocytes and early embryos. (**A**) Alluvial diagram of human-mouse homologous genes. (**B**) Violin plots showing CPE density (count per nucleotide) and papCPE number for human OET up-regulated/mouse OET down-regulated gene pairs, human OET down-regulated/mouse OET up-regulated gene pairs, and all homologous genes in human (blue) and mouse (red). (**C**) Line plots showing h*DROSHA* and m*Drosha* RPF and mRNA dynamics (mean with range). The CPE, PAS elements and their distances in h*DROSHA*/m*Drosha* 3′UTRs are shown below. (**D**) Normalized GFP intensity of h*DROSHA*/ m*Drosha* 3′UTR reporter in mouse oocytes. (**E**) Line plots showing h*NUP62* and m*Nup62* RPF and mRNA dynamics (mean with range). The CPE, PAS elements, and their distances in h*NUP62*/m*Nup62* 3′UTRs are shown below. The negative distance indicates that the two elements overlap. (**F**) Normalized GFP intensity of h*NUP62*/m*Nup62* 3′ UTR reporter in mouse oocytes. (**G**) Schematic of the CPE, PAS elements, and their distances for WT and mutant h*NUP62* or m*Nup62* 3′UTR. (**H**) Normalized GFP intensity of WT and mutant h*NUP62*/m*Nup62* 3′UTR reporters in mouse oocytes. *P* values (unpaired two-sample *t* test) for (B) to (H): **P* < 0.05; ***P* < 0.01; ****P* < 0.001; *****P* < 0.0001; ns, not significant.

and mouse OET down-regulated genes. Consistently, h*DROSHA*, but not m*Drosha*, contained papCPEs in its 3′UTR (Fig. 2C). This is consistent with DROSHA and its processing of target microRNA being dispensable for mouse oocyte development but possibly being functional in human oocytes (*37*–*39*). Of note, h*DROSHA* was already translated in FGOs, and its up-regulation of translation upon meiotic resumption was also accompanied by an increase in mRNA levels, likely caused by enhanced polyadenylation, leading to insignificant TE changes (fig. S10A). Nevertheless, when introducing 3′UTR-GFP reporters to FGOs followed by in vitro maturation to MII oocytes (fig. S10C), h*DROSHA*-reporter, but not m*Drosha*-reporter, was highly translated in mouse oocytes (Fig. 2D and fig. S10D). Conversely, h*NUP62* and m*Nup62* belonged to the human OET down-regulated and mouse OET up-regulated class, as confirmed by their TE dynamics (fig. S10B). Accordingly, three papCPEs are present in mouse *Nup62* 3′UTR but not in its human counterpart (Fig. 2E). *NUP62* encodes a central channel nucleoporin component and also functions in transcription and chromatin organization (*40*). When introducing a GFP reporter to mouse FGOs, followed by in vitro maturation toward the MII oocytes, the reporter with the m*NUP62* 3′UTR (m*NUP62*-reporter), but not the h*NUP62* 3′UTR (h*NUP62*-reporter), was highly translated in MII oocytes, suggesting again that the 3′UTRs explain the species-specific translation (Fig. 2F and fig. S10E). Moreover, the h*NUP62*-reporter became highly translated in mouse MII oocytes when it added papCPEs from the m*Nup62* 3′UTR. Conversely, deleting papCPEs from m*Nup62* 3′UTR impaired its translation activation (Fig. 2, G and H, and fig. S10E). Therefore, these data support the notion that distinct CPE usages underlie the differential translation between human and mouse.

Finally, we also identified genes that were translated in one species but silenced throughout OET in the other (fig. S11A). These genes were enriched in homeobox TFs and immunity. We also analyzed genes that were present and translated in one species but absent in the other (fig. S11B), among which a large number (N = 117 in human and N = 21 in mouse) were zinc finger proteins. KRAB zinc finger proteins are known to repress transposable elements and are highly divergent among species (*41*).

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

Another notable family is the PRD-like TF family, many members of which are present in human but not in mouse (*42*, *43*).

Pre-ZGA translation is essential for human ZGA

We then investigated how translation activity regulated embryonic development and ZGA. Consistent with the results in mouse (*18*, *44*–*46*), inhibition of translation by cycloheximide (CHX) in human 3 pronuclei (3PN) embryos (discarded during clinical application) starting from the 1C and 4C stage led to developmental arrest at the 1C to 4C stage and the 5C to 9C stage, respectively (Fig. 3A and fig. S12A, right). We cultured the treated embryos to day 5 (with the time of in vitro fertilization as day 0) and subjected them to RNA-seq. We observed severely impaired ZGA and maternal RNA clearance in embryos with CHX treatment from the 1C stage (Fig. 3B and fig. S12A, left). Although these defects were alleviated when the embryos were treated with CHX from the 4C stage, a substantial portion (24.2%) of major ZGA genes still failed to be properly activated (Fig. 3B and fig. S12A, left). Minor ZGA in CHX-treated 4C embryos largely recovered, consistent with these genes being transcribed earlier in embryos. Their expression became even higher than that of control on day 5 (Fig. 3B and fig. S12A, left), suggesting defects in timely silencing or decay of their transcripts. Thus, pre-ZGA translation is required for both human ZGA and maternal RNA clearance. Such translation activation from MII oocyte to 4C embryos mainly comes from maternal transcripts rather than from minor ZGA, because most of them were insensitive to α-amanitin treatment (*24*) (Fig. 3C).

PRD-like homeobox TFs are highly translated around ZGA

To identify specific ZGA regulators, we searched for TFs that were highly translated before and

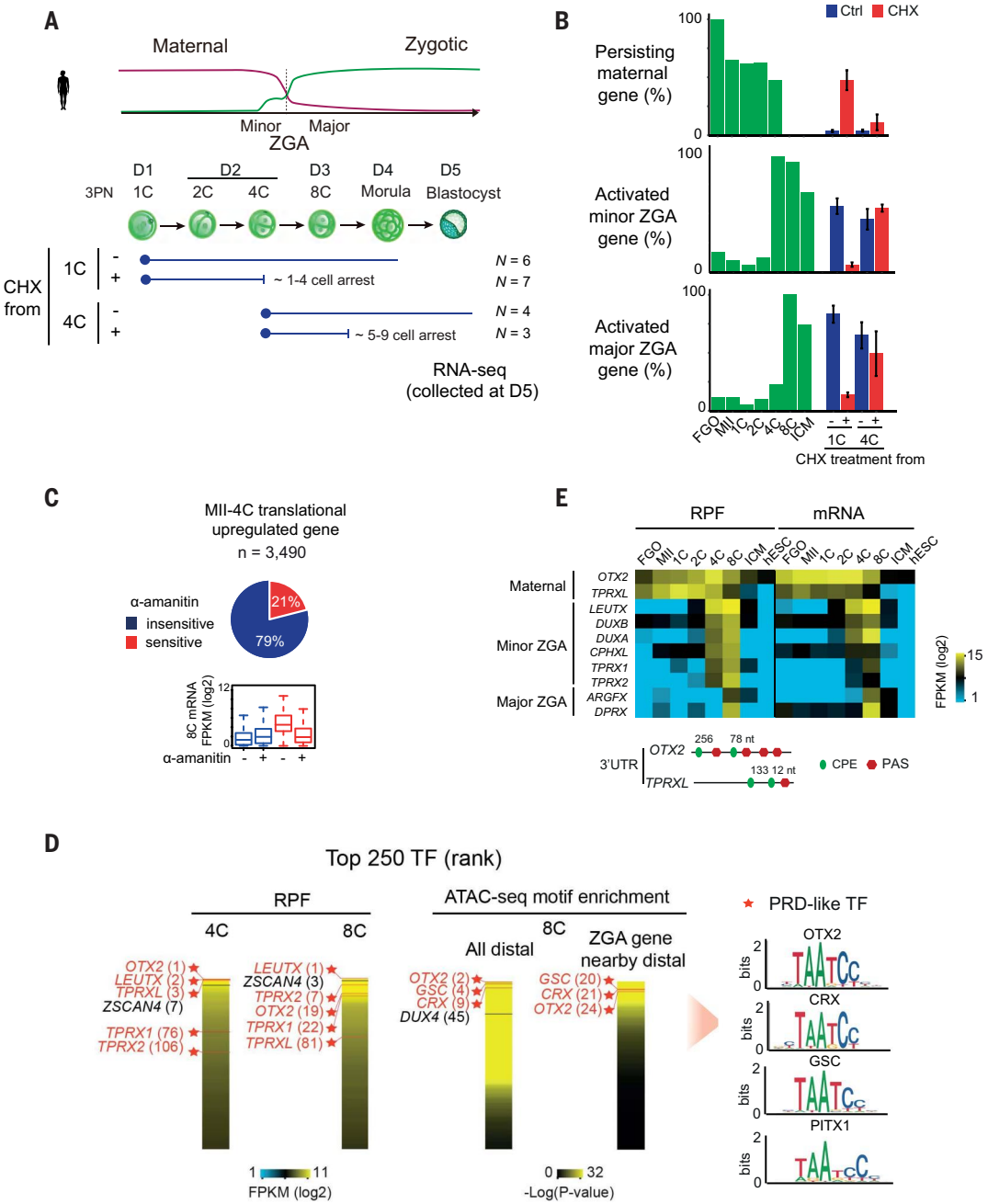


Fig. 3. Pre-ZGA translation is essential for human ZGA. (**A**) Schematic of the CHX treatment experiments in human early embryos. The red and green curves represent maternal and zygotic transcripts, respectively. (**B**) Percentages of persisting maternal genes (mRNA FPKM > 50% of FGO, top), activated minor ZGA genes (mRNA FPKM > 20% of 4C, middle), and major ZGA genes (mRNA FPKM > 20% of 8C, bottom) for the reference data (no treatment) (green), the DMSO group (blue), and the CHX group (red) on day 5 are shown. Some minor ZGA and major ZGA genes have nonzero FPKM in oocytes. (**C**) Pie chart (top) showing the proportions of α-amanitin sensitive and -insensitive genes in the MII-4C translational activated genes, based on the reported human 8C RNA-seq data (*24*). Boxplots show the corresponding mRNA levels (bottom). (**D**) Top 250 TFs based on the RPF at the 4C or 8C stage, and top 250 TF motifs enriched at all distal accessible regions or those near ZGA genes based on human 8C ATAC-seq data (*24*). TFs marked in red are PRD-like family proteins reported or predicated to bind the PRD-like motif TAATCC, as exemplified by the OTX2/CRX/GSC/PITX1 motifs. (**E**) Heatmap of RPF and mRNA levels for the three groups of PRD-like homeobox TFs. CPE and PAS in *OTX2* and *TPRXL* 3′UTRs are illustrated below.

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

around major ZGA. The top three TFs translated at the 4C stage were *OTX2*, *LEUTX*, and *TPRXL* (Fig. 3D), all belonging to the PRD-like homeobox TF family (*42*, *43*, *47*), which have a PRD class homeodomain but lack the PAIRED domain (*43*, *48*). At the 8C stage, two additional PRD-like homeobox TFs, *TPRX1* and *TPRX2*, became highly translated, consistent with their reported mRNA levels (*42*, *47*, *49*). There were at least 10 PRD-like homeobox TFs highly translated around ZGA (Fig. 3E). They can be classified into three groups on the basis of the timing of their translation. Maternal TFs included *OTX2* and *TPRXL*, which started to be translated upon oocyte meiotic resumption and began to decline after major ZGA (Fig. 3E). Consistently, *OTX2* and *TPRXL* both contain a papCPE in their 3′UTRs (Fig. 3E, bottom). Minor ZGA TFs included *LEUTX*, *DUXA*, *DUXB*, *TPRX1*, *TPRX2*, and *CPHXL*, because both their mRNA and their RPF levels started to increase at the 4C stage and culminated at the 8C stage. DUXA and DUXB, as DUX family members, share similar DNA-binding motifs with DUX4 (*50*). Finally, major ZGA TFs, including *ARGFX* and *DPRX*, were activated at the 8C stage (Fig. 3E). Overexpression studies in cell lines suggested a role for PRD-like homeobox TFs in regulating a subset of ZGA genes (*47*, *49*, *51*), although the exact roles in vivo remain largely unknown. They arose through gene duplication from an ancestor gene, *CRX*, followed by asymmetric sequence evolution (*47*). Moreover, the putative enhancers [distal accessible regions based on the assay for transposase-accessible chromatin sequencing (ATAC-seq) (*24*)] in human 8C embryos near major ZGA genes are enriched for TAATCC, a PRD-like TF-binding motif (*42*, *52*) [represented by OTX2, CRX, GSC, and PITX1 (table S4) in the HOMER motif database (*53*)] (Fig. 3D). These data are consistent with previous studies proposing that PRD-like homeobox TFs are involved in human ZGA (*42*, *43*, *47*, *49*).

TPRX factors regulate human preimplantation development

We then investigated whether PRD-like homeobox TFs play a role in human ZGA and early development. We prioritized the maternal TFs OTX2 and TPRXL and the minor ZGA TFs LEUTX (most translated at 4C to 8C) and TPRX1/2 (related to TPRXL), because they are more likely to function upstream of ZGA. We attempted to knock down these TFs in 3PN human embryos. Although they exhibit a lower developmental rate toward blastocysts (*54*), 3PN embryos have similar accessible chromatin and histone mark landscapes (*24*, *35*) and ZGA timing (*24*) (also see below) as 2PN embryos. Because we failed to efficiently deplete *OTX2*, and knocking down *LEUTX* had minor effects on ZGA and embryonic development (fig. S12B), we mainly focused on *TPRX1/2/L*.

TPRXL was suggested to arise from the reverse transcription product of *TPRX1* with further sequence acquisition (*43*) (Fig. 4A). *TPRXL* was once annotated as a noncoding gene with a predicted open reading frame (ORF) (*55*). However, Ribo-lite signals in human embryos were enriched on a shorter ORF (Fig. 4B), consistent with a separate study (*43*). To confirm its presence, we generated an antibody against TPRXL and confirmed its specificity (fig. S12C). Immunostaining showed that TPRXL was translated in oocytes and became strongly enriched in the nucleus in 2C to 8C embryos until the blastocyst stage (Fig. 4C). The small interfering RNA (siRNA) (morpholino was also added for *TPRXL*) against target genes or negative controls was injected into human 1C embryos in parallel (Fig. 4D). RNA-seq analysis confirmed highly efficient KD for all factors (mRNAs dropped to 3.3 to 25.8% of control on average) (Fig. 4E). We also validated the depletion of TPRX1 and TPRXL by immunofluorescence (TPRX2 antibody is not available) (fig. S12, D and E). On day 3 (the 8C stage), development was slightly or moderately delayed for *TPRX1* KD embryos, *TPRX2* KD embryos, *TPRX1/2* double KD embryos, and *TPRXL* KD embryos (Fig. 4F and fig. S13, A and B). Simultaneous KD of *TPRX1/2/L* [*TPRX* triple KD (TKD)] led to a severe developmental delay on day 3 (Fig. 4F and fig. S13B). Although 3PN embryos exhibited asynchronized development, such delay was reproduced in four batches (fig. S13C) and persisted to day 4 (morula stage) (Fig. 4F and fig. S13B) (two batches). We did not assess development beyond day 4 because of the low blastocyst rate even for the control 3PN embryos. Our data suggest that TPRXL and TPRX1/2 together regulate human early development.

TPRX factors are required for human ZGA

Single-embryo RNA-seq analysis showed that the global transcriptome of *TPRX* TKD embryos on day 3 was altered, because it fell between the 4C and 8C stages, lagging behind the control embryos along the developmental trajectory in a principal components analysis (Fig. 4G). By contrast, only small effects of global transcriptome were observed when knocking down *TPRX1*, *TPRX2*, or *LEUTX* alone, and moderate effects were observed when knocking down *TPRXL* or *TPRX1/2* (Fig. 4G and figs. S12B and S14, A and B). Detailed analysis showed that ZGA genes were predominantly down-regulated in *TPRX* TKD embryos, with ~31% of 8C-specifically activated genes and 36% of morula-specifically activated genes down-regulated at days 3 and 4, respectively (Fig. 5A). *TPRX1/2* double KD and *TPRXL* KD also led to activation failure of ~15 to 21% of major ZGA genes, whereas KD of *TPRX1* or *TPRX2* alone only down-regulated ~2% of major ZGA genes, suggesting functional redundancy among TPRX members (Fig. 5B and fig. S14, A and B). For the subsequent analyses, we mainly focused on the *TPRX* TKD group. Down-regulated ZGA genes (both minor and major) included *ZSCAN4*, *DUXB*, *DUXA*, *NANOGNB*, *DPPA4*, *GATA6*, *DPRX*, *ARGFX*, *DPRX*, *RBP7*, and *KLF5* (Fig. 5C and fig. S14B). DPPA4 was shown to function in activating 2C-like genes and developmental genes in mouse (*56*, *57*). GATA6 plays a critical role in regulating early lineage specification (*58*). Conversely, maternal transcripts were slightly up-regulated (fig. S14B), consistent with human ZGA playing a critical role in maternal RNA decay (*31*). These data suggest that TPRXL and TPRX1/2 both contribute to ZGA and may activate downstream TFs to initiate the transcription cascade.

In *TPRX* TKD embryos, the down-regulated, but not up-regulated, genes were enriched for the PRD-like binding motifs in their nearest distal open chromatin regions (putative enhancers) (Fig. 5D and fig. S14C), as determined based on the ATAC-seq dataset (*24*). Major ZGA genes that failed to be activated contain more PRD-like TF-binding motifs, especially in the nearest distal open chromatin regions, than those that were not affected (Fig. 5, A, right, compare red and blue bars, and E). Conversely, ZGA genes with PRD-like TF motifs presented in nearby distal open chromatin were more likely to be down-regulated compared with those without motifs (Fig. 5F). Such features were also evident in *TPRXL* and *TPRX1/2* KD embryos, and to a much lesser extent in *TPRX1* and *TPRX2* single KD embryos (Fig. 5F). As a control, the PRD-like motif enrichment was not evident for either affected or unaffected morula-specifically activated genes in *TPRX* TKD embryos (Fig. 5A, right), suggesting that the morula program defects might be indirectly caused by the aberrant 8C program. Furthermore, among genes down-regulated at 8C, 57% (124 of 217) were still not properly activated on day 4 (<50% of the expression level of both day 3 and day 4 control embryos), with some showing little expression as exemplified by *NANOGNB*, *GATA6*, *ARGFX*, *DPRX*, *DPPA4*, etc. (Fig. 5C). As controls, *LEUTX* (a minor ZGA gene) and *CTCF* [a major ZGA gene (*59*)] were activated normally. Furthermore, although the global transcriptome had clearly progressed from day 3 to day 4 toward morula for control embryos, *TPRX* TKD embryos appeared to be stuck near the 8C stage (Fig. 4G). In fact, even though two morula-like embryos appeared in *TPRX* TKD on day 4, their transcriptomes more resembled the 8C stage (Fig. 4G, arrows). Therefore, genes with the TPRX motif in nearby regions were preferentially affected in *TPRX* TKD embryos, suggesting that the ZGA defects were unlikely to be due to a general transcription defect or developmental delay.

To identify the possible TPRX-binding target genes, we overexpressed Flag-tagged TPRX factors in H9 hESCs (fig. S15A). *TPRX1* or *TPRX2*

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

overexpression up-regulates 679 and 597 genes [with a fold change ≥1.5 (adjusted $P < 0.05$) or a fold change ≥3]. These up-regulated genes were enriched for PRD-like motif in their associated distal open chromatin in human early embryos compared with down-regulated genes and random genes (fig. S15B). We observed strong binding of TPRX1 and TPRX2 in hESCs, but not in enhanced GFP (EGFP)–overexpressing control cells, using small-scale Tn5-associated chromatin cleavage with sequencing (Stacc-seq) (*60*) (Fig. 6A and fig. S15C). However, we failed to detect chromatin binding of TPRXL in hESCs under the same conditions. Both the TPRX1- and TPRX2-binding sites were strongly enriched for the expected PRD-like TF motifs (Fig. 6B and table S4). The binding of TPRX1 and TPRX2 correlated with each other genome wide, suggesting similar targeting preferences (fig. S15D). Genes down-regulated upon *TPRX* TKD in human embryos were preferentially bound by TPRX1/2 in hESCs compared with those that were up-regulated or not affected (Fig. 6C). TPRX1 and TPRX2 binding was found near the promoters of *DPPA3*, *ZSCAN4*, *KLF5*, and *DUXB* (Fig. 6A). In sum, TPRXs preferentially regulate ZGA genes harboring their binding motifs.

Ectopic expression of TPRXs activates key human ZGA marker genes

RNA-seq analysis in *TPRX*-overexpressing hESCs revealed that genes activated by TPRX1 ($N = 679$) and TPRX2 ($N = 597$) showed significant overlap ($N = 324$, $P < 2.2 \times 10^{-16}$) (Fig. 6D and fig. S16, A and B), again suggesting that TPRX1 and TPRX2 may function redundantly. Among these genes activated by both proteins, a small but significant number of genes ($N = 24$, $P = 3.0 \times 10^{-5}$) were minor or major ZGA genes such as *DUXB*, *HESX1*, *KLF5*, *RBP7*, *ZSCAN4*, etc. (Fig. 6, D and E). Half of these ZGA genes were also down-regulated in *TPRX* TKD embryos ($N = 12$, $P = 4.5 \times 10^{-3}$) (including *RBP7*, *KLF5*, and *ZSCAN4*). *DPPA3*, a marker for naïve hESCs (*61*) and a highly expressed maternal gene that may drive DNA hypomethylation in embryos (*62*, *63*), was strongly bound and activated by TPRX1 and TPRX2 (Fig. 6, A and E). As a control, *CTCF* was neither bound nor activated by TPRX1 or TPRX2 in hESCs (Fig. 6A). *DUX4* was also not activated (Fig. 6E). TPRX2 could also activate *TPRX1* but not vice versa (Fig. 6E). Weak but evident binding of both TPRX1 and TPRX2 was found at the promoter and putative enhancers of *TPRX1* (fig. S16C, shaded). TPRXL barely activated genes in primed hESCs, and overexpressing *TPRXL* in naïve hESCs resulted in increased but still limited gene activation ($N = 23$), which might have been caused by the lack of necessary cofactors in hESCs (Fig. 4A and fig. S16B). In sum, these data indicate that ectopic TPRXs can activate a subset of ZGA genes in hESCs.

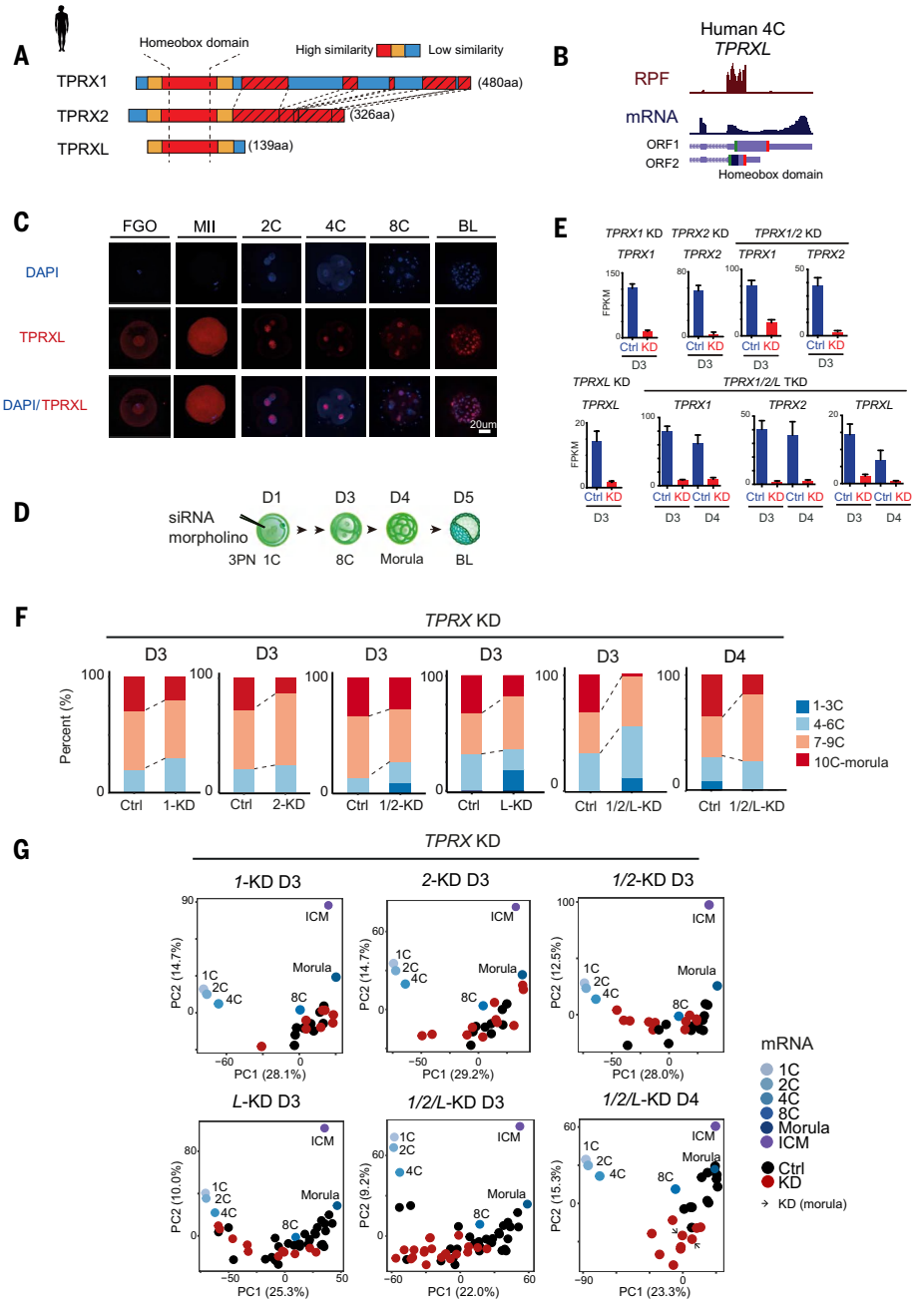


Fig. 4. TPRX factors are required for human embryonic development. (**A**) Schematic of protein sequence alignment of TPRXs based on the European Molecular Biology Laboratory–European Bioinformatics Institute multiple sequence alignment tool. Regions in red, yellow, and blue indicate high, medium, and low similarity, respectively. Slashed regions show high similarity between TPRX1 and TPRX2. (**B**) The University of California Santa Cruz (UCSC browser views of RPF and mRNA for *TPRXL* at the 4C stage. Two predicted ORFs are shown below. Dark blue shows homeobox domain. Green and red bars mark the start and stop codons, respectively. (**C**) TPRXL immunostaining in FGO; MII oocytes; 2C, 4C, 8C, and 3PN embryos; and blastocysts (BL). (**D**) Schematic of gene KD experiments. 3PN zygotes were injected by siRNA (for *LEUTX/TPRX1/2/L*) and morpholino (for *TPRXL*) or control siRNA and morpholino. RNA-seq samples were collected on day 3 (D3) or day 4 (D4). (**E**) Bar charts showing mRNA levels of *TPRX1/2/L* in control and KD embryos. The error bar denotes the SEM. (**F**) Developmental rates of *TPRX1* KD ($N = 10$), *TPRX2* KD ($N = 13$), *TPRX1/2* KD ($N = 12$), *TPRXL* KD ($N = 11$), and *TPRX1/2/L* TKD embryos collected on D3 ($N = 26$) or D4 ($N = 12$) after microinjection. *TPRX1* KD and *TPRX2* KD shared the same control ($N = 10$), and *TPRX1/2* KD shared the same control ($N = 15$) with *LEUTX* KD. *TPRXL* KD and *TPRX1/2/L* TKD shared the same control ($N = 34$) on D3. *TPRX1/2/L* TKD had its own control ($N = 14$) on D4. (**G**) PCA analysis of the transcriptome of 1C/2C/4C/8C/morula/ICM (without injection), KD and control embryos on D3 or D4. Arrested embryos with ZGA completely failed, which indicates abnormal development of low-quality embryos, were removed in transcriptome analysis. Two morulae appeared in *TPRX1/2/L* TKD embryos, yet their transcriptomes were still similar to those of 8C embryos (arrows).

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

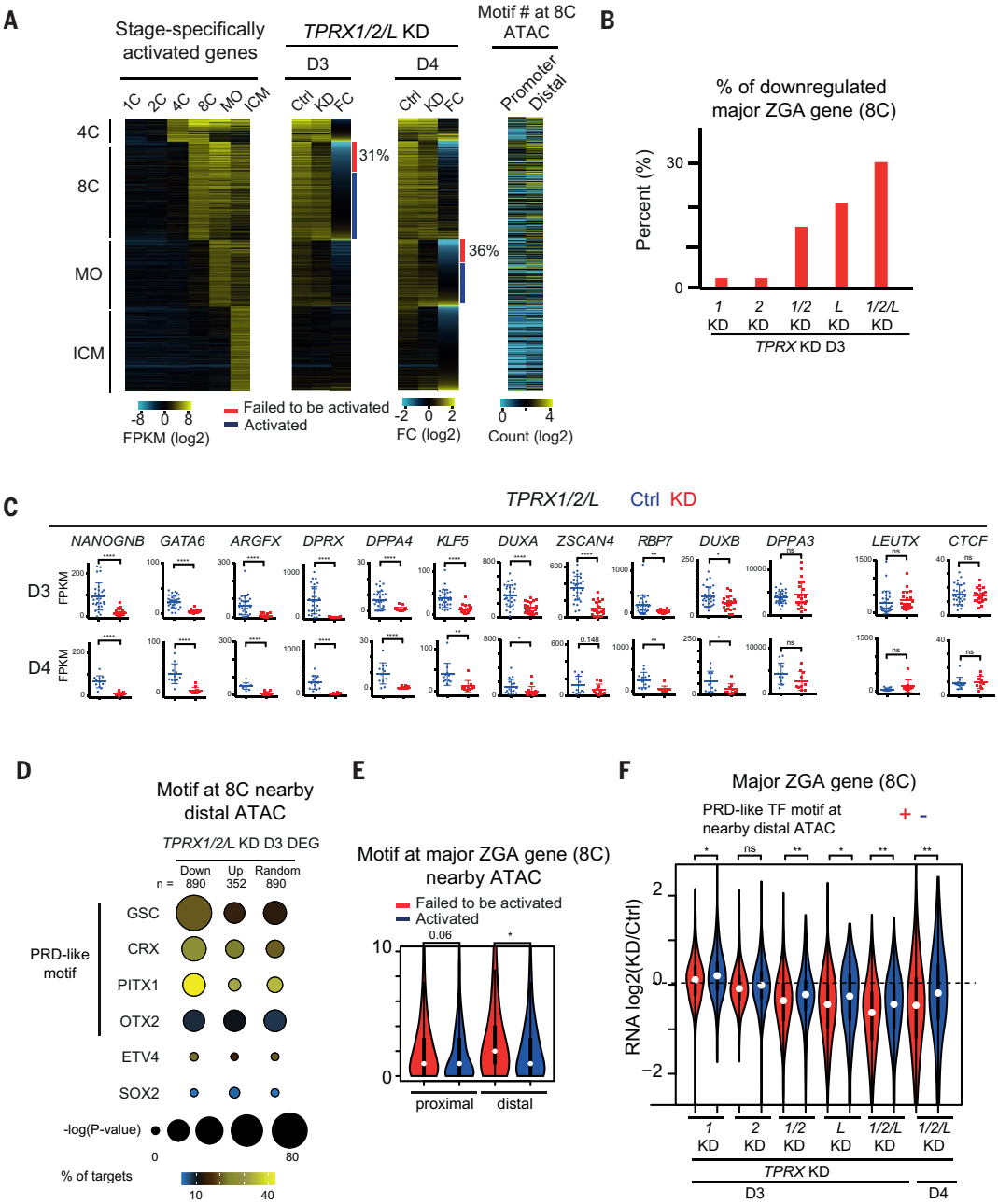


Fig. 5. TPRX factors are required for human ZGA. (**A**) Heatmaps showing the average mRNA levels of 4C to ICM stage-specifically activated ZGA genes in *TPRX1/2/L* TKD and control embryos on D3 and D4. PRD-like motif numbers at the 8C proximal or distal open chromatin region near each gene promoter are also shown. Genes failed to be activated and normally activated are labeled by the red and blue bars, respectively. MO, morula. FC, fold change. (**B**) Percentages of 8C activated major ZGA genes that failed to be activated in KD embryos on D3. (**C**) mRNA levels (single embryo) of example genes significantly down-regulated upon *TPRX1/2/L* TKD in control or TKD embryos. Genes not affected are shown as controls (*LEUTX* and *CTCF*). (**D**) Motif enrichment in the 8C nearby distal open regions [based on 8C ATAC-seq (*24*)] of DEGs upon *TPRX1/2/L* TKD on D3. (**E**) Violin plots showing the number of PRD-like motifs in 8C nearby proximal or distal open chromatin regions for genes that failed to be activated or were normally activated in *TPRX1/2/L* TKD embryos on D3. (**F**) Violin plots showing the mRNA fold changes (log2 ratio) for major ZGA genes with (red) or without (blue) a PRD-like TF motif at nearby putative enhancers in KD embryos on D3 and D4. *P* values (unpaired two-sample *t* test) for (C), (E), and (F): $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$; ns, not significant.

Discussion

Nearly half of human-mouse homologous genes showed distinct translation patterns during OET, highlighting species-specific regulation and cautioning against direct extrapolation of data from model organisms to human. Such differences likely reflect the evolutionary adaptation to species-specific needs. Host defensive responses to the transposon activity appear to be a major source of the differences between species, which is in agreement with the notion that the domestication of transposons varies among different species, and diverse ZNFs evolve in each species during "arms races" with transposons (*41*, *64*). Mouse, but not human, exhibits widespread maternal RNA destabilization in MII oocytes, which correlates with the more-extensive deadenylation in mouse (*32*, *33*). Given that these genes generally contain fewer CPEs and shorter 3′UTRs in mouse than in human, we propose that such early destabilization in mouse MII oocytes may be intrinsically coded in anticipation of an unusually early ZGA compared with other mammals. A substantial fraction of these genes undergo retranslation around ZGA in both human and mouse embryos, underscoring a potentially underappreciated role for recycling these transcripts for embryonic development. CPE configuration (especially papCPEs) appears to contribute to the differential translation between mouse and human oocytes. These data emphasize a critical role of 3′UTR diversity in regulating species-specific RNA translation and stability. Such variability may stem from gene rearrangements, local nucleotide changes, transposition of insertion sequences, or alternative polyadenylation (*65*).

PRD-like TFs have been implicated in human ZGA and preimplantation development based on expression timing, phylogenetic analysis, and overexpression experiments in hESCs (*42*, *43*, *47*, *49*). *TPRX1* was recently recognized as a marker of 8C-like cells in hESCs (*66*, *67*), and regulates transcription during the 8C-like cell conversion (*67*). Moreover, TPRXs do not appear to regulate *DUX4* in embryos and hESCs. Although ectopically expressed DUX4 was shown

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

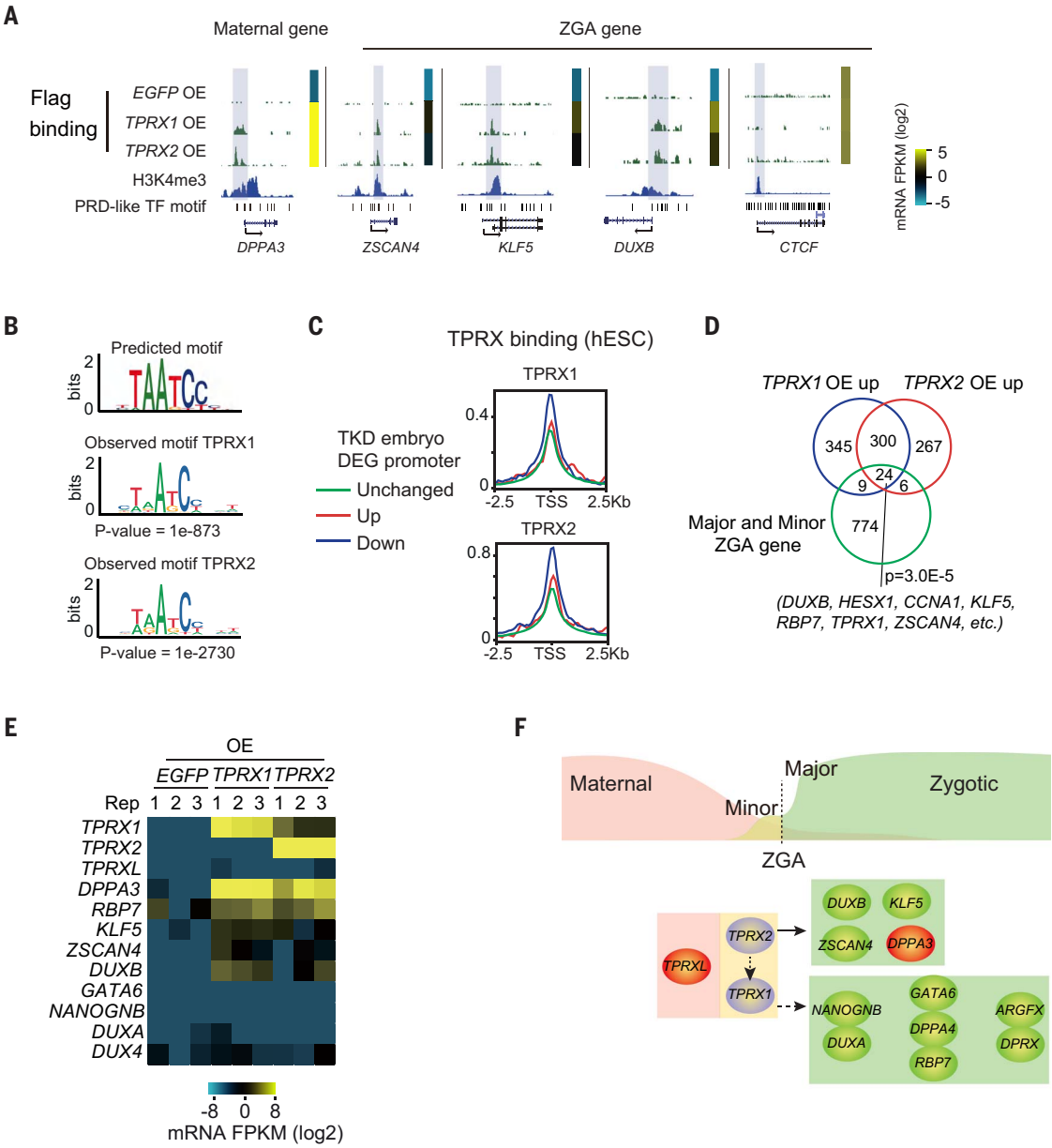


Fig. 6. Ectopic TPRXs activate key ZGA marker genes. (**A**) Stacc-seq signals near example genes in hESCs with ectopic expression of 3×Flag-EGFP, 3×Flag-TPRX1, and 3×Flag-TPRX2. H3K4me3 in the 8C embryo is also shown (*35*). The promoter regions are shaded. *CTCF* is a negative control that is neither bound nor activated by TPRX1/2. (**B**) Sequence logos of predicted and observed motifs for TPRX1/2. (**C**) Enrichment of TPRX1/2 binding in hESCs at the promoters of *TPRX1/2/L* TKD embryo DEGs. (**D**) Venn diagram showing the overlapping of ZGA genes (minor and major) and *TPRX1/2* OE up-regulated genes. The *P* value (Fisher's exact test) for the overlap between genes activated by both *TPRX1* and *TPRX2* overexpression (N = 324) and ZGA genes (minor and major) (N = 813) is shown. Examples of overlapped genes are shown below. (**E**) Heatmaps showing the mRNA levels of example genes upon *EGFP*, *TPRX1*, and *TPRX2* overexpression in hESCs. (**F**) Model illustrating the role of TPRXs in human ZGA. The pink, yellow, and green curves represent profiles of destabilized maternal transcripts and minor and major waves of ZGA transcripts, respectively. The maternal factor TPRXL and minor ZGA factors TPRX1/2 regulate human ZGA and early embryonic development. The solid arrow indicates the possible direct target genes that were bound and activated by TPRXs in hESCs and down-regulated in *TPRX* KD embryos, including *DUXB*, *ZSCAN4* (minor ZGA genes), *KLF5* (major ZGA genes), and *DPPA3*. The dotted arrows indicate potential target genes that were down-regulated in *TPRX1/2/L* TKD embryos but not found to be activated or bound by ectopic TPRX1/2 in hESCs, including *NANOGNB* and *DUXA* (minor ZGA genes) and *ARGFX*, *DPRX*, *GATA6*, *DPPA4*, and *RBP7* (major ZGA genes). Ectopic TPRX2 could bind and activate *TPRX1* in hESCs.

to increase the percentage of TPRX1-positive cells in the hESC population (*66*), it does not bind directly at *TPRX1* (*68*). We postulate that different TFs may work in parallel to activate genes with corresponding motifs. Future studies are warranted to understand the molecular mechanisms underlying their actions and interplay during ZGA. We envision that our study will pave the way for understanding the mechanisms underlying the cascade of transcription at the beginning of life, which will benefit human infertility research.

Methods summary

A detailed materials and methods section is provided in the supplementary materials. In brief, human studies were approved by the ethics boards of participating institutions. Human immature oocytes from in vitro fertilization or intracytoplasmic sperm injection treatments were clinically discarded and donated by 20- to 38-year-old women after signing informed consent. In vitro maturation of oocytes was conducted following protocols described previously (*69*). The oocyte and embryo vitrifications were performed as described previously (*70*). High-quality embryos were selected for this study according to morphological criteria (*71*).

R2-lite integrated Ribo-lite (*18*) and Smart-seq2 (*19*), as described previously. Briefly, oocytes and embryos were lysed in 20 μl of ice-cold Ribo-lite lysis buffer. Next, 2 μl of the lysed samples was subjected to Smart-seq2, and the remaining sample was subjected to Ribo-lite profiling. mRNA-seq data and Ribo-lite data were analyzed separately, as described previously with several modifications (*18*).

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

For the 3′UTR reporter assay, the 3′UTR sequences for *DROSHA*/*Drosha* and *NUP62*/*Nup62* were cloned from the human or mouse oocyte cDNA library. Reporter plasmid construction, mutation, and mRNA preparation were conducted as described previously with several modifications (*18*).

CHX treatment experiments were performed on human 3PN zygotes. Embryos were cultured to day 5 in the presence of dimethyl sulfoxide (DMSO) or CHX. The developmental phenotype was recorded with microscopy, and each embryo was subjected to Smart-seq2 library preparation. To investigate the function of *LEUTX* and *TPRX1/2/L* in human ZGA, the siRNAs and morpholinos targeting these genes were microinjected into 3PN zygotes. The injected embryos with normal morphology were harvested on days 3 and 4 after fertilization for Smart-seq2 library preparation.

Primed hESCs were maintained on Matrigel-coated plates in TeSR-E8 medium (STEMCELL Technologies). Naive-state hESCs were generated as described previously (*72*) with some modifications. For the construction of vectors containing the human *TPRX1*, *TPRX2*, and *TPRXL* genes, the coding sequences were synthesized or cloned from human 4C and 8C embryo cDNAs and inserted into the PiggyBac vector or the FUW lentivirus vector for doxycycline-inducible expression. Stacc-seq was used for Flag-tagged TPRX1/TPRX2/EGFP in hESCs with the wash steps, and the data were analyzed as described previously (*60*).

REFERENCES AND NOTES

1. R. M. Schultz, P. Stein, P. Svoboda, The oocyte-to-embryo transition in mouse: Past, present, and future. *Biol. Reprod.* **99**, 160–174 (2018). doi: 10.1093/biolre/ioy013; pmid: 29462259
2. V. Yartseva, A. J. Giraldez, The maternal-to-zygotic transition during vertebrate development: A model for reprogramming. *Curr. Top. Dev. Biol.* **113**, 191–232 (2015). doi: 10.1016/bs.ctdb.2015.07.020; pmid: 26358874
3. R. Mendez, J. D. Richter, Translational control by CPEB: A means to the end. *Nat. Rev. Mol. Cell Biol.* **2**, 521–529 (2001). doi: 10.1038/35080081; pmid: 11433366
4. J. D. Richter, CPEB: A life in translation. *Trends Biochem. Sci.* **32**, 279–285 (2007). doi: 10.1016/j.tibs.2007.04.004; pmid: 17481902
5. C. A. Fox, M. D. Sheets, M. P. Wickens, Poly(A) addition during maturation of frog oocytes: Distinct nuclear and cytoplasmic activities and regulation by the sequence UUUUUAU. *Genes Dev.* **3** (12b), 2151–2162 (1989). doi: 10.1101/gad.3.12b.2151; pmid: 2628165
6. M. Piqué, J. M. López, S. Foissac, R. Guigó, R. Méndez, A combinatorial code for CPE-mediated translational control. *Cell* **132**, 434–448 (2008). doi: 10.1016/j.cell.2007.12.038; pmid: 18267074
7. W. J. Racki, J. D. Richter, CPEB controls oocyte growth and follicle development in the mouse. *Development* **133**, 4527–4537 (2006). doi: 10.1242/dev.02651; pmid: 17050619
8. Q. Q. Sha *et al*., A MAPK cascade couples maternal mRNA translation and degradation to meiotic cell cycle progression in mouse oocytes. *Development* **144**, 452–463 (2017). pmid: 27993988
9. M. T. Lee, A. R. Bonneau, A. J. Giraldez, Zygotic genome activation during the maternal-to-zygotic transition. *Annu. Rev. Cell Dev. Biol.* **30**, 581–613 (2014). doi: 10.1146/annurev-cellbio-100913-013027; pmid: 25150012
10. P. G. Hendrickson *et al*., Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017). doi: 10.1038/ng.3844; pmid: 28459457
11. A. De Iaco *et al*., DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017). doi: 10.1038/ng.3858; pmid: 28459456
12. J. L. Whiddon, A. T. Langford, C. J. Wong, J. W. Zhong, S. J. Tapscott, Conservation and innovation in the DUX4-family gene network. *Nat. Genet.* **49**, 935–940 (2017). doi: 10.1038/ng.3846; pmid: 28459454
13. M. Guo *et al*., Precise temporal regulation of Dux is important for embryo development. *Cell Res.* **29**, 956–959 (2019). doi: 10.1038/s41422-019-0238-4; pmid: 31591446
14. Z. Chen, Y. Zhang, Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nat. Genet.* **51**, 947–951 (2019). doi: 10.1038/s41588-019-0418-7; pmid: 31133747
15. D. Bosnakovski, M. D. Gearhart, S. Ho Choi, M. Kyba, Dux facilitates post-implantation development, but is not essential for zygotic genome activation†. *Biol. Reprod.* **104**, 83–93 (2021). doi: 10.1093/biolre/ioaa179; pmid: 32997106
16. A. De Iaco, S. Verp, S. Offner, D. Grun, D. Trono, DUX is a non-essential synchronizer of zygotic genome activation. *Development* **147**, dev177725 (2020). pmid: 31806660
17. S. Vuoristo *et al*., DUX4 regulates oocyte to embryo transition in human. BioRxiv 732289 [Preprint] (2020); https://doi.org/10.1101/732289.
18. Z. Xiong *et al*., Ultrasensitive Ribo-seq reveals translational landscapes during mammalian oocyte-to-embryo transition and pre-implantation development. *Nat. Cell Biol.* **24**, 968–980 (2022). doi: 10.1038/s41556-022-00928-6; pmid: 35697785
19. S. Picelli *et al*., Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014). doi: 10.1038/nprot.2014.006; pmid: 24385147
20. N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, J. S. Weissman, The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012). doi: 10.1038/nprot.2012.086; pmid: 22836135
21. C. Zhang, M. Wang, Y. Li, Y. Zhang, Profiling and functional characterization of maternal mRNA translation during mouse maternal-to-zygotic transition. *Sci. Adv.* **8**, eabj3967 (2022). doi: 10.1126/sciadv.abj3967; pmid: 35108058
22. Y. Atlasi *et al*., The translational landscape of ground state pluripotency. *Nat. Commun.* **11**, 1617 (2020). doi: 10.1038/s41467-020-15449-9; pmid: 32238817
23. I. Virant-Klun, S. Leicht, C. Hughes, J. Krijgsveld, Identification of maturation-specific proteins by single-cell proteomics of human oocytes. *Mol. Cell. Proteomics* **15**, 2616–2627 (2016). doi: 10.1074/mcp.M115.056887; pmid: 27215607
24. J. Wu *et al*., Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* **557**, 256–260 (2018). doi: 10.1038/s41586-018-0080-8; pmid: 29720659
25. S. A. Harvey *et al*., Identification of the zebrafish maternal and paternal transcriptomes. *Development* **140**, 2703–2710 (2013). doi: 10.1242/dev.095091; pmid: 23720042
26. D. Jukam, S. A. M. Shariati, J. M. Skotheim, Zygotic genome activation in vertebrates. *Dev. Cell* **42**, 316–332 (2017). doi: 10.1016/j.devcel.2017.07.026; pmid: 28829942
27. X. X. Dai *et al*., A combinatorial code for mRNA 3′-UTR-mediated translational control in the mouse oocyte. *Nucleic Acids Res.* **47**, 328–340 (2019). doi: 10.1093/nar/gky971; pmid: 30335155
28. X. G. Luong, E. M. Daldello, G. Rajkovic, C. R. Yang, M. Conti, Genome-wide analysis reveals a switch in the translational program upon oocyte meiotic resumption. *Nucleic Acids Res.* **48**, 3257–3276 (2020). doi: 10.1093/nar/gkaa010; pmid: 31970406
29. Y. Mishima, Y. Tomari, Codon usage and 3′ UTR length determine maternal mRNA stability in zebrafish. *Mol. Cell* **61**, 874–885 (2016). doi: 10.1016/j.molcel.2016.02.027; pmid: 26990990
30. Q. Q. Sha *et al*., Characterization of zygotic genome activation-dependent maternal mRNA clearance in mouse. *Nucleic Acids Res.* **48**, 879–894 (2020). doi: 10.1093/nar/gkz1111; pmid: 31777931
31. Q. Q. Sha *et al*., Dynamics and clinical relevance of maternal mRNA clearance during the oocyte-to-embryo transition in humans. *Nat. Commun.* **11**, 4917 (2020). doi: 10.1038/s41467-020-18680-6; pmid: 33004802
32. Y. Liu, H. Nie, C. Zhang, Z. Hou, J. Wang, F. Lu, Poly(A) tail length is a major regulator of maternal gene expression during the mammalian oocyte-to-embryo transition. BioRxiv 2021.08.29.458052 [Preprint] (2021); https://doi.org/10.1101/2021.08.29.458052.
33. Y. Liu, K. Wu, F. Shao, H. Nie, J. Zhang, C. Li, Z. Hou, J. Wang, B. Zhou, H. Zhao, F. Lu, Dynamics of poly(A) tail length and non-A residues during the human oocyte-to-embryo transition. bioRxiv 2021.08.29.458075 [Preprint] (2021); https://doi.org/10.1101/2021.08.29.458075
34. H. Zheng *et al*., Resetting epigenetic memory by reprogramming of histone modifications in mammals. *Mol. Cell* **63**, 1066–1079 (2016). doi: 10.1016/j.molcel.2016.08.032; pmid: 27635762
35. W. Xia *et al*., Resetting histone modifications during human parental-to-zygotic transition. *Science* **365**, 353–360 (2019). doi: 10.1126/science.aaw5118; pmid: 31273069
36. Z. Chen, Y. Zhang, Maternal H3K27me3-dependent autosomal and X chromosome imprinting. *Nat. Rev. Genet.* **21**, 555–571 (2020). doi: 10.1038/s41576-020-0245-9; pmid: 32514155
37. J. Ma *et al*., MicroRNA activity is suppressed in mouse oocytes. *Curr. Biol.* **20**, 265–270 (2010). doi: 10.1016/j.cub.2009.12.042; pmid: 20116252
38. P. Paloviita *et al*., Small RNA expression and miRNA modification dynamics in human oocytes and early embryos. *Genome Res.* **31**, 1474–1485 (2021). doi: 10.1101/gr.268193.120; pmid: 34340992
39. N. Suh *et al*., MicroRNA function is globally suppressed in mouse oocytes and early embryos. *Curr. Biol.* **20**, 271–277 (2010). doi: 10.1016/j.cub.2009.12.044; pmid: 20116247
40. B. Kalverda, H. Pickersgill, V. V. Shloma, M. Fornerod, Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm. *Cell* **140**, 360–371 (2010). doi: 10.1016/j.cell.2010.01.011; pmid: 20144760
41. G. Ecco, M. Imbeault, D. Trono, KRAB zinc finger proteins. *Development* **144**, 2719–2729 (2017). doi: 10.1242/dev.132605; pmid: 28765213
42. V. Töhönen *et al*., Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nat. Commun.* **6**, 8207 (2015). doi: 10.1038/ncomms9207; pmid: 26360614
43. H. A. Booth, P. W. Holland, Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. *Gene* **387**, 7–14 (2007). doi: 10.1016/j.gene.2006.07.034; pmid: 17005330
44. F. Aoki, K. T. Hara, R. M. Schultz, Acquisition of transcriptional competence in the 1-cell mouse embryo: Requirement for recruitment of maternal mRNAs. *Mol. Reprod. Dev.* **64**, 270–274 (2003). doi: 10.1002/mrd.10227; pmid: 12548659
45. Q. Wang, K. E. Latham, Requirement for protein synthesis during embryonic genome activation in mice. *Mol. Reprod. Dev.* **47**, 265–270 (1997). doi: 10.1002/(SICI)1098-2795(199707)47:3<265::AID-MRD5>3.0.CO;2-J; pmid: 9170106
46. Q. Wang, Y. G. Chung, W. N. deVries, M. Struwe, K. E. Latham, Role of protein synthesis in the development of a transcriptionally permissive state in one-cell stage mouse embryos. *Biol. Reprod.* **65**, 748–754 (2001). doi: 10.1095/biolreprod65.3.748; pmid: 11514337
47. I. Maeso *et al*., Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. *BMC Biol.* **14**, 45 (2016). doi: 10.1186/s12915-016-0267-0; pmid: 27296695
48. T. R. Bürglin, M. Affolter, Homeodomain proteins: An update. *Chromosoma* **125**, 497–521 (2016). doi: 10.1007/s00412-015-0543-8; pmid: 26464018
49. E. Madissoon *et al*., Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Sci. Rep.* **6**, 28995 (2016). doi: 10.1038/srep28995; pmid: 27412763
50. A. Leidenroth, J. E. Hewitt, A family history of DUX4: Phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. *BMC Evol. Biol.* **10**, 364 (2010). doi: 10.1186/1471-2148-10-364; pmid: 21110847
51. E. M. Jouhilahti *et al*., The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation. *Development* **143**, 3459–3469 (2016). doi: 10.1242/dev.134510; pmid: 27578796
52. S. Katayama *et al*., Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis. *Sci. Rep.* **8**, 17421 (2018). doi: 10.1038/s41598-018-35547-5; pmid: 30479355
53. S. Heinz *et al*., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010). doi: 10.1016/j.molcel.2010.05.004; pmid: 20513432
54. H. Y. Zhu *et al*., Histone demethylase KDM4A overexpression improved the efficiency of corrected human tripronuclear zygote development. *Mol. Hum. Reprod.* **27**, gaab012 (2021). doi: 10.1093/molehr/gaab012; pmid: 33599278

Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

55. T. Ota *et al*., Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36**, 40–45 (2004). doi: 10.1038/ng1285; pmid: 14702039
56. M. Eckersley-Maslin *et al*., Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes Dev.* **33**, 194–208 (2019). doi: 10.1101/gad.321174.118; pmid: 30692203
57. A. De Iaco, A. Coudray, J. Duc, D. Trono, DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. *EMBO Rep.* **20**, e47382 (2019). doi: 10.15252/embr.201847382; pmid: 30948459
58. K. K. Niakan, K. Eggan, Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev. Biol.* **375**, 54–64 (2013). doi: 10.1016/j.ydbio.2012.12.008; pmid: 23261930
59. X. Chen *et al*., Key role for CTCF in establishing chromatin structure in human embryos. *Nature* **576**, 306–310 (2019). doi: 10.1038/s41586-019-1812-0; pmid: 31801998
60. B. Liu *et al*., The landscape of RNA Pol II binding reveals a stepwise transition during ZGA. *Nature* **587**, 139–144 (2020). doi: 10.1038/s41586-020-2847-y; pmid: 33116310
61. W. A. Pastor *et al*., Naive human pluripotent cells feature a methylation landscape devoid of blastocyst or germline memory. *Cell Stem Cell* **18**, 323–329 (2016). doi: 10.1016/j.stem.2016.01.019; pmid: 26853856
62. Y. Li *et al*., Stella safeguards the oocyte methylome by preventing de novo methylation mediated by DNMT1. *Nature* **564**, 136–140 (2018). doi: 10.1038/s41586-018-0751-5; pmid: 30487604
63. C. B. Mulholland *et al*., Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. *Nat. Commun.* **11**, 5972 (2020). doi: 10.1038/s41467-020-19603-1; pmid: 33235224
64. M. Bruno, M. Mahgoub, T. S. Macfarlan, The arms race between KRAB–Zinc finger proteins and endogenous retroelements and its impact on mammals. *Annu. Rev. Genet.* **53**, 393–416 (2019). doi: 10.1146/annurev-genet-112618-043717; pmid: 31518518
65. P. Menendez-Gil *et al*., Differential evolution in 3'UTRs leads to specific gene expression in Staphylococcus. *Nucleic Acids Res.* **48**, 2544–2563 (2020). doi: 10.1093/nar/gkaa047; pmid: 32016395
66. J. Taubenschmid-Stowers *et al*., 8C-like cells capture the human zygotic genome activation program in vitro. *Cell Stem Cell* **29**, 449–459.e6 (2022). doi: 10.1016/j.stem.2022.01.014; pmid: 35216671
67. M. A. Mazid *et al*., Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. *Nature* **605**, 315–324 (2022). doi: 10.1038/s41586-022-04625-0; pmid: 35314832
68. Z. Yao *et al*., DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. *Hum. Mol. Genet.* **23**, 5342–5352 (2014). doi: 10.1093/hmg/ddu251; pmid: 24861551
69. H. Shirasawa *et al*., Retrieval and in vitro maturation of human oocytes from ovaries removed during surgery for endometrial carcinoma: A novel strategy for human oocyte research. *J. Assist. Reprod. Genet.* **30**, 1227–1230 (2013). doi: 10.1007/s10815-013-0040-z; pmid: 23842746
70. T. K. Roy, C. K. Bradley, M. C. Bowman, S. J. McArthur, Single-embryo transfer of vitrified-warmed blastocysts yields equivalent live-birth rates and improved neonatal outcomes compared with fresh transfers. *Fertil. Steril.* **101**, 1294–1301 (2014). doi: 10.1016/j.fertnstert.2014.01.046; pmid: 24582521
71. T. Baczkowski, R. Kurzawa, W. Głabowski, Methods of embryo scoring in in vitro fertilization. *Reprod. Biol.* **4**, 5–22 (2004). pmid: 15094792
72. T. W. Theunissen *et al*., Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 471–487 (2014). doi: 10.1016/j.stem.2014.07.002; pmid: 25090446

ACKNOWLEDGMENTS

We thank members of the Xie laboratory and the Chen laboratory for discussions and comments during the preparation of the manuscript; the Animal Center and Biocomputing Facility at Tsinghua University for support; R. Méndez for kindly providing scripts for the CPE analysis; J. Na for generously providing the primed H9 and H1 hESCs; and W. Liu for sharing the naïve hESC culture protocol. **Funding:** This work was supported by the National Natural Science Foundation of China (grant 31988101 to W.X. and Z.C., grant 31725018 to W.X., and grants 82192874 and 31871509 to H.Z.); the National Key R&D Program of China (grants 2021YFA1100102 and 2019YFA0508901 to W.X. and grant 2021YFC2700300 to K.W.); the Research Unit of Gametogenesis and Health of ART-Offspring, Chinese Academy of Medical Sciences (grant 2020RU001 to Z.C.); Innovative Research Team of High-Level Local Universities in Shanghai (grant SHSMU-ZLCX20210200 to Z.C. and grant SHSMU-ZLCX20210201 to K.W.); the Tsinghua-Peking Center for Life Sciences (W.X.); Fundamental Research Funds of Shandong University; and the Beijing Municipal Science and Technology Commission (grant Z181100001318006 to W.X.). W.X. is the recipient of a Howard Hughes Medical Institute International Research Scholar award. **Author contributions:** H.Z., Z.C., and W.X. conceived and designed the project. Z.H., C.Z., and K.W. collected human oocytes and embryos with help from W.T., X.M., and M.L. K.W. and C.Z. performed human 3PN zygote injection, embryo IF, and CHX treatment. Z.Z. and Z.X. developed R2-lite. Z.Z. performed R2-lite experiments with help from Z.X. and Q.W. Z.Z., Q.W., and C.Z. performed RNA-seq. Q.W. performed TF overexpression and Stacc-seq in hESCs. Z.Z., Q.W., and F.K. conducted the reporter experiments. Z.Z., Q.W., and C.Z. analyzed the data with help from Z.X. and B.L. Z.Z., Q.W., C.Z., and W.X. prepared the figures and wrote the manuscript with help from all authors. H.Z., Z.C., and W.X. supervised related experiments and the project. **Competing interests:** A patent for R2-lite has been filed by W.X., Z.X., and Z.Z. The remaining authors declare no competing interests. **Data and materials availability:** The sequencing data have been deposited to the Gene Expression Omnibus (GEO) with accession number GSE197265 and to the Genome Sequence Archive for human (GSA-Human) with accession number HRA002463.

License information: Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abo7923
Materials and Methods
Supplementary Text
Figs. S1 to S16
Tables S1 to S4
References (*73*–*93*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 28 February 2022; accepted 25 August 2022
Published online 8 September 2022
10.1126/science.abo7923

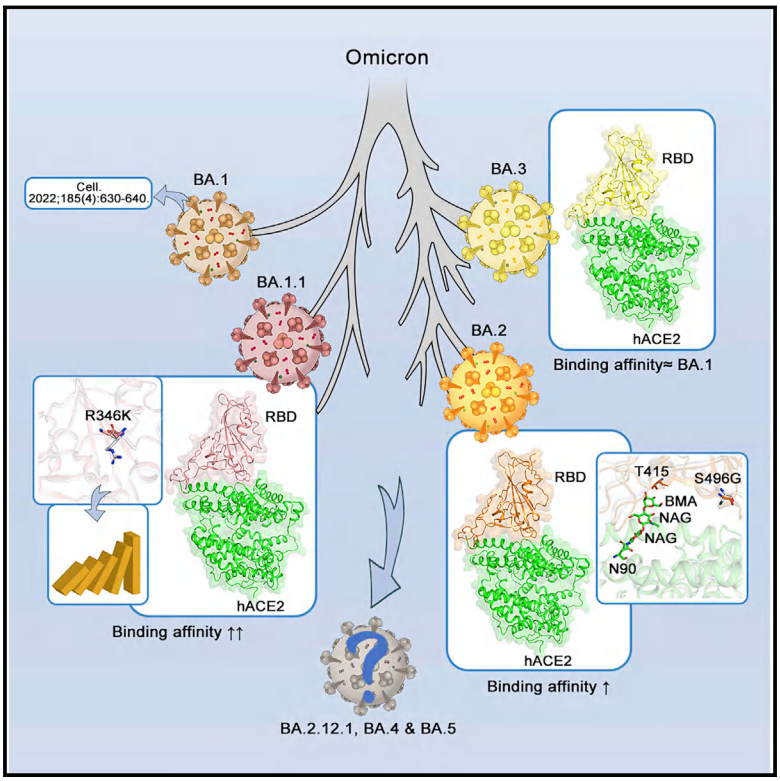
Downloaded from https://www.science.org at Institute of Biophysics,CAS on October 09, 2025

Cell

Article

Structural basis of human ACE2 higher binding affinity to currently circulating Omicron SARS-CoV-2 sub-variants BA.2 and BA.1.1

Graphical abstract



Highlights

- Omicron BA.1.1 and BA.2 show higher binding strength to hACE2 than the prototype and BA.1
- Details in the binding interface of BA.1.1, BA.2, and BA.3.RBD with hACE2 are deciphered
- R346K in BA.1.1 RBD enhances the interaction with hACE2 through long-range alterations

Authors

Linjie Li, Hanyi Liao, Yumin Meng, ..., Niu Huang, Jianxun Qi, George Fu Gao

Correspondence

zhaoxin@im.ac.cn (X.Z.), sunyeping@im.ac.cn (Y.S.), huangniu@nibs.ac.cn (N.H.), gaof@im.ac.cn (G.F.G.)

In brief

The biochemical and structural analysis of the human angiotensin-converting enzyme-2 (ACE2, the receptor for SARS-CoV-2 viral entry) and the receptor-binding domain (RBD) of four Omicron sub-variants—BA.1, BA.1.1, BA.2, and BA.3—helps to reveal the structural basis of differences in sub-variant binding affinities and the impact of RBD mutations.

Li et al., 2022, Cell *185*, 2952–2960
August 4, 2022 © 2022 Elsevier Inc.
https://doi.org/10.1016/j.cell.2022.06.023

CellPress

CellPress

Cell

Article

Structural basis of human ACE2 higher binding affinity to currently circulating Omicron SARS-CoV-2 sub-variants BA.2 and BA.1.1

Linjie Li,[1,2,6] Hanyi Liao,[1,2,6] Yumin Meng,[1,2,6] Weiwei Li,[1,2,6] Pengcheng Han,[1,3,6] Kefang Liu,[1] Qing Wang,[4] Dedong Li,[1] Yanfang Zhang,[1] Liang Wang,[1] Zheng Fan,[1] Yuqin Zhang,[1] Qiyue Wang,[1] Xin Zhao,[1,*] Yeping Sun,[1,*] Niu Huang,[4,5,*] Jianxun Qi,[1,2] and George Fu Gao[1,2,7,*]

[1]CAS Key Laboratory of Pathogen Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]School of Medicine, Zhongda Hospital, Southeast University, Nanjing 210009, China

[4]National Institute of Biological Sciences, No. 7 Science Park Road, Zhongguancun Life Science Park, Beijing 102206, China

[5]Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing 102206, China

[6]These authors contributed equally

[7]Lead contact

*Correspondence: zhaoxin@im.ac.cn (X.Z.), sunyeping@im.ac.cn (Y.S.), huangniu@nibs.ac.cn (N.H.), gaof@im.ac.cn (G.F.G.)

https://doi.org/10.1016/j.cell.2022.06.023

SUMMARY

The currently circulating Omicron sub-variants are the SARS-CoV-2 strains with the highest number of known mutations. Herein, we found that human angiotensin-converting enzyme 2 (hACE2) binding affinity to the receptor-binding domains (RBDs) of the four early Omicron sub-variants (BA.1, BA.1.1, BA.2, and BA.3) follows the order BA.1.1 > BA.2 > BA.3 ≈ BA.1. The complex structures of hACE2 with RBDs of BA.1.1, BA.2, and BA.3 reveal that the higher hACE2 binding affinity of BA.2 than BA.1 is related to the absence of the G496S mutation in BA.2. The R346K mutation in BA.1.1 majorly affects the interaction network in the BA.1.1 RBD/hACE2 interface through long-range alterations and contributes to the higher hACE2 affinity of the BA.1.1 RBD than the BA.1 RBD. These results reveal the structural basis for the distinct hACE2 binding patterns among BA.1.1, BA.2, and BA.3 RBDs.

INTRODUCTION

Omicron (B.1.1.529), which is the SARS-CoV-2 variant with most mutations, was defined as the fifth variant of concern (VOC) by the World Health Organization (WHO) on November 26, 2021 (Kandeel et al., 2022). The Omicron variants were recently divided into four sub-variants, i.e., BA.1, BA.1.1, BA.2, and BA.3 (PAHO and WHO, 2022). The BA.2 sub-variant was suggested to have a growth advantage over other circulating variants and is more transmissible than the BA.1 sub-variant. Statistical analysis shows that the effective reproduction number of BA.2 is 1.4-fold higher than that of BA.1 (Yamasoba et al., 2022). During the revision of this manuscript, some more Omicron sub-variants are emerging, e.g., BA.2.12.1, BA.4, BA.5, etc.

The growth advantage of BA.2 was thought to be related to its higher capacity for immune escape. However, the difference between the level of neutralizing antibodies against BA.1 and BA.2 in individuals who accepted vaccination or experienced SARS-CoV-2 infection was small, so it is difficult to explain the higher transmissibility of BA.2 than BA.1 (Bruel et al., 2022; Callaway, 2022; Iketani et al., 2022; Yu et al., 2022).

SARS-CoV-2 binding to its receptor, human angiotensin-converting enzyme 2 (hACE2), is a critical step for initiating the infection. The binding of SARS-CoV-2 depends on the interactions between hACE2 and the receptor-binding domain (RBD) in the spike (S) protein of the virus. There are 15–16 amino acid (aa) mutations in the RBD of Omicron sub-variants compared with the prototype (PT) virus (Lu et al., 2020b; Tan et al., 2020; Wei et al., 2020).

The RBD of the S protein is also the major target of neutralizing antibodies (Dai and Gao, 2021; Peng et al., 2021; Xu et al., 2021). Previously, we solved both the crystal and cryo-EM structures of the Omicron BA.1 RBD in complex with hACE2, revealing that although some mutations in the BA.1 RBD increase the affinity to hACE2, the others attenuate it. Consequently, the affinity of Omicron BA.1 RBD is similar to that of the PT strain RBD (Han et al., 2022). In this report, we evaluated the binding properties of BA.1.1, BA.2, and BA.3 RBDs to hACE2 and explored the structural basis for the difference in their binding affinities.

 © 2022 Elsevier Inc.

CellPress

RESULTS

Affinity of RBDs from the SARS-CoV-2 variants to hACE2 and entry efficiency of the pseudoviruses

Among the SARS-CoV-2 RBDs, there are 15 aa mutations in BA.1 and 16 in BA.1.1 with an extra mutation at R346K. BA.2 and BA.3 contain 16 and 15 aa mutations, correspondingly (Figure 1A). As of April 28, 2022, there have been 75 Pango Omicron sub-variants. Among these sub-variants, BA.1.1 is the globally dominant Omicron sub-variant, accounting for approximately 26.62% (874,168) of the total Omicron genomes. The other top 10 sub-variants are BA.2.9, BA.2, BA.1.17.2, BA.1.15, BA.1.16, BA.1, BA.1.1.1, BA.2.10, and BA.1.14, accounting for approximately 10.55% (346,329), 8.83% (290,135), 6.50% (213,581), 6.34% (208,159), 4.80% (157,468), 4.59% (150,661), 2.73% (89,652), 2.46% (80,763), and 2.37% (77,890) of the total Omicron genomes, correspondingly (Figure 1B) (outbreak.info, 2022).

We assessed the binding affinities between hACE2 and RBDs from the four Omicron sub-variants (BA.1, BA.1.1, BA.2, and BA.3), together with the other earlier four VOCs (Alpha, Beta, Gamma, and Delta), and the PT virus (Lu et al., 2020a; Wei et al., 2020) by surface plasmon resonance (SPR) (Figure 1C). Our results showed that the K_D values for the binding of the PT RBD to hACE2 is 24.4 nM, similar to what was previously reported (Han et al., 2022; Wang et al., 2020; Zhang et al., 2021), whereas those of Omicron sub-variants, BA.1, BA.1.1, BA.2, and BA.3, were 19.5, 5.9, 10.0, and 22.1 nM, correspondingly. Compared with BA.1 RBD, the K_D values of BA.1.1 and BA.2 were 3.3- and 2.0-fold lower, respectively, indicating that their affinities to hACE2 were a bit higher than that of the PT RBD. However, the affinity of BA.3 to hACE2 was very close to BA.1. The K_D values of the RBDs binding to hACE2 in Alpha, Beta, Gamma, and Delta variants were 6.7, 19.7, 16.0, and 25.1 nM, correspondingly, suggesting that only Alpha RBD showed a comparable affinity with BA.1.1. The increased affinity of the BA.1.1 and BA.2 compared with BA.1 is consistent with the high frequencies of these two sub-variants in GISAID. We further detected the binding capacity of the above RBDs to hACE2 expressed on the surface of BHK-21 cells by flow cytometry (Figure S1). These results suggested that all the RBDs tested could efficiently bind to hACE2 on the cell surface. In order to detect the entry efficiency of the Omicron sub-variants, we prepared the VSV-backbone pseudoviruses expressing the S proteins from the SARS-CoV-2 PT and the VOCs Alpha, Beta, Gamma, Delta, and the four Omicron sub-variants. The entry efficiency of the pseudoviruses into Vero cells was evaluated (Figure 1D). All Omicron sub-variants were able to enter the cells efficiently. The relative entry efficiencies of Omicron sub-variants are consistent with the recent research (Saito et al., 2022).

Structural comparison of hACE2 in complex with BA.1, BA.1.1, BA.2, and BA.3 RBDs

The crystal structures of BA.1.1, BA.2, and BA.3 RBDs in the complex with hACE2 have been solved at resolutions of 3.0, 2.9, and 2.7 Å, correspondingly, in this study (Table S1).

We performed an interface analysis of the three complex structures with the cutoff for the atom contact distance being 4.5 Å (the cutoff distance for hydrogen bonds and salt bridges being 3.5 Å) and made a comparison with the interaction interface in the BA.1 RBD/hACE2 structure we previously solved (Han et al., 2022). The results show that the same 21 hACE2 residues are involved in constituting the binding interface with 20 BA.1 RBD residues, 21 BA.1.1 RBD residues, 21 BA.2 RBD residues, and 19 BA.3 RBD residues in the four complexes, correspondingly. The interaction atom pairs and numbers of interface atom contacts, hydrogen bonds, and salt bridges are listed in Tables S2 and S3.

Among the BA.1 RBD/hACE2, BA.2 RBD/hACE2, and BA.3 RBD/hACE2 complexes, the BA.2 RBD/hACE2 complex interface has the largest total number of atom contacts and hydrogen bonds, which might correlate with the higher hACE2 affinity of the BA.2 RBD compared with BA.1 and BA.3. However, the BA.1.1 RBD with the highest hACE2 affinity forms a less total number of atom contacts and hydrogen bonds with hACE2 than BA.2 RBD, but it has two more salt bridges than BA.2 RBD (Figure S2; Table S3). Therefore, the descending order of BA.1.1, BA.2, BA.1, and BA.3 RBD affinity to hACE2 can be partly explained by the different number of total atomic interactions, hydrogen bonds, and salt bridges, although other factors including the strength of individual interactions, especially those of critical mutation sites also play a role. This is consistent with our previous report that the lower affinity of dog ACE2 (dACE2) to the SARS-CoV-2 RBD than hACE2 can be explained by fewer interface atom contacts, hydrogen bonds, and salt bridges of dACE2 with RBD (Zhang et al., 2021).

The contact interfaces between hACE2 and SARS-CoV-2 or the Omicron sub-variant BA.1 RBD are distributed over two patches (Han et al., 2022; Wang et al., 2020). As Patch 1 in the BA.1 RBD/hACE2 complex, which we previously reported (Han et al., 2022), all Patch 1 of the three complexes solved here contains 5 hACE2 residues (S19, Q24, H34, E35, and Y83), which form hydrogen bonds or salt bridges with RBD residues, but Patch 1 of the BA.1.1 complex has an extra hACE2 D38, forming a hydrogen bond and a salt bridge with BA.1.1 RBD S496 and R493, correspondingly (Figures 2A–2D, upper panels).

Unlike the BA.1.1 complex, hACE2 D38 in the BA.1 RBD, BA.2 RBD, and BA.3 RBD complexes forms a hydrogen bond and a salt bridge with RBD Y449 and R498, correspondingly, in all the three complexes. In addition to hACE2 D38, Patch 2 of the BA.1.1 complex contains Y41, K353, and D355, whereas that of the three complexes contains hACE2 Y41, Q42, and K353. The hydrogen bonds network in Patch 2 of the BA.2 RBD complex is the most extensive, whereas that of the BA.1.1 RBD complex is the weakest (Figures 2A–2D, lower panels).

The binding interface residues of RBDs in the four complexes shown in (Figures 2E–2H) are evidently different. The two patches of the BA.1 and BA.1.1 RBDs are connected, whereas the boundaries of those in the BA.2 and BA.3 are distinguishable. Among all the mutations in the four RBDs, only the G496S is located in the binding interface. Notably, T415 of BA.2 RBD forms a hydrogen bond with hACE2 N90-glycan, which is discussed later (Figure 2G).

Correlation of the S496G substitution in the RBDs with hACE2 binding

The affinity of BA.2 RBD to hACE2 is a little bit higher, if any, than that of the BA.1/BA.3 RBDs. The difference between the

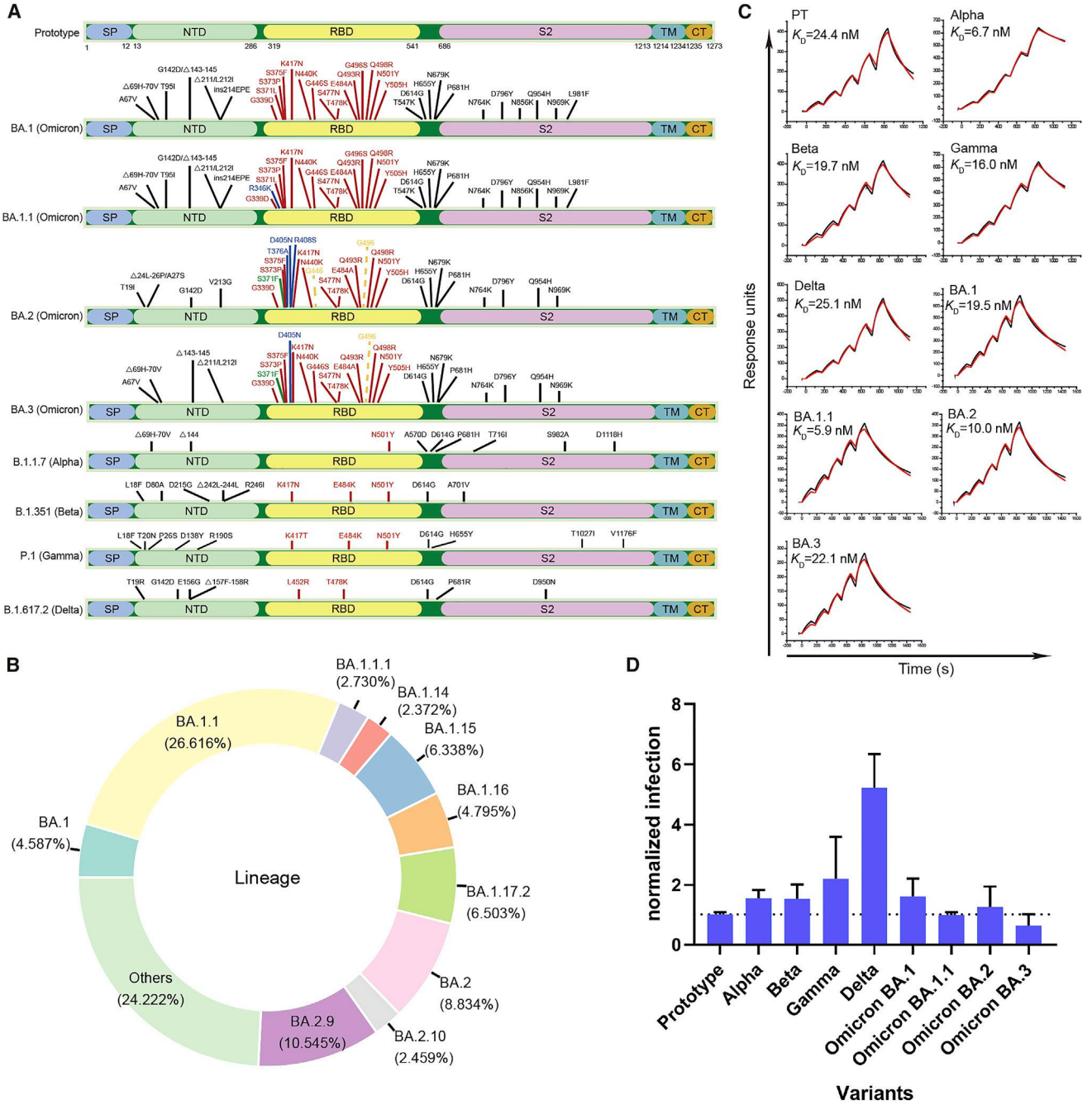


Figure 1. Mutation distribution in VOC S proteins and SPR for VOC RBD binding to hACE2 and VOC pseudovirus entry assay

(A) Schematic diagram for mutation distribution in the S proteins from the five VOCs, including the four Omicron sub-variants (BA.1, BA.1.1, BA.2, and BA.3). The mutations unique to BA.1.1, BA.2, or BA.3 are indicated in blue. The mutations unique to BA.2 and BA.3 are indicated in green. The mutations present in BA.1 and BA.1.1 but absent in BA.2 and BA.3 are shown with dashed lines.

(B) Donut chart for the global composition of Omicron sub-variants (data were collected from GISAID on April 25th, 2022).

(C) The SPR curves for the VOC RBDs binding to hACE2.

(D) Pseudovirus entry assay for the five VOCs, including the four Omicron sub-variants (BA.1, BA.1.1, BA.2, and BA.3). The error bars indicate the standard deviations (SD) for three or four independent experiments.

See also Figure S1.

CellPress

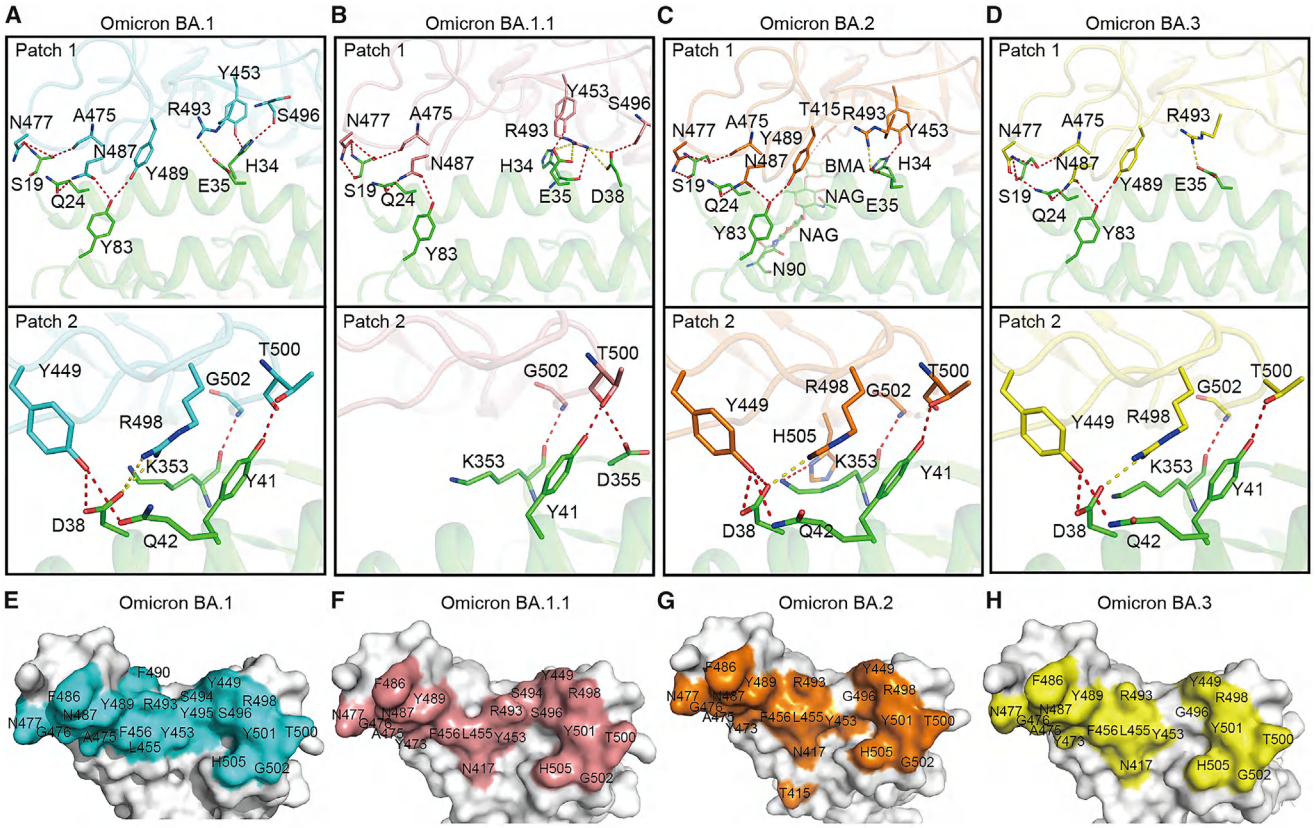


Figure 2. The complex structures of SARS-CoV-2 sub-variants BA1.1, BA.2, and BA.3 RBDs bound to hACE2

(A–D) The hydrogen bond (red) and salt bridge (yellow) networks of BA.1 (A), BA1.1 (B), BA.2 (C), and BA.3 (D) RBDs contact with hACE2. The complex structures are shown as a cartoon, and residues involved in hydrogen bond or salt bridge formation are shown as sticks. The RBDs of BA.1, BA1.1, BA.2, and BA.3 are shown as cyan, salmon, orange, and yellow, correspondingly. hACE2 is shown as green.

(E–H) The binding surface of hACE2 with the BA.1 (E), BA1.1 (F), BA.2 (G), and BA.3 (H) RBDs.

See also Figure S2.

BA.2 RBD and the BA.1 RBD is that BA.2 contains S371F, T376A, D405N, and R408S mutations, but BA.1 RBD does not, whereas BA.1 contains the G446S and G496S mutations, but BA.2 does not. Among these mutations, only the G496S mutation is in the RBD/hACE2 interface. In the BA.1 RBD/hACE2, BA.2 RBD/hACE2, and BA.3 RBD/hACE2 complexes, hACE2 D38 forms a salt bridge and hydrogen bond networks with RBD R498 and Y449. However, the side chains of these residues appear less compact in the BA.1 RBD/hACE2 complex with the G496S mutation, whereas in the BA.1.1 complex with the same mutation, this interaction network is disrupted (Figure 3A). Therefore, we speculate that S496 may disturb the polar interaction network formed between hACE2 D38 and RBD R498/Y449.

To verify the effects of G496S on receptor binding, we first evaluated the reverse mutation of S496G in BA.1.1 RBD on the ACE2 affinity. Our results indicate that the S496G site mutation induces an approximately 3.6-fold affinity increase, whereas the G496S mutation in BA.2 RBD induces an approximately 1.5-fold decrease in hACE2 binding affinity (Figure S3). Therefore, the SPR measurements support that G496S attenuates RBD binding.

To further explore the mechanism for the reduced binding associated with the G496S mutation, we performed molecular dynamics (MDs) simulations and examined the effect of the G496S mutation on the free energy of BA.1 RBD binding. Our results indicate that the molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) free energy of the BA.1 RBD with G496 binding to hACE2 is 5.34 kcal/mol ($\Delta\Delta E_{Binding}$) lower than that with S496 (Table S4). Representative snapshots clustered from the MD trajectories suggested that the more compact intra-chain hydrogen bonding networks formed by the hydroxyl side chain of S496 within RBD reduce the inter-chain interactions between BA.1 RBD and ACE2 (Figure S4). For control, we also performed the simulations of S446 BA.1 versus G446 BA.1 binding, showing that $\Delta\Delta E_{Binding}$ between S446 and G446 is only 1.15 kcal mol^{-1}, which is much lower than the $\Delta\Delta E_{Binding}$ between S496 and G496. These results also support that the G496S mutation attenuates the binding affinity of RBDs to hACE2.

Interestingly, the visibility of the electronic densities of the hACE2 N90-linked glycans was very different in the four complexes. The densities for the first three glycan residues (*N*-acetyl-β-d-glucosamine [NAG], NAG, and Beta-D-mannose

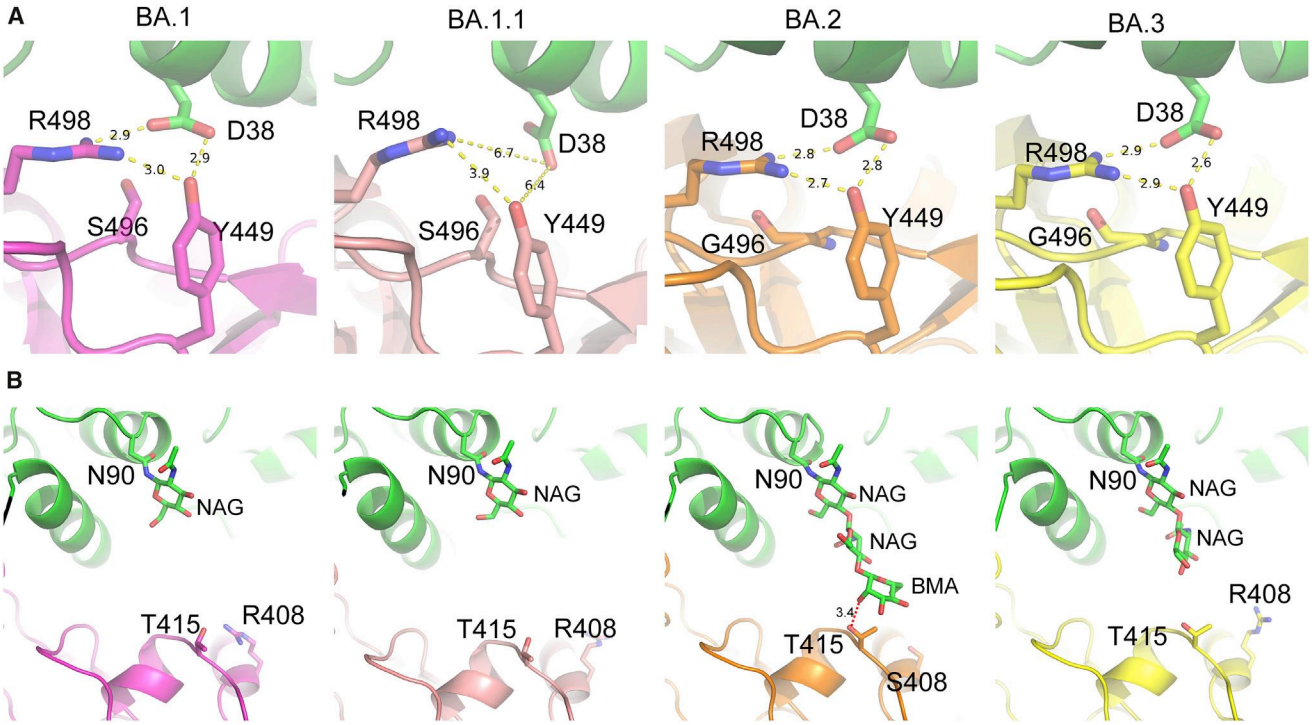


Figure 3. The G496S mutation in RBD and N90-glycan in hACE2 modulate the RBDs/hACE2 interaction

(A) The G496S mutation affects the local polar interaction networks at the RBD/hACE2 interfaces.

(B) The interaction between hACE2 N90-glycan may be related to the R408S mutations in the BA.2 RBD/hACE2 complex. The four panels from the leftmost one in (A) and (B) are BA.1 RBD/hACE2 (PDB: 7WBP), BA.1.1 RBD/hACE2, BA.2 RBD/hACE2, and BA.3 RBD/hACE2, correspondingly. The numbers on the dash lines are distances (Å) between the two atoms they link.

See also Figures S3–S6.

[BMA]) of the hACE2 N90-linked glycan in the BA.2 RBD/hACE2 complex were sharp and clear, but only the densities of the first residues (NAG) of ACE2 N90-glycan in the BA.1 RBD/hACE2 and BA.1.1 RBD/hACE2 complexes and the first two residues (NAG and NAG) of ACE2 N90-glycan in the BA.3 RBD/hACE2 complex were visible (Figure 3B). More importantly, in the BA.2 RBD/hACE2 complex, the third residue (BMA) of the hACE2 N90-glycan chain forms a hydrogen bond with BA.2 T145, which stabilizes the glycan chain (Figures 3B and S5) and presumably increases the affinity of the BA.2 RBD to hACE2. The unique mutation of BA.2 is R408S. We hypothesized that the large side chain of R408 in RBDs of BA.1, BA.1.1, and BA.3 has steric clashes with the N90-glycan chain. With the R408S mutation in BA.2, the steric clashes decrease, and therefore, the N90-glycan has a higher probability to get close and bind to T415. The glycosylation of viral envelope proteins together with receptors has a wide range of functions, including regulating viral tropism (Tian et al., 2021). Here, we present a molecular insight into how glycosylation regulates virus receptor binding.

To test the role of the R408S mutation on RBD/hACE2 interaction, we assessed the hACE2 affinity of BA.2 RBD with the S408R mutation, showing that the K_D values of the BA.2 RBD (15.67 nM) and BA.2 RBD with the S408R mutation (16.37 nM) were also similar (Figure S6). These results suggest that the R408S mutation has very limited RBD binding to hACE2.

Correlation of the enhanced binding of the BA.1.1 RBD to hACE2 compared with the BA.1 RBD with the long-distance "conduction" interactions induced by R346K

Our SPR data showed that the affinity of the BA.1.1 RBD to hACE2 is higher than the BA.1 RBD. Sequence alignment indicated that there is only one substitution residue (R346K) located far away from the receptor-binding motif (RBM) between BA.1 RBD and BA.1.1 RBD. Although this residue does not contact with hACE2 directly (Figure 4A), the binding affinity of BA.1.1 RBD with hACE2 was enhanced compared with BA.1 RBD with hACE2 (Figure 1C). To explore the underlying molecular mechanism, we determined the structure of the BA.1.1 RBD/hACE2 complex.

In the BA.1 RBD/hACE2 complex, R346 "pushes" N450 and changes the conformation of the corresponding loop to hACE2. Then, Y449 on the same loop contacts with D38 of hACE2 (Figures 4A, 4B, and 4D). When R346 was substituted by K346, it points to the opposite side and does not influence N450, the corresponding loop springs back, and the Y449 no longer contacts with D38 of hACE2. D38 of hACE2 points to the other side and loses the interactions with Y449 and R498 (Figures 4A, 4C, and 4E) and contacts with R493 of BA.1.1 RBD instead. Consequently, BA.1.1 RBD R493 forms salt bridges with E35 and D38 of hACE2, and it also forms contacts with H34 of hACE2. In addition, BA.1.1 RBD S496 forms a hydrogen bond with hACE2 D38. Therefore, these residues form a hub that generates strong interface

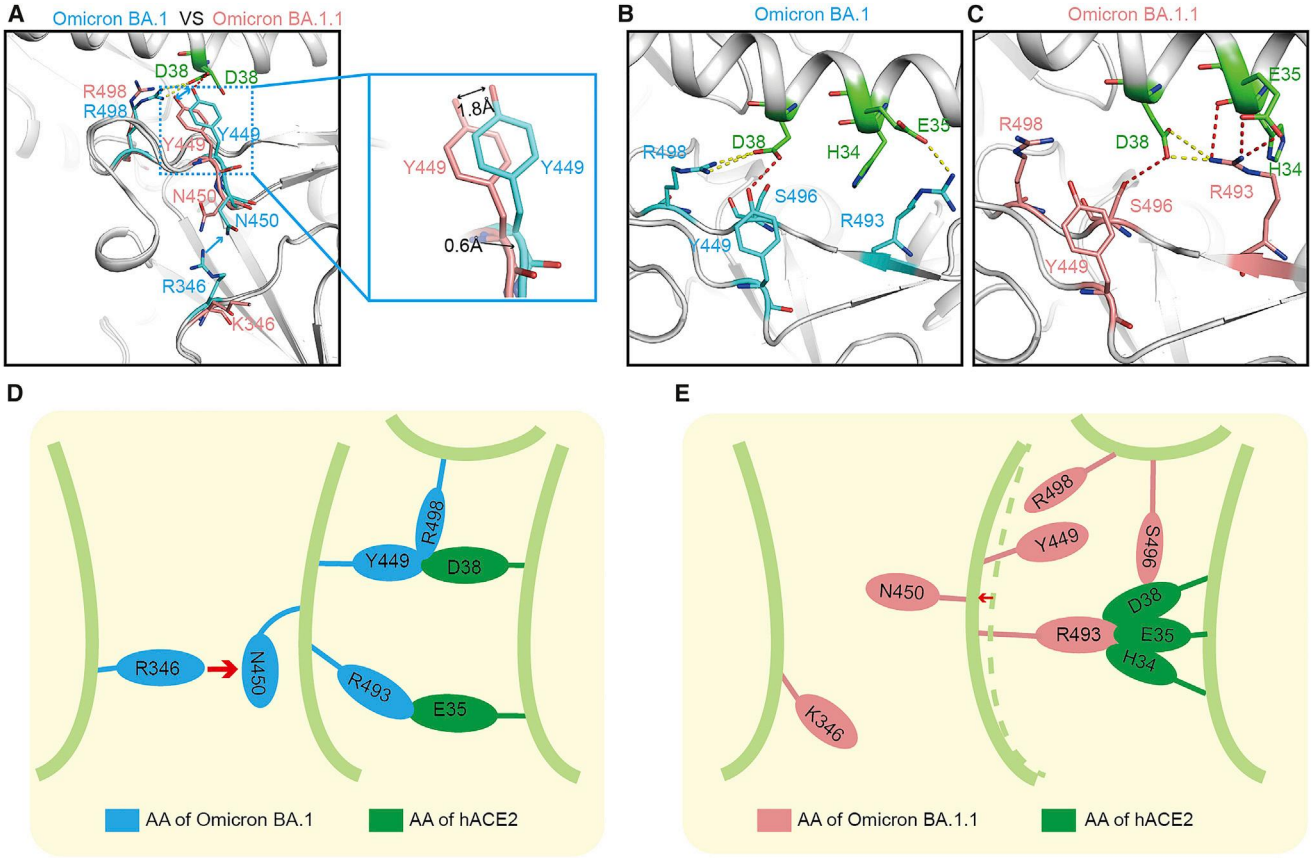


Figure 4. A single substitution far away from the binding motif of BA.1.1 RBD changes the interaction network with hACE2

(A–C) Structural comparison between BA.1 RBD/hACE2 complex and BA.1.1 RBD/hACE2 complex. The complex structures are shown as a cartoon and the key amino acids are shown as sticks. The key amino acids in BA.1 RBD and BA.1.1 RBD are colored by cyan and salmon, correspondingly. The key residues in the hACE2 are colored green.

(D and E) Pattern diagram of R346K substitution in BA.1.1 RBD changing the interaction network. (D) Schematic of the conformation of the key amino acids in BA.1 RBD.

(E) Schematic of the conformation change of the key amino acids after R346K substitution in BA.1.1 RBD.

interactions (Figures 4C and 4E). In summary, although R346K is far away from the RBM region, it also enhances the binding affinity of BA.1.1 RBD with hACE2 through long-distance "conduction" interactions like the "butterfly effect" (Figure 4E).

DISCUSSION

Since its first appearance, SARS-CoV-2 Omicron disseminated rapidly and became the dominant circulating variant globally. Although only four Omicron sub-variants (BA.1, BA.1.1, BA.2, and BA.3) were defined as of February 3, 2022 (PAHO and WHO, 2022), the Omicron family of viruses continues to grow in complexity. Over the past few months, several new Omicron sub-variants are gaining attention, including BA.2.12.1, which is spreading widely in the Northeast United States, and BA.4 and BA.5, which are growing rapidly to dominate South Africa (Haseltine, 2022; Maxmen, 2022). These sub-variants can evolve mutations to specifically evade humoral immunity (Cao et al., 2022). Previous studies showed that hACE2 affinities of Omicron (B.1.1.529 or BA.1) S proteins/RBDs are comparable with those of PT S proteins/RBDs (Han et al., 2022; Mannar et al., 2022) or higher than them (Cameroni et al., 2022; Cui et al., 2022; Lan et al., 2022; Yin et al., 2022). In this study, we showed that RBDs from other two Omicron sub-variants, BA.1.1 and BA.2, have higher hACE2 affinities than that of BA.1 and gave structural explanations for the difference in binding ability.

First, we propose that the absence of the G496S mutation is related to the enhanced affinity of BA.2 RBD. Here, we observed that S496 disturbs the local interaction networks in hACE2 D38, RBD R498, and Y449. Our SPR data and MD simulations of BA.1.1 RBD S496G mutants confirmed that the G496S mutation reduces RBD affinity. Our results are consistent with previous findings showing that G496S mutation attenuates the affinity of the RBD to hACE2 in deep mutational scanning studies (Starr et al., 2020; Haseltine, 2021).

An intriguing finding of this study is that hACE2 N90-glycan forms a hydrogen bond with BA.2 T415. The explicit binding of BA.2 T415 to hACE2 N90-glycan and the enhanced hACE2 affinity of BA.2 RBD led us to propose that N90-glycan/T415 interaction is favorable to the binding of BA.2 RBD to hACE2.

However, Chan et al. reported that hACE2 N90-glycosylation inhibits PT RBD binding to hACE2 (Chan et al., 2020). Solving this paradox, we propose that the binding of hACE2 N90-glycan to BA.2 RBD T415 alleviates the inhibitory effect of N90-glycan on RBD/hACE2 binding. Because of the adjacency of the R408S mutation to T415, we hypothesized that the binding of hACE2 N90-glycan to T415 is related to the R408S mutation in BA.2. However, the SPR results showed that the R408S mutation in BA.1.1 RBD and the S408R reverse mutation in BA.2 RBD had little effect on the hACE2 affinity. Therefore, either the R408S mutation has indeed no effect on the glycan/T415 binding, or the effect is too weak to be detected by SPR, i.e., the limitation of the method used. Anyway, here we provide evidence showing that hACE2 glycan is involved in SARS-CoV-2 RBD binding.

Another interesting finding is that the R346K mutation induces major alterations at the RBD/hACE2 interface through long-distance conduction interactions. Mutations located outside of the binding interface affecting the affinity via long-range conduction effects have been reported previously (Joughin et al., 2005; Unal et al., 2013). Actually, the "push" configuration induced by the R346K mutation captured by crystallography may represent one of the dynamic states for BA.1.1. Collectively, the available evidence warns against the misleading concept suggesting that only interface residues determine affinity and explain the role of SARS-CoV-2 RBD mutations.

Limitations of the study

Receptor binding is the initial step for SARS-CoV-2 infection. In this study, we focused on the interaction between the RBDs of four Omicron sub-variants and hACE2. The transmissibility of a virus is affected by multiple factors of complex nature, such as viral stability, receptor-binding affinity, fusion efficiency, replication activity, host immune responses, etc. To provide a deeper understanding of how the Omicron variants are transmitted, these aspects need to be further investigated.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cells
- METHOD DETAILS
  - Gene cloning
  - Protein expression and purification
  - SPR assay
  - Complex preparation and Crystallization
  - X-ray diffraction data collection and structure determination
  - Flow cytometry assay
  - Production and quantification of pseudoviruses
  - Pseudovirus infection assays
  - Molecular dynamics (MD) simulations
  - Molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) calculations
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Flow cytometry analysis
  - Binding affinity analysis
  - Pseudovirus infection assays

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2022.06.023.

ACKNOWLEDGMENTS

We are particularly grateful to the Pathogen Microbiology and Immunology Public Technology Service Center for its support. We thank Hua Tao and Tong Zhao from the Institute of Microbiology, Chinese Academy of Sciences, for their technical support in the FACS assay. We acknowledge the staff of beamline BL02U1, BL10U2, and BL19U1 at the Shanghai Synchrotron Radiation Facility for assistance during the data collection. This work was supported by the National Key R&D Program of China (2021YFC2301401, 2020YFA0907102, 2020YFA0509202, and 2021YFC0863300), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB29010202 and XDB37030204), the Queensland-Chinese Academy of Sciences Collaborative Science Fund 2020 (153211KYSB20200001), and the National Natural Science Foundation of China (grant 92169208).

AUTHOR CONTRIBUTIONS

N.H., J.Q., and G.F.G. initiated and designed the project. J.Q. and N.H. supervised and coordinated the experiments for crystal preparations, diffraction data collection, and structure determinations. G.F.G. supervised the structural analysis, SPR and FACS assays, and pseudoviruses transductions. L.L., H.L., Y.M., and W.L. purified the proteins and grew the crystals with the help of P.H., Q.W., Y.Z., Y.Z., and Q.W.; L.L. and H.L. performed the SPR analysis with the help of Z.F.; and Y.M. and W.L. conducted the flow cytometry assay. J.Q. collected the crystal structural data and solved the structures. D.L. and X.Z. performed the pseudotyped virus-related assays. L.W. collected and analyzed the sequence data from GISAID. K.L., X.Z., Y.S., N.H., J.Q., and G.F.G. wrote and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 1, 2022
Revised: May 15, 2022
Accepted: June 9, 2022
Published: June 16, 2022

REFERENCES

Abraham, M.J., Murtolad, T., Schulzbc, R., Pálla, S., Smithbc, J.C., Hessa, B., and Lindahlad, E. (2015). GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). Phenix: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. *66*, 213–221. https://doi.org/10.1107/S0907444909052925.

Best, R.B., Zhu, X., Shim, J., Lopes, P.E., Mittal, J., Feig, M., and Mackerell, A.D., Jr. (2012). Optimization of the additive CHARMM all-atom protein force

field targeting improved sampling of the backbone φ, ψ and side-chain χ(1) and χ(2) dihedral angles. J. Chem. Theory Comput. *8*, 3257–3273. https://doi.org/10.1021/ct300400x.

Bruel, T., Hadjadj, J., Maes, P., Planas, D., Seve, A., Staropoli, I., Guivel-Benhassine, F., Porrot, F., Bolland, W.H., Nguyen, Y., et al. (2022). Serum neutralization of SARS-CoV-2 Omicron sublineages BA.1 and BA.2 in patients receiving monoclonal antibodies. Nat. Med. *28*, 1297–1302. https://doi.org/10.1038/s41591-022-01792-5.

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. J. Chem. Phys. *126*, 014101. https://doi.org/10.1063/1.2408420.

Callaway, E. (2022). Why does the Omicron sub-variant spread faster than the original? Nature *602*, 556–557. https://doi.org/10.1038/d41586-022-00471-2.

Cameroni, E., Bowen, J.E., Rosen, L.E., Saliba, C., Zepeda, S.K., Culap, K., Pinto, D., VanBlargan, L.A., De Marco, A., di Iulio, J., et al. (2022). Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. Nature *602*, 664–670. https://doi.org/10.1038/s41586-021-04386-2.

Cao, Y., Yisimayi, A., Jian, F., Song, W., Xiao, T., Wang, L., Du, S., Wang, J., Li, Q., Chen, X., et al. (2022). BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. Nature. https://doi.org/10.1038/s41586-022-04980-y.

Chan, K.K., Dorosky, D., Sharma, P., Abbasi, S.A., Dye, J.M., Kranz, D.M., Herbert, A.S., and Procko, E. (2020). Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. Science *369*, 1261–1265. https://doi.org/10.1126/science.abc0870.

Cui, Z., Liu, P., Wang, N., Wang, L., Fan, K., Zhu, Q., Wang, K., Chen, R., Feng, R., Jia, Z., et al. (2022). Structural and functional characterizations of infectivity and immune evasion of SARS-CoV-2 Omicron. Cell *185*. 860.e13–871.e13. https://doi.org/10.1016/j.cell.2022.01.019.

Dai, L., and Gao, G.F. (2021). Viral targets for vaccines against COVID-19. Nat. Rev. Immunol. *21*, 73–82. https://doi.org/10.1038/s41577-020-00480-0.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an *N*.log(*N*) method for Ewald sums in large systems. J. Chem. Phys. *98*, 10089–10092. https://doi.org/10.1063/1.464397.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132. https://doi.org/10.1107/S0907444904019158.

Han, P., Li, L., Liu, S., Wang, Q., Zhang, D., Xu, Z., Han, P., Li, X., Peng, Q., Su, C., et al. (2022). Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2. Cell *185*. 630.e10–640.e10. https://doi.org/10.1016/j.cell.2022.01.001.

Haseltine, W.A. (2021). How omicron evades natural immunity, vaccination, and monoclonal antibody treatments. https://www.forbes.com/sites/williamhaseltine/2021/12/17/how-omicron-evades-natural-immunity-vaccination-and-monoclonal-antibody-treatments/?sh=3c9a3f2460e0.

Haseltine, W.A. (2022). New members of the omicron family of viruses: BA.2.12.1, BA.4, and BA.5. https://www.forbes.com/sites/williamhaseltine/2022/04/20/new-members-of-the-omicron-family-of-viruses-ba2121-ba4-and-ba5/?sh=255752617ff8.

Hess, B., Bekker, H., Berendsen, H.J.C., and Fraaije, J.G.E.M. (1997). LINCS: a linear constraint solver for molecular simulations. J. Comput. Chem. *18*, 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

Hole, K., Velazquez-Salinas, L., and Clavijo, A. (2010). Improvement and optimization of a multiplex real-time reverse transcription polymerase chain reaction assay for the detection and typing of vesicular stomatitis virus. J. Vet. Diagn. Invest. *22*, 428–433. https://doi.org/10.1177/104063871002200315.

Iketani, S., Liu, L., Guo, Y., Liu, L., Chan, J.F., Huang, Y., Wang, M., Luo, Y., Yu, J., Chu, H., et al. (2022). Antibody evasion properties of SARS-CoV-2 Omicron sublineages. Nature *604*, 553–556. https://doi.org/10.1038/s41586-022-04594-4.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. *79*, 926–935. https://doi.org/10.1063/1.445869.

Joughin, B.A., Green, D.F., and Tidor, B. (2005). Action-at-a-distance interactions enhance protein binding affinity. Protein Sci. *14*, 1363–1369. https://doi.org/10.1110/ps.041283105.

Kandeel, M., Mohamed, M.E.M., Abd El-Lateef, H.M., Venugopala, K.N., and El-Beltagi, H.S. (2022). Omicron variant genome evolution and phylogenetics. J. Med. Virol. *94*, 1627–1632. https://doi.org/10.1002/jmv.27515.

Kumari, R., Kumar, R., and Open Source Drug Discovery Consortium, and Lynn, A. (2014). g_mmpbsa—a GROMACS tool for high-throughput MM-PBSA calculations. J. Chem. Inf. Model. *54*, 1951–1962. https://doi.org/10.1021/ci500020m.

Lan, J., He, X., Ren, Y., Wang, Z., Zhou, H., Fan, S., Zhu, C., Liu, D., Shao, B., Liu, T.Y., et al. (2022). Structural insights into the SARS-CoV-2 Omicron RBD-ACE2 interaction. Cell Res. *32*, 593–595. https://doi.org/10.1038/s41422-022-00644-8.

Lu, R., Niu, P., Zhao, L., Wang, H., Wang, W., and Tan, W. (2020a). Sequencing the complete genome of COVID-19 virus from clinical samples using the sanger method. China CDC Wkly. *2*, 447–452. https://doi.org/10.46234/ccdcw2020.088.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. (2020b). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet *395*, 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8.

Mannar, D., Saville, J.W., Zhu, X., Srivastava, S.S., Berezuk, A.M., Tuttle, K.S., Marquez, A.C., Sekirov, I., and Subramaniam, S. (2022). SARS-CoV-2 Omicron variant: antibody evasion and cryo-EM structure of spike protein-ACE2 complex. Science *375*, 760–764. https://doi.org/10.1126/science.abn7760.

Maxmen, A. (2022). Why call it BA.2.12.1? A guide to the tangled Omicron family. Nature *606*, 446–447. https://doi.org/10.1038/d41586-022-01466-9.

Miyamoto, S., and Kollman, P.A. (1992). Settle: an analytical version of the shake and rattle algorithm for rigid water models. J. Comput. Chem. *13*, 952–962. https://doi.org/10.1002/jcc.540130805.

Muik, A., Wallisch, A.K., Sänger, B., Swanson, K.A., Mühl, J., Chen, W., Cai, H., Maurus, D., Sarkar, R., Türeci, Ö., et al. (2021). Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine-elicited human sera. Science *371*, 1152–1153. https://doi.org/10.1126/science.abg6105.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. *276*, 307–326.

outbreak.info (2022). Omicron variant report. https://outbreak.info/situation-reports/omicron?loc=ZAF&loc=GBR&loc=USA&selected=ZAF.

PAHO; WHO (2022). Sublineages of SARS-CoV-2 VOC omicron. https://www.paho.org/en/documents/sublineages-sars-cov-2-voc-omicron.

Peng, R.C., Wu, L.A., Wang, Q.L., Qi, J.X., and Gao, G.F. (2021). Cell entry by SARS-CoV-2. Trends Biochem. Sci. *46*, 848–860. https://doi.org/10.1016/j.tibs.2021.06.001.

Saito, A., Irie, T., Suzuki, R., Maemura, T., Nasser, H., Uriu, K., Kosugi, Y., Shirakawa, K., Sadamasu, K., Kimura, I., et al. (2022). Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. Nature *602*, 300–306. https://doi.org/10.1038/s41586-021-04266-9.

Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., et al. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell *182*. 1295.e20–1310.e20. https://doi.org/10.1016/j.cell.2020.08.012.

Tan, W., Zhao, X., Ma, X., Wang, W., Niu, P., Xu, W., Gao, G.F., and Wu, G. (2020). A novel coronavirus genome identified in a cluster of pneumonia cases — Wuhan, China 2019–2020. China CDC Wkly. *2*, 61–62. https://doi.org/10.1016/S0140-6736(20)30154-9.

Tian, W., Li, D., Zhang, N., Bai, G., Yuan, K., Xiao, H., Gao, F., Chen, Y., Wong, C.C.L., and Gao, G.F. (2021). O-glycosylation pattern of the SARS-CoV-2 spike protein reveals an "O-Follow-N" rule. Cell Res. *31*, 1123–1125. https://doi.org/10.1038/s41422-021-00545-2.

Unal, H., Jagannathan, R., Bhatnagar, A., Tirupula, K., Desnoyer, R., and Karnik, S.S. (2013). Long range effect of mutations on specific conformational

changes in the extracellular loop 2 of angiotensin II type 1 receptor. J. Biol. Chem. *288*, 540–551. https://doi.org/10.1074/jbc.M112.392514.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.Y., et al. (2020). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell *181*. 894–904.e9. https://doi.org/10.1016/j.cell.2020.03.045.

Wei, Q., Wang, Y., Ma, J., Han, J., Jiang, M., Zhao, L., Ye, F., Song, J., Liu, B., Wu, L., et al. (2020). Description of the first strain of 2019-nCoV, C-Tan-nCoV wuhan strain – National Pathogen Resource Center, China, 2020. China CDC Wkly *2*, 81–82. https://doi.org/10.46234/ccdcw2020.023.

Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., et al. (2018). MolProbity: more and better reference data for improved all-atom structure validation. Protein Sci. *27*, 293–315. https://doi.org/10.1002/pro.3330.

Xu, K., Dai, L., and Gao, G.F. (2021). Humoral and cellular immunity and the safety of COVID-19 vaccines: a summary of data published by 21 May 2021. Int. Immunol. *33*, 529–540. https://doi.org/10.1093/intimm/dxab061.

Yamasoba, D., Kimura, I., Nasser, H., Morioka, Y., Nao, N., Ito, J., Uriu, K., Tsuda, M., Zahradnik, J., Shirakawa, K., et al. (2022). Virological characteristics of the SARS-CoV-2 Omicron BA.2 spike. Cell *185*. 2103.e19–2115.e19. https://doi.org/10.1016/j.cell.2022.04.035.

Yin, W., Xu, Y., Xu, P., Cao, X., Wu, C., Gu, C., He, X., Wang, X., Huang, S., Yuan, Q., et al. (2022). Structures of the Omicron spike trimer with ACE2 and an anti-Omicron antibody. Science *375*, 1048–1053. https://doi.org/10.1126/science.abn8863.

Yu, J., Collier, A.Y., Rowe, M., Mardas, F., Ventura, J.D., Wan, H., Miller, J., Powers, O., Chung, B., Siamatu, M., et al. (2022). Neutralization of the SARS-CoV-2 omicron BA.1 and BA.2 Variants. N. Engl. J. Med. *386*, 1579–1580. https://doi.org/10.1056/NEJMc2201849.

Zhang, Z., Zhang, Y., Liu, K., Li, Y., Lu, Q., Wang, Q., Zhang, Y., Wang, L., Liao, H., Zheng, A., et al. (2021). The molecular basis for SARS-CoV-2 binding to dog ACE2. Nat. Commun. *12*, 4195. https://doi.org/10.1038/s41467-021-24326-y.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial strains		
Escherichia coli strain DH5α	Vazyme	Cat# C502-02
Escherichia coli strain BL21 (DE3)	Vazyme	Cat# C504-02
Escherichia coli strain DH10Bac	Gibco	Cat# 10361012
Chemicals, antibodies and recombinant proteins		
Polyethylenimine (PEI)	Polysciences	Cat# 24885-2
Anti-His/APC	Miltenyi Biotec	Cat# 130-119-820; RRID:SCR_008984
FuGENE 6 Transfection Reagent	Promega	Cat# E2691
BaseMuncher endonuclease	Abcam	Cat# ab270049
SARS-CoV-2 prototype RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_402119
SARS-CoV-2 Alpha RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_683466
SARS-CoV-2 Beta RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_678615
SARS-CoV-2 Gamma RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_833172
SARS-CoV-2 Delta RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_2020954
SARS-CoV-2 Omicron BA.1 RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_6640916
SARS-CoV-2 Omicron BA.1.1 RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_ 496747
SARS-CoV-2 Omicron BA.2 RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_9845731
SARS-CoV-2 Omicron BA.3 RBD protein with His-tag, spike residues 319-541	This paper	accession number: EPI_ISL_7605713
Human ACE2 protein, residues 19-615	This paper	accession number: NP_001358344
Critical commercial assays		
HisTrap HP 5 mL column	GE Healthcare	Cat# 17524802
Superdex 200 Increase 10/300 GL	GE Healthcare	Cat# 28990944
HiLoad 16/600 Superdex 200 pg	GE Healthcare	Cat# 28989335
Series S Sensor Chip SA	GE Healthcare	Cat# 29104992
Membrane concentrator	Millipore	Cat# UFC901096
Membrane concentrator	Millipore	Cat# SLHP033RB
Deposited data		
Omicron BA.1.1 RBD/hACE2 complex (X-RAD)	This paper	PDB: 7XAZ
Omicron BA.2 RBD/hACE2 complex (X-RAD)	This paper	PDB: 7XB0
Omicron BA.3 RBD/hACE2 complex (X-RAD)	This paper	PDB: 7XB1
Experimental models: Cell lines		
HEK293F cells	Gibco	Cat# 11625-019
HEK293T cells	ATCC	ATCC CRL-3216
BHK-21 cells	ATCC	CCL-10
Sf9 cells	Gibco	Cat# 11496015

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Recombinant DNA		
pEGFP-N1	MiaoLingPlasmid	Cat# P0133
pCAGGS	MiaoLingPlasmid	Cat# P0165
pFastBac1	Invitrogen	Cat# 10359-016
Software and algorithms		
PyMOL software	Molecular Graphics System, Version 1.8 Schrö dinger	https://pymol.org/2/
Graphpad Prism 8	GraphPad Software	https://www.graphpad.com/scientific-software/prism/
BIAcore 8K Evaluation software	GE Healthcare	N/A
FlowJo V10	FLOWJO	https://www.flowjo.com/solutions/flowjo/downloads
HKL2000	HKL Research	https://www.hkl-xray.com/hkl-2000
COOT	(Emsley and Cowtan, 2004)	http://www.mrc-lmb.cam.ac.uk/personal/peemsley/coot/
Phenix	(Adams et al., 2010)	http://www.phenix-online.org/
MolProbity	Duke Biochemistry	http://molprobity.biochem.duke.edu/index.php
GROMACS	(Abraham et al., 2015)	http://www.gromacs.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, George F. Gao (gaof@im.ac.cn).

Materials availability

All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement.

Data and code availability

- The crystal structures of Omicron BA.1.1 RBD/hACE2 complex (PDB: 7XAZ), Omicron BA.2 RBD/hACE2 complex (PDB: 7XB0), and Omicron BA.3 RBD/hACE2 complex (PDB: 7XB1) have been deposited in the Protein Data Bank (www.rcsb.org).
- This study did not generate custom computer code.
- Any additional information required to reanalyze the data reported in this work is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cells

HEK293F suspension-cultured cells (Gibco, Cat#11625-019) were cultured at 37 °C in SMM 293-TII Expression Medium (Sino Biological, Cat# M293TII), and BHK-21 adherent cells (ATCC CCL-10) were cultured at 37 °C in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS).

METHOD DETAILS

Gene cloning

The extracellular domain of hACE2 (residues 19-615, GenBank: NP_001358344) was cloned into the Baculovirus transfection vector pFastBac1 (Invitrogen) with the gp67 signal peptide sequence added to the N-terminus of the hACE2 gene and the Hexa-His tag sequence at the C-terminus. The coding sequences of RBD of SARS-CoV-2 (residue 319-541), RBDs of VOCs (residue 319-541) and hACE2 (residues 19-615, GenBank:NP_001358344) including the Hexa-His tag sequence at the C-terminus were inserted into the pCAGGS vector.

Protein expression and purification

The pFastBac1-hACE2 recombinant bacmids produced by transforming the plasmid into DH10Bac *E.coli*. were transfected into Sf9 cells using FuGENE 6 Transfection Reagent (Promega). The virus was amplified, and the High Five cells were used for protein expression. Supernatants were collected after 48 h-infection. The pCAGGS-hACE2 and pCAGGS-RBDs were expressed in HEK293F cells. Cell culture supernatants were collected after a 5-day infection.

The supernatants containing the proteins mentioned before were purified using His-Trap HP columns (GE Healthcare) and the SuperdexTM 200 Increase 10/300 GL column (GE Healthcare). Purified proteins were stored in a buffer containing 20 mM Tris-HCl and 150 mM NaCl (pH 8.0).

The hACE2 protein expressed via the Bac-to-Bac Baculovirus expression system was used for crystallization. The hACE2 protein produced by HEK293F cells was used for SPR assay. The RBD proteins produced by HEK293F cells were used for crystallization and SPR. The proteins for SPR assay were stored in PBST buffer (1.8 mM KH2PO4, 10 mM Na_2HPO4 (pH 7.4), 137 mM NaCl, 2.7 mM KCl, and 0.005% (v/v) Tween 20).

SPR assay

The biotinylated SARS-CoV-2 and VOCs RBD proteins were transferred into PBST buffer (1.8 mM KH2PO4, 10 mM Na2HPO4 (pH 7.4), 137 mM NaCl, 2.7 mM KCl, and 0.05% (v/v) Tween 20) and immobilized on SA chip. Serially diluted hACE2s were then flowed over the chip in PBST buffer. Binding affinities were measured using a BIAcore 8K (GE Healthcare) at 25 °C in the single-cycle mode. Binding kinetics were analyzed with Biacore Insight software (GE Healthcare) using a 1:1 Langmuir binding model.

K_D values of SPR experiments were obtained with BIAcore 8K Evaluation Software (GE Healthcare), using a 1:1 binding model. The values indicate the mean ± SD of three independent experiments.

Complex preparation and Crystallization

Purified hACE2 were mixed and incubated with BA.1.1 RBD, BA.2 RBD and BA.3 RBD on ice for about 2 h. The three mixtures were then purified on HiLoad 16/600 Superdex 200PG column in a buffer containing 20 mM Tris (pH 8.0) and 50 mM NaCl. Purified complex proteins (BA.1.1, RBD-hACE2, BA.2 RBD-hACE2 and BA.3 RBD-hACE2) were concentrated to 7 ~ 10 mg/mL for crystallization using a vapour-diffusion sitting-drop method. In general, 1 μL protein was mixed with 1 μL reservoir solutions, and the mixtures were sealed and equilibrated against 100 μL of reservoir solutions at 18°C. After 48h, high-resolution complex crystals of BA.2 RBD-hACE2 were obtained in 0.1 M ammonium acetate, 0.1 M Bis-Tris pH5.5, 17% w/v PEG10,000 and complex crystals of BA.3 RBD-hACE2 were grown in 0.1 M Imidazole pH7.0, 12% w/v Polyethylene glycol 20,000, whereas the complex crystals of BA.1.1 RBD-hACE2 were grown in 0.1 M Potassium chloride, 0.1 M Sodium HEPES pH7.5, 15% w/v PEG 6000.

X-ray diffraction data collection and structure determination

To collect the data, complex crystals were briefly soaked in the corresponding reservoir solutions containing 20% (v/v) glycerol for cryo-protection and then flash-cooling in liquid nitrogen. The diffraction data were collected at the Shanghai Synchrotron Radiation Facility (SSRF) BL02U1. The datasets were processed using HKL2000 software (Otwinowski and Minor, 1997). The structures of these three complexes were determined via the molecular replacement method using Phaser (Adams et al., 2010) with a search model (previously reported complex structure: SARS-CoV-2 RBD binding with hACE2 (PDB: 6LZG)). The atomic models were built using Coot (Emsley and Cowtan, 2004) and refined with phenix.refine (Adams et al., 2010). MolProbity tool (Williams et al., 2018) was used to assess the stereochemical qualities of the final models. Finally, all structures were generated using Pymol software (https://pymol.org/2/).

Flow cytometry assay

The plasmids containing hACE2 fused with eGFP were transfected into BHK-21 cells. Then stable cells were obtained by antibiotic screening. A mixture containing RBDs from SARS-CoV-2 PT (0.3 μg/mL) or VOCs (0.3μg/mL) was incubated with the hACE2-expressing BHK-21 stable cells at 4°C for 0.5 h. Subsequently, cells were washed with PBS thrice and stained with APC mouse anti-his secondary antibody for 0.5 h before being analyzed using BD FACS Canto FlowCytometer (BD Biosciences). BHK-21 cells were used as negative controls. The data of all samples were analyzed using FlowJo 10.8.1 (TreeStar Inc., Ashland, OR, USA).

All experiments were performed three times; one representative of each experiment is shown in the Figure S1.

Production and quantification of pseudoviruses

To obtain SARS-CoV-2 PT and VOC pseudoviruses, a replication-deficient vesicular stomatitis virus (VSV) vector backbone (VSV-ΔG-GFP) expressing the corresponding spike proteins were constructed as described protocols (Muik et al., 2021). Overall, 30 μg of spike protein expression plasmids were transfected into HEK293T cells. After 24h, The VSV-ΔG-GFP pseudoviruses were added to the transfected cell supernatant. After incubation for 1 h at 37°C, the inoculum was removed. Then, the cells were washed with PBS and cultured in DMEM containing both 10% FBS and anti-VSV-G antibody produced by I1Hybridoma ATCC® CRL2700™. The pseudoviruses were obtained 30 h post-infection. After being filtered by 0.45 μm filters (Millipore, Cat#SLHP033RB), the pseudoviruses were aliquoted and stored at −80 °C.

Prior to pseudovirus quantification, 0.5 U/μL BaseMuncher endonuclease (Abcam) was used to remove unpackaged RNA (37 °C for 1.5 h). Viral RNA was extracted using an RNA extraction kit (Bioer Technology) and quantified by quantitative RT–PCR (qPCR) using a 7500 fast real-time PCR system (Applied Biosystems). The primers and probes used to detect the L gene of the VSV virus are as described in the previous literature (Hole et al., 2010).

Pseudovirus infection assays

The numbers of the pseudovirus particles for SARS-CoV-2 and its variants were normalized by the VSV L gene qPCR, and then diluted to have the same number of viral particles per microliter *via* quantitative RT-PCR. Then, 100 μL of pseudovirus was added to each well of 96-well plates containing Vero cells. After 15 h, each whole well was scanned using a CQ1 confocal image cytometer (Yokogawa) and the total numbers of GFP-positive cells were determined using the software bundled with the instrument (Yokogawa). Each group included 6 biological replicates, and the analysis was repeated 3-4 times. Statistical analysis was performed using Graphpad Prism 8.

Molecular dynamics (MD) simulations

The structure of the BA.1 RBD/hACE2 complex without the RBD G496S mutation (G496 BA.1 RBD/hACE2) was modeled based on the BA.1 RBD/hACE2 complex structure. After stripping N-acetyl-β-glucosaminide glycans and water molecules from the crystal structure and adding TIP3P water (Best et al., 2012; Jorgensen et al., 1983) and counter ions, both G496 RBD-ACE2 and S496 RBD-ACE2 systems were subjected to five parallel MD simulations with different random seeds. MD simulations were performed using a GPU cluster with a GROMACS package (version 2020.5) (Abraham et al., 2015) at the temperature of 300 K and with the time step of 2 fs. The CHARMM36 force field was applied. All bonds involving hydrogen atoms were constrained using the LINCS algorithm (Hess et al., 1997) and water molecules were restrained using the SETTLE algorithm (Miyamoto and Kollman, 1992). Atom-based truncation scheme was applied, with a non-bond cutoff of 12 Å, and with the force switching function initiated at 10 Å for van-der-Waals interactions. Long-range electrostatic interactions were calculated using the Particle Mesh Ewald method with fourth-order cubic interpolation and 1.6 Å grid spacing (Darden et al., 1993). The dynamics of each system was modeled with a step-by-step equilibration. Energy was minimized using the steepest descent algorithm with the maximum force less than 1000 kJ/mol/nm, 200 ps of number of particles, volume, and temperature (NVT) equilibration with 1000 kJ/mol·nm^2 harmonic positional restraints on the heavy atoms of the protein, and three 200 ps steps of number of particles, pressure, and temperature (NPT) equilibration with 1000, 1000 and 500 kJ/mol·nm^2 harmonic positional restraints on the protein backbone heavy atoms, correspondingly. Then each equilibrated system was subjected to 200 ns NPT MD production using a Velocity rescaling thermostat without any restraints on the protein atoms (Bussi et al., 2007).

Molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) calculations

For both the G496 RBD-ACE2 and S496 RBD-ACE2 systems, 100 snapshots were extracted from the 200 ns trajectories of each simulation. A total of 500 snapshots of each system was subjected to MM/PBSA calculations using the g_mmpbsa tool (Kumari et al., 2014). The entropy contribution was ignored and the solute dielectric constant was set to 2. The binding energy between RBD and ACE2 was calculated as $\Delta E_{Binding} = E_{RBD/hACE2} - E_{RBD} - E_{hACE2}$. Besides, the binding energy was decomposed on each residue.

QUANTIFICATION AND STATISTICAL ANALYSIS

Flow cytometry analysis

All experiments were performed three times independently; one representative of each experiment is shown in this article.

Binding affinity analysis

K_D values of SPR experiments were obtained with BIAcore 8K Evaluation Software (GE Healthcare), using a 1:1 binding model. The values indicate the mean ± SD of three replicated experiments.

Pseudovirus infection assays

Each group contained 6 replicates. And the analysis was repeated 3-4 times.

Supplemental figures

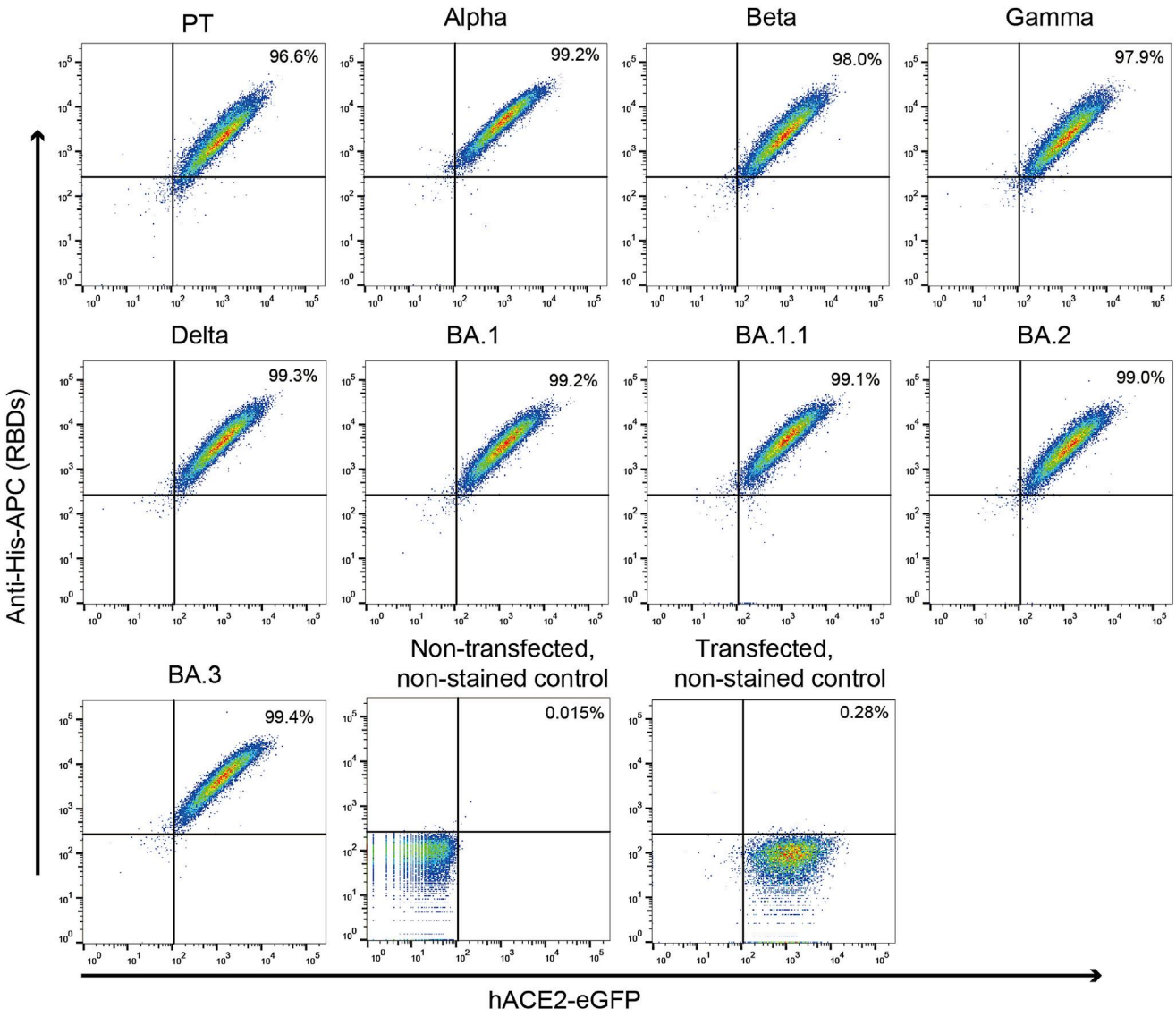


Figure S1. Flow cytometry assay for VOC RBDs binding to hACE2-expressing BHK-21 cells, related to Figure 1

The BHK-21 transfected with hACE2 fused with eGFP were incubated with the indicated His-tagged RBDs and stained with the anti-His-APC antibody. Non-transfected, non-stained control: the cells were neither transfected with hACE2 nor stained with the RBDs and the anti-His-APC antibody. Transfected, non-stained control: the cells were transfected with hACE2 but not stained with the RBDs and the anti-His-APC antibody.

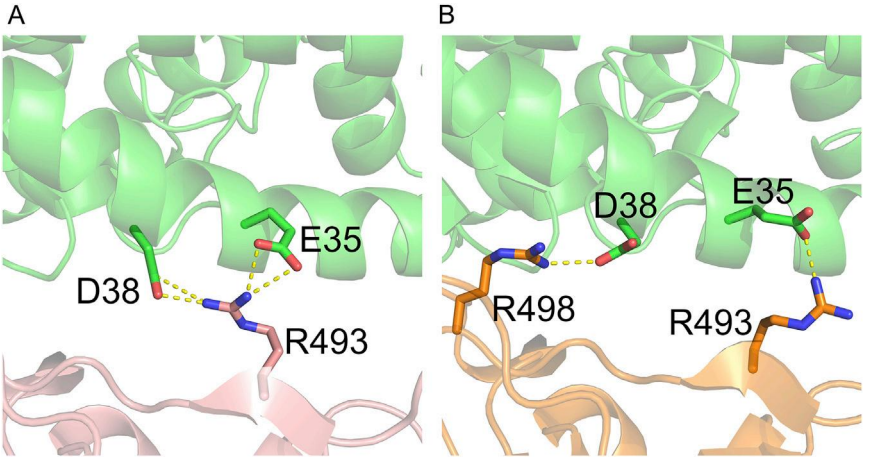


Figure S2. Salt bridges formed between BA.1.1 RBD/hACE2 and BA.2 RBD/hACE2 interfaces, related to Figure 2

(A) BA.1.1 RBD R493 forms four salt bridges with hACE2 E35 and D38.

(B) BA.2 RBD R493 forms a salt bridge with hACE2 E35 and BA.2 RBD R498 forms a salt bridge with hACE2 D38.

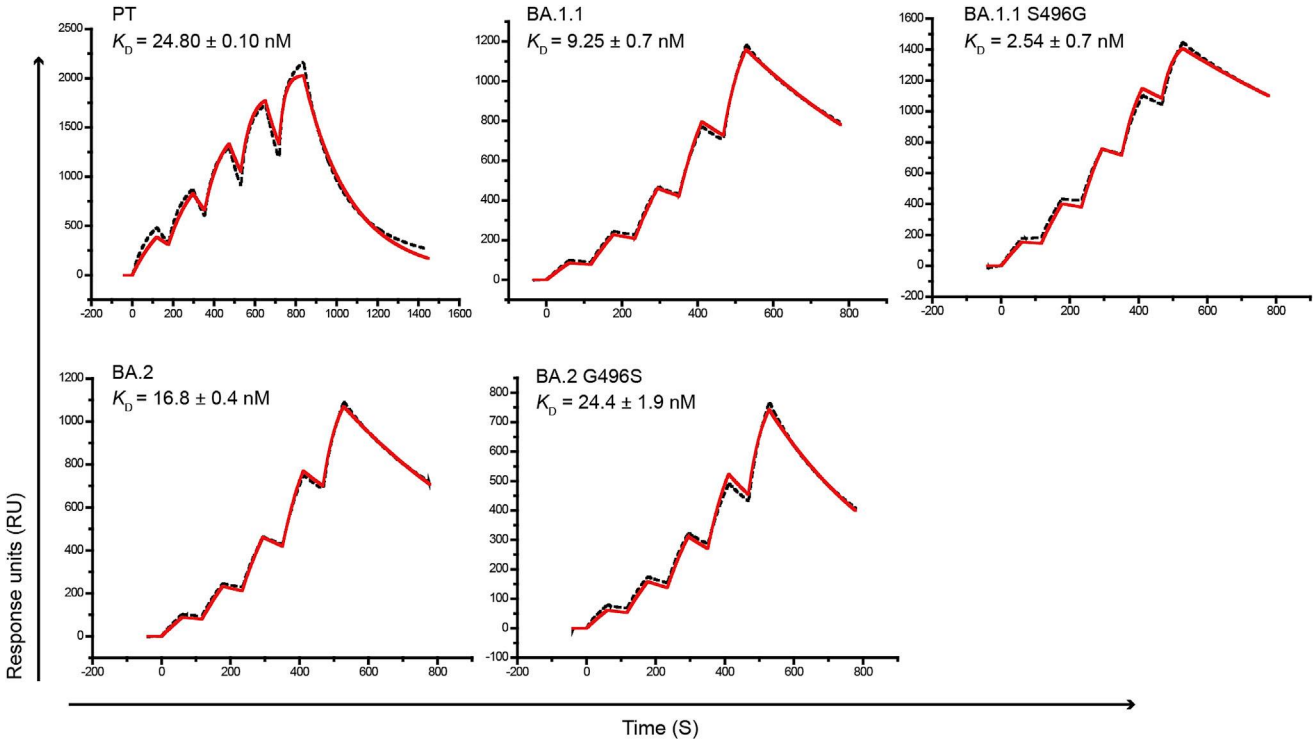


Figure S3. The SPR curves for the PT RBD, BA.1.1 RBD, BA.1.1 with the S496G mutation, BA.2 RBD, and BA.2 with the G496S mutation binding to hACE2, related to Figure 3

The mean and standard deviation of K_D values representing three independent experiments are shown.

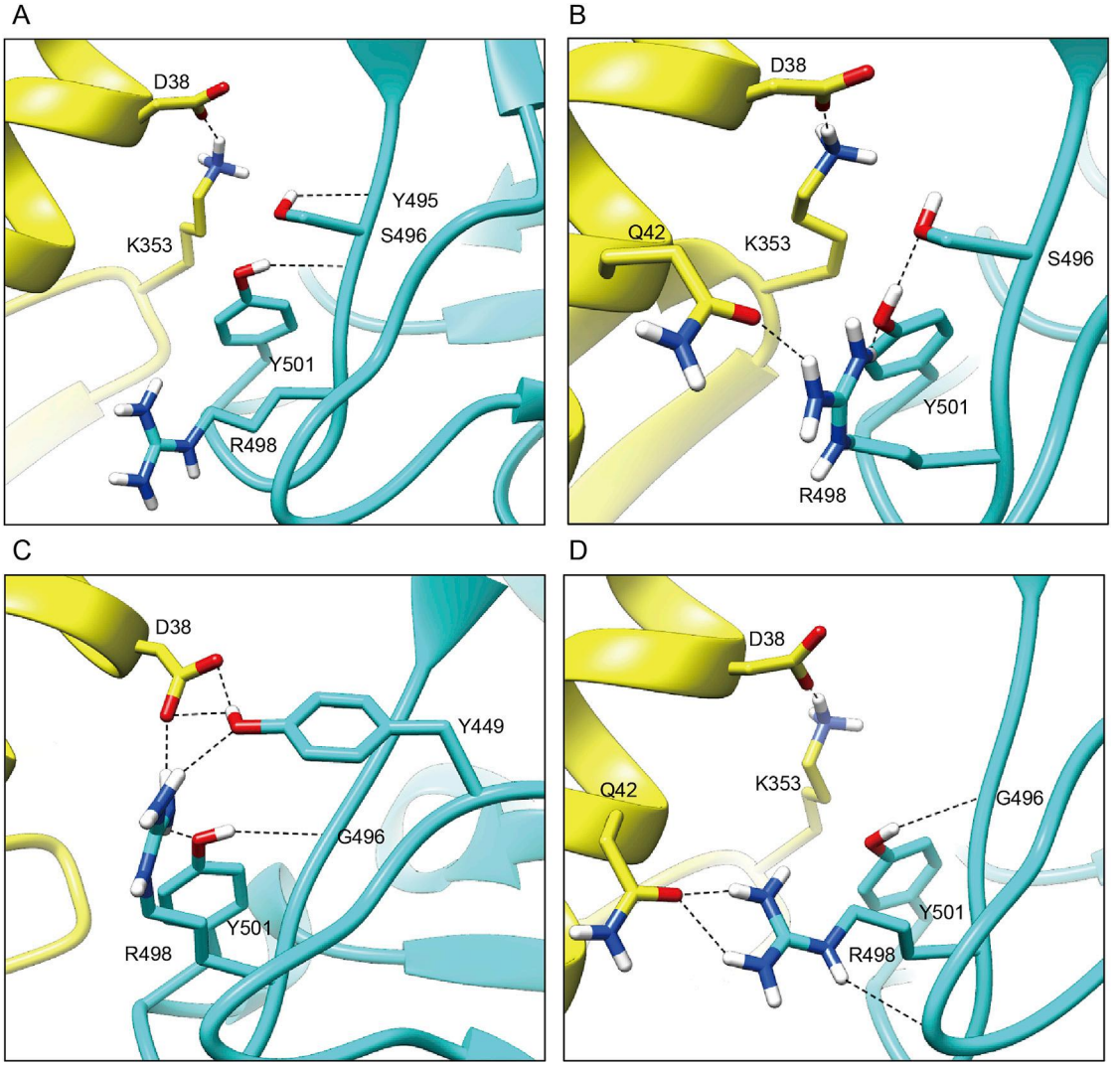


Figure S4. Representative snapshots clustered from MD trajectories of the (A and B) S496 BA.1 RBD-ACE2 system and the (C and D) G496 BA.1 RBD-ACE2 system, related to Figure 3

RBD domains are colored in cyan, and ACE2 is colored in yellow. H-bonds are displayed as black dotted lines.

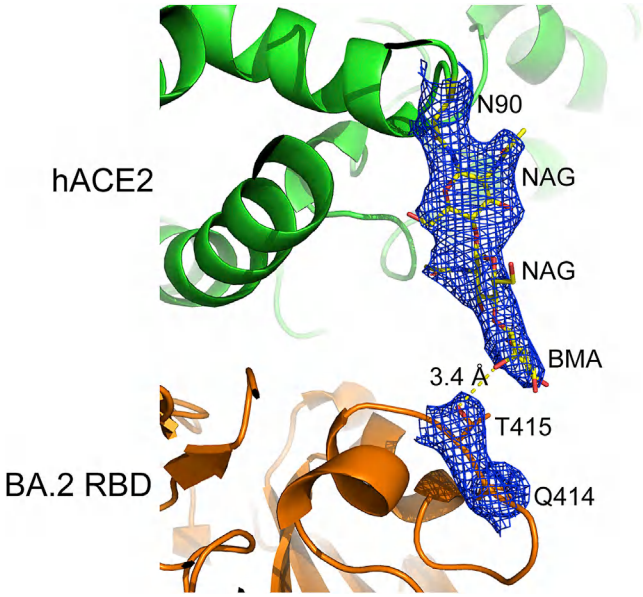


Figure S5. Electron densities of the hydrogen bond between BA.2 T415 and hACE2 N90-glycan, related to Figure 3

The local 2Fo-Fc electronic density map at 1.0 σ for hACE2 N90-glycan and BA.2 T415 is shown as mesh lines. The hydrogen bond is shown as the yellow dash line and its length is indicated.

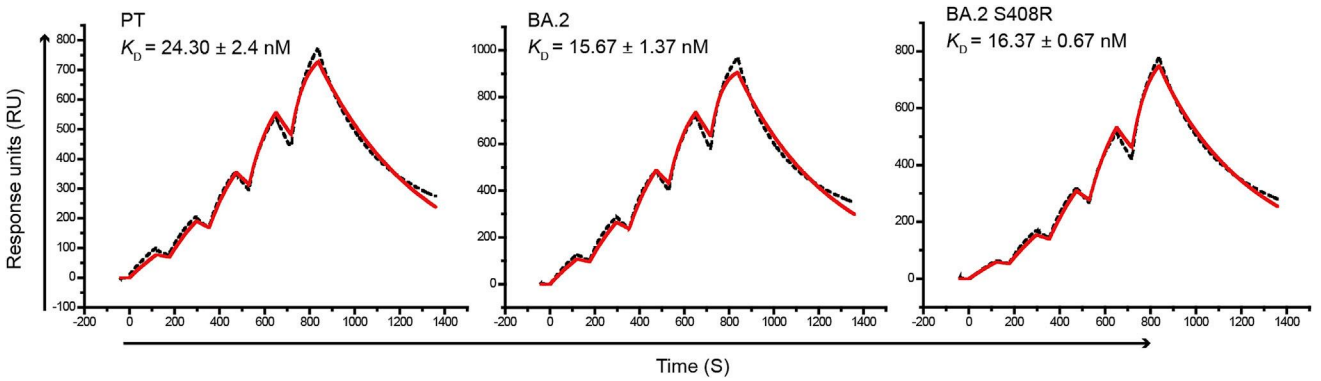


Figure S6. The SPR curves for the PT RBD, BA.2, and BA.2 with the S408R mutation, related to Figure 3

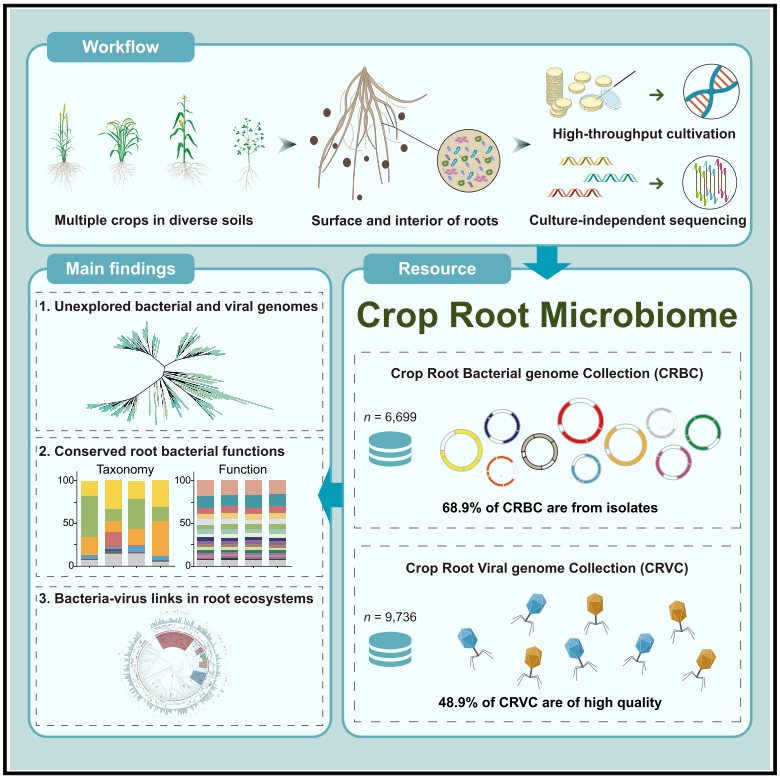
The mean and standard deviation of K_D values representing three independent experiments are shown.

Cell

Resource

Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns

Graphical abstract



Highlights

- Crop root bacterial (CRBC) and viral (CRVC) genome categories were established
- We found 1,817 undefined bacterial species and 1,572 unreported viral genera
- Despite varied taxa, root microbiome functions remain conserved across plants and soils
- A total of 27% of abundant bacteria in crop root microbiomes are linked to phages

Authors

Rui Dai, Jingying Zhang, Fang Liu, ..., Asaf Levy, Klaus Schlaeppi, Yang Bai

Correspondence

ybai@pku.edu.cn

In brief

Comprehensive bacterial and viral genome collections from crop roots uncover a large number of unexplored species and highlight conserved patterns within root microbiomes. Genomes and isolates are accessible at www.cropmicrobiome.com.

Dai et al., 2025, Cell *188*, 2521–2539
May 1, 2025 © 2025 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2025.02.013

CellPress

Cell

CellPress
OPEN ACCESS

Resource

Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns

Rui Dai,[1,2,3,4,11] Jingying Zhang,[2,3,11] Fang Liu,[1,11] Haoran Xu,[1,3,4] Jing-Mei Qian,[1,3,4] Shani Cheskis,[5] Weidong Liu,[1,3,4] Binglei Wang,[2] Honghui Zhu,[6] Lotte J.U. Pronk,[7] Marnix H. Medema,[7] Ronnie de Jonge,[8,9] Corné M.J. Pieterse,[8] Asaf Levy,[5] Klaus Schlaeppi,[10] and Yang Bai[2,3,12,*]

[1]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[2]Peking-Tsinghua Center for Life Sciences, State Key Laboratory of Gene Function and Modulation Research, Peking-Tsinghua-NIBS Graduate Program, School of Life Sciences, Peking University, Beijing 100871, China
[3]CAS-JIC Centre of Excellence for Plant and Microbial Science, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[4]College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing 100101, China
[5]Department of Plant Pathology and Microbiology, Institute of Environmental Science, The Faculty of Agriculture, Food, and Environment, The Hebrew University of Jerusalem, Rehovot, Israel
[6]State Key Laboratory of Applied Microbiology Southern China, Institute of Microbiology, Guangdong Academy of Sciences, Guangzhou 510070, China
[7]Bioinformatics Group, Wageningen University & Research, 6708 PB Wageningen, the Netherlands
[8]Plant-Microbe Interactions, Department of Biology, Science for Life, Utrecht University, 3584 CH Utrecht, the Netherlands
[9]AI Technology for Life, Department of Information and Computing Sciences, Science for Life, Utrecht University, 3584 CC Utrecht, the Netherlands
[10]Department of Environmental Sciences, University of Basel, Basel 4056, Switzerland
[11]These authors contributed equally
[12]Lead contact
*Correspondence: ybai@pku.edu.cn
https://doi.org/10.1016/j.cell.2025.02.013

SUMMARY

Reference genomes of root microbes are essential for metagenomic analyses and mechanistic studies of crop root microbiomes. By combining high-throughput bacterial cultivation with metagenomic sequencing, we constructed comprehensive bacterial and viral genome collections from the roots of wheat, rice, maize, and *Medicago*. The crop root bacterial genome collection (CRBC) significantly expands the quantity and phylogenetic diversity of publicly available crop root bacterial genomes, with 6,699 bacterial genomes (68.9% from isolates) and 1,817 undefined species, expanding crop root bacterial diversity by 290.6%. The crop root viral genome collection (CRVC) contains 9,736 non-redundant viral genomes, with 1,572 previously unreported genus-level clusters in crop root microbiomes. From these, we identified conserved bacterial functions enriched in root microbiomes across soils and host species and uncovered previously unexplored bacteria-virus connections in crop root ecosystems. Together, the CRBC and CRVC serve as valuable resources for investigating microbial mechanisms and applications, supporting sustainable agriculture.

INTRODUCTION

Plant roots act as a hub for recruiting a diverse array of microorganisms from the surrounding soil.[1] These microorganisms colonize both the surface and interior of roots and, along with their microbial genomes, are collectively known as the root microbiome, which plays a critical role in influencing the growth and health of host plants, especially agronomic crops.[2–7] Recently, functional investigations and community profiling of root bacteria have emerged as a cutting-edge area in crop root microbiome research,[7–10] highlighting the urgent need for a comprehensive collection of bacterial genomes specific to crop root ecosystems. Such genome collections, particularly those derived from cultivated isolates, are invaluable resources for exploring the functional and biological potential of crop root microbiomes.

The absence of a comprehensive bacterial genome collection presents a significant challenge for in-depth research on the crop root microbiome. Marker gene-based amplicon sequencing technology captures the taxonomic diversity of microbiomes[11–14] but does not address their functional diversity. Bacterial genomes encode massive functional diversity, which can vary even between strains of the same species.[15–17] While shotgun metagenomic techniques can identify functional gene diversity within root microbes and allow for more precise taxonomic classification, they rely on comprehensive, high-quality reference genomes to be effective.[18,19]

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Strain and genome collections of root bacteria are limited for agronomically relevant crops. Pioneering studies in model plants, such as *Arabidopsis thaliana* and *Lotus japonicus*, have generated nearly 800 bacterial genomes.[20–22] However, these resources from model plants are poorly suitable for crop research, as bacterial colonization exhibits host specificity.[22,23] Bacteria derived from one plant species can trigger strong immune responses in other plants, reducing their colonization.[22] To fully understand these microbial functions, genomic data derived from crop root microbiomes are crucial.[6,15,24,25] However, existing genomic resources related to the crop root microbiome primarily rely on small-scale datasets derived from divergent methodologies, and dominated with traditional agricultural species like *Rhizobium* and *Bacillus*.[4,26,15,24,27–29] In addition, limited access to isolates and inadequately organized metadata have hindered the full potential of these resources for the broader scientific community. Given the growing commercial interest in developing probiotics for agricultural applications—such as biocontrol agents, biofertilizers, and biostimulant inocula—the absence of a systematic collection of crop root bacteria represents a missed opportunity for advancing crop microbiome research.

Cultivation methods capture only a limited range of root bacterial taxa.[5,20] Metagenome-assembled genomes (MAGs) provide an alternative and complementary approach to accurately acquire genome information from uncultured root microbes.[30,31] However, obtaining MAGs for the root microbiomes of crops and model plants presents significant challenges. These include the high proportion of host plant genome sequences in root metagenomic data and the vast differences in microbial abundances within the root microbiome, both of which significantly increase sequencing costs, often 10–20 times higher than for environmental samples. These factors make it less feasible to acquire MAGs for microbes colonizing root surfaces and interiors.[6,25] As a result, comprehensive bacterial genome collections for crop root microbiome research remain scarce, and functional information for both cultivable and uncultivable bacteria is limited. This gap hinders mechanistic studies and genome-level analyses of the crop root microbiome.

In addition to research on root bacterial microbiomes, the study of root viromes has attracted considerable interest, given that viruses are widespread and play significant roles in microbial communities.[32,33] Numerous viral genomes have been discovered in the human gut, soils, and marine environments, often exhibiting distinct ecological specificities.[34–37] Across ecosystems, viruses, especially phages that infect bacteria, show intrinsic interaction with bacteria, profoundly influencing the population and life activities of their hosts. However, the distribution and prevalence of these viruses within plant root microbiomes, particularly in crops, remain unclear. Phages shape the genomes of bacteria and have led to the evolution of many antiviral systems.[38,39] Some phages possess the ability to regulate host behaviors, which not only benefits phage survival but also assists the host in, e.g., resisting infection by other phages.[40] Due to the limited genomic and metagenomic data available for crop root microbiomes, a dedicated viral genome collection for crop roots has yet to be developed. Establishing a systematic and comprehensive genome collection of crop root viruses is indispensable for advancing research in crop root microbiome studies.

In this work, we combined high-throughput bacterial cultivation with shotgun metagenomic sequencing to establish the crop root bacterial genome collection (CRBC) and the crop root viral genome collection (CRVC), comprehensive databases of bacterial and viral genomes from roots of multiple crop species. Using these resources, we evaluated the phylogenetic diversity and genome novelty of CRBC and CRVC, explored conserved characteristics of crop root microbiomes, and examined the interactions between bacteria and viruses within the crop root microbiome. The resources will advance research on crop root microbiomes, enable strain-level mechanistic research, and provide a foundation for developing microbial applications to support sustainable agricultural practices, and are accessible on our website (www.cropmicrobiome.com).

RESULTS

The CRBC comprises 6,699 genomes of root bacteria

To infer the genome sequences and functional repertoire of the bacterial root microbiome, we established a systematic and comprehensive collection of bacterial genomes from crop roots. This collection includes genomes of both cultivated bacterial strains and uncultured bacteria obtained through metagenomic assembly (Table S1A). The CRBC is derived from the roots of wheat, rice, maize, and *Medicago* grown in agricultural soils. Our sampling strategy enriched for root-associated microbiota, capturing microbes residing on the rhizoplane and within the endosphere (STAR Methods). To retrieve the culturable fraction of the crop root bacterial microbiome, we conducted extensive isolation of bacterial strains for whole-genome sequencing. To complement this and cover the unculturable fraction, we included MAGs by performing deep, culture-independent metagenomic sequencing on crop root samples (Figure 1).

The CRBC dataset comprises 6,699 high-quality bacterial genomes. We obtained 4,618 bacterial genomes (5.5 Tbp in raw reads) from pure root bacterial cultures. Following high-throughput bacterial isolation and selection of representative bacterial strains (STAR Methods), we identified 1,496 root bacterial genomes from wheat, 1,287 from rice, 1,056 from maize, and 779 from *Medicago* (Table S1A). Additionally, 2,081 MAGs of root bacteria were assembled from 332 root metagenomic samples across 14 datasets, spanning various soils and crop species (Tables S1A–S1D). These MAGs, representing a range of genomes of uncultured CRBC bacteria, complement the cultured bacterial genomes, adding 21 phyla and 147 genera not found in cultured isolates (Table S1E). Notably, the taxonomic diversity within the 6,699 genomes (68.9% from isolates) is extensive, spanning 27 phyla, 49 classes, and 113 orders. This diversity includes numerous genomes within prominent root bacterial phyla, such as Proteobacteria (3,659 genomes), Actinobacteriota (1,720 genomes), Bacteroidota (395 genomes), and Firmicutes (433 genomes). An impressive 79.1% of these genomes are of high quality, with an average completeness of 98.9% and an average contamination of 0.9%, aligning with standards in the Genome Taxonomy Database (GTDB),[41] the most

CellPress
OPEN ACCESS

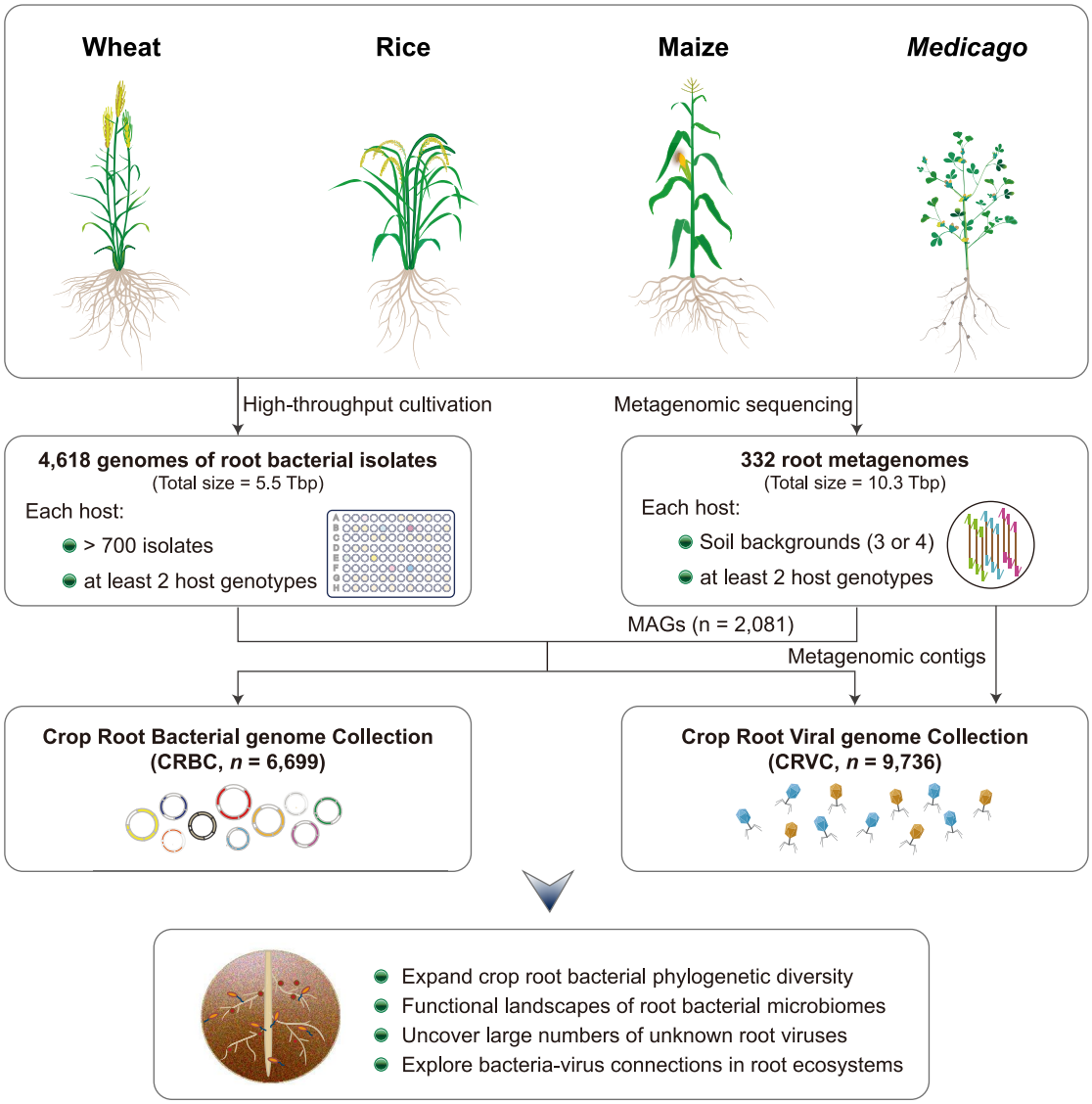


Figure 1. Design and overview of the CRBC and CRVC

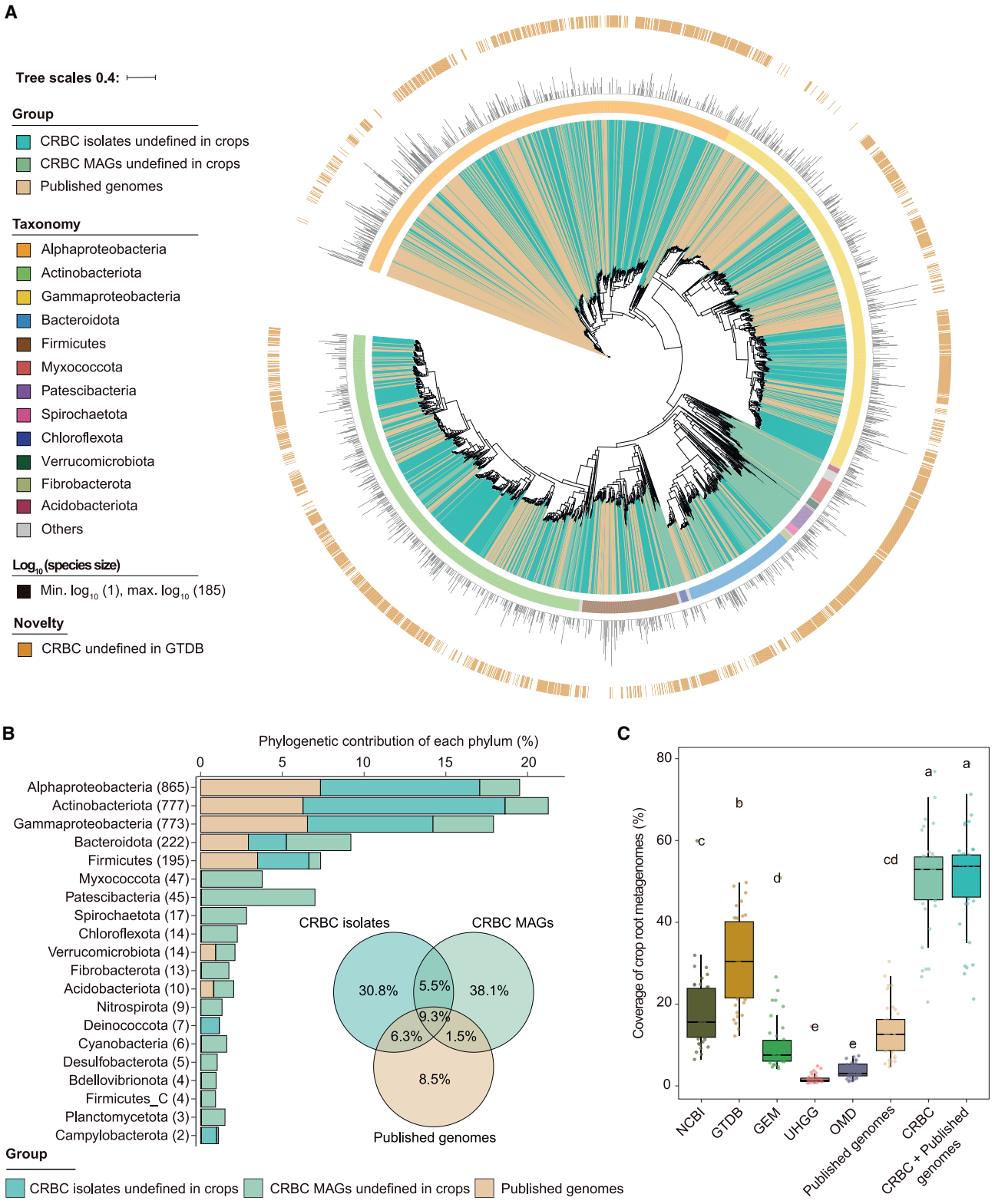
Based on comprehensive root-derived bacterial isolates and root metagenomic data, we established the crop root bacterial genome collection (CRBC) and the crop root viral genome collection (CRVC) from wheat, rice, maize, and *Medicago*. The CRBC encompasses 6,699 root bacterial genomes. The database includes 4,618 genomes (5.5 Tbp of raw reads) from bacterial isolates from crop roots. Each crop comprises over 700 isolates. The database also includes 2,081 MAGs (10.3 Tbp of raw reads) obtained from 332 root metagenomic samples. Each crop species is represented by at least two genotypes and spans three or four soil backgrounds. We also reconstructed 9,736 viral genomes by using bacterial genomes and root metagenomic contigs and defined them as the CRVC.
See also Figure S1.

comprehensive database of over 300,000 prokaryotic genomes from diverse ecosystems (Figures S1A–S1C; Table S1A; STAR Methods). With its high-quality and expansive genomic representation, the resources of the CRBC will serve as a valuable foundation for both microbiological and genomic studies on the crop root microbiome.

The CRBC enhances crop root bacterial genome diversity and metagenomic read coverage

To uncover novel bacterial species within the CRBC, we systematically surveyed crop root bacterial genomes across different public databases, including NCBI,[42] the Integrated Microbial Genomes and Microbiomes (IMG/M),[43] European Nucleotide Archive (ENA),[44] and GTDB.[41] We found that fewer than 1% of the bacterial genomes were derived from crop roots (Table S1F), with 3,073 genomes having a quality score (QS) > 50, representing only 6 phyla and 33 orders (Figures S1D and S1E; Table S1G; STAR Methods).

We compared the CRBC genomes with publicly available crop root genomes and found that the CRBC substantially increased the number and diversity of crop root bacterial species. Using an average nucleotide identity (ANI) threshold of 95%, the combined set of genomes (6,699 + 3,073) clustered into a total of



(legend on next page)

3,044 distinct species (Figure 2A). Remarkably, the CRBC contributed significantly to this diversity, encompassing 2,318 species, 2.8 times the species richness of publicly available crop root bacterial genomes. The CRBC included 2,212 bacterial species not previously reported in crop roots (Figure S1F). Compared with genomes in GTDB, CRBC genomes contained 1,817 undefined bacterial species, including 1,329 isolates and 488 MAGs, which exhibited an average ANI of less than 95% with GTDB genomes (Figures 2A and S1G). Of these species, the following ranks could not be assigned, phylum (1), class (1), order (5), family (22), and 177 species with an unknown genus compared with the GTDB (Table S2A). Similar trends were obtained by using the relative evolutionary divergence (RED) value to infer the taxonomic ranks (Table S2A). On average, each CRBC genome contains 14.9% uncharacterized genes, and these undefined bacterial species substantially expand the tree of life, spanning 26 phyla, 47 classes, and 182 families (Tables S1A and S2A).

The CRBC expands the phylogenetic diversity (PD) of crop root bacteria, increasing that of publicly available crop root bacteria by 290.6% (Figure 2B; Table S2B; STAR Methods). Genomes derived from CRBC isolates notably expanded the phylogeny of bacterial phyla commonly found in the root microbiome, such as Alphaproteobacteria, Actinobacteriota, Gammaproteobacteria, Bacteroidota, and Firmicutes. Specifically, four essential root bacterial families—Burkholderiaceae, Sphingomonadaceae, Microbacteriaceae, and Nocardioidaceae—each containing over 100 species, offer valuable resources for targeted genomic investigations (Table S2C). In contrast, the CRBC MAGs predominantly expanded the diversity of phyla associated with uncultured bacteria in crop roots, such as Myxococcota, Patescibacteria, and Fibrobacterota. Intriguingly, a combined analysis of bacterial genomes and crop root metagenomic data revealed that bacteria corresponding to the MAGs were estimated to have lower growth rates than cultured bacterial isolates (Figure S1H; Table S2D; STAR Methods), underscoring the importance of MAGs in capturing less abundant bacterial species.

Then, we assessed the read coverage in root metagenomic data by comparing the CRBC with other public databases, including NCBI RefSeq,[42] GTDB,[41] Genomes from Earth's Microbiomes catalog (GEM),[45] Unified Human Gastrointestinal Genome collection (UHGG),[46] Ocean Microbiomics Database (OMD),[31] and public genomes from crop root bacteria. Using 42 metagenome samples from crop roots and 37 samples from the crop rhizosphere, which were not used in the CRBC construction (Table S2E; STAR Methods), we aligned the reads from these metagenomes to each reference database to determine read coverage. The CRBC achieved significantly higher coverage of metagenomic data (mean coverage: 50.2%) compared with the other reference databases (Figure 2C). Notably, root metagenomes had higher coverage with CRBC than rhizosphere data (Figure S1I). Reference genomes from human and ocean ecosystems covered only a small fraction (below 4%) of root metagenomic sequences (Table S2E). Although GEM included rhizosphere microbial genomes and aimed to cover diverse soil ecosystems, its mean coverage was only 10.6%, highlighting the distinct microbial compositions between rhizosphere soils and root-associated microbiomes. Finally, the most comprehensive collections of reference genomes, NCBI and GTDB, achieved mean coverages of 18.2% and 31.1%, respectively, both considerably lower than the CRBC. Taken together, these findings indicate the importance of niche-specific genome reference databases and highlight the CRBC's substantial improvement in supporting crop root metagenome analyses.

CRBC genomes harbor diverse functions and metabolite genes for crop growth benefits

Taking advantage of the genomic resource, we systematically examined all high-quality, non-redundant genomes in the CRBC and published crop root bacteria for the coexistence pattern of their capacity to encode genes associated with plant growth-promoting (PGP) functions. We observed a widespread distribution of genomic traits related to nutrient utilization,[47–51] the biosynthesis of plant growth hormones,[52–56] and resistance to biotic and abiotic stresses[57–59] among crop root bacteria

Figure 2. The CRBC dramatically expands the PD of publicly available crop root bacterial genomes

(A) The CRBC contains 1,817 undefined bacterial species. The maximum-likelihood phylogenetic tree contains 3,044 representative species of the CRBC and crop root bacteria in public databases (NCBI, IMG/M, and ENA; STAR Methods). The clades associated with phylogenetic branches of 2,212 CRBC unique bacterial species, which show ANI below 95% when compared with published genomes derived from crop roots, are designated as undefined in crops and highlighted in blue (n = 1,702, contributed by CRBC isolates) or green (n = 510, contributed by CRBC MAGs). Published bacterial species from crop roots (n = 832) are indicated in beige. The initial inner circle illustrates the taxonomy of representative species at the phylum level. Proteobacteria are displayed at the class level due to excessive species numbers. Black bars represent $\log_{10}$-transformed numbers of non-redundant genomes (ANI of <99.9%) within each representative bacterial species. The outermost circle indicates 1,817 undefined bacterial species (green, n = 1,817) in the CRBC when compared with bacterial genomes from GTDB with the threshold of 95% ANI.

(B) The CRBC dramatically extends the PD of crop root bacterial genomes in public databases. The bar plot illustrates the proportional contribution of the CRBC and published crop root bacteria to the PD of crop root bacterial genomes, with the top 20 most abundant phyla displayed. Bars are colored by the contributions of genomes, with CRBC isolates undefined in crops (blue), CRBC MAGs undefined in crops (green), and published species from crop roots in beige. The numbers of representative species in each phylum are specified in brackets following the phylum names. The inset Venn diagram illustrates the percentage of total PD of all genomes in (A). See also Table S2B. Note that the CRBC expands the PD of crop root bacterial genomes in public databases by 290.6% (30.8% + 5.5% + 38.1%)/(6.3% + 9.3% + 1.5% + 8.5%).

(C) The CRBC shows a higher coverage of crop root metagenomic reads than public genome resources. The boxplot shows proportions of metagenomic reads of 42 root samples unrelated with the CRBC construction (wheat, n = 12; rice, n = 12; maize, n = 6; *Medicago*, n = 12) aligned to genomes in public databases and the CRBC. The abbreviations are: NCBI, NCBI RefSeq database; GTDB, Genome Taxonomy Database; GEM, the Genome from Earth's Microbiomes catalog; UHGG, the Unified Human gastrointestinal Genome collection; OMD, the Ocean Microbiomics Database (adjusted $p < 0.05$, Kruskal-Wallis rank-sum test and Dunn's test). See also Table S2E.

See also Figure S1.

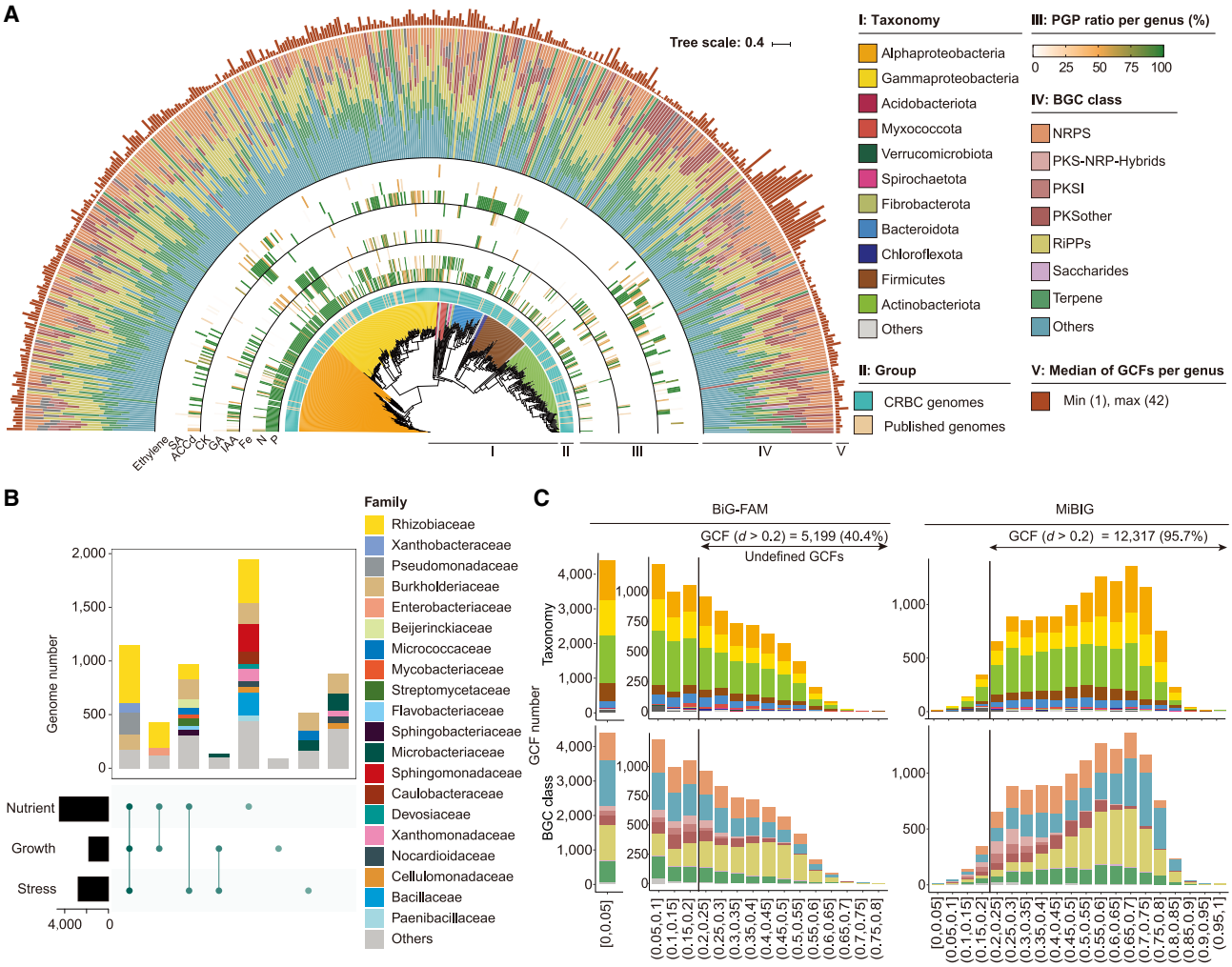


Figure 3. The CRBC harbors diverse beneficial functions and numerous metabolite gene clusters

(A) Distribution of PGP functions and BGCs. The phylogenetic tree is constructed using the high-quality representative genomes of each bacterial genus in the CRBC and public databases (STAR Methods). Phylum level taxonomy is colored (I). Genera containing CRBC genomes are labeled in blue (II). For each genus, plant growth-promoting (PGP) functions are illustrated as the percentage of genomes containing all the necessary genes for each function (III; STAR Methods). Bacterial PGP functions are categorized into three groups: nutrient utilization, growth promotion, and stress tolerance, according to the KEGG database and prior literatures. Nutrient utilization includes phosphorus nutrition (P), nitrogen fixation (N), and siderophore biosynthesis (Fe). Growth promotion includes biosynthesis of indole-3-acetic acid (IAA), gibberellin (GA), and cytokinin (CK). Stress tolerance includes the biosynthesis of 1-aminocyclopropane-1-carboxylic acid deaminase (ACCd), salicylic acid (SA), and ethylene. The BGC compositions are presented as the average of genomes at the genus level (IV). BGC classes are shown with the following abbreviations: non-ribosomal peptide synthetases (NRPSs); polyketide synthase (PKS); ribosomally synthesized and post-translationally modified peptide (RiPP). The outermost bar chart (section V) shows the median number of BGCs in all genomes of each genus. See also Table S3C.

(B) Coexistence patterns of PGP functions within individual bacterial genomes. The intersection scheme with vertical lines illustrates co-existent functions within individual genomes. The number and taxonomy of genomes for each coexistence pattern are shown in the stacked bar plot in the above panel. The number of genomes for each PGP group is shown in the lower left.

(C) The CRBC encodes a large number of undefined BGCs. All 48,643 BGCs identified in genomes of the CRBC, and public databases are clustered into 12,865 GCFs. Comparison with computationally predicted (BiG-FAM, left) and experimentally validated (MIBiG, right) BGCs revealed 5,199 undefined GCFs with an average cosine distance greater than 0.2, 81.1% of which are uniquely contributed by CRBC genomes. The upper panels are colored according to the taxonomy of genomes encoding BGCs, whereas the bottom panels are colored according to BGC classes.

See also Figure S2.

(Figure 3A; Tables S3A–S3C). Among the 6,109 high-quality bacteria genomes, 5,231 encoded at least one of the tested PGP functions, implying a significant role for root bacteria in supporting crop growth (Figure S2A; Table S3C). More than 43.8% of these genomes encode at least two categories of PGP functions, and 18.7% encode all three categories, with dominant representation from the Rhizobiaceae, Xanthobacteraceae, Pseudomonadaceae, and Burkholderiaceae (Figure 3B). Further analysis identified over 1,000 bacterial genomes with both nitrogen-fixation and phosphorus solubilization capabilities, especially in

CellPress
OPEN ACCESS

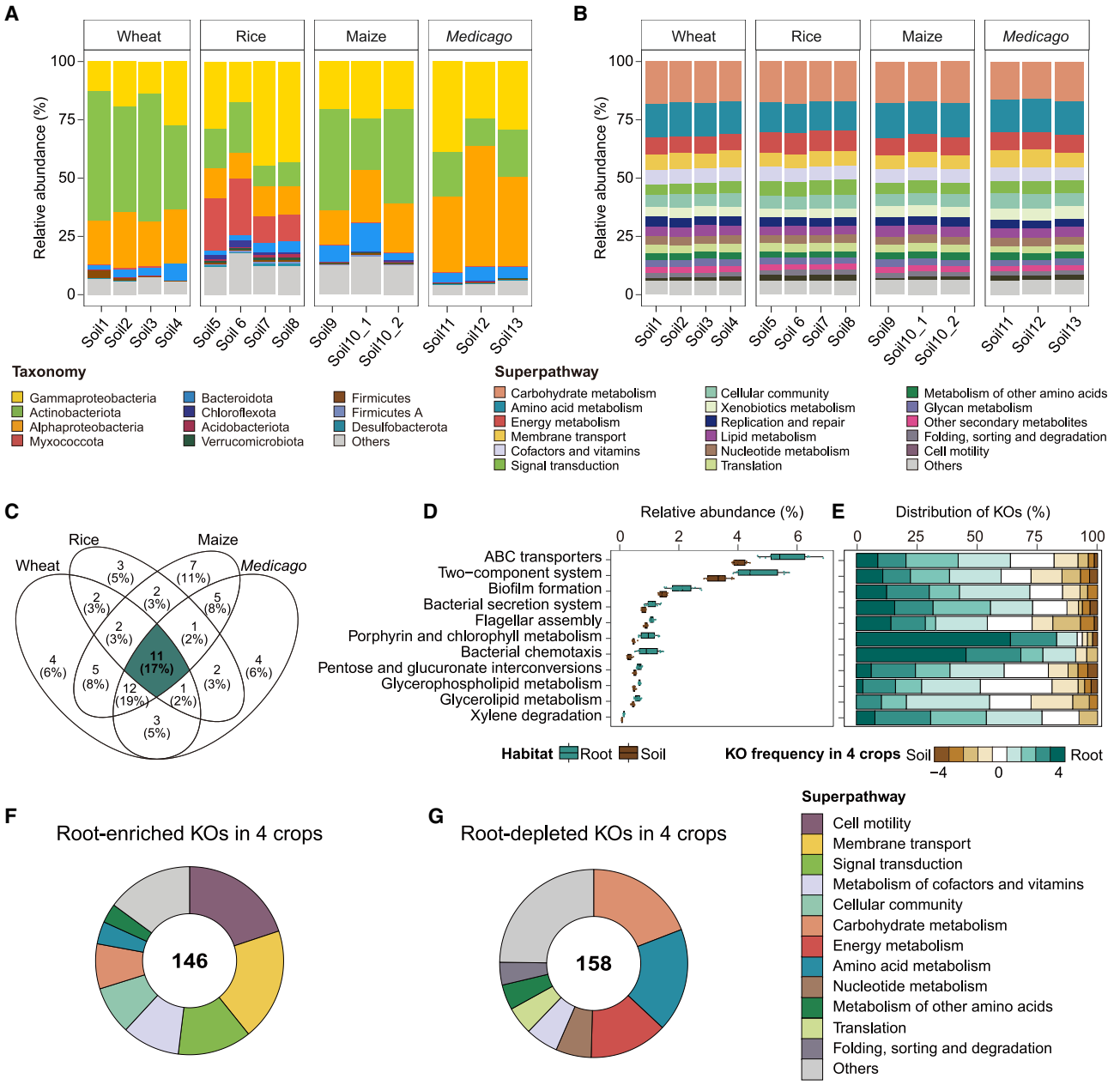


Figure 4. Conserved genetic features of the root bacterial microbiomes across multiple crop species grown in diverse soils

(A) Taxonomic composition of root bacterial metagenomes shows dramatic variation among 14 metagenomic datasets. The taxonomy of bacterial genes in 332 root metagenomic samples was classified by Kraken2 with the GTDB database implemented with crop root bacterial genomes (STAR Methods). Bacteria are displayed at the phylum level. The top 11 most abundant phyla are displayed, while phyla of lower abundance are shown as others. Proteobacteria are displayed at the class level due to excessive abundance.

(B) Convergent functions of root bacterial metagenomes among 14 metagenomic datasets. Bacterial genes are annotated according to the KEGG database and are summarized at the KEGG pathway level 2 (superpathway). The top 17 most abundant superpathways are displayed, while the functions of lower abundance annotated by the KEGG are shown as others.

(C) Consistent enrichment patterns of genetic functions in the root bacterial metagenomes compared with soils across multiple crops grown in diverse soils. The Venn diagrams show the overlap of KEGG bacterial functions that are enriched in root microbiomes. Functions enriched in the roots of all four crop species are highlighted in green (root, $n = 332$; soil, $n = 75$, adjusted $p < 0.05$, Wilcoxon rank-sum test). See also Table S4I.

(D and E) Abundance and KO distribution of functions consistently enriched in the root microbiomes. Boxplots (D) show the cumulative abundance of top 11 abundant bacterial functions that are consistently enriched in roots of wheat, rice, maize, and *Medicago* compared with corresponding soils. The distribution (E) of KOs in each function are colored according to their frequency of enrichment (green) or depletion (brown) in the roots of crop species. The colors from light to dark

(legend continued on next page)

Rhizobiaceae and Burkholderiaceae (Figure S2B). Bacillaceae were notable for their dual capabilities in phosphorus solubilization and siderophores biosynthesis. Notably, 18 genomes encode functions related to phosphorus, nitrogen, and siderophore utilization, with 8 genomes from the *Klebsiella* genus (Table S3C). In contrast, functions related to plant growth hormone biosynthesis or resistance to biotic and abiotic stresses rarely co-occurred in a single bacterium (Figures S2C and S2D). This bacterial resource offers a valuable opportunity to investigate the genomic coexistence of these functions and to identify promising candidates for future agricultural applications.

We next identified biosynthetic gene clusters (BGCs) within crop root bacteria, which are related with crop growth and health.[60–62] In total, we identified 48,643 BGCs across the CRBC and publicly available crop root bacterial genomes (Figure 3A; Table S3D). The predominant categories of BGCs were ribosomally synthesized and post-translationally modified peptides (RiPPs), non-ribosomal peptide synthetase (NRPS), and terpenes, representing 26.1%, 15.4%, and 14.3% of the clusters, respectively (Figure S2E). These bacteria exhibited diverse biosynthetic capabilities (Figures S2F–S2H), averaging 8 BGCs per genome. Next, we clustered the identified BGCs into 12,865 gene cluster families (GCFs) (Table S3E; STAR Methods). Strikingly, 5,199 (40.4%) of the identified GCFs did not show close similarity to any computationally predicted BGCs from the biosynthetic gene cluster families (BiG-FAM) database,[63] while 12,317 (95.7%) of GCFs showed no close similarity to experimentally validated BGCs from the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database[64] (Figure 3C; STAR Methods). These undefined GCFs were mainly contributed by Proteobacteria and Actinobacteriota, encoded functions related to RiPPs, NRPS, polyketide synthase (PKS), and terpenes. The most significant differences in our GCFs compared with known databases were in biosynthetic families related to RiPPs, predominantly originating from the genera *Dyadobacter*, *Paenibacillus*, and *Arthrobacter* (Tables S3D and S3E). These BGCs present an untapped reservoir of secondary metabolites with presumably specialized functions in crop root bacteria, potentially offering new avenues for agricultural applications.

Conserved genetic characteristics of root microbiomes across multiple crop species grown in diverse soils

The scarcity of bacterial genomes and the interference of host DNA hinder our understanding of root ecosystems. To investigate the genomic patterns of root microbes, we systematically explored conserved features within 332 deeply sequenced root metagenomes (median depth of 30.4 Gbp) from 14 datasets across diverse crop species and soil backgrounds, including wheat, rice, maize, and *Medicago*. Each crop included at least two genotypes and three to four soils to ensure diverse sample representation. On average, 82.0% of the sequencing reads were host derived (Table S1D). After removing host reads, we annotated microbial genes using our genome resources and publicly available microbial reference genomes. Rarefaction analysis showed that the non-redundant microbial gene reservoirs for each crop had reached saturation (Figure S3A; Table S3F). The number of non-redundant genes reproducibly detected in the root microbiomes was 13.2–55.4 times of those observed in their respective hosts (Figure S3B; Table S3G; STAR Methods). On average, 23.8% of root microbial genes could not be annotated by Kyoto Encyclopedia of Genes and Genomes (KEGG) (Table S3H), underscoring the extensive genetic potential of the root microbiome. Notably, we found a high abundance of bacteria in root microbiomes, comprising an average of 95.7% of the overall microbial relative abundance (Figure S3C; Table S3I). In contrast, fungi, archaea, viruses, and protists represented much smaller fractions of the root microbiome, averaging 0.74%, 0.15%, 3.21%, and 0.16%, respectively. These findings highlight the dominance of bacteria in root microbiomes and underpin the importance of developing systematic genome reference collections for crop root bacteria.

We next identified the conserved features of bacteria within the crop root microbiome using our root metagenomic datasets across multiple host plant species grown in diverse soil backgrounds (Tables S4A–S4F). For each root dataset, we collected 4–6 soil samples from unplanted soils alongside root sampling. Notably, although the taxonomic compositions of root microbiomes were strongly influenced by soil backgrounds and host species, the compositions of superpathway (2nd level of KEGG pathway) functions in root microbiomes were similar across all 14 datasets (Figures 4A, 4B, and S3D), suggesting the presence of conserved genetic features in multiple crop root ecosystems. Further, we compared the microbiomes of roots and surrounding soils in each dataset and found that, within each crop species, a higher proportion of microbial taxa was influenced by soil backgrounds compared with the proportion of microbial functions affected by soils (Figures S3E–S3H; Tables S4G–S4J). A total of 11 microbial functions were conservatively enriched in root microbiomes of all four crops grown in various soils (Figure 4C; Table S4I). These functions are related to ATP-binding cassette (ABC) transporters, two-component systems, biofilm formation, bacterial chemotaxis, and flagellar assembly (Figure 4D). Their enrichment patterns were confirmed by analysis at the KEGG functional orthologs (KOs) level (Figure 4E; Table S4J). Notably, compared with KOs depleted in roots, most root-enriched KOs were associated with functions related to cell motility, membrane transport, and signal transduction (Figures 4F and 4G), genes belonging to which have been previously identified in random barcoded transposon mutant sequencing in the laboratory in single

correspond to the number of crop species enriched or depleted, ranging from 0 to 4. The horizontal bars within the boxes represent median values. The tops and bottoms of the boxes represent the 75th and 25th percentiles, respectively. The upper and lower whiskers extend to data no more than 1.5× the interquartile range from the upper edge and lower edge of the box, respectively (root, $n = 332$; soil, $n = 75$, adjusted $p < 0.05$, Wilcoxon rank-sum test). See also Table S4J.
(F and G) Functional proportion of KOs enriched or depleted in roots of four host species. Doughnut charts indicate KOs consistently enriched (F) or depleted (G) in the roots of wheat, rice, maize, and *Medicago* compared with corresponding soils. KOs are presented at the superpathway level (root, $n = 332$; soil, $n = 75$, adjusted $p < 0.05$, Wilcoxon rank-sum test).
See also Figures S3 and S4.

plant species,[65,66] or genomic exploration of root microbes.[21,67] Our metagenomic findings underscore the significance and widespread presence of these functions in the root ecosystem across diverse soil backgrounds and crop species.

Genome references are essential for linking functional genes to specific taxa. Therefore, we searched for the crop root bacterial taxa that predominantly harbor the identified root-enriched functional KOs (Table S4K). Overall, phylogenetically related genera exhibited similar functional patterns (Figure S4A). Proteobacteria were particularly enriched in root-specific functions, aligning with their dominance in the crop root microbiome. Genomes from genera such as *Rhizobacter*, *Variovorax*, *Acidovorax*, and *Rubrivivax* encoded over 130 root-enriched KOs (Table S4L), and Gammaproteobacteria were enriched in functions related to cellular community (Figure S4B). In addition, KOs related to cell motility, including bacterial chemotaxis and flagellar assembly, were abundant in Proteobacteria, Firmicutes, and several Actinobacteriota, such as Microbacteriaceae and Nocardioidaceae. However, these functions were less prevalent in Bacteroidota and most other Actinobacteriota (Table S4L). This analysis identified key determinants of root-associated bacteria and highlights essential taxon-to-function links, offering potential targets to enhance bacterial competence within crop root ecosystems.

The CRVC reveals unreported viral genus-level clusters and enhanced viral genetic diversity in crop root ecosystems

Viruses intrinsically interact with other microbes and play pivotal roles in natural ecosystems.[68] However, a systematic collection of viral genomes is currently lacking in crop root microbiome research. Based on CRBC and crop bacterial genomes in published databases, along with 29.5 million contigs derived from root metagenomes in this study, we systematically retrieved viruses within crop root ecosystems. Using a hybrid analysis involving geNomad,[69] VirSorter,[70] and DeepVirFinder,[71] we identified 9,736 non-redundant viral genomes with completeness of over 50%, collectively referred to as the CRVC (Figure S5A; Table S5A; STAR Methods). Most (95.3%) of these viral genomes are identified as bacteriophages, with 71.0% exhibiting lysogenic behavior (Figure S5B; Table S5A), reflecting our pipeline's focus on discovering viral genomes from bacterial genomes and metagenomes. Nearly half of the genomes within the CRVC are of high quality, with 723 complete viral genomes identified based on direct terminal repeats (DTRs) or inverted terminal repeats (ITRs), and most of the CRVC genomes belonged to *Caudoviricetes* (Figures S5C and S5D; Table S5A). The virus sizes vary extensively, ranging from 1,501 to 477,160 bp, with a median length of 40,590 bp (Figure S5C). Notably, 65.5% of the viral genomes in the CRVC were derived from the CRBC and our root metagenomes (Figure S5D).

To identify novel viral taxa within the CRVC, we clustered all CRVC viral genomes with those from public databases (NCBI RefSeq,[42] Metagenomic Gut Virus catalog [MGV],[35] Global Ocean Viromes 2.0 [GOV 2.0],[37] and viruses from Integrated Microbial Genomes/Virus [IMG/VR][36] root and soil ecosystems) using the Minimum Information about an Uncultivated Virus Genome (MIUViG) proposed standard thresholds of 95% ANI over 85% alignment fraction.[72] We identified 7,653 species-level clusters (viral operational taxonomic units [vOTUs]) in the CRVC, of which 92.8% are not reported in existing databases (Figure S5E; Table S5B). Between 12.3% and 36.1% of CRVC proteins were successfully annotated using KEGG,[51] Pfam,[73] and VOGDB[74] databases, while 57.6% of proteins exhibiting unknown functions (Figures S5F and S5G; Table S5C), underscoring CRVC's vast genetic potential. To assign higher taxonomic ranks, we grouped genomes into genus-level categories based on whole-genome gene-sharing profiles (STAR Methods). Notably, 1,572 (50.8%) of the 3,097 viral genus-level clusters in the CRVC did not cluster with viral genomes from other published databases (Figure 5B; Table S5D; STAR Methods). To minimize technical biases from bacterial isolation when comparing databases, we compared metagenome-derived viruses in the CRVC with those in public databases. We found that the CRVC shared more viral clusters with IMG/VR root and soil environments than with human and ocean environments (Figure S5H), suggesting an ecosystem-specific nature of viral distribution.

Using genomes from the CRVC, we investigated viral genetic diversity within crop root metagenomes. High intra-population genetic diversity (microdiversity) enhances a viral population's ability to adapt to environmental changes, ensuring its evolutionary potential.[75–77] Microdiversity was assessed by aligning metagenomic sequencing reads to a representative genome and calculating the coverage diversity of each nucleotide position. Quantitative analysis revealed 2,690 viral species-level clusters in crop root metagenomes (Table S5E). These viruses were categorized into four groups based on distribution patterns: rare, single-crop regional, single-crop multizonal, and multi-crops multizonal (STAR Methods). Most viruses (82.4%) were consistently detected, with 31.8%–58.5% appearing across multiple locations but within a single crop species (Figure S5I). Notably, 123 viruses were found across different locations and crop species, accounting for an average of 13.8% of the total viral abundance (Table S5E). Microdiversity analysis showed higher genetic diversity in viruses with broader distribution, especially those infecting diverse crop species and spanning geographic locations (Figure 5C), suggesting that ecological niche differentiation drives viral selective variation. Lytic viruses were more abundant in root environments than temperate viruses, but temperate viruses exhibited greater microdiversity (Figure S5J; Tables S5E and S5F), suggesting the interactions with bacteria may shape their evolutionary pathways.

Phage-bacteria interactions in crop root ecosystems

Utilizing genome and metagenome resources, we first investigated the abundance of viruses in crop root ecosystems. Metagenomic data from roots and their corresponding soils were annotated using the CRVC and IMG/VR databases (STAR Methods). We observed a significantly higher relative abundance of phages in root microbiomes than in the corresponding soils (Figure 6A; Table S5E; relative abundance, $mean_{root}$ = 0.00645%, $mean_{soil}$ = 0.00002%). This pattern was consistent across 14 datasets representing diverse soil sources and crop species. Similar results were obtained when using only the IMG/VR database as reference (Figure S6A). This analysis

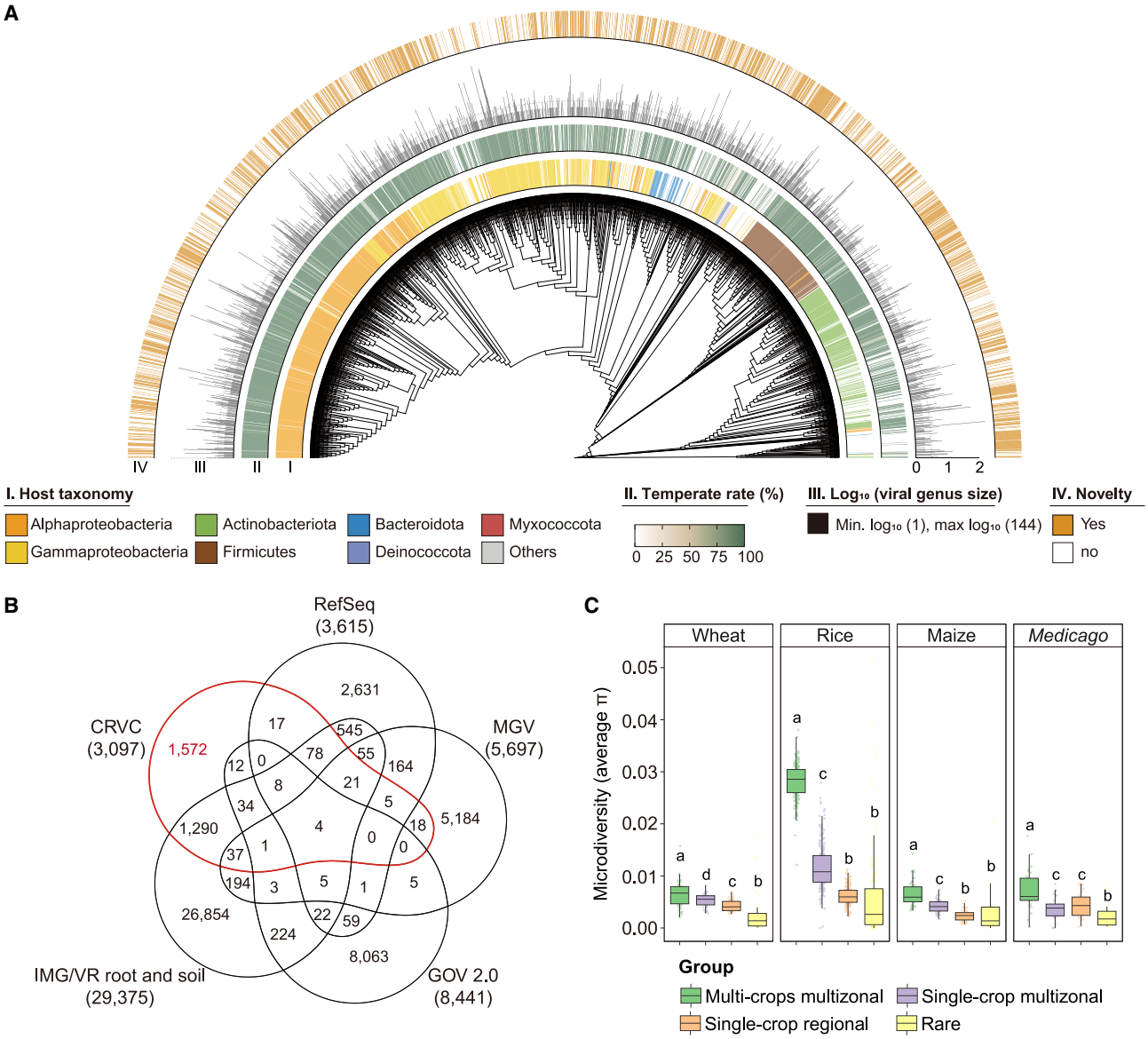


Figure 5. The CRVC unveils 1,572 unreported viral genus-level clusters and unique patterns in crop root microbiomes

(A) The CRVC reveals a striking diversity of unreported viral genus-level clusters and connections with crop root bacterial genomes. The hierarchical clustering dendrogram was built based on the clustered viral protein families in the CRVC using the Jaccard distance with the average clustering model (STAR Methods). The first circle (I) indicates the taxonomy of host bacteria for each viral cluster at the genus level, showing that 2,199 (71.0%) of CRVC clusters exhibit connections with crop root bacteria. Connections were determined according to viral contigs and CRISPR spacers detected in bacterial genomes. The second inner circle represents the proportion of temperate phages within each viral genus (II). Black bars show the $\log_{10}$-transformed genome numbers in each viral genus, ranging from 1 to 144 (III). In the outermost circle (IV), the gene-sharing network analysis reveals 1,572 unreported viral genus-level clusters in the CRVC compared with publicly available databases (STAR Methods).

(B) Half of the CRVC genus-level clusters are not reported in public databases. The Venn diagram illustrates overlaps of the CRVC genus-level clusters with the public viral databases including RefSeq, MGV, GOV2, and IMG/VR root and soil (STAR Methods).

(C) Viruses identified in root metagenomes across various crop species and geographical locations exhibit a higher microdiversity. Boxplots illustrate the microdiversity of viruses detected in roots of multiple crops and locations are higher compared with those detected in a single crop or location (adjusted $p < 0.05$, Kruskal-Wallis rank-sum test and Dunn's test; STAR Methods). Using prevalence characteristics, phages are categorized into four prevalence groups: stable presence across roots of multiple crops in multiple locations (multi-crops multizonal, green), within roots of a single crop species in multiple locations (single-crop multizonal, purple), within roots of a single crop species in a single location (single-crop regional, orange), and viruses with prevalence lower than 10% of samples (rare, yellow). Wheat, $n = 47$; rice, $n = 212$; maize, $n = 45$; *Medicago*, $n = 28$.

See also Figure S5.

CellPress
OPEN ACCESS

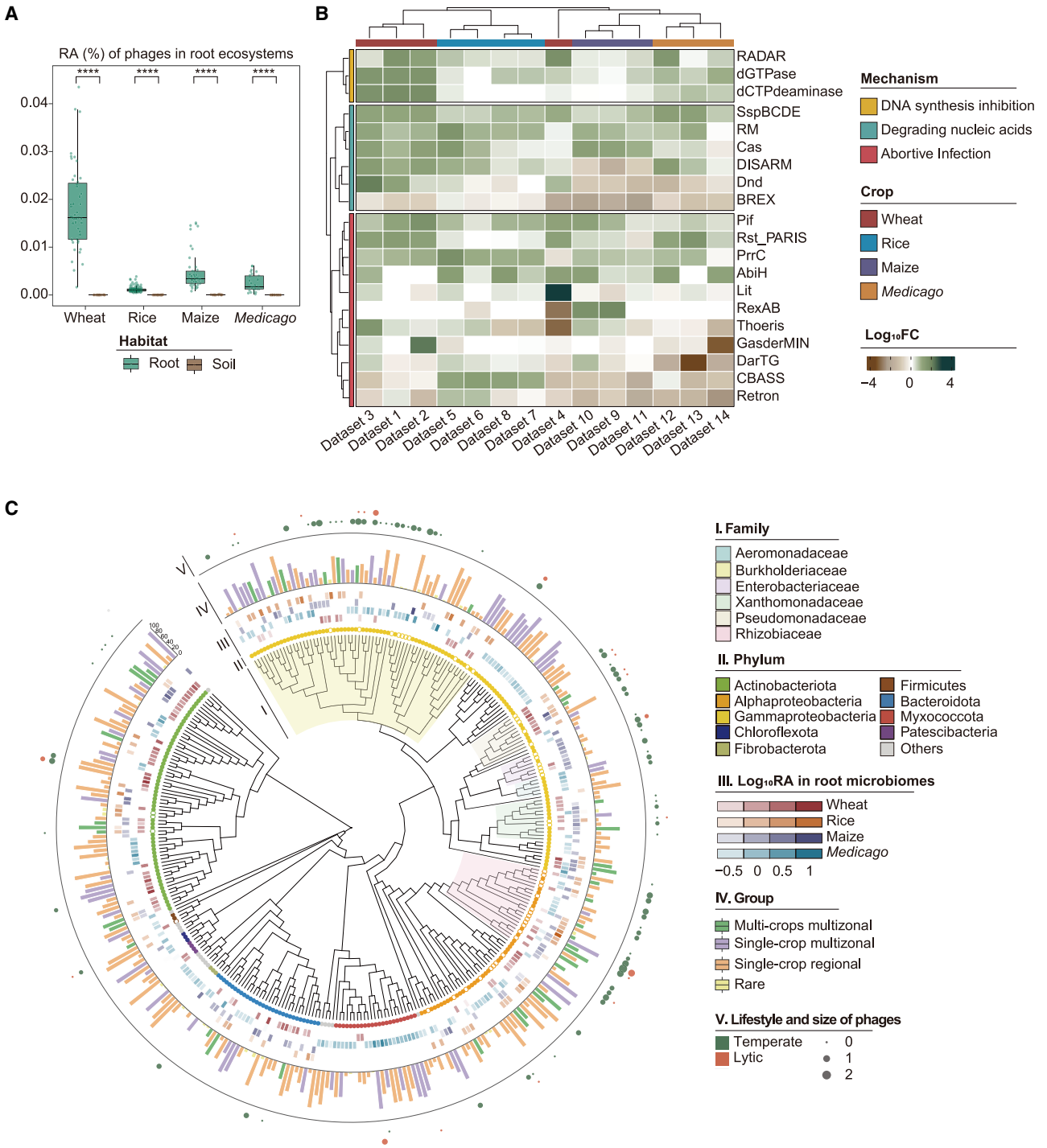


Figure 6. Phage-bacteria connections within crop root ecosystems

(A) Phage abundances in the root microbiomes are higher than those in corresponding soils. The box plot shows the relative abundances of phages in root and soil metagenomic samples from wheat, rice, maize, and *Medicago* (root, $n = 332$; soil, $n = 75$. ****adjusted $p < 0.0001$, Wilcoxon rank-sum test).

(B) The bacterial antiviral systems are enriched in the root microbiomes of multiple crops grown in diverse soil backgrounds. The heatmap illustrates the normalized $\log_{10}$-transformed fold changes in enrichment (green) and depletion (brown) of abundance concerning bacterial antiviral systems within the root microbiomes of wheat, rice, maize, and *Medicago* compared with their corresponding soils. Bacterial antiviral systems are categorized into three groups based on mechanisms: DNA synthesis inhibition, degrading nucleic acids, and abortive infection. Metagenomic datasets were clustered using the ward.D2 model.

(legend continued on next page)

highlights the adaptability of viruses colonizing the surfaces and interiors of roots.

Consistently, we observed that 94.0% of root bacterial genomes possessed at least one bacterial antiviral defense system, with a median of five defense families per genome (Figure S6B; Table S6A). In crop root microbiomes, most of the bacterial antiviral systems were more enriched compared with those in soils. Specifically, systems that inhibit DNA synthesis and degrade nucleic acids were consistently enriched in the roots across 14 datasets. For example, the R-M systems and CRISPR-Cas systems, two widespread and typical immune systems, exhibited convergent enrichment patterns in root microbiomes across crop species and soil sources (Figures 6B and S6C; Table S6B), suggesting the adaptability of root bacteria to phage-rich environments. In summary, the enrichment of both phages and bacterial antiviral defense systems within crop root microbiomes highlights the presence of complex and largely unexplored interactions between phages and root bacteria.

Next, we systematically investigated phage-bacteria interactions across root microbiomes from multiple crops and soil sources. We found that a total of 5,885 (60.2%) of the 9,772 crop root bacterial genomes had connections with phages, with phylum-specific proportions ranging from 50.9% to 83.5% (Figure S6D). These connections were identified through CRISPR spacer matches or the presence of phages (Table S6C), with most phages showing specific connections with distinct bacterial species (Figure S6E). Additionally, 27.0% of abundant and prevalent root bacteria exhibited connections with phages, as confirmed by both genomic matches and metagenomic co-occurrence (Figure 6C; Table S6D). These phage connections exhibited a preference for specific bacterial families, including Burkholderiaceae, Rhizobiaceae, Xanthomonadaceae, and Pseudomonadaceae (Figure 6C; Tables S6E and S6F; STAR Methods). Interestingly, phage-bacteria pairs confirmed by genomic matches exhibited significantly stronger associations within root ecosystems than those identified solely by co-occurrence analysis (Figure S6F; Table S6G), indicating that pairs confirmed by genomic matches are more reliable. Together, these findings underscore the pivotal role of the CRBC and CRVC resources in unraveling the intricate network of connections within the crop root microbiome.

DISCUSSION

Bacteria colonizing the surface and, more intimately, the interior of roots have significant impacts on crop growth and health.[1,2] Obtaining the genomic content of these bacteria and the corresponding isolates is essential for deciphering the mechanisms behind their interactions with host plants and for developing applications that benefit crop ecosystems. In this study, we established a systematic CRBC, comprising 6,699 genomes (68.9% from cultivated isolates) from three major grain crops wheat, rice, and maize, and the important forage legume *Medicago*, using two complementary methods: high-throughput bacterial cultivation and metagenome-derived MAGs (Figure 1). The CRBC contains 1,817 undefined bacterial species compared with all current public databases (Figure 2A), not only addressing the shortage of publicly available crop root bacterial genomes but also massively extending the diversity of bacterial genomes from all ecosystems.[31,46,45] Interestingly, 43.8% of crop root bacterial genomes encode at least two distinct PGP functional categories including nutrient utilization, biosynthesis of plant growth hormones, and stress resistance (Figure 3B)—functions that have garnered significant attention in recent years.[78,79] These genomic resources and extensive isolates provide unique opportunities to explore the roles and molecular mechanisms by which the root microbiome influences crop growth and health.[80–82]

Root metagenomics, specifically in the rhizoplane and endosphere, presents unique challenges compared with sequencing rhizosphere soil and gut microbiomes due to the high proportion of host DNA in root samples, which significantly raises the costs of capturing microbial taxa. These constraints have resulted in a shortage of crop root microbiome data, with limited samples in small-scale studies,[83] thus limiting our understanding of root metagenome patterns. In this study, we performed deep sequencing of 332 root metagenomic samples from 14 datasets and found that on average, 82.0% are host reads. Notably, bacteria dominate root microbiomes, constituting an average of 95.7% of the total microbial abundance, while fungi, archaea, viruses, and protists represented much smaller fractions (Figure S3C; Table S3I). This pattern is consistent with the bacterial dominance in the gut microbiome,[84,85] highlighting the importance of bacteria in the interactions between hosts and their associated microbiomes.

The CRBC represents a significant step forward in accurately quantifying and characterizing bacterial functional patterns within and across crop root ecosystems. Marker-gene-based profiling has already demonstrated that the taxonomic composition of the root microbiome is shaped by environmental factors and host plants,[2,6,11,12,14] such as the increase of Actinobacteriota in low-water soil environments.[9,80] Our comparison between metagenomes of roots and soils confirmed this taxonomic variation (Figure 4A). However, we were surprised to observe that root metagenome functions were conserved across 14 datasets, regardless of soil sources, host species, and genotypes

(C) The genome-level analysis reveals the broad connections of bacteria and phages in crop root ecosystems. The maximum-likelihood phylogenetic tree contains 330 bacterial species that are in the top ten relative abundances in the root microbiome of each metagenomic sample (n = 332; STAR Methods). The bacterial families showing intensive connections with phages are colored in the phylogenetic branches (I). The phylum information of each bacterial species is color-coded (II). Undefined crop root bacterial species are in solid points while published crop bacterial species are in hollow points. Log_{10}-transformed mean relative abundance of each bacterial species in the root microbiomes of each crop is shown in the heatmap, with each crop species color-coded (III). The prevalence of each bacterial species in crop root microbiomes are shown in the bar plot and categorized and color-coded into four groups (IV). The out layers (V) illustrate the genome-level connections of phages with each bacterial species in crop root microbiomes. The size of points reflects the log_{10}-transformed number of phage species interacting with each bacterial species. Temperate phages and lytic phages are color-coded in green and red, respectively. See also Table S6E.
See also Figure S6.

(Figure 4B). Genes associated with specific conserved functions, such as cell motility, secretion systems, and signaling transduction, have previously been identified in random barcoded transposon mutant sequencing in the laboratory in single plant species,[65,66] or genomic exploration of root microbes.[21,67] Our metagenomic results from diverse soil backgrounds, host plant species, and genotypes provide an ecological support for the significance and prevalence of these functions in crop root ecosystems. Given that similar conserved patterns are observed in the human gut microbiome across individuals with different backgrounds,[86,87] we hypothesize that microbiome-host interactions share conserved molecular principles. Our work provides valuable insights into potential targets for microbiome manipulation in agricultural applications.

Viruses have intrinsic connections with bacteria and are increasingly recognized for their roles in agriculture. However, their genomes and ecological patterns remain understudied in the crop root microbiome compared with other ecosystems.[88–91] Bacteria-phage connections represent an underexplored aspect in the crop root ecosystem, primarily due to a lack of genomic resources that limit systematic analysis. Our metagenomic samples from diverse crops and soil backgrounds revealed that nearly 30% of the prevalent root bacterial species across multiple crop species grown in various soils contained phage genomes or CRISPR spacer matches with phage genomes (Figure 6C). These connections showed a preference for specific bacterial families, including the abundant Rhizobiaceae, Xanthomonadaceae, and Pseudomonadaceae, which are known to play roles in crop growth and health.[92–95] Interestingly, viruses found across multiple hosts or locations exhibited significantly greater genetic diversity (Figure 5C). A similar pattern was observed in ocean viral data,[37] supporting the interpretation that the genetic diversity may contribute to viral speciation and adaptability in diverse environments.

Our genomes are valuable resources for exploring novel mechanisms and functional potentials within the crop root microbiome. For example, we have identified a type II-C CRISPR-Cas system promoting factor based on a *Chryseobacterium* genome in the CRBC.[96] The PcrIIC1 protein within the *CbCas9* (*Chryseobacterium* sp. Cas9, CbCas9) gene cluster mediates the dimerization of CbCas9, allowing it to tolerate mismatches in the spacer and protospacer adjacent motif (PAM) sequences, thereby improving bacterial immunity against phages.[96] This finding is in line with our observation that bacterial antiviral systems and viruses are both enriched in root microbiomes compared with soils (Figures 6A and 6B). In addition, by leveraging complete genome sequences, we examined the presence and coexistence of microbial functions at the genome level, insights that fragmented metagenomic sequencing cannot provide, thereby expanding our knowledge of potential beneficial bacteria. We also identified a significant number of undefined BGCs from diverse phylogenetic origins (Figure 3C), highlighting the genetic potential of the CRBC. Based on our bacterial and viral resources, we are establishing an international crop microbiome repository, which other scientists can continuously expand and benefit. This repository will serve as a valuable resource of genomes and isolates for the global crop root microbiome research community.

Together, the CRBC and CRVC genomic resources will serve an important role similar to that of well-known and widely used human and environmental microbiome genome resources,[31,46,88] advancing crop root microbiome research into mechanistic, genome-level understanding. The availability of a large number of isolates with accompanying genomic information will particularly stimulate functional and mechanistic studies of the crop root microbiome, ultimately benefiting applied agricultural sciences.

Limitations of the study

While the CRBC and CRVC provide comprehensive coverage of the dominant bacteria and viruses in the crop root microbiome, additional microbial genomes are needed particularly for the low abundant members of crop root ecosystems. The undefined species in this study are based solely on genome comparison; therefore, further investigation is still needed on their phenotypic and chemotaxonomic characteristics as well as ecological roles. These species also represent some rare cases of previously published isolates that lack genomes or associated metadata. Many bacterial and viral species-level clusters currently contain only one or two genomes, limiting the ability to identify their pangenomes and intraspecies diversity. Expanding microbial genomes to include crops grown in more diverse soil types and from various continents will also be crucial. Increased global cooperation to store and share root microbial genomes and isolates, with the standardized procedure, will be important and required in the future to advance root microbiome research to the genome and strain level.

RESOURCE AVAILABILITY

Lead contact

For further information concerning the genomes and bacterial isolates, please contact Dr. Yang Bai from Peking University (ybai@pku.edu.cn).

Materials availability

Details on CRBC bacterial isolate strains can be accessed through the Guangdong Microbial Culture Collection Center (GDMCC; https://gdmcc.net/#/index). Accession IDs for each isolate are listed in Table S1A. Furthermore, all cultivated bacterial strains are available upon request by contacting Dr. Yang Bai (www.cropmicrobiome.com; ybai@pku.edu.cn).

Data and code availability

The computational scripts for analysis in this work are available on GitHub (https://github.com/Bailab-pku/CRBC-and-CRVC). The bacterial and viral genomes, along with annotation files, and other related data are accessible on Zenodo (accessions: 14091751, 14095420, 13918137, and 13939322). Raw sequencing data for bacterial genomes and root metagenomes are available on NCBI (accession PRJNA1183633 and PRJNA1184367), and maize rhizosphere metagenome data are available on ENA (accession PRJEB77048). The genomic data in this work are also accessible through our website (http://www.cropmicrobiome.com/). The genomic datasets used in this manuscript are all publicly accessible. For details, please refer to the "deposited data" section in the key resources table.

ACKNOWLEDGMENTS

We thank Prof. Paul Schulze-Lefert (Max Planck Institute for Plant Breeding Research, Cologne, Germany) and Dr. Simon Roux (DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, USA) for their suggestions for improving the work. We thank Luisa Arias Giraldo and Nico Louwen

for their help with the design of the web portal. This work was supported by the National Key Research and Development Program of China (grant nos. 2022YFF1001800 to Y.B.), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant nos. XDA24020000 to Y.B. and XDA28030202 and XDA0450000 to J.Z.), The National Natural Science Foundation of China (grant nos. 32430003 to Y.B. and 32322002 to J.Z.), the Joint CAS-MPG Research Project (grant no. HZXM20225001MI to Y.B.), the CAS Project for Young Scientists in Basic Research (grant no. YSBR-078 to Y.B.), the Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (grant no. CI2024C007YN to Y.B.), WISH-ROOTS (2022-EJPSOILS-WISHROOTS-CAS to Y.B.), the Hainan Excellent Talent Team, Disciplinary Breakthrough Project of MOE, and the New Cornerstone Science Foundation through the XPLORER PRIZE to Y.B.

AUTHOR CONTRIBUTIONS

Y.B. designed and supervised the project. R.D. and F.L. performed the informatic analysis. S.C., A.L., and R.d.J. checked and supported the analysis. J.Z., H.X., J.-M.Q., and W.L. cultivated bacterial strains and collected crop root metagenomic samples. B.W., H.Z., L.J.U.P., and M.H.M. participated in the website construction. K.S. provided crop rhizosphere metagenomes in Switzerland. R.D., J.Z., F.L., A.L., R.d.J., C.M.J.P., K.S., and Y.B. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Plant materials and growth conditions
- METHOD DETAILS
  - Bacterial isolation, cultivation and genomic DNA extraction
  - Whole-genome sequencing and *de novo* assembly of bacterial isolates
  - Harvesting and metagenomic sequencing of root microbiome samples
  - Quality control, assemble and gene prediction of metagenomic samples
  - Binning and refinement of MAGs
  - Collecting genomes of crop root bacteria from public databases
  - Quality control, dereplication, and species clustering of bacterial genomes
  - Comparison of bacterial species within the CRBC and representative species in the GTDB database
  - Taxonomic classification and phylogenetic analysis of root bacterial genomes
  - Reads coverage of root and rhizosphere metagenomic samples using different genome databases
  - Functional annotation and BGC and antiviral system identification of bacterial genomes
  - PGP functions in bacterial genomes
  - Comparison of BGCs in crop root bacterial genomes and public databases
  - Taxonomic classification and functional annotation of crop root metagenomic samples
  - Prediction and quality control of viral genomes
  - Gene prediction, functional annotation, and protein clustering of viral genomes
  - Clustering of viral genomes at the species and genus level
  - Taxonomic classification of the CRVC
  - Host prediction of the CRVC
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quantification of bacterial species and evaluation of bacterial growth rate
  - Quantification of metagenomic reads and feature table generation
  - Rarefaction analysis and evaluation of microbial genetic potential
  - Quantification of viruses and antiviral defense systems in metagenomic samples
  - Microdiversity calculation of viral genomes
  - Definition of prevalence groups for bacterial and viral genomes
  - Statistic analysis and data visualization

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.02.013.

Received: June 29, 2024
Revised: November 14, 2024
Accepted: February 16, 2025
Published: March 12, 2025

REFERENCES

1. Bakker, P.A.H.M., Pieterse, C.M.J., de Jonge, R., and Berendsen, R.L. (2018). The soil-borne legacy. Cell *172*, 1178–1180. https://doi.org/10.1016/j.cell.2018.02.024.
2. Russ, D., Fitzpatrick, C.R., Teixeira, P.J.P.L., and Dangl, J.L. (2023). Deep discovery informs difficult deployment in plant microbiome science. Cell *186*, 4496–4513. https://doi.org/10.1016/j.cell.2023.08.035.
3. Trivedi, P., Leach, J.E., Tringe, S.G., Sa, T., and Singh, B.K. (2020). Plant-microbiome interactions: from community assembly to plant health. Nat. Rev. Microbiol. *18*, 607–621. https://doi.org/10.1038/s41579-020-0412-1.
4. Wang, F., Zhang, H., Liu, H., Wu, C., Wan, Y., Zhu, L., Yang, J., Cai, P., Chen, J., and Ge, T. (2024). Combating wheat yellow mosaic virus through microbial interactions and hormone pathway modulations. Microbiome *12*, 200. https://doi.org/10.1186/s40168-024-01911-z.
5. Zhang, J., Liu, Y.X., Zhang, N., Hu, B., Jin, T., Xu, H., Qin, Y., Yan, P., Zhang, X., Guo, X., et al. (2019). NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. Nat. Biotechnol. *37*, 676–684. https://doi.org/10.1038/s41587-019-0104-4.
6. Bulgarelli, D., Garrido-Oter, R., Münch, P.C., Weiman, A., Dröge, J., Pan, Y., McHardy, A.C., and Schulze-Lefert, P. (2015). Structure and function of the bacterial root microbiota in wild and domesticated barley. Cell Host Microbe *17*, 392–403. https://doi.org/10.1016/j.chom.2015.01.011.
7. He, X., Wang, D., Jiang, Y., Li, M., Delgado-Baquerizo, M., McLaughlin, C., Marcon, C., Guo, L., Baer, M., Moya, Y.A.T., et al. (2024). Heritable microbiome variation is correlated with source environment in locally adapted maize varieties. Nat. Plants *10*, 598–617. https://doi.org/10.1038/s41477-024-01654-7.
8. Carrión, V.J., Perez-Jaramillo, J., Cordovez, V., Tracanna, V., De Hollander, M., Ruiz-Buck, D., Mendes, L.W., van Ijcken, W.F.J., Gomez-Exposito, R., Elsayed, S.S., et al. (2019). Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science *366*, 606–612. https://doi.org/10.1126/science.aaw9285.
9. Xu, L., Naylor, D., Dong, Z., Simmons, T., Pierroz, G., Hixson, K.K., Kim, Y.M., Zink, E.M., Engbrecht, K.M., Wang, Y., et al. (2018). Drought delays development of the sorghum root microbiome and enriches for monoderm bacteria. Proc. Natl. Acad. Sci. USA *115*, E4284–E4293. https://doi.org/10.1073/pnas.1717308115.
10. Yu, P., He, X., Baer, M., Beirinckx, S., Tian, T., Moya, Y.A.T., Zhang, X., Deichmann, M., Frey, F.P., Bresgen, V., et al. (2021). Plant flavones enrich rhizosphere Oxalobacteraceae to improve maize performance under nitrogen deprivation. Nat. Plants *7*, 481–499. https://doi.org/10.1038/s41477-021-00897-y.

11. Edwards, J., Johnson, C., Santos-Medellín, C., Lurie, E., Podishetty, N.K., Bhatnagar, S., Eisen, J.A., and Sundaresan, V. (2015). Structure, variation, and assembly of the root-associated microbiomes of rice. Proc. Natl. Acad. Sci. USA *112*, E911–E920. https://doi.org/10.1073/pnas.1414592112.

12. Peiffer, J.A., Spor, A., Koren, O., Jin, Z., Tringe, S.G., Dangl, J.L., Buckler, E.S., and Ley, R.E. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc. Natl. Acad. Sci. USA *110*, 6548–6553. https://doi.org/10.1073/pnas.1302837110.

13. Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., Rauf, P., Huettel, B., Reinhardt, R., Schmelzer, E., et al. (2012). Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature *488*, 91–95. https://doi.org/10.1038/nature11336.

14. Lundberg, D.S., Lebeis, S.L., Paredes, S.H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., del Rio, T.G., et al. (2012). Defining the core Arabidopsis thaliana root microbiome. Nature *488*, 86–90. https://doi.org/10.1038/nature11237.

15. Garrido-Oter, R., Nakano, R.T., Dombrowski, N., Ma, K.W., AgBiome, T., McHardy, A.C., and Schulze-Lefert, P. (2018). Modular traits of the Rhizobiales root microbiota and their evolutionary relationship with symbiotic rhizobia. Cell Host Microbe *24*, 155–167.e155. https://doi.org/10.1016/j.chom.2018.06.006.

16. Rosconi, F., Rudmann, E., Li, J., Surujon, D., Anthony, J., Frank, M., Jones, D.S., Rock, C., Rosch, J.W., Johnston, C.D., et al. (2022). A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. Nat. Microbiol. *7*, 1580–1592. https://doi.org/10.1038/s41564-022-01208-7.

17. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. *5*, e1000344. https://doi.org/10.1371/journal.pgen.1000344.

18. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell *176*, 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

19. Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D., Dunn, M., Mkandawire, T.T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat. Biotechnol. *37*, 186–192. https://doi.org/10.1038/s41587-018-0009-7.

20. Bai, Y., Müller, D.B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P.C., Spaepen, S., Remus-Emsermann, M., et al. (2015). Functional overlap of the Arabidopsis leaf and root microbiota. Nature *528*, 364–369. https://doi.org/10.1038/nature16192.

21. Levy, A., Gonzalez, I.S., Mittelviefhaus, M., Clingenpeel, S., Paredes, S.H., Miao, J., Wang, K., Devescovi, G., Stillman, K., Monteiro, F., et al. (2018). Genomic features of bacterial adaptation to plants. Nat. Genet. *50*, 138–150. https://doi.org/10.1038/s41588-017-0012-9.

22. Wippel, K., Tao, K., Niu, Y., Zgadzaj, R., Kiel, N., Guan, R., Dahms, E., Zhang, P., Jensen, D.B., Logemann, E., et al. (2021). Host preference and invasiveness of commensal bacteria in the Lotus and Arabidopsis root microbiota. Nat. Microbiol. *6*, 1150–1162. https://doi.org/10.1038/s41564-021-00941-9.

23. Thoenen, L., Kreuzer, M., Pestalozzi, C., Florean, M., Mateo, P., Züst, T., Wei, A., Giroud, C., Rouyer, L., Gfeller, V., et al. (2024). The lactonase BxdA mediates metabolic specialisation of maize root bacteria to benzoxazinoids. Nat. Commun. *15*, 6535. https://doi.org/10.1038/s41467-024-49643-w.

24. Grubbs, K.J., Bleich, R.M., Santa Maria, K.C., Allen, S.E., Farag, S., AgBiome, T., Shank, E.A., and Bowers, A.A. (2017). Large-scale bioinformatics analysis of bacillus genomes uncovers conserved roles of natural products in bacterial physiology. mSystems *2*, e00040–e00017. https://doi.org/10.1128/mSystems.00040-17.

25. Nayfach, S., and Pollard, K.S. (2016). Toward accurate and quantitative comparative metagenomics. Cell *166*, 1103–1116. https://doi.org/10.1016/j.cell.2016.08.007.

26. Cavassim, M.I.A., Moeskjær, S., Moslemi, C., Fields, B., Bachmann, A., Vilhjálmsson, B.J., Schierup, M.H., W Young, J.P., and Andersen, S.U. (2020). Symbiosis genes show a unique pattern of introgression and selection within a Rhizobium leguminosarum species complex. Microb. Genom. *6*, e000351. https://doi.org/10.1099/mgen.0.000351.

27. Coombs, J.T., and Franco, C.M.M. (2003). Isolation and identification of Actinobacteria from surface-sterilized wheat roots. Appl. Environ. Microbiol. *69*, 5603–5608. https://doi.org/10.1128/AEM.69.9.5603-5608.2003.

28. Greenlon, A., Chang, P.L., Damtew, Z.M., Muleta, A., Carrasquilla-Garcia, N., Kim, D., Nguyen, H.P., Suryawanshi, V., Krieg, C.P., Yadav, S.K., et al. (2019). Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. Proc. Natl. Acad. Sci. USA *116*, 15200–15209. https://doi.org/10.1073/pnas.1900056116.

29. Epstein, B., Abou-Shanab, R.A.I., Shamseldin, A., Taylor, M.R., Guhlin, J., Burghardt, L.T., Nelson, M., Sadowsky, M.J., and Tiffin, P. (2018). Genome-wide association analyses in the model rhizobium Ensifer meliloti. mSphere *3*, e00386–e00318. https://doi.org/10.1128/mSphere.00386-18.

30. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504. https://doi.org/10.1038/s41586-019-0965-1.

31. Paoli, L., Ruscheweyh, H.J., Forneris, C.C., Hubrich, F., Kautsar, S., Bhushan, A., Lotti, A., Clayssen, Q., Salazar, G., Milanese, A., et al. (2022). Biosynthetic potential of the global ocean microbiome. Nature *607*, 111–118. https://doi.org/10.1038/s41586-022-04862-3.

32. Wang, X., Wei, Z., Yang, K., Wang, J., Jousset, A., Xu, Y., Shen, Q., and Friman, V.P. (2019). Phage combination therapies for bacterial wilt disease in tomato. Nat. Biotechnol. *37*, 1513–1520. https://doi.org/10.1038/s41587-019-0328-3.

33. Kauffman, K.M., Hussain, F.A., Yang, J., Arevalo, P., Brown, J.M., Chang, W.K., VanInsberghe, D., Elsherbini, J., Sharma, R.S., Cutler, M.B., et al. (2018). A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. Nature *554*, 118–122. https://doi.org/10.1038/nature25474.

34. Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering earth's virome. Nature *536*, 425–430. https://doi.org/10.1038/nature19094.

35. Nayfach, S., Páez-Espino, D., Call, L., Low, S.J., Sberro, H., Ivanova, N.N., Proal, A.D., Fischbach, M.A., Bhatt, A.S., Hugenholtz, P., et al. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat. Microbiol. *6*, 960–970. https://doi.org/10.1038/s41564-021-00928-6.

36. Camargo, A.P., Nayfach, S., Chen, I.A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S.J., Reddy, T.B.K., Mukherjee, S., Schulz, F., et al. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res. *51*, D733–D743. https://doi.org/10.1093/nar/gkac1037.

37. Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., et al. (2019). Marine DNA viral macro- and microdiversity from pole to pole. Cell *177*, 1109–1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

38. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense

systems in the microbial pangenome. Science *359*, eaar4120. https://doi.org/10.1126/science.aar4120.

39. Georjon, H., and Bernheim, A. (2023). Publisher Correction: The highly diverse antiphage defence systems of bacteria. Nat. Rev. Microbiol. *21*, 833. https://doi.org/10.1038/s41579-023-00987-y.
40. Al-Shayeb, B., Sachdeva, R., Chen, L.X., Ward, F., Munk, P., Devoto, A., Castelle, C.J., Olm, M.R., Bouma-Gregson, K., Amano, Y., et al. (2020). Clades of huge phages from across earth's ecosystems. Nature *578*, 425–431. https://doi.org/10.1038/s41586-020-2007-4.
41. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794. https://doi.org/10.1093/nar/gkab776.
42. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.
43. Chen, I.A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J.R., Seshadri, R., et al. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res. *47*, D666–D677. https://doi.org/10.1093/nar/gky901.
44. Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., Gupta, D., Gupta, V., Haseeb, M., Ihsan, M., et al. (2024). The European nucleotide archive in 2023. Nucleic Acids Res. *52*, D92–D97. https://doi.org/10.1093/nar/gkad1067.
45. Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.M., Huntemann, M., et al. (2021). A genomic catalog of earth's microbiomes. Nat. Biotechnol. *39*, 499–509. https://doi.org/10.1038/s41587-020-0718-6.
46. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.
47. Kuypers, M.M.M., Marchant, H.K., and Kartal, B. (2018). The microbial nitrogen-cycling network. Nat. Rev. Microbiol. *16*, 263–276. https://doi.org/10.1038/nrmicro.2018.9.
48. McGrath, J.W., Chin, J.P., and Quinn, J.P. (2013). Organophosphonates revealed: new insights into the microbial metabolism of ancient molecules. Nat. Rev. Microbiol. *11*, 412–419. https://doi.org/10.1038/nrmicro3011.
49. Zeng, J., Tu, Q., Yu, X., Qian, L., Wang, C., Shu, L., Liu, F., Liu, S., Huang, Z., He, J., et al. (2022). PCycDB: a comprehensive and accurate database for fast analysis of phosphorus cycling genes. Microbiome *10*, 101. https://doi.org/10.1186/s40168-022-01292-1.
50. Dai, Z., Liu, G., Chen, H., Chen, C., Wang, J., Ai, S., Wei, D., Li, D., Ma, B., Tang, C., et al. (2020). Long-term nutrient inputs shift soil microbial functional profiles of phosphorus cycling in diverse agroecosystems. ISME J. *14*, 757–770. https://doi.org/10.1038/s41396-019-0567-9.
51. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.
52. Nett, R.S., Montanares, M., Marcassa, A., Lu, X., Nagel, R., Charles, T.C., Hedden, P., Rojas, M.C., and Peters, R.J. (2017). Elucidation of gibberellin biosynthesis in bacteria reveals convergent evolution. Nat. Chem. Biol. *13*, 69–74. https://doi.org/10.1038/nchembio.2232.
53. Salazar-Cerezo, S., Martinez-Montiel, N., Garcia-Sanchez, J., Perez-y-Terron, R., and Martinez-Contreras, R.D. (2018). Gibberellin biosynthesis and metabolism: a convergent route for plants, fungi and bacteria. Microbiol. Res. *208*, 85–98. https://doi.org/10.1016/j.micres.2018.01.010.
54. Kieber, J.J., and Schaller, G.E. (2018). Cytokinin signaling in plant development. Development *145*, dev149344. https://doi.org/10.1242/dev.149344.
55. Frébortová, J., and Frébort, I. (2021). Biochemical and structural aspects of cytokinin biosynthesis and degradation in bacteria. Microorganisms *9*, 1314. https://doi.org/10.3390/microorganisms9061314.
56. Zhang, P., Jin, T., Kumar Sahu, S., Xu, J., Shi, Q., Liu, H., and Wang, Y. (2019). The distribution of tryptophan-dependent indole-3-acetic acid synthesis pathways in bacteria unraveled by large-scale genomic analysis. Molecules *24*, 1411. https://doi.org/10.3390/molecules24071411.
57. Glick, B.R., Cheng, Z., Czarny, J., and Duan, J. (2007). Promotion of plant growth by acc deaminase-producing soil bacteria. Eur. J. Plant Pathol. *119*, 329–339. https://doi.org/10.1007/s10658-007-9162-4.
58. Ding, P., and Ding, Y. (2020). Stories of salicylic acid: a plant defense hormone. Trends Plant Sci. *25*, 549–565. https://doi.org/10.1016/j.tplants.2020.01.004.
59. Kerbarh, O., Ciulli, A., Howard, N.I., and Abell, C. (2005). Salicylate biosynthesis: overexpression, purification, and characterization of Irp9, a bifunctional salicylate synthase from Yersinia enterocolitica. J. Bacteriol. *187*, 5061–5066. https://doi.org/10.1128/JB.187.15.5061-5066.2005.
60. Bhattacharyya, P.N., and Jha, D.K. (2012). Plant growth-promoting rhizobacteria (PGPR): emergence in agriculture. World J. Microbiol. Biotechnol. *28*, 1327–1350. https://doi.org/10.1007/s11274-011-0979-9.
61. Hayat, R., Ali, S., Amara, U., Khalid, R., and Ahmed, I. (2010). Soil beneficial bacteria and their role in plant growth promotion: a review. Ann. Microbiol. *60*, 579–598. https://doi.org/10.1007/s13213-010-0117-1.
62. Wang, X., Zhang, J., Lu, X., Bai, Y., and Wang, G. (2024). Two diversities meet in the rhizosphere: root specialized metabolites and microbiome. J. Genet. Genomics *51*, 467–478. https://doi.org/10.1016/j.jgg.2023.10.004.
63. Kautsar, S.A., Blin, K., Shaw, S., Weber, T., and Medema, M.H. (2021). BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res. *49*, D490–D497. https://doi.org/10.1093/nar/gkaa812.
64. Terlouw, B.R., Blin, K., Navarro-Muñoz, J.C., Avalon, N.E., Chevrette, M.G., Egbert, S., Lee, S., Meijer, D., Recchia, M.J.J., Reitz, Z.L., et al. (2023). MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. *51*, D603–D610. https://doi.org/10.1093/nar/gkac1049.
65. Cole, B.J., Feltcher, M.E., Waters, R.J., Wetmore, K.M., Mucyn, T.S., Ryan, E.M., Wang, G., Ul-Hasan, S., McDonald, M., Yoshikuni, Y., et al. (2017). Genome-wide identification of bacterial plant colonization genes. PLoS Biol. *15*, e2002860. https://doi.org/10.1371/journal.pbio.2002860.
66. Georgoulis, S.J., Shalvarjian, K.E., Helmann, T.C., Hamilton, C.D., Carlson, H.K., Deutschbauer, A.M., and Lowe-Power, T.M. (2021). Genome-wide identification of tomato xylem sap fitness factors for three plant-pathogenic Ralstonia species. mSystems *6*, e0122921. https://doi.org/10.1128/mSystems.01229-21.
67. Vannier, N., Mesny, F., Getzke, F., Chesneau, G., Dethier, L., Ordon, J., Thiergart, T., and Hacquard, S. (2023). Genome-resolved metatranscriptomics reveals conserved root colonization determinants in a synthetic microbiota. Nat. Commun. *14*, 8274. https://doi.org/10.1038/s41467-023-43688-z.
68. Piel, D., Bruto, M., Labreuche, Y., Blanquart, F., Goudenège, D., Barcia-Cruz, R., Chenivesse, S., Le Panse, S., James, A., Dubert, J., et al. (2022). Phage-host coevolution in natural populations. Nat. Microbiol. *7*, 1075–1086. https://doi.org/10.1038/s41564-022-01157-1.
69. Camargo, A.P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P.S.G., Nayfach, S., and Kyrpides, N.C. (2024). Identification of mobile genetic elements with geNomad. Nat. Biotechnol. *42*, 1303–1312. https://doi.org/10.1038/s41587-023-01953-y.

70. Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ *3*, e985. https://doi.org/10.7717/peerj.985.

71. Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., Poplin, R., and Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. Quant. Biol. *8*, 64–77. https://doi.org/10.1007/s40484-019-0187-4.

72. Roux, S., Adriaenssens, E.M., Dutilh, B.E., Koonin, E.V., Kropinski, A.M., Krupovic, M., Kuhn, J.H., Lavigne, R., Brister, J.R., Varsani, A., et al. (2019). Minimum information about an uncultivated virus genome (MIUViG). Nat. Biotechnol. *37*, 29–37. https://doi.org/10.1038/nbt.4306.

73. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539. https://doi.org/10.1038/msb.2011.75.

74. Trgovec-Greif, L., Hellinger, H.J., Mainguy, J., Pfundner, A., Frishman, D., Kiening, M., Webster, N.S., Laffy, P.W., Feichtinger, M., and Rattei, T. (2024). VOGDB-database of virus orthologous groups. Viruses *16*, 1191. https://doi.org/10.3390/v16081191.

75. Siranosian, B.A., Tamburini, F.B., Sherlock, G., and Bhatt, A.S. (2020). Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. Nat. Commun. *11*, 280. https://doi.org/10.1038/s41467-019-14103-3.

76. Shah, S.A., Deng, L., Thorsen, J., Pedersen, A.G., Dion, M.B., Castro-Mejía, J.L., Silins, R., Romme, F.O., Sausset, R., Jessen, L.E., et al. (2023). Expanding known viral diversity in the healthy infant gut. Nat. Microbiol. *8*, 986–998. https://doi.org/10.1038/s41564-023-01345-7.

77. Mavrich, T.N., and Hatfull, G.F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. Nat. Microbiol. *2*, 17112. https://doi.org/10.1038/nmicrobiol.2017.112.

78. Compant, S., Clément, C., and Sessitsch, A. (2010). Plant growth-promoting bacteria in the rhizo- and endosphere of plants: their role, colonization, mechanisms involved and prospects for utilization. Soil Biol. Biochem. *42*, 669–678. https://doi.org/10.1016/j.soilbio.2009.11.024.

79. Finkel, O.M., Salas-González, I., Castrillo, G., Conway, J.M., Law, T.F., Teixeira, P.J.P.L., Wilson, E.D., Fitzpatrick, C.R., Jones, C.D., and Dangl, J.L. (2020). A single bacterial genus maintains root growth in a complex microbiome. Nature *587*, 103–108. https://doi.org/10.1038/s41586-020-2778-7.

80. Santos-Medellín, C., Liechty, Z., Edwards, J., Nguyen, B., Huang, B., Weimer, B.C., and Sundaresan, V. (2021). Prolonged drought imparts lasting compositional changes to the rice root microbiome. Nat. Plants *7*, 1065–1077. https://doi.org/10.1038/s41477-021-00967-1.

81. Wagner, M.R., Tang, C., Salvato, F., Clouse, K.M., Bartlett, A., Vintila, S., Phillips, L., Sermons, S., Hoffmann, M., Balint-Kurti, P.J., et al. (2021). Microbe-dependent heterosis in maize. Proc. Natl. Acad. Sci. USA *118*, e2021965118. https://doi.org/10.1073/pnas.2021965118.

82. Zhang, L., Zhang, M., Huang, S., Li, L., Gao, Q., Wang, Y., Zhang, S., Huang, S., Yuan, L., Wen, Y., et al. (2022). A highly conserved core bacterial microbiota with nitrogen-fixation capacity inhabits the xylem sap in maize plants. Nat. Commun. *13*, 3361. https://doi.org/10.1038/s41467-022-31113-w.

83. Ofek-Lalzar, M., Sela, N., Goldman-Voronov, M., Green, S.J., Hadar, Y., and Minz, D. (2014). Niche and host-associated functional signatures of the root surface microbiome. Nat. Commun. *5*, 4950. https://doi.org/10.1038/ncomms5950.

84. Yan, Q., Li, S., Yan, Q., Huo, X., Wang, C., Wang, X., Sun, Y., Zhao, W., Yu, Z., Zhang, Y., et al. (2024). A genomic compendium of cultivated human gut fungi characterizes the gut mycobiome and its relevance to common diseases. Cell *187*, 2969–2989.e24. https://doi.org/10.1016/j.cell.2024.04.043.

85. Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. *37*, 179–185. https://doi.org/10.1038/s41587-018-0008-8.

86. Human; Microbiome; Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. https://doi.org/10.1038/nature11234.

87. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. Nature *489*, 220–230. https://doi.org/10.1038/nature11550.

88. Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D., and Lawley, T.D. (2021). Massive expansion of human gut bacteriophage diversity. Cell *184*, 1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.

89. Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., et al. (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science *348*, 1261498. https://doi.org/10.1126/science.1261498.

90. Ma, B., Wang, Y., Zhao, K., Stirling, E., Lv, X., Yu, Y., Hu, L., Tang, C., Wu, C., Dong, B., et al. (2024). Biogeographic patterns and drivers of soil viromes. Nat. Ecol. Evol. *8*, 717–728. https://doi.org/10.1038/s41559-024-02347-2.

91. Kwak, M.-J., Kong, H.G., Choi, K., Kwon, S.-K., Song, J.Y., Lee, J., Lee, P.A., Choi, S.Y., Seo, M., Lee, H.J., et al. (2018). Rhizosphere microbiome structure alters to enable wilt resistance in tomato. Nat. Biotechnol. *36*, 1100–1109. https://doi.org/10.1038/nbt.4232.

92. Compant, S., Nowak, J., Coenye, T., Clément, C., and Ait Barka, E. (2008). Diversity and occurrence of Burkholderia spp. in the natural environment. FEMS Microbiol. Rev. *32*, 607–626. https://doi.org/10.1111/j.1574-6976.2008.00113.x.

93. Mercado-Blanco, J., and Bakker, P.A.H.M. (2007). Interactions between plants and beneficial Pseudomonas spp.: exploiting bacterial traits for crop protection. Antonie Leeuwenhoek *92*, 367–389. https://doi.org/10.1007/s10482-007-9167-1.

94. Masson-Boivin, C., Giraud, E., Perret, X., and Batut, J. (2009). Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? Trends Microbiol. *17*, 458–466. https://doi.org/10.1016/j.tim.2009.07.004.

95. Ryan, R.P., Vorhölter, F.-J., Potnis, N., Jones, J.B., Van Sluys, M.-A., Bogdanove, A.J., and Dow, J.M. (2011). Pathogenomics of Xanthomonas: understanding bacterium–plant interactions. Nat. Rev. Microbiol. *9*, 344–355. https://doi.org/10.1038/nrmicro2558.

96. Zhang, S., Sun, A., Qian, J.M., Lin, S., Xing, W., Yang, Y., Zhu, H.Z., Zhou, X.Y., Guo, Y.S., Liu, Y., et al. (2024). Pro-CRISPR PcrIIC1-associated Cas9 system for enhanced bacterial immunity. Nature *630*, 484–492. https://doi.org/10.1038/s41586-024-07486-x.

97. Basenko, E.Y., Pulman, J.A., Shanmugasundram, A., Harb, O.S., Crouch, K., Starns, D., Warrenfeltz, S., Aurrecoechea, C., Stoeckert, C.J., Jr., Kissinger, J.C., et al. (2018). FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J. Fungi (Basel) *4*, 39. https://doi.org/10.3390/jof4010039.

98. Quiza, L., Tremblay, J., Greer, C.W., Hemmingsen, S.M., St-Arnaud, M., Pozniak, C.J., and Yergeau, E. (2021). Rhizosphere shotgun metagenomic analyses fail to show differences between ancestral and modern wheat genotypes grown under low fertilizer inputs. FEMS Microbiol. Ecol. *97*, fiab071. https://doi.org/10.1093/femsec/fiab071.

99. Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. *42*, D633–D642. https://doi.org/10.1093/nar/gkt1244.

100. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361. https://doi.org/10.1093/nar/gkw1092.

101. Siddell, S.G., Smith, D.B., Adriaenssens, E., Alfenas-Zerbini, P., Dutilh, B.E., Garcia, M.L., Junglen, S., Krupovic, M., Kuhn, J.H., Lambert, A.J., et al. (2023). Virus taxonomy and the role of the international committee on taxonomy of viruses (ICTV). J. Gen. Virol. *104*, 001840. https://doi.org/10.1099/jgv.0.001840.

102. Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H. (2016). Linking virus genomes with host taxonomy. Viruses *8*, 66. https://doi.org/10.3390/v8030066.

103. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

104. Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., and Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat. Commun. *13*, 2561. https://doi.org/10.1038/s41467-022-30269-9.

105. Zhang, J., Liu, Y.X., Guo, X., Qin, Y., Garrido-Oter, R., Schulze-Lefert, P., and Bai, Y. (2021). High-throughput cultivation and identification of bacteria from the plant root microbiota. Nat. Protoc. *16*, 988–1012. https://doi.org/10.1038/s41596-020-00444-7.

106. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

107. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

108. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes de novo assembler. Curr. Protoc. Bioinformatics *70*, e102. https://doi.org/10.1002/cpbi.102.

109. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

110. Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome *6*, 158. https://doi.org/10.1186/s40168-018-0541-1.

111. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

112. Lu, J., Rincon, N., Wood, D.E., Breitwieser, F.P., Pockrandt, C., Langmead, B., Salzberg, S.L., and Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. Nat. Protoc. *17*, 2815–2839. https://doi.org/10.1038/s41596-022-00738-y.

113. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

114. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

115. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

116. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

117. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

118. Wu, Y.-W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics *32*, 605–607. https://doi.org/10.1093/bioinformatics/btv638.

119. Chan, P.P., Lin, B.Y., Mak, A.J., and Lowe, T.M. (2021). TRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. *49*, 9077–9096. https://doi.org/10.1093/nar/gkab688.

120. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868. https://doi.org/10.1038/ismej.2017.126.

121. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

122. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. PLoS Comput. Biol. *14*, e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

123. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics *36*, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

124. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. https://doi.org/10.1038/nmeth.4197.

125. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

126. Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., and Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. *49*, W29–W35. https://doi.org/10.1093/nar/gkab335.

127. Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L.C., Yeong, M., Cruz-Morales, P., Abubucker, S., et al. (2020). A computational framework to explore large-scale biosynthetic diversity. Nat. Chem. Biol. *16*, 60–68. https://doi.org/10.1038/s41589-019-0400-9.

128. Kautsar, S.A., van der Hooft, J.J.J., de Ridder, D., and Medema, M.H. (2021). BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience *10*, giaa154. https://doi.org/10.1093/gigascience/giaa154.

129. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

130. Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol. *39*, 727–736. https://doi.org/10.1038/s41587-020-00797-0.

131. Gao, Y., and Li, H. (2018). Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples. Nat. Methods *15*, 1041–1044. https://doi.org/10.1038/s41592-018-0182-0.

132. Madeira, F., Madhusoodanan, N., Lee, J., Eusebi, A., Niewielska, A., Tivey, A.R.N., Lopez, R., and Butcher, S. (2024). The EMBL-EBI job bispatcher sequence analysis tools framework in 2024. Nucleic Acids Res. *52*, W521–W525. https://doi.org/10.1093/nar/gkae241.

133. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

134. Putri, G.H., Anders, S., Pyl, P.T., Pimanda, J.E., and Zanini, F. (2022). Analysing high-throughput sequencing data in python with HTSeq 2.0. Bioinformatics *38*, 2943–2945. https://doi.org/10.1093/bioinformatics/btac166.

135. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585. https://doi.org/10.1038/s41587-020-00774-7.

136. Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 90. https://doi.org/10.1186/s40168-020-00867-0.

137. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

138. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584. https://doi.org/10.1093/nar/30.7.1575.

139. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. *46*, W246–W251. https://doi.org/10.1093/nar/gky425.

140. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

141. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37. https://doi.org/10.1093/nar/gkr367.

142. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Software *4*, 1686. https://doi.org/10.21105/joss.01686.

143. Pedersen, T.L. (2025). patchwork: the composer of plots. https://github.com/thomasp85/patchwork.

144. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

145. Garnier, S., Ross, N., Rudis, R., Camargo, A.P., Sciaini, M., and Scherer, C. (2024). viridis(Lite) – colorblind-friendly color maps for R. https://sjmgarnier.github.io/viridis/.

146. Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szoecs, E., et al. (2025). vegan: community ecology package. https://github.com/vegandevs/vegan.

147. Kassambara, A. (2023). rstatix: pipe-friendly framework for basic statistical tests. https://rpkgs.datanovia.com/rstatix/.

148. Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., Dai, Z., Lam, T.T., Guan, Y., and Yu, G. (2022). Ggtree: a serialized data object for visualization of a phylogenetic tree and annotation data. Imeta *1*, e56. https://doi.org/10.1002/imt2.56.

149. Xu, S., Dai, Z., Guo, P., Fu, X., Liu, S., Zhou, L., Tang, W., Feng, T., Chen, M., Zhan, L., et al. (2021). ggtreeExtra: compact visualization of richly annotated phylogenetic data. Mol. Biol. Evol. *38*, 4039–4042. https://doi.org/10.1093/molbev/msab166.

150. Revell, L.J. (2024). phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ *12*, e16505. https://doi.org/10.7717/peerj.16505.

151. Graves, S., Piepho, H.-P., Selzer, L., and Dorai-Raj, S. (2024). multcompView: visualizations of paired comparisons. https://github.com/lselzer/multcompview.

152. Gao, C.H., Chen, C., Akyol, T., Dusa, A., Yu, G., Cao, B., and Cai, P. (2024). ggVennDiagram: intuitive venn diagram software extended. Imeta *3*, e177. https://doi.org/10.1002/imt2.177.

153. Pohlert, T. (2024). PMCMRplus: calculate pairwise multiple comparisons of mean rank sums extended. https://CRAN.R-project.org/package=PMCMRplus.

154. Neuwirth, E. (2022). RColorBrewer: colorbrewer palettes. https://CRAN.R-project.org/package=RColorBrewer.

155. Wang, H., Xu, X., Vieira, F.G., Xiao, Y., Li, Z., Wang, J., Nielsen, R., and Chu, C. (2016). The power of inbreeding: NGS-based GWAS of rice reveals convergent evolution during rice domestication. Mol. Plant *9*, 975–985. https://doi.org/10.1016/j.molp.2016.04.018.

156. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731. https://doi.org/10.1038/nbt.3893.

157. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26. https://doi.org/10.1093/nar/gkab1112.

158. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. *2*, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7.

159. Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. Biol. Conserv. *61*, 1–10. https://doi.org/10.1016/0006-3207(92)91201-3.

160. Hu, L., Robert, C.A.M., Cadot, S., Zhang, X., Ye, M., Li, B., Manzo, D., Chervet, N., Steinger, T., van der Heijden, M.G.A., et al. (2018). Root exudate metabolites drive plant-soil feedbacks on growth and defense by shaping the rhizosphere microbiota. Nat. Commun. *9*, 2738. https://doi.org/10.1038/s41467-018-05122-7.

161. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288. https://doi.org/10.1093/bioinformatics/btm098.

162. UniProt Consortium (2007). The universal protein resource (UniProt). Nucleic Acids Res. *35*, D193–D197. https://doi.org/10.1093/nar/gkl929.

163. Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. *37*, 632–639. https://doi.org/10.1038/s41587-019-0100-8.

164. Dion, M.B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. Nat. Rev. Microbiol. *18*, 125–138. https://doi.org/10.1038/s41579-019-0311-5.

165. Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., Matot, E., Jona, G., Harmelin, A., Cohen, N., et al. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science *349*, 1101–1106. https://doi.org/10.1126/science.aac4812.

166. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60. https://doi.org/10.1038/nature11450.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial strains		
CRBC isolates	This paper	http://www.cropmicrobiome.com; https://gdmcc.net/#/index
Biological samples		
Root and soil samples of *Triticum aestivum* L.	This paper	N/A
Root and soil samples of *Oryza sativa* L.	This paper	N/A
Root and soil samples of *Zea mays* L.	This paper	N/A
Root and soil samples of *Medicago truncatula*	This paper	N/A
Deposited data		
Bacterial genomes of CRBC	This paper	http://www.cropmicrobiome.com; Zenodo: 14091751
Viral genomes of CRVC	This paper	http://www.cropmicrobiome.com; Zenodo: 14091751
Gene catalogs and KEGG annotations	This paper	http://www.cropmicrobiome.com; Zenodo: 14095420
BGCs of CRBC	This paper	http://www.cropmicrobiome.com; Zenodo: 13918137 and 13939322
Raw sequencing of root metagenomes and genomes from cultured bacteria	This paper	NCBI: PRJNA1183633 and PRJNA1184367
Maize rhizosphere metagenomic samples	This paper	ENA: PRJEB77048
NCBI bacterial genomes	NCBI	https://ftp.ncbi.nlm.nih.gov/genomes/genbank
NCBI RefSeq bacterial and viral genomes	NCBI	https://ftp.ncbi.nlm.nih.gov/genomes/refseq/
IMG/M bacterial genome dataset	JGI[43]	https://img.jgi.doe.gov/cgi-bin/m/main.cgi
IMG/VR Viral Resources v4	JGI[36]	https://img.jgi.doe.gov/cgi-bin/vrer/main.cgi
FungiDB	Basenko et al.[97]	https://fungidb.org/fungidb/app
Chickpea root bacterial symbionts	Greenlon et al.[28]	NCBI: PRJNA453501
Rhizobiales genome collection	Garrido-Oter et al.[15]	ENA: PRJEB26998
Lotus japonicus root culture collection	Wippel et al.[22]	ENA: PRJEB37696
Rhizobium leguminosarum collection	Cavassim et al.[26]	NCBI: PRJNA510726
Ensifer spp collection	Epstein et al.[29]	NCBI: PRJNA401434
Bacillus genomes collection	Grubbs et al.[24]	NCBI: PRJNA400804
Wheat rhizosphere metagenomic samples	Quiza et al.[98]	NCBI: PRJNA643787
Genome Taxonomy Database (GTDB)	GTDB team[41]	https://gtdb.ecogenomic.org/
GEM database	Nayfach et al.[45]	https://portal.nersc.gov/GEM
UHGG database	Almeida et al.[46]	http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/
OMG database	Paoli et al.[31]	https://microbiomics.io/ocean/
RDP database	Cole et al.[99]	https://mothur.org/wiki/rdp_reference_files/
BiG-FAM database	Kautsar et al.[63]	https://bigfam.bioinformatics.nl
MIBiG database	Terlouw et al.[64]	https://mibig.secondarymetabolites.org
KEGG database	Kanehisa et al.[100]	https://www.genome.jp/kegg/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MGV database	Nayfach et al.[35]	https://portal.nersc.gov/MGV/
GOV 2.0 database	Gregory et al.[37]	https://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/GOV2.0
ICTV database	Siddell et al.[101]	https://ictv.global/
Virus-Host Dtabase	Mihara et al.[102]	https://www.genome.jp/virushostdb/
Pfam database	Mistry et al.[103]	http://pfam.xfam.org/
VOGDB database	Trgovec-Greif et al.[74]	http://vogdb.org
DefenseFinder Structures DB	Tesson et al.[104]	https://defensefinder.mdmlab.fr/wiki/structure
Oligonucleotides		
27F (AGAGTTTGATCCTGGCTCAG)	Zhang et al.[105]	16S primers for Sanger sequencing
1492R (TACGGCTACCTTGTTACGACTT)	Zhang et al.[105]	16S primers for Sanger sequencing
799F (AACMGGATTAGATACCCKG)	Zhang et al.[105]	First step primers for high throughput identification
1193R (ACGTCATCCCCACCTTCC)	Zhang et al.[105]	First step primers for high throughput identification
Software and algorithms		
USEARCH	Edgar et al.[106]	http://www.drive5.com/usearch/
Trimmomatic v0.39	Bolger et al.[107]	https://github.com/usadellab/Trimmomatic
SPAdes v3.14.0	Prjibelski et al.[108]	https://github.com/ablab/spades
CheckM v1.1.8	Parks et al.[109]	https://github.com/Ecogenomics/CheckM
MetaWRAP v1.3.2	Uritskiy et al.[110]	https://github.com/bxlab/metaWRAP
KneadData v0.7.6	The huttenhower lab	https://github.com/biobakery/kneaddata
Kraken2 v2.1.1	Wood et al.[111]	https://github.com/DerrickWood/kraken2
Kraken tool extract_kraken_reads.py	Lu et al.[112]	https://github.com/jenniferlu717/KrakenTools/blob/master/extract_kraken_reads.py
SeqKit	Shen et al.[113]	https://bioinf.shenwei.me/seqkit/
MEGAHIT v1.2.9	Li et al.[114]	https://github.com/voutcn/megahit
Prodigal v2.6.3	Hyatt et al.[115]	https://github.com/hyattpd/Prodigal
CD-HIT v.4.8.1	Fu et al.[116]	https://sites.google.com/view/cd-hit
MetaBAT2	Kang et al.[117]	https://bitbucket.org/berkeleylab/metabat
MaxBin2	Wu et al.[118]	http://sourceforge.net/projects/maxbin/
tRNAscan-SE v 2.0.12	Chan et al.[119]	http://trna.ucsc.edu/tRNAscan-SE/
barrnap v0.9	Torsten Seemann	https://github.com/tseemann/barrnap
dRep v3.2.2	Olm et al.[120]	https://github.com/MrOlm/drep
Mash v2.3	Ondov et al.[121]	https://github.com/marbl/mash
MUMmer v4.0.0	Marcais et al.[122]	https://github.com/mummer4/mummer
GTDB-Tk v2.0.0	Chaumeil et al.[123]	https://github.com/Ecogenomics/GTDBTk
Salmon v1.9.0	Patro et al.[124]	https://combine-lab.github.io/salmon/
DIAMOND v2.0.15.153	Buchfink et al.[125]	https://github.com/bbuchfink/diamond
antiSMASH v6.1.0	Blin et al.[126]	https://antismash.secondarymetabolites.org/#!/start
BiG-SCAPE v1.1.5	Navarro-Muñoz et al.[127]	https://github.com/medema-group/BiG-SCAPE
DefenseFinder v1.2.2	Tesson et al.[104]	https://defensefinder.mdmlab.fr/
BiG-SLICE v1	Kautsar et al.[128]	https://github.com/medema-group/bigslice
Bowtie 2 v2.4.5	Langmead and Salzberg[129]	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
InStrain v1.8.0	Olm et al.[130]	https://github.com/MrOlm/inStrain/
DEMIC	Gao and Li[131]	https://sourceforge.net/projects/demic/
EMBOSS Transeq v6.6.0.0	Madeira et al.[132]	https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq
SAMtools v1.9	Li et al.[133]	https://www.htslib.org/
HTseq v2.0.2	Putri et al.[134]	https://htseq.readthedocs.io/en/latest/
geNomad v1.5.2	Camargo et al.[69]	https://portal.nersc.gov/genomad
VirSorter v1.0.6	Roux et al.[70]	https://github.com/simroux/VirSorter
DeepVirFinder v1.0	Ren et al.[71]	https://github.com/jessieren/DeepVirFinder
CheckV v1.0.1	Nayfach et al.[135]	https://bitbucket.org/berkeleylab/checkv/src/master/
Prodigal v2.11.0-gv	Camargo et al.[69]	https://portal.nersc.gov/genomad
VIBRANT v1.2.1	Kieft et al.[136]	https://github.com/AnantharamanLab/VIBRANT
MMSeqs2 v13.45111	Steinegger and Söding[137]	https://github.com/soedinglab/MMseqs2
mcl v14-137	Enright et al.[138]	https://micans.org/mcl/
CRISPRCasFinder v2.0.3	Couvin et al.[139]	https://github.com/dcouvin/CRISPRCasFinder
BLAST	Camacho et al.[140]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
CoverM v0.6.1	Aroney et al.	https://github.com/wwood/CoverM
hmmsearch v3.3.2	Finn et al.[141]	https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch
tidyverse (R package)	Wickham et al.[142]	https://www.tidyverse.org/
patchwork (R package)	Thomas Lin Pedersen[143]	https://github.com/thomasp85/patchwork
ComplexHeatmap (R package)	Gu et al.[144]	https://github.com/jokergoo/ComplexHeatmap
viridis (R package)	Garnier et al.[145]	https://cran.r-project.org/web/packages/viridis/
vegan (R package)	Oksanen et al.[146]	https://github.com/vegandevs/vegan
rstatix (R package)	Alboukadel Kassambara[147]	https://rpkgs.datanovia.com/rstatix/
ggtree (R package)	Xu et al.[148]	https://www.bioconductor.org/packages/ggtree
ggtreeExtra (R package)	Xu et al.[149]	https://bioconductor.org/packages/release/bioc/html/ggtreeExtra.html
phytools (R package)	Liam J. Revell[150]	https://cran.r-project.org/web/packages/phytools/index.html
multcompView (R package)	Graves et al.[151]	https://github.com/lselzer/multcompview
ggVennDiagram (R package)	Gao et al.[152]	https://github.com/gaospecial/ggVennDiagram/
PMCMRplus (R package)	Thorsten Pohlert[153]	https://cran.r-project.org/web/packages/PMCMRplus/index.html
RColorBrewer (R package)	Erich Neuwirth[154]	https://cran.r-project.org/web/packages/RColorBrewer/index.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Plant materials and growth conditions

Healthy wild-type cultivars of wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), maize (*Zea mays* L.), and *Medicago* (*Medicago truncatula*) were used for bacterial isolation and cultivation. Wheat cultivars, including BenDiHuangHuaMai, China Spring, Jing 411, RHT, and Xiaoyan 54, were cultivated at Changping Farm in Beijing, China. Rice cultivars, including Nipponbare and IR24, were grown on

the same farm as wheat but in a separate field. Maize cultivars, including Ji 53 and Zong 3, were cultivated in Dongying, Shandong, China. *Medicago* cultivars, including Jemalong A17 and R108, were grown in the greenhouse using agricultural soil collected from Changping Farm in Beijing, China (Tables S1A–S1C).

For field experiments, wheat cultivars, including Fangnong 16, Jimai 78, Jimai 22, Xiaoyan 54, and Zhoumai 18, were cultivated in fields at Changping, Beijing, and Yangling, Shanxi. Rice cultivars, including Nipponbare, IR24, and materials from the USDA rice mini-core collection,[155] were grown in fields at Hainan, Hefei and Beijing in China. Maize cultivars, including inbred lines Chang 7-2, Jing 724, Zheng 58, Zi 330, 5237, as well as hybrid varieties Jingke 968, HG968, Jingnongke 728, Ximeng 1919, Ximeng 3358, and Ximeng 6, were cultivated in fields at Shangzhuang, Beijing and Quzhou, Hebei. *Medicago* cultivars, including Jemalong A17 and R108, were grown in the greenhouse with agricultural soil collected from Shangzhuang, Changping Farm, Beijing, and Sipin, Jilin (Table S1D).

For both bacterial isolation and field sampling, fresh roots were collected from healthy crops harvested at the vegetative stage of growth, see following details.

METHOD DETAILS

Bacterial isolation, cultivation and genomic DNA extraction

Bacteria were isolated from the roots of wheat, rice, maize, and *Medicago* (Table S1A). Fresh roots were collected from healthy crops grown in agricultural soils. Three independent plants for each crop genotype were harvested at the vegetative stage. Roots were immersed in 30 mL of sterile 1× PBS to remove loosely adhering soil particles and vortexed at 180 rpm for 15 min; this process was repeated three times. The roots were cut into 2 mm sections and mixed to maximize bacterial species diversity. Visible nodules and nodule initials in *Medicago* roots were removed. A total of 0.02 g of mixed tissues was immersed in 200 μL of 10 mM $MgCl_2$ and ground into a homogeneous mixture. The homogenate was diluted to different concentrations using 10% TSB and R2A liquid media. The diluted homogenate was distributed into 96-well, sterile cell culture plates and cultivated for 16–20 days at room temperature.

Cultivated bacterial DNA was extracted to amplify the V5-V7 regions of the 16S rRNA genes via two-sided barcoded PCR.[5,105] The DNA extraction from bacterial cultures was carried out through a lysis procedure: 6 μL of bacterial culture was mixed with 10 μL of Buffer I, which is composed of 25 mM NaOH and 0.2 mM EDTA at a pH of 12, and the mixture was then incubated at 95°C for 30 minutes, then the pH was neutralized by adding 10 μL of Buffer II (containing 40 mM Tris-HCl at pH 7.5). To identify and track bacterial isolates in 96-well plates, a two-step PCR approach was employed, utilizing degenerate primers 799F and 1193R that were customized with unique barcodes for each well and plate to amplify the V5-V7 variable regions of the bacterial 16S rRNA genes. In the first PCR phase, 3 μL of the lysed bacterial DNA was amplified in a 30 μL reaction mixture that included 0.75 U of HS-Taq DNA polymerase, 10× buffer, 0.2 mM dNTPs (supplied by Takara), and 0.1 μM each of the forward (799F) and reverse (1193R) primers (obtained from Life Technologies). The PCR conditions for this stage consisted of an initial denaturation at 94°C for 2 minutes, followed by 25 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute, with a final extension step at 72°C for 5 minutes. The PCR products from each well were then diluted 40-fold to prepare for the second PCR stage. In the second PCR stage, 3 μL of the diluted first-stage PCR product was amplified in a 30 μL reaction mixture that contained 0.75 U of HS-Taq DNA Polymerase, 10× buffer, 0.2 mM dNTPs, 0.1 μM of one of the 96 barcoded forward primers,[105] and 0.1 μM of the reverse barcoded primer. The PCR cycling conditions for this stage were: an initial denaturation at 94°C for 2 minutes, followed by 25 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute, and a final extension at 72°C for 5 minutes. After amplification, the PCR products were purified using the Agencourt AMPure XP Kit from Beckman Coulter GmbH and the Wizard® SV Gel and PCR Clean-Up System from Promega. The DNA concentrations were measured using the PicoGreen dsDNA Assay Kit from Life Technologies, and the samples were pooled in equal volumes.

Purified PCR products were subjected to Illumina sequencing. After bacterial identification based on 16S rRNA marker genes using USEARCH,[99,106] bacteria were selected from target wells in the 96-well cell culture plates and further purified on 1/2 TSB or R2A agar plates. After three continuous purifications, single colonies were transferred to a liquid medium and cultured for 3–5 days. Liquid cultures were then validated by Sanger sequencing with 27F and 1492R primers. To select taxonomically diverse isolates, we chose bacterial strains with a threshold of $\leq$ 97% 16S rRNA gene identity within each crop species for whole-genome sequencing. At the same time and to capture intraspecies genetic diversity, we also included replicated bacteria that shared $\geq$ 97% 16S rRNA gene identity but were isolated from independent crop roots or different soils. Key criteria for replicated strains were that they must represent independent host colonization events. For bacterial genomic DNA extraction, 500 μL of bacterial solution was added to a Lysing Matrix E tube and homogenized twice at 7,200 rpm for 30 s with a Precellys Evolution homogenizer (Bertin Technologies). DNA was extracted using a FastDNA Spin kit (MP Biomedicals), according to the manufacturer's instructions. DNA concentrations were measured using a PicoGreen dsDNA Assay kit (Life Technologies). In total, 1 μg of DNA from each bacterial strain was used for whole-genome sequencing.

Whole-genome sequencing and *de novo* assembly of bacterial isolates

For each bacterial strain, genomic DNA was sequenced on an Illumina NovaSeq 6000 platform using 150-base pair (bp) paired-end reads, with 1 Gbp of raw data sequenced per genome to ensure ~20× sequencing depth. In total, 5.5 Tbp of raw reads were collected. Clean reads with adapters removed were filtered by quality analysis using Trimmomatic v0.39[107] with options 'SLIDINGWINDOW:4:20' and 'MINLEN:100' and assembled using SPAdes v3.14.0[108] with options '–isolate, –careful, –cov-cutoff auto' to form

contigs, and only contigs longer than 1,000 bp were retained for subsequent analysis. After assembly, genome purities were evaluated by using the lineage_wf workflow in CheckM v1.1.8.[109] Multi-isolates with contamination of > 5% were split into separate genomes using the binning module of MetaWRAP v1.3.2.[110] In total, we obtained 4,620 bacterial genomes with less than 5% contamination from pure isolates (wheat, 1,497; rice, 1,287; maize, 1,057; *Medicago*, 779). The majority (99.1%) of these genomes are of high quality, surpassing the 90.0% completeness threshold, and having contamination levels lower than 5.0% (Table S1A; STAR Methods).

Harvesting and metagenomic sequencing of root microbiome samples

In total, 332 root samples were *in situ* collected from wheat, rice, maize, and *Medicago* grown in 9 field soils. To ensure the representativeness and diversity of the crop root microbiome, at least two plant genotypes of each crop species were grown in three to four different soil backgrounds and were categorized into 14 datasets (Table S1D). All plants were harvested at the vegetative stage. For wheat, rice, and maize plants, individual root tissue was dug out from the field and gently shaken to remove loosely adhering soil. Root microbiome samples were then washed and collected as described previously.[5] *Medicago* plants were grown in a greenhouse in agricultural soils from geographically different locations. Visible nodules and nodule initials in *Medicago* roots were removed. Roots were then washed and collected the same way as described before. As a control, for each dataset, four to six bulk soil samples were collected from corresponding fields or pots without plants. In total, 75 bulk soil samples were collected. Both root and bulk soil samples were frozen in liquid nitrogen and stored at –80 °C. Root and soil samples were smashed for DNA extraction using a FastDNA SPIN kit (MP Biomedicals). DNA was fragmented by sonication following Illumina's instructions to generate PCR-free libraries for shotgun sequencing.

Quality control, assemble and gene prediction of metagenomic samples

In total, 10.3 Tbp of raw reads were generated across 407 root and soil metagenomic samples. After removing adapter, barcode sequences, and low-quality reads, host crop reads were depleted using the KneadData v0.7.6 profile (https://github.com/biobakery/kneaddata) with the options 'SLIDINGWINDOW:4:20, MINLEN:50, and –very-sensitive'. To further remove host crop reads, reads were taxonomically classified by Kraken2 v2.1.1[112] with the default database, and reads classified as host crop species were filtered using the Kraken tool extract_kraken_reads.py.[111] For root samples, the percentage of clean microbial reads varied across datasets, with an average of 17.9% (Table S1D). To achieve relatively balanced coverage of microbial diversity, cleaned reads were normalized to no more than 6 Gbp or 8 Gbp for root and soil samples using SeqKit tool,[113] respectively. For each dataset, clean reads from roots or soil were *de novo* co-assembled into contigs using MEGAHIT v1.2.9[114] under meta large mode with a minimum contig length threshold of 500 bp. Prokaryotic coding sequences were predicted from the assembled contigs using Prodigal v2.6.3[115] with the option '-p meta'. To remove redundant gene sequences, predicted sequences were clustered into non-redundant (NR) genes at 95% identity and 90% coverage of the shorter gene using CD-HIT v.4.8.1[116] with options '-c 0.95, -as 0.9, -G 0, -g 1, -d 0, and -l 150'. Considering their limited genetic information, NR genes shorter than 150 bp were removed. The resulting NR genes for both root and soil subsets were then further merged by clustering to generate an integrated NR gene reference for later use as the reference for metagenomic read quantification.

Binning and refinement of MAGs

MAGs were reconstructed from microbial contigs of 332 root metagenomic samples using MetaWRAP v1.3.2.[110] *De novo* metagenomic assembly was also performed on individual samples according to the previously mentioned methods to identify unique microbes. Assembled contigs with lengths of over 1,000 bp were binned into individual draft genomes based on sequence composition and coverage information using MetaBAT2[117] and MaxBin2[118] as implemented in the binning module of MetaWRAP with default parameters. To improve the quality of the assembled genomes, the bin_refinement module was used on the assembled and binned genomes with a contamination cut-off below 10% and a completeness cut-off over 50%.[156]

Collecting genomes of crop root bacteria from public databases

To uncover novel bacterial species within the CRBC, we systematically collected crop root bacterial genomes across different public databases, including NCBI,[157] IMG/M,[43] and ENA[44] databases. A total of 5,441 plant root-related bacterial genomes were obtained from public databases using metadata from April 2023. For details, in NCBI, we scanned the NCBI Datasets genome table (https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/) and BioProject and BioSample information from NCBI plant-associated environmental packages. Keywords such as root, rhizosphere, rhizoplane, endophyte, endosphere, and nodule were used to filter the root-associated genomes. As a result, we obtained 3,168 root-related bacterial genomes from NCBI. For the IMG/M system (https://img.jgi.doe.gov/), 2,273 root-related bacterial genomes were downloaded based on the criterion of ecosystem type belonging to "Root", which also includes genomes from rhizosphere and nodules. Additionally, 294 bacterial genomes from lotus roots under accession number PRJEB37696 were downloaded from the ENA database. We then screened the crop hosts of these root-related bacterial genomes by both their English names and scientific names, e.g. rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), maize (*Zea mays*), and barley (*Hordeum vulgare*), according to the World Crops Database (https://world-crops.com). In total 3,073 quality-checked crop root-related bacterial genomes were obtained from public databases (Table S1G).

Quality control, dereplication, and species clustering of bacterial genomes

The quality of all 10,973 crop root bacterial genomes was assessed by CheckM v1.1.8[109] using the lineage_wf workflow with default parameters. In total, 9,772 genomes passed the quality criterion of a score exceeding 50 (QS = completeness – 5 × contamination)[158] and were retained for further analysis. These genomes consisted of 4,618 isolates and 2,081 MAGs from the CRBC and 3,073 bacterial genomes from public databases (Tables S1A and S1G). Prediction of tRNA and rRNA in bacterial genomes was conducted using tRNAscan-SE v 2.0.12[119] and barrnap v0.9 (https://github.com/tseemann/barrnap) with default parameters, respectively. High-quality bacterial genomes met the following specific criteria: completeness > 90%, contamination < 5%, and the presence of *23S*, *16S*, and *5S rRNA* genes along with at least *18 tRNAs*.[156] Yet, the challenge of assembling *rRNA* genes from short-read sequencing data has led to difficulty in obtaining the *5S rRNA* gene. For subsequent functional analysis, we defined genomes with completeness > 90% and contamination < 5% as high quality.

We dereplicated the 9,772 genomes using an ANI ≥ 99.9% and coverage ≥ 30% using dRep v3.2.2.[120] The options used for dereplication were '-strW 0, -centW 0, -nc 0.3, and -sa 0.999'. Subsequently, clustering at the species level was performed with an ANI of ≥ 95% and coverage of ≥ 30%. For species-level clustering, ANI was set to -sa 0.95. To designate representatives for NR genomes and species, the genome with the highest quality was selected, prioritizing cultured genomes over MAGs. The scoring criteria were based on the following formula: completeness – 5 × contamination + 0.5 × log (N_{50}). In total, 2,212 species lacking published crop root genomes were identified as undefined bacterial species from crop roots within the CRBC (Table S2A).

Comparison of bacterial species within the CRBC and representative species in the GTDB database

To explore the novelty of CRBC genomes compared with bacterial genomes across different habitats, we compared all the genomes within the CRBC to 62,291 reference prokaryotic species in GTDB Release 207 (8 April 2022).[41] We first converted the GTDB reference species and CRBC into sketch files by Mash v2.3[121] with default parameters. We then used mash dist to estimate the pairwise Mash distances between CRBC genomes and GTDB representative species. For each CRBC genome, we aligned the genome to the best-matched reference GTDB genome with the lowest Mash distance using DNAdiff v1.3 from MUMmer v4.0.0[122] to calculate the genome coverage (AlignedBases) and ANI (AvgIdentity). We considered a CRBC genome to match the reference GTDB if the coverage was ≥ 30% and ANI was ≥ 95%. CRBC species lacking a match to the GTDB were defined as undefined species across various habitats (Table S2A).

To infer the higher taxonomic ranks of the undefined CRBC genomes, we employed a strategy based on relative evolutionary divergence (RED) to delineate taxa above the species level, as recommended by GTDB-tk.[123] First, we used the de_novo_wf workflow to construct a *de novo* phylogenetic tree for the CRBC and GTDB reference genomes. The tree was then annotated, and the taxonomic ranks were inferred using GTDB-tk. Potential unknown taxa were labeled according to the rank inference results. The analysis revealed that these species encompass the following potential unique taxa: 1 phylum, 1 class, 3 orders, 15 families, and 101 genera (Table S2A).

Taxonomic classification and phylogenetic analysis of root bacterial genomes

Taxonomic classification of all bacteria was conducted using GTDB-Tk v2.0.0,[123] the classify_wf workflow, and the unsplit bacterial tree with the option '–full_tree'. The GTDB-Tk infer method was used to construct a maximum-likelihood phylogenetic tree *de novo* using representatives of root bacterial species or genera. For tree construction, we used a concatenated multiple sequence alignment of 120 bacterial phylogenetically informative bacterial markers, analyzed under the WAG model for amino-acid substitution. Phylogenetic diversity of the species tree was calculated by the sum of total branch lengths,[159] and PD of each phylum was calculated by the sum of branch lengths of the representative genomes belonging to the phylum (Table S2B). The phylogenetic gain (PG) of the CRBC compared with that of publicly available root genomes was calculated by the formula[158] $PG_{CRBC} = PD_{total} - PD_{Pub}$. Phylogenetic analysis revealed the limited species diversity present in public genomes of crop root bacteria, contributing merely 25.6% to the overall phylogenetic diversity (PD) of all crop root bacteria.

Reads coverage of root and rhizosphere metagenomic samples using different genome databases

Read coverage analysis of metagenomic data was conducted using Salmon v1.9.0[124] with options '–libType A, –meta, and –validateMappings', and mapping rates were used to evaluate read utilization efficiency of diverse genome databases. We focused on root-related metagenomic samples including 42 root metagenome samples not used for CRBC construction (wheat, $n = 12$; rice, $n = 12$; maize, $n = 6$; *Medicago*, $n = 12$) and 37 rhizosphere metagenome samples sourced from previous studies[98,160] (wheat, $n = 29$; maize grown in Europe [PRJEB77048], $n = 8$, two independent sequencing from four rhizosphere samples; Table S2E). To maximize comparability, host reads were removed and the metagenomic data were normalized to 1 Gbp per sample. Bacterial genome reference databases included NCBI RefSeq (September, 2022), GTDB Release 207,[41] GEM,[45] UHGG[46] and OMD.[31] Coverages of crop bacterial genomes from the CRBC and published databases was evaluated individually and then integrated for analysis (Table S2E).

Functional annotation and BGC and antiviral system identification of bacterial genomes

Coding sequences for all crop root bacterial genomes were predicted using Prodigal v2.6.3[115] with options '-p single, -f gff, and -a, -d'. These predicted protein sequences were compared with prokaryotic orthologs of the KEGG database[100] (v.102 release April 2021) using the DIAMOND BLASTP v2.0.15.153[125] with options '–outfmt 6, –max-target-seqs 1, –evalue 1e-5, –sensitive, –block-size 6,

and –index-chunks 1', and the identity threshold was set to 50% (Table S4L). Secondary metabolite BGCs were identified using antiSMASH[126] v6.1.0 with the option '–cb-knownclusters'. Because fragmented BGCs are likely to introduce bias for genome comparison, BGCs were subsequently screened to retain those predicted from contigs longer than 5 kb,[31] resulting in a total of 48,643 BGCs, with 67% predicted to be complete by antiSMASH (Table S3D). These BGCs were divided into BGC classes according to BiG-SCAPE v1.1.5.[127] For individual genomes, BGC composition was calculated as the percentage of BGCs that belong to each class. The prediction of bacterial antiviral defense systems within bacterial genomes was performed using DefenseFinder v1.2.2[104] with default parameters (Table S6B). Defense system families in each genome were calculated by summing the number of unique subsystems.

PGP functions in bacterial genomes

To elucidate the PGP functions encoded within the CRBC and published crop root bacterial genomes, bacterial PGP biological processes and genes/orthologs involved were compiled from previous studies. These included genes involved in nutrient utilization, such as phosphorus nutrition,[48–50] nitrogen fixation,[47] siderophore biosynthesis,[51] biosynthesis of plant growth hormones including IAA[52,53,56], GA,[52,53,56] and CK,[54,55] and resistance to biotic and abiotic stresses by biosynthesis of ACC deaminase,[57] salicylic acid,[58,59] and ethylene[51] (Tables S3A and S3B). For PGP functional annotation, we collected and curated PGP-related proteins within the KEGG database, encompassing those encoded across eukaryotes, prokaryotes, and viruses. Additionally, proteins involved in bacterial GA biosynthesis, such as experimentally-validated copalyl diphosphate synthase and kaurene synthase[52,53] that were not incorporated by KEGG, were manually downloaded and added into this PGP-related protein reference. To ensure robustness and accuracy in profiling PGP features, we used high-quality, non-redundant genomes within crop root bacterial genomes (n = 6,109). Based on BLAST results, as mentioned earlier, PGP genes encoded in each genome were summarized at the KO level as 0/1 binary data. A PGP feature was considered present when all the orthologs involved in a specific PGP process were encoded within a genome (Table S3C). The conservation of specific PGP functions at the genus level was assessed by determining the percentage of NR genomes encoding that particular function.

Comparison of BGCs in crop root bacterial genomes and public databases

To quantify BGC novelty and richness across phylogenies, BGCs were further clustered into 12,865 GCFs by computing all-against-all cosine distance using BiG-SLICE v1[128] and a 0.2 distance threshold. We compared these BGCs with the public BGC databases, including the computationally predicted BGCs in the BiG-FAM database[63] and experimentally validated BGCs from the MIBiG v3.1 database.[64] For each of these BGCs, the minimum cosine distance (d_{min}) to public BGCs was defined as its distance against the public database. The average distance of a BGC within a GCF was considered its final cosine distance ($\bar{d}_{min}$). A GCF was defined as undefined when its $\bar{d}_{min}$ against the public database was greater than 0.2. To illustrate the extent of GCF in terms of BGC classes and taxonomy, the $\bar{d}_{min}$ calculated between crop root GCFs and each of these public databases was displayed along the gradient distance from 0 to 1.00 with an interval of 0.05 (Table S3E). GCFs containing BGCs that belong to several different BGC classes were annotated as "Multiple classes". At the genus level, BGC composition and GCF richness were summarized as median values across all high-quality NR genomes under a specific genus. When a genus included BGCs from genomes of the CRBC, the genus was annotated as CRBC present.

Taxonomic classification and functional annotation of crop root metagenomic samples

To optimize microbial gene identification for metagenomic analysis, NR genes were taxonomically classified with Kraken2 v2.1.1[112] using the default database incorporated with the CRBC, published crop root bacteria, and publicly available reference databases including GTDB,[41] IMG/VR v4,[36] FungiDB,[97] and Ensemble protists in addition to the default Kraken built-in database. Briefly, the reference genome fasta files were downloaded and formatted to fit with Kraken2 format requirements and were integrated into Kraken2 default libraries (Tables S3F–S3I). For functional annotation, the NR gene set was first translated into protein sequences using EMBOSS Transeq v6.6.0.0[132] and annotated by DIAMOND BLASTP against the KEGG protein databases.

Prediction and quality control of viral genomes

We predicted viral sequences from 9,772 bacterial genomes and assembled contigs of 332 crop root metagenomes. To achieve this, we used geNomad v1.5.2,[69] which uses a hybrid framework that integrates both alignment-free and gene-based models to identify viral contigs. Additionally, to ensure comprehensive predictions, we used two different algorithms for viral sequence identification, i.e., VirSorter v1.0.6[70] and DeepVirFinder v1.0,[71] which rely on reference genome datasets and deep learning for viral sequence prediction, respectively. For geNomad (v1.3), we used end-to-end execution with the conservative preset and the default database. For VirSorter, Viromedb (–db 2) was used as a reference to predict prophage and viral fragments from genomes, and viral contigs classified as category 1, 2, 4, or 5 were retained. Contigs were scored by DeepVirFinder with default parameters, and those with a P value < 0.05 were considered potential viral sequences. Predicted viral contigs were further filtered using a length threshold of 1 kb.[35] Genome quality, gene classification and identification, and trimming of host regions were conducted with CheckV v1.0.1[135] using the end-to-end workflow. Finally, 16,546 viral genomes with host gene content below 30% and genome quality classified as medium or above (medium-quality, high-quality, and complete) were retained. To prevent redundant sequences predicted by the

different algorithms, we deduplicated the viral sequences based on 100% sequence similarity and 100% coverage of short sequences. This process resulted in a final set of predicted 9,736 unique crop root viral genomes (Table S5A).

Gene prediction, functional annotation, and protein clustering of viral genomes

For a comprehensive comparison, we obtained viral genomes from public databases of representative ecosystems, including NCBI RefSeq, the Metagenomic Gut Virus catalogue (MGV),[35] Global Ocean Viromes 2.0 (GOV2.0),[37] and IMG/VR v4.[36] Within IMG/VR, viruses originating from roots and soil were considered representative of plant root and soil environments. To ensure comparability, only genomes with CheckV quality classified as medium or above were retained, with the exception of those from NCBI RefSeq with confirmed completeness. Viral proteins from the NCBI virus were directly downloaded, whereas for other databases we used Prodigal v2.11.0-gv[69] to predict coding sequences with default parameters, followed by translation into proteins. In total, 587,992 genes were predicted from the CRVC (Table S5C). Viral life cycle prediction and functional annotation were conducted using VIBRANT v1.2.1[136] with option '-f prot -virome' using reference databases including KEGG, Pfam,[103] and VOGDB[74] (http://vogdb.org/). Protein clustering was conducted using MMSeqs2 v13.45111[137] with thresholds of 50% average amino acid identity (AAI) and 80% coverage against UniRef50[161,162] and options '–cov-mode 1, -c 0.8, –cluster-mode 2, –min-seq-id 0.5, and –kmer-per-seq 80'. Finally, the CRVC dataset contained 200,069 protein clusters, of which only 33,323 were clustered with previously published viral sequences.

Clustering of viral genomes at the species and genus level

To assess the diversity and novelty of our viral genomes, all viruses were clustered at the species and genus levels (Tables S5B and S5D). According to MIUViG standards[72] a threshold of 95% ANI and coverage of 85% of the shorter sequence were used for species-level (vOTU) clustering. Similarity and coverage between genomes were estimated using scripts from the CheckV database.[135] All-vs-all local comparisons between viral sequences were performed using the blastn package from BLAST v2.5.0[140] with options '-max_target_seqs 30000 and -perc_identity 90'. The script cluster.py from the MGV database was used to perform species-level clustering using a greedy, centroid-based algorithm. A total of 7,653 species-level clusters were identified in the CRVC.

At the genus level, clustering was performed based on protein similarity and the protein share network between genomes.[163,164] Viruses from the CRVC and public databases were represented by representative genomes at the species level. Initially, all-vs-all protein alignments were conducted using DIAMOND BLASTP v2.0.15.153[125] with options '–evalue 1e-5, –max-target-seqs 10,000, –query-cover 50, and –subject-cover 50'. Subsequently, we calculated the number of shared proteins between genomes, the proportion of proteins shared, and the average AAI of shared proteins between each genome pair. The score between genome pairs was calculated based on $min_{cov} \times mean_{aai}$, and the score was used as the network edge between the genome pairs. Finally, genome pairs were filtered based on the set threshold, and mcl v14-137[138] was used to perform clustering with the option '–abc -I'. We benchmarked combinations of different filtering thresholds, including protein sharing ratios of 10%, 15%, 20%, 25%, and 30%, and note that when the viral genome was too large, at least 20 proteins needed to be shared between them. We also tested average AAI values of 20%, 30%, 40%, 50%, 60%, and 70% as well as MCL inflation factors of 1.1, 1.2, 1.4, 2.0, 4.0, and 6.0. We evaluated their combinations by comparing them against the taxonomy and clustering results derived from the RefSeq viruses, which served as our mock dataset, using Adjusted Mutual Information scores. Finally, we selected a protein network sharing threshold at average AAI of 30% and an inflation factor of 4.0 for genus-level clustering, with precision and recall thresholds set to 0.85 and 0.88, respectively. Viral species- or genus-level clusters that could not be clustered with other viruses were defined as unreported.

Taxonomic classification of the CRVC

We performed taxonomic classification on the CRVC by integrating protein annotation and genus-level clustering results. First, following the taxonomy of the ICTV Master Species List MSL38 v1 (number 22),[101] we curated corresponding RefSeq representative virus sequences to serve as classification standards. For taxonomic classification, if the genome in the CRVC could be clustered at the genus level with known reference viruses and the taxonomic consistency of reference viruses within the cluster exceeded 50%, we assigned the taxonomic label to that cluster. For viruses that did not cluster with any reference viral genomes, we used MMSeqs2[137] based on the lowest common ancestor algorithm to classify the taxonomy of the genomes. The custom reference viral databases used by MMSeq2 were built on the proteins of RefSeq viral genomes. This construction yielded a classification rate of 94.9% at the family level with an accuracy of 99.9% by comparing the results against the mock taxonomy derived from RefSeq data.

Host prediction of the CRVC

For predicting the domain-level hosts of the CRVC, we defined the viral host domain based on taxonomic statistics from the Virus-Host DB.[102] We established connections between crop root bacterial genomes and phages based on hits of predicted phage sequences and CRISPR spacers identified within bacterial genomes. The predicted phage sequences included explicit prophages and phage fragments in 9,772 bacterial genomes. Prophages were defined by VirSorter categories 4 and 5 and geNomad's provirus labels, and we excluded contigs in MAGs if more than 50% genes were predicted viral according to CheckV. For phage fragments without prophage structure, we only included those predicted from isolates and excluded MAGs to eliminate false positives caused by binning errors. Spacer sequences within bacteria were predicted using CRISPRCasFinder v2.0.3[139] with default parameters, retaining spacer sequences in CRISPR arrays with evidence level 4 (high confidence). In total, we obtained 18,417 phage sequences

from 5,624 bacterial genomes and 27,952 spacer sequences from 644 bacterial genomes. To establish phage–bacterial connections at the within the bacterial and viral genomes, we used BLAST to match these sequences to viral genomes of the CRVC and IMG/VR sequences originating from root or soil environments. The alignment options for phage sequences were '-max_target_seqs 30000 and -perc_identity 95', and for spacer sequences, the alignment options were 'blastn-short, -dust no, -word_size 7, -perc_identity 100, -max_hsps 1, and -max_target_seqs 100000'. For phage sequences, a BLAST length of $\geq$ 1 kb and an ANI of $\geq$ 96% were considered positive BLAST hits. For CRISPR spacers, only spacers matched across the whole length of spacers were considered as positive BLAST hits (Table S6A). The definition of temperate phages was based on the classification from VIBRANT and the identification of prophages.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of bacterial species and evaluation of bacterial growth rate

To ensure the accuracy of read mapping, 3,044 species-level representative genomes were used as references. Reads of 332 root metagenomic samples were aligned to this genome reference using Bowtie 2 v2.4.5[129] under 'end-to-end and –very sensitive' mode. Subsequently, InStrain v1.8.0[130] was used to calculate the coverage and breadth of each sample against the bacterial genomes. A breadth value of > 0.5 was considered to indicate the presence of a genome in the sample (Table S6D). The top ten most abundant bacterial species in each crop root metagenomic sample were considered to be highly abundant. The peak-to-trough ratio[165] of the copy numbers of DNA near the replication origins and copy numbers of bacteria terminus points in a sample was calculated to measure genome growth dynamics. Evaluation of bacterial growth rate was conducted using Dynamic Estimator of Microbial Communities (DEMIC)[131] with default parameters (Table S2D).

Quantification of metagenomic reads and feature table generation

Host-depleted, high-quality reads of metagenomic root samples were mapped against corresponding NR gene references using Bowtie 2. A positive match was defined by an alignment identity of $\geq$ 95% and an alignment length $\geq$ 60% of read length. All positive matches from the Bowtie 2 SAM output were next extracted and the resulting filtered SAM files were sorted and converted into BAM format using SAMtools v1.9.[133] A non-redundant gene count table was generated using HTseq v2.0.2[134] with options '–mode=union, –stranded=no, –type=CDS, –idattr=gene_id, –nonunique=none, –secondary-alignment=ignore, and –supplementary-alignment =ignore'. This gene count table was then correlated to the microbiome according to genes assignment to microbes by Kraken 2, and further normalized by the gene length and count.[166] The relative abundance table was scaled to 10^9 per sample (Tables S4A–S4F). Taxonomic and functional feature count tables were generated based on non-redundant gene annotation. The function feature tables were further filtered to exclude unannotated features and were normalized within each sample. Considering the presence of some features (e.g., KOs) with extremely low abundance, accurate repeats between technical replicates were not feasible. To reduce the interference of these features on differential abundance analysis, we defined the measurable KOs by evaluating the repetition of deeply sequenced root metagenomic samples. We sequenced five root samples three times each and compared the consistency of the relative abundance tables by calculating the Pearson correlation coefficients between each pair of technical repeats at different abundance thresholds. The correlation coefficient curve indicated that the repeatability of KO abundance was satisfactory when the threshold of relative abundance reached 3.0×10^{-5} (normalized to 1). In terms of the functional pathway table, the technical repeatability was high (r = 0.98), so no abundance cutoff was set. To ensure accuracy, functions that were exclusively from animals or plants were excluded (Table S4E).

Rarefaction analysis and evaluation of microbial genetic potential

To evaluate the representativeness of non-redundant genes for crop root metagenomic samples, we calculated the gene richness as the number of included root samples increased. For each specific sample number gradient, a random shuffle of the root samples was repeated 20 times (Figure S3A; Tables S3F). To evaluate the microbial genetic diversity for each crop species, we filtered non-redundant microbial genes in at least 70% of the root samples within each dataset, integrated the gene counts across each crop species, and defined them as representative non-redundant microbial genes (Figure S3B; Table S3G). The ratio of representative NR microbial gene entries to the gene number of the crop genome was used to measure microbial genetic diversity for each crop.

Quantification of viruses and antiviral defense systems in metagenomic samples

The representative genomes at the species level in the CRVC and IMG/VR were selected as references. Host-removed metagenomic sequences were aligned to the reference genomes using Bowtie 2 in end-to-end mode with options '–non-deterministic and –very-sensitive'. SAMtools was used to sort and convert the files to BAM format. The filter mode of CoverM v0.6.1 (https://github.com/wwood/CoverM) was used to filter reads with an alignment identity threshold of 95%. Metagenomic aligned bases and coverage of viral genomes were calculated using CoverM with options '-m covered_bases and -m covered_fraction', respectively. Viruses with coverage of $\geq$ 70% or aligned bases of $\geq$ 5 kb were considered present in the metagenomic sample. Virus abundance was computed using CoverM with option '-m trimmed_mean', which calculated the average read depth after removing the 5% of bases with the highest and lowest genome coverage (tpmeans). The obtained tpmean value was divided by the number of reads in the sample, yielding the relative virus abundance. The relative abundance of phages in each sample was calculated as the sum of the relative

abundances of all bacteriophages in the sample. For quantification of defense systems in crop root and soil samples, the protein profiles from DefenseFinder[104] were used to annotate translated NR genes using hmmsearch v3.3.2[141] with options '–cut_ga and noali'. Based on the annotations and previously generated relative abundance table of NR genes, a feature table of defense-related proteins was generated. The relative abundance of each antiviral defense system was calculated as the sum of the abundances of mandatory proteins[104] within that system.

Microdiversity calculation of viral genomes

Metagenomic data were used to analyze the intrapopulation genetic diversity (microdiversity) of viral genomes within sampled environments. We used InStrain v1.8.0[130] to calculate the average nucleotide diversity (π) of each viral genome in each sample.[37] For accuracy, only genomes with coverage of at least 5× across all base pairs were used for calculation. Microdiversity within each prevalence group described in the next section was defined as the mean π value of all viruses belonging to that prevalence group in the sample.

Definition of prevalence groups for bacterial and viral genomes

Genomes were grouped based on the prevalence characteristics of bacterial and viral genomes in crop root metagenomic samples. Prevalence refers to the frequency at which a genome was detected in samples from a single host crop at a single location. Genomes with a prevalence of less than 10% or those found in only a single sample if the sample number was below 10 were defined as rare. The remaining genomes were considered stably detected in root ecosystems. Further categorization included the following three groups: stable presence across roots of multiple crop species at multiple locations (multi-crop multizonal), stable presence in roots of a single crop species across multiple locations (single-crop multizonal), and stable presence in roots of a single crop species at a single location (single-crop regional). For bacteria, there were 60 species in the multi-crop multizonal group, 99 in the single-crop multizonal group, 455 in the single-crop regional group, and 331 in the rare group. For viruses, there were 123, 281, 1,311, and 975 species in each group, respectively.

Statistic analysis and data visualization

Comparisons between two groups were conducted using a two-tailed Wilcoxon rank-sum tests. For comparisons among multiple groups, a Kruskal–Wallis rank sum test was used, followed by a Dunn's test for pairwise comparisons. Multiple comparisons were adjusted using the false discovery rate (FDR) method, with significance considered when the adjusted P value was less than 0.05. Asterisks were used to denote the level of statistical significance (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$). Multiple groups were labeled with letters based on adjusted P values. A correlation analysis was conducted using the cor.test() function in R with Spearman's rho, which calculated correlation coefficients and P values. Similar to the previous procedure, multiple comparisons were adjusted using the FDR method, with significance defined as an adjusted P value of less than 0.05. Downstream data processing and visualization were performed using R 4.3.1 and packages including tidyverse, reshape2, rstatix, vegan, PMCMRplus, and multicompView for data processing and statistical analysis and ggforce, ggpubr, ggVennDiagram, UpSetR, ComplexHeatmap, viridis, RcolorBrewer, and patchwork for visualization. Visualization of phylogenetic trees was conducted using ape, picant, pytools, ggtree, ggtreeExtra, and ggnewscale. Note that tree-branch lengths in Figures 5A and 6C do not reflect taxonomic distance.

CellPress
OPEN ACCESS

Supplemental figures

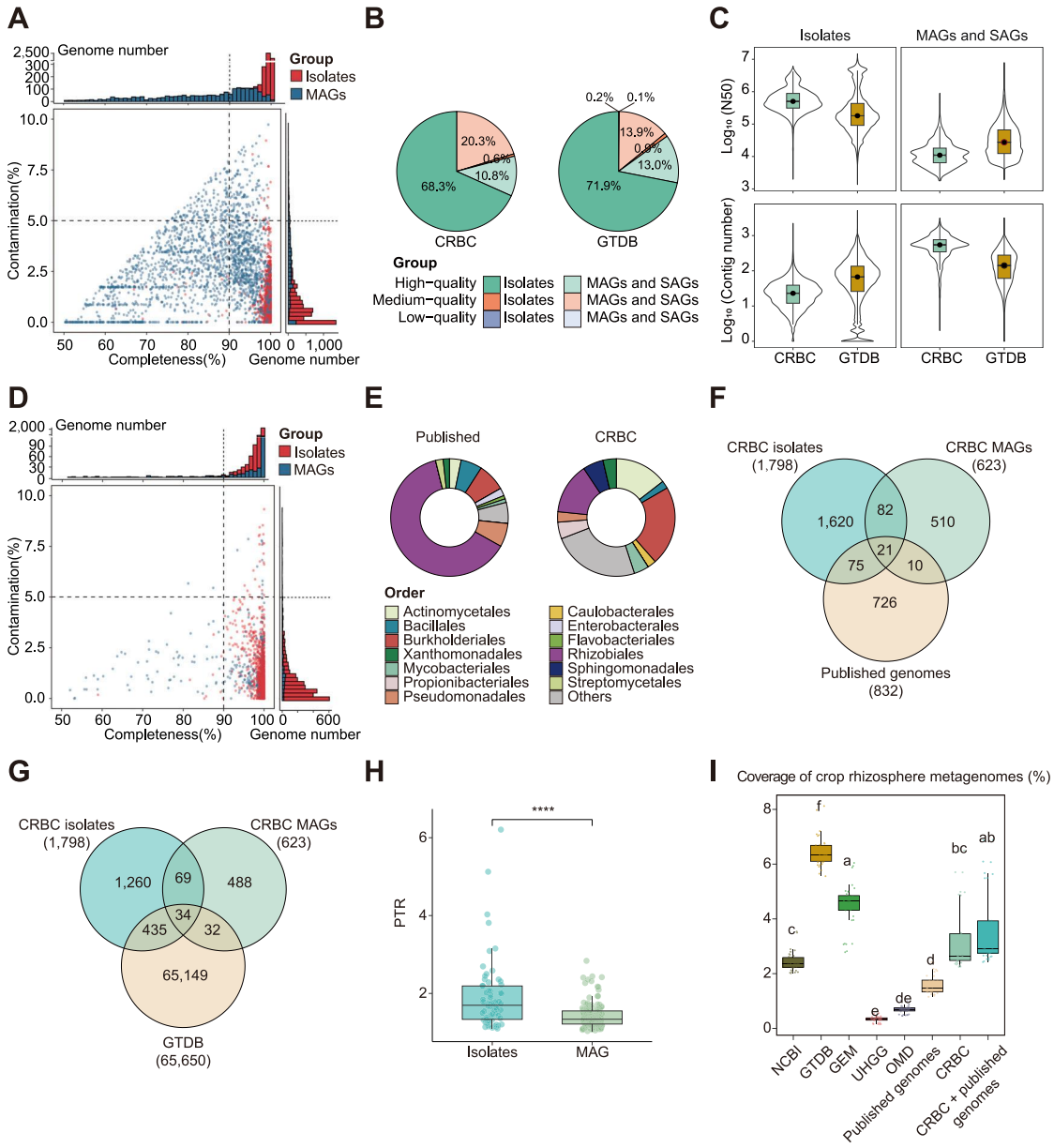


Figure S1. Characteristics of the CRBC and publicly available genomes, related to Figures 1 and 2

(A) Genome quality of the CRBC. The 6,699 genomes in the CRBC are displayed according to the genome completeness and contamination rates. Genome numbers are color-coded based on their sources, with isolates represented in red and MAGs in dark blue.

(B) Distribution of genome quality in the CRBC and GTDB. The pie chart shows the composition of genomes at different quality levels.

(C) Genomes in the CRBC show comparable quality to those in the GTDB. The boxplots show $\log_{10}$-transformed N50 length (top) and contig number (bottom) in each genome of the CRBC and GTDB. Isolates are shown in the left and MAGs are shown in the right, respectively.

(D) Quality of published crop root bacterial genomes. The 3,073 published genomes of crop root bacteria are displayed according to the genome completeness and contamination rates.

(E) The taxonomic composition of published crop root bacterial genomes (left) and the CRBC (right) at the order level. The bacterial taxonomy is displayed at the order level.

(legend continued on next page)

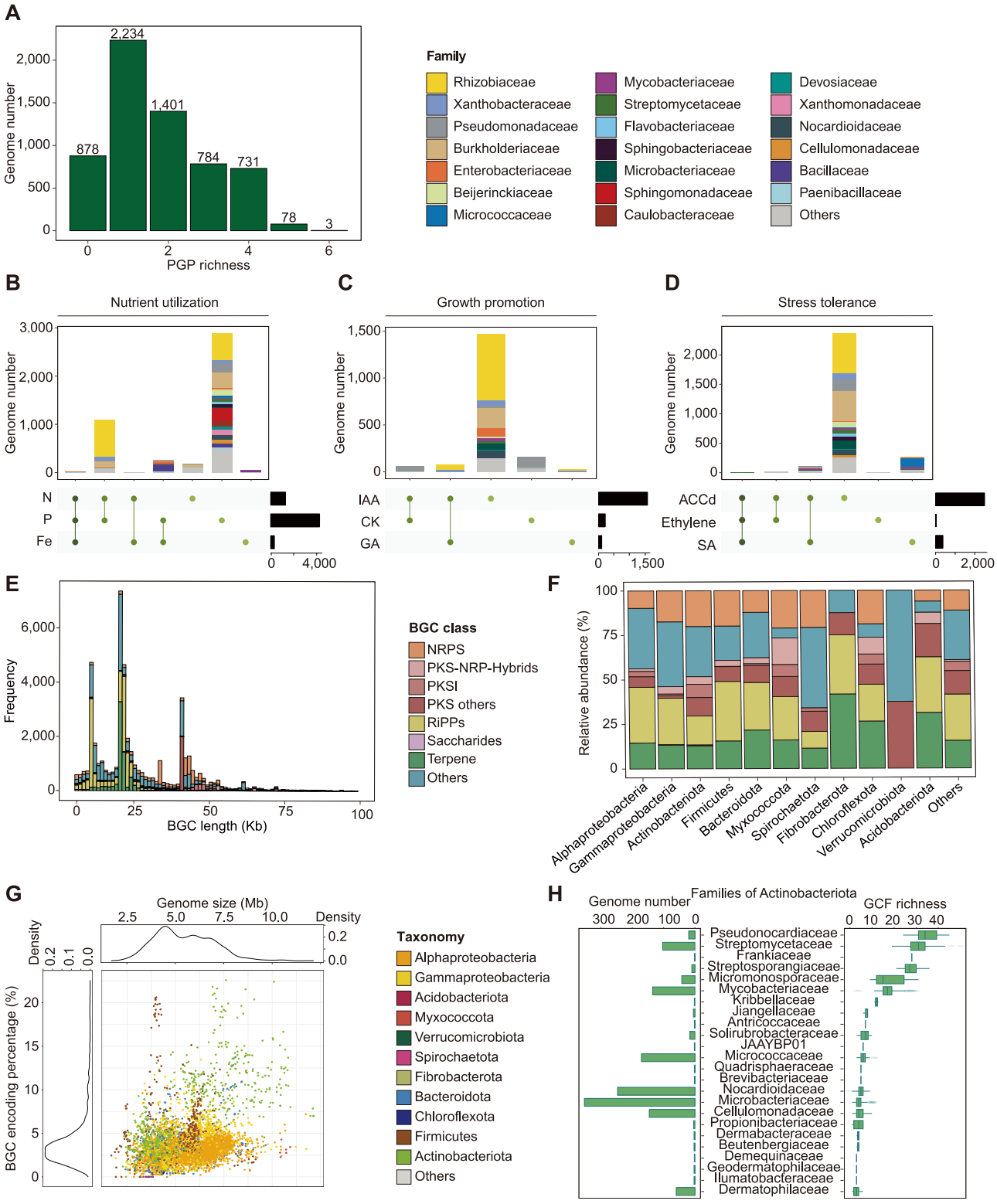
(F) Overlap of bacterial species in the CRBC and published crop root bacteria. The species are color-coded based on their genomic sources, with CRBC isolates represented in blue, CRBC MAGs in green and published crop root genomes in beige. Note that 76.9% ([1,620+82]/2,212) of genomes not reported before in crop roots are from the CRBC isolates.

(G) Overlap of CRBC genomes with GTDB. The comparison between 6,699 CRBC genomes and over 60,000 GTDB representative genomes shows a total of 1,817 undefined CRBC species, with 1,329 identified from CRBC isolates and 448 from CRBC MAGs.

(H) The CRBC MAGs showed lower growth rates than the CRBC isolates within and between taxa. The bacterial isolates were analyzed by DEMIC based on the ratio of the coverage at the origin and terminus (peak-to-trough ratio [PTR]) of bacterial genomes within crop root microbial populations (STAR Methods, adjusted $p < 0.05$, Wilcoxon rank-sum test).

(I) The reference genomes showed source specificity in rhizosphere metagenomic data. The boxplot shows proportions of metagenomic reads of crop rhizosphere samples in maize ($n = 8$) and wheat ($n = 29$) aligned to genomes in public databases and the CRBC. The abbreviations are: NCBI, NCBI RefSeq database; GTDB, Genome Taxonomy Database; GEM, the Genome from Earth's Microbiomes catalog; UHGG, the Unified Human Gastrointestinal Genome collection; OMD, the Ocean Microbiomics Database (adjusted $p < 0.05$, Kruskal-Wallis rank-sum test and Dunn's test).

CellPress
OPEN ACCESS



(legend on next page)

Figure S2. Characteristics of PGP functions and BGCs in the CRBC and published crop root bacterial genomes, related to Figure 3

(A) Distribution of genomes across their PGP functions. Bar plots show the number of PGPs in each genome. Note that among the 6,109 high-quality genomes of crop root bacteria, 5,231 genomes were found to encode at least one tested PGP function.

(B) Coexistence patterns of nutrient utilization functions within individual bacterial genomes. The number and taxonomy of genomes for each coexistence pattern are shown in the stacked bar plot in the upper panel. The lower panel with vertical lines illustrates co-existent functions (nitrogen fixation, phosphorus and siderophore biosynthesis) within individual genomes. The number of genomes for each PGP group is shown in the lower right panel.

(C) Coexistence patterns of crop growth functions within individual bacterial genomes. The number and taxonomy of genomes for each coexistence pattern are shown in the stacked bar plot in the upper panel. The lower panel with vertical lines illustrates co-existent functions (biosynthesis of IAA, CK, and GA) within individual genomes. The number of genomes for each PGP group is shown in the lower right panel.

(D) Coexistence patterns of stress tolerance functions within individual bacterial genomes. The number and taxonomy of genomes for each coexistence pattern are shown in the stacked bar plot in the upper panel. The lower panel with vertical lines illustrates co-existent functions (biosynthesis of ACCd, ethylene and SA) within individual genomes. The number of genomes for each PGP group is shown in the lower right panel.

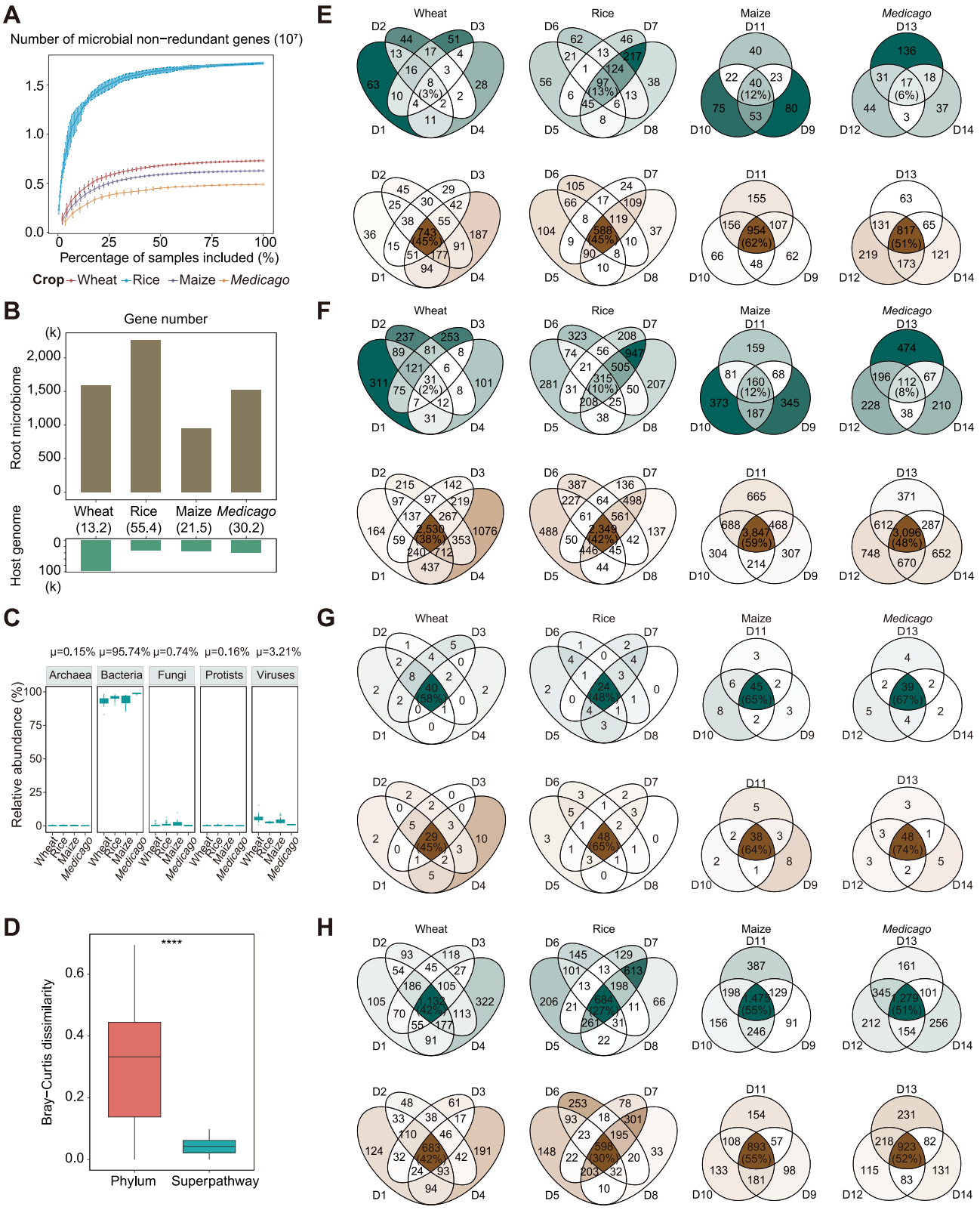
(E) Distribution of BGC classes according to the length of gene clusters. The bars are color-coded based on BGC classes. The lengths of NRPS are relatively higher than other BGC classes.

(F) Composition of BGC classes in each bacterial phylum. Proteobacteria are displayed at the class level due to their excessive species number.

(G) Relationship between the BGCs percentage in the genome and the corresponding genome size. Each dot represents a bacterial genome. The scatter plot illustrates the BGC percentage and genome size in each genome. The density of data in each axis is shown on the left or top of the scatter plot. Note that Actinobacteriota and Firmicutes have a higher proportion of genes encoding BGCs.

(H) Distribution of genome number and GCF richness of families in Actinobacteriota. The bar plot and boxplot show the genome number (left) and GCF number (right) in each family of Actinobacteriota. Families are arranged in descending order according to the median of GCF richness.

CellPress
OPEN ACCESS



(legend on next page)

Figure S3. Taxonomic and functional patterns of root metagenomes from multiple crops in diverse soils, related to Figure 4

(A) Accumulation curves indicate that the number of non-redundant microbial genes in the root metagenomes of each crop species is approaching saturation. Data from wheat, rice, maize, and *Medicago* are presented in different colors. Each observation was sampled 20 times without replacement.

(B) The gene count comparison between the root metagenomes and host crop genomes. The bar plot shows gene count of non-redundant genes within the root metagenomes (upper) compared with corresponding crops (lower). The number with bracket under the host name represents the ratio of gene numbers between the root microbiome and the corresponding host.

(C) Relative abundance of archaea, bacteria, fungi, protists, and viruses in the root metagenomes of wheat, rice, maize, and *Medicago*. Note that bacteria contribute an average of 95.7% of overall microbial relative abundance within the root metagenomes.

(D) Taxonomic and functional composition of root metagenomic samples. Boxplot showing pairwise Bray-Curtis dissimilarity with root metagenomes of multiple crops grown in diverse soils based on taxonomic and functional compositions (n = 332, adjusted $p < 0.05$, Wilcoxon rank-sum test).

(E) Families enriched or depleted in root metagenomes of each crop species derived from diverse soils. The Venn diagram illustrates overlaps of bacterial families enriched (upper, green) and depleted (bottom, brown) in roots of wheat, rice, maize, and *Medicago* from multiple datasets. Each area is color-coded according to the number of functions in the area.

(F) Genera enriched or depleted in root metagenomes of each crop species derived from diverse soils. The Venn diagram illustrates overlaps of bacterial genera enriched (upper, green) and depleted (bottom, brown) in roots of wheat, rice, maize, and *Medicago* from multiple datasets. Each area is color-coded according to the number of functions in the area.

(G) Microbial functions enriched or depleted in root metagenomes of each crop species derived from diverse soils. The Venn diagram illustrates overlaps of independent datasets reflecting the functions consistently enriched (upper, green) and depleted (bottom, brown) in roots of wheat, rice, maize, and *Medicago*. Each area is color-coded according to the number of functions in the area.

(H) Functional KOs enriched or depleted in root metagenomes of each crop species derived from diverse soils. The Venn diagram illustrates overlaps of independent datasets reflecting the functional KOs consistently enriched (upper, green) and depleted (bottom, brown) in roots of wheat, rice, maize, and *Medicago*. Each area is color-coded according to the number of functions in the area.

CellPress
OPEN ACCESS

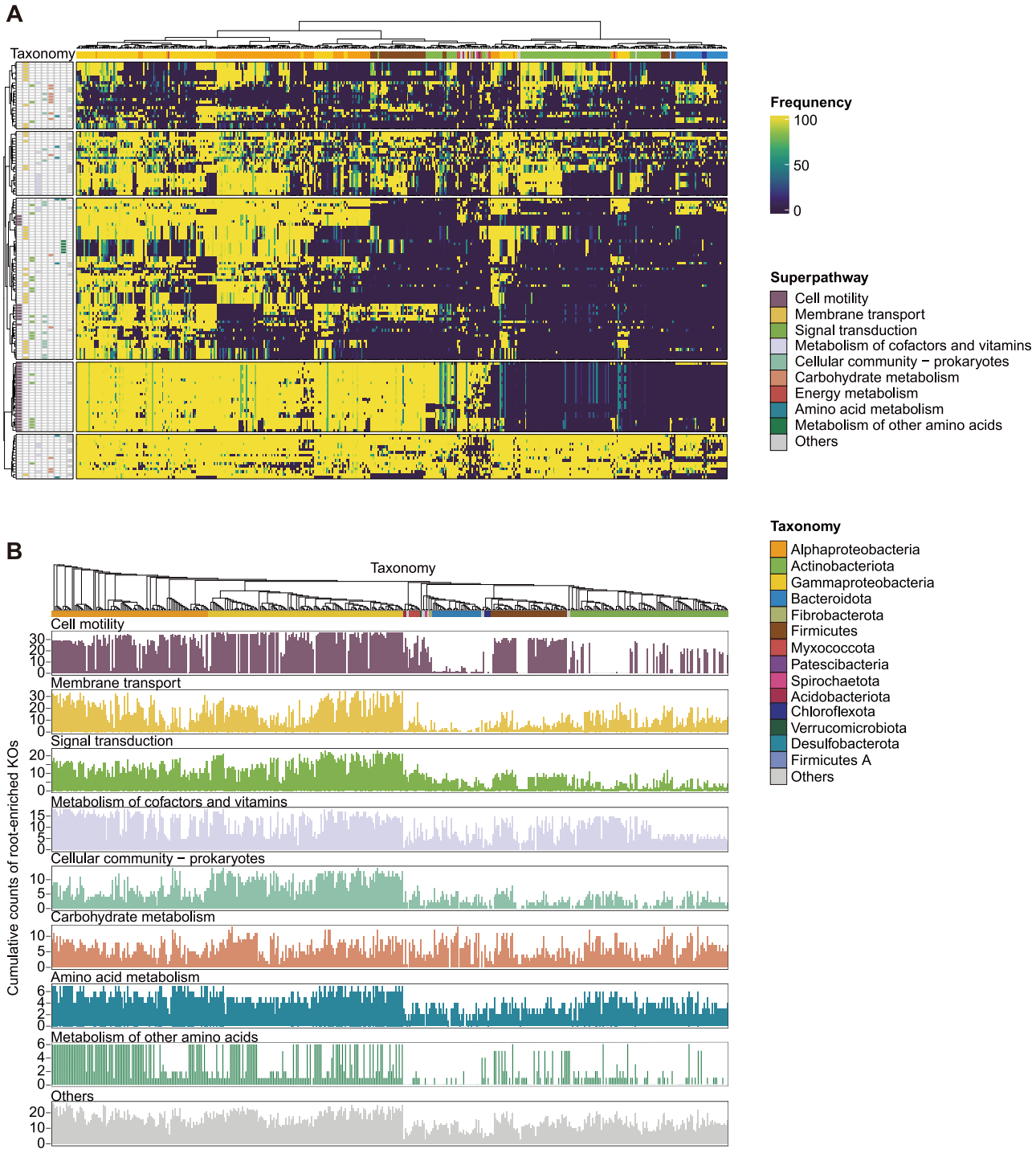


Figure S4. Distribution of 146 root-enriched KOs in the crop root bacterial genomes, related to Figure 4

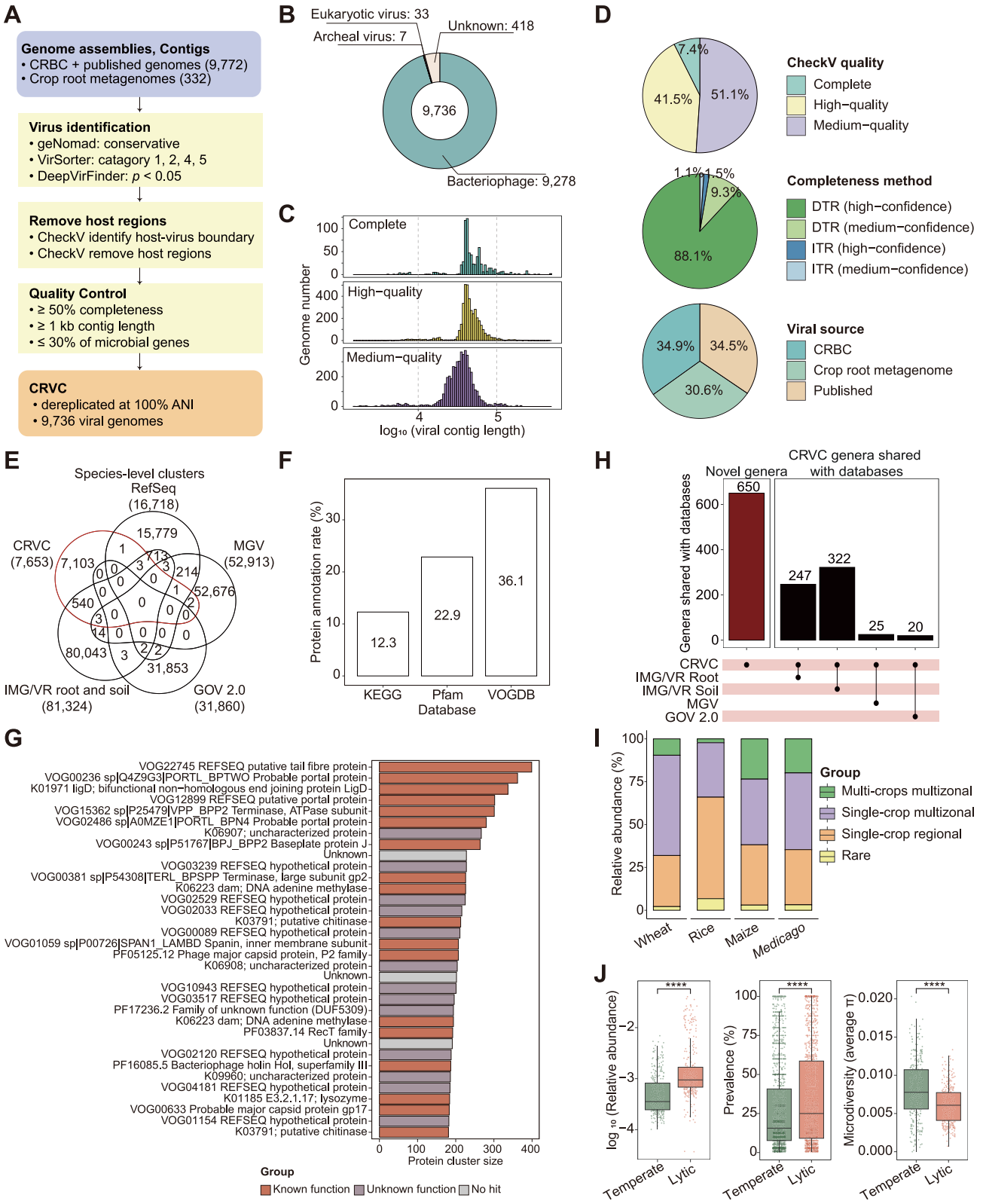
(A) Clustering of 146 root-enriched KOs in the crop root bacterial genomes. The central heatmap shows the frequency of root-enriched KOs across crop root bacterial genomes. The frequency is calculated based on the percentage of genomes containing the KO in each genus. The distributions of KOs and genomes are clustered separately using the ward.D2 method. KOs are clustered into 5 groups according to their distribution in crop root bacteria and colored according to the

(legend continued on next page)

superpathways to which they belong. The taxonomy information of genomes is presented at the phylum level, with Proteobacteria at the class level due to their excessive abundance. Note that phylogenetically related genera exhibited similar functional patterns of root enrichment.

(B) Functional landscape of 146 root-enriched KOs in the genomes of crop root bacteria. Bar plots show the cumulative counts of root-enriched KOs that are prevalent in over 50% of genomes within each bacterial genus. KOs are grouped and color-coded based on their KEGG superpathways. Bacteria genera are arranged according to their phylogeny.

CellPress
OPEN ACCESS



(legend on next page)

Figure S5. Quality, protein content, and distribution of the CRVC, related to Figure 5

(A) Workflow of CRVC construction. Viral contigs were predicted from 9,772 bacterial genomes in the CRBC and published databases and 332 crop root metagenomes, using a combination of software tools including geNomad, VirSorter, and DeepVirFinder. The removal of host regions and estimation of viral completeness were conducted using CheckV. Viral genomes underwent quality control and dereplication at 100% ANI, yielding 9,736 non-redundant microbial viral genomes designated as the CRVC. Note that 65.5% of CRVC genomes are derived from CRBC genomes and our crop root metagenomes.

(B) The composition of the CRVC. The doughnut chart illustrates the proportion and number of viruses in the CRVC according to the domain of hosts. The CRVC comprises 9,736 non-redundant viruses, including 9,278 bacteriophages, 7 archaeal viruses, 33 eukaryotic viruses, and 418 entries with unclear hosts (STAR Methods).

(C) Genome quality and length distribution of the CRVC. The bar plots show the length distribution of CRVC genomes. CRVC genome completeness is categorized into three quality groups: 723 complete genomes, 4,041 high-quality genomes (>90% completeness), and 4,972 medium-quality genomes (50%–90% completeness). Note that nearly half (48.9%) of the genomes in the CRVC are over 90% completeness (high-quality and complete). The median length of viral contigs is 40.6 kbp.

(D) Overview of CRVC quality. The upper pie chart represents the distribution of the CRVC genomes across different quality categories as evaluated by CheckV. Note all of the CRVC genomes exceed the medium-quality level with completeness over 50%. The middle pie chart illustrates the distribution of 723 complete CRVC genomes acquired through the direct terminal repeats (DTRs) or inverted terminal repeats (ITRs) methods, with methods and confidence levels color-coded for clarity. The bottom pie chart illustrates the proportion of the CRVC genomes derived from the CRBC genomes, our crop root metagenomes, and published crop root bacteria, each color-coded for clarity.

(E) Majority of the CRVC species-level clusters are not reported in public databases. The Venn diagram illustrates overlaps of the CRVC species-level clusters (vOTUs) with the public viral databases including, RefSeq, MGV, GOV2, and IMG/VR. In the IMG/VR database, viral genomes sourced from both root and soil environments were incorporated into the analysis.

(F) Annotation rate of the CRVC proteins by different reference databases. The bar plot shows the percentage of viral proteins in the CRVC annotated using the KEGG, Pfam, VOGDB databases, respectively. In total, we predicted 587,992 proteins within the CRVC, of which the annotation rates for these databases ranged from 12.3% to 36.1%. Note that the low annotation rates across various reference databases suggest that a significant portion of the CRVC comprises proteins with unknown functions.

(G) Functional annotations of the largest 34 protein clusters within CRVC. The bar plot shows the number of proteins in each abundant protein cluster, and only protein clusters with numbers greater than 180 are shown. The annotation of each protein cluster is color-coded according to annotation categories: known, unknown functions, and no hit. Note that a substantial proportion of protein clusters are either with unknown functions or lacking matches in databases.

(H) Comparison between the CRVC and public viral databases using genomes derived from metagenomes. The bar plot shows the number of metagenome-derived CRVC viral clusters that are shared with publicly available databases at the genus level. Bar plots showing that 650 metagenome-derived unreported CRVC clusters whose genomes are distinct from those in published viral databases, including viruses from IMG/VR, MGV, and GOV2. The CRVC exhibits a higher shared presence of viruses with IMG/VR soil and roots than other ecosystems, suggesting the ecosystem specificity of viral distribution.

(I) Viral distribution in crop root habitats. The stacked bar plot illustrates the composition of viral categories in the root microbiomes of each crop species. Using prevalence characteristics, viruses are categorized into four prevalence groups: stable presence across roots of multiple crops grown in multiple locations (multi-crops multizonal, green), within roots of a single crop species in multiple locations (single-crop multizonal, purple), within roots of a single crop species in a single location (single-crop regional, orange), and viruses with prevalence lower than 10% of samples (rare, yellow).

(J) The characteristics of temperate and lytic viruses in crop root microbiomes. The bar plots illustrate the $\log_{10}$-transformed relative abundance (left), prevalence (middle), and microdiversity (right) of temperate and lytic viruses in 332 crop root metagenomic samples (adjusted $p < 0.05$, Wilcoxon rank-sum test).

CellPress
OPEN ACCESS

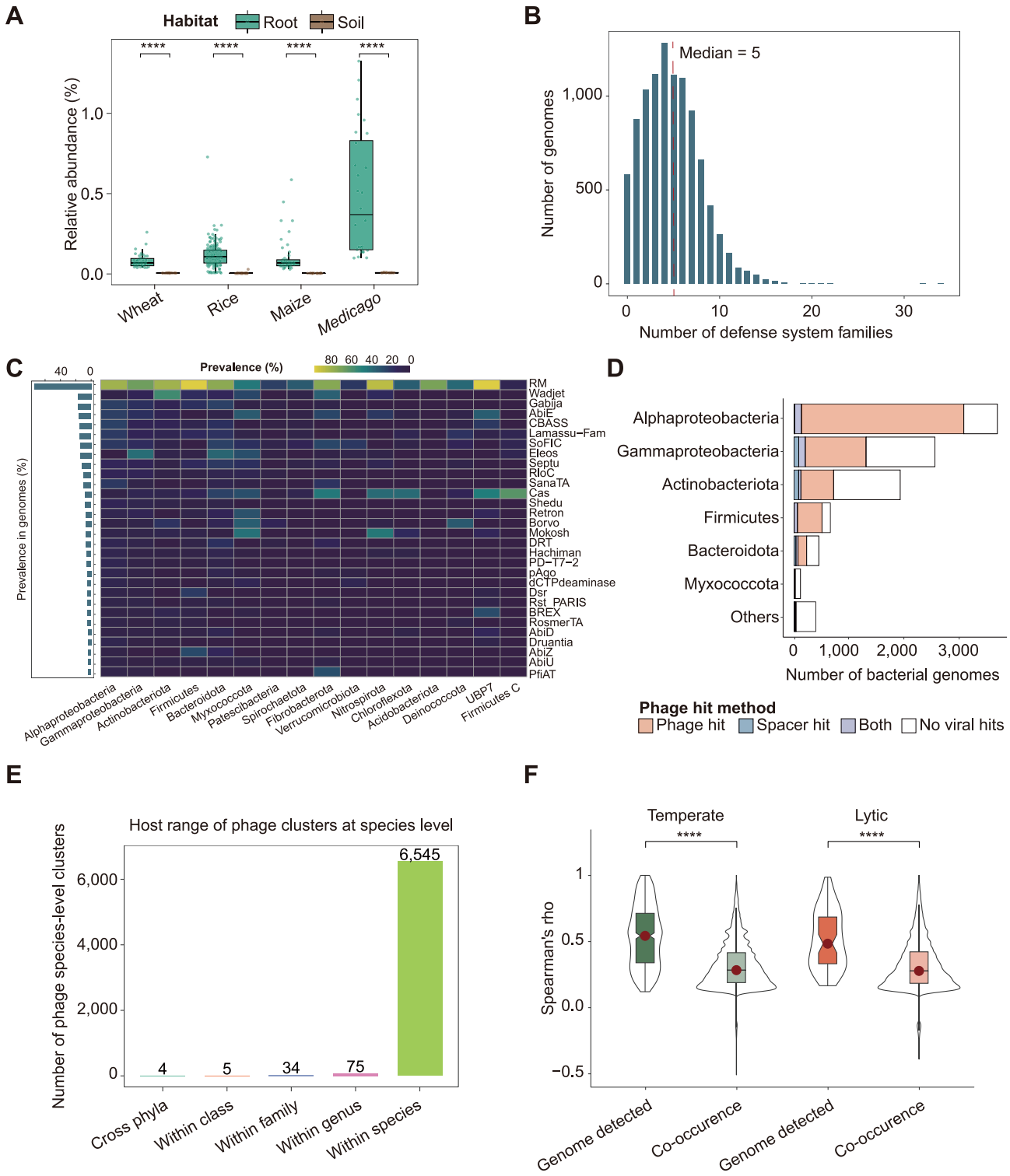


Figure S6. Phage-bacteria connections within bacterial genomes and crop root ecosystems, related to Figure 6

(A) Phage abundances in the root microbiomes of multiple crops are higher than those in corresponding soils based on the IMG/VR version 4 database. The boxplot shows the median and quartiles of phage relative abundance in root metagenomic samples from wheat, rice, maize, *Medicago*, and corresponding soils. Statistic differences are assessed using the Wilcoxon rank-sum test, root, $n = 332$; soil, $n = 75$. **** represents adjusted $p < 0.0001$.

(legend continued on next page)

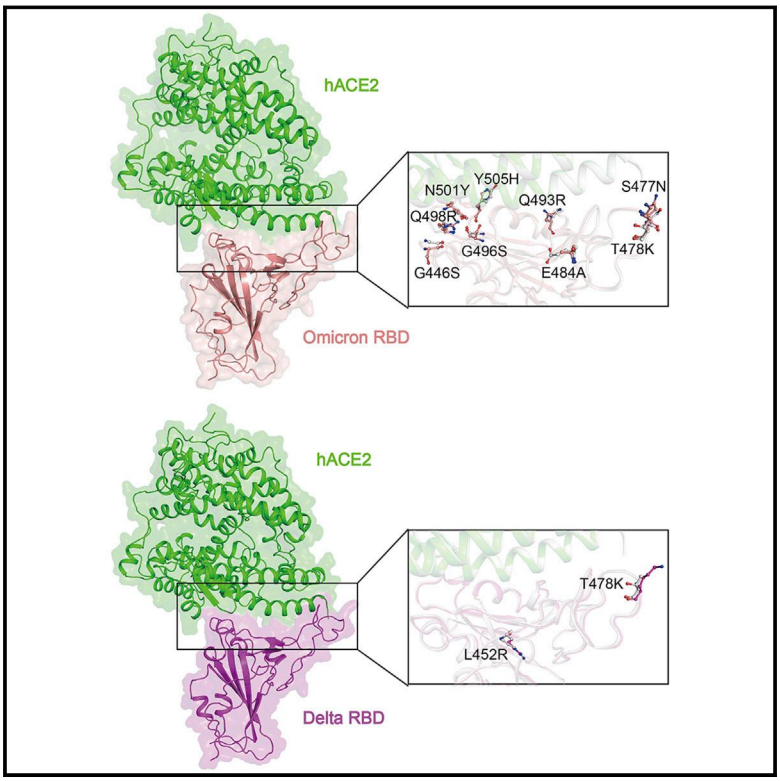
(B) The distribution of antiviral defense system families in the bacterial genomes. The histogram shows the distribution of the number of defense system families in each bacterial genome.
(C) The prevalence of bacterial defense systems in crop root bacterial genomes. The heatmap illustrates the prevalence of each bacterial defense system at the phylum level. Proteobacteria are shown at the class level due to their excessive abundance. The bar plot (left) showing the prevalence of each bacterial defense system in all crop root bacteria.
(D) Substantial proportion of genomes within major crop root bacteria phyla exhibit connections with phage genomes. Bar plot shows the number and proportion of crop root bacterial genomes with phage hits according to phage contigs and CRISPR spacers detected in bacterial genomes. Each bar represents the genome number within each bacterial phylum (class level for Proteobacteria). Bacterial genomes with phage hits are colored orange, those with CRISPR spacer hits are colored blue, and genomes with phage hits detected by both methods are shown in purple.
(E) Large majority of phage species-level clusters show specific connections with bacteria. The bar plot illustrates the host range of phage species-level clusters at different taxonomic levels. Note that most phage clusters have a narrow host range and can only interact with bacteria within a single genus or species.
(F) The phage-bacteria connections identified by genomes show higher correlation efficiency than links only determined by co-occurrence in crop root ecosystems. The boxplot illustrates the Spearman correlation coefficients of temperate phage-bacteria connections detected via genomes (dark green) or via co-occurrence (light green), as well as the Spearman correlation coefficients of lytic phage-bacteria connections detected via genomes (dark red) or via co-occurrence (light red). Statistic differences are assessed using the Wilcoxon rank-sum test, root, $n = 332$. **** represents adjusted $p < 0.0001$.

Cell

Article

Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2

Graphical abstract



Highlights

- Omicron, delta, and prototype SARS-CoV-2 RBDs show similar binding strength to hACE2
- The complexes of SARS-CoV-2-RBD with hACE2 for omicron and delta variants were resolved
- The roles of key residues in the omicron RBD for receptor recognition were identified

Authors

Pengcheng Han, Linjie Li, Sheng Liu, ..., Jianxun Qi, George F. Gao, Peiyi Wang

Correspondence

jxqi@im.ac.cn (J.Q.),
gaof@im.ac.cn (G.F.G.),
wangpy@sustech.edu.cn (P.W.)

In brief

Structural analysis of the complexes of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) RBD with the human ACE2 receptor for omicron and delta reveals variant-specific binding features.

Han et al., 2022, Cell *185*, 630–640
February 17, 2022 © 2022 Elsevier Inc.
https://doi.org/10.1016/j.cell.2022.01.001

CellPress

CellPress

Cell

Article

Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2

Pengcheng Han,[1,2,10] Linjie Li,[1,3,10] Sheng Liu,[4,10] Qisheng Wang,[5,10] Di Zhang,[1,6,10] Zepeng Xu,[1,6] Pu Han,[1] Xiaomei Li,[7] Qi Peng,[1] Chao Su,[8] Baihan Huang,[1] Dedong Li,[1] Rong Zhang,[1] Mingxiong Tian,[9] Lutang Fu,[4] Yuanzhu Gao,[4] Xin Zhao,[1] Kefang Liu,[1] Jianxun Qi,[1,3,*] George F. Gao,[1,3,*] and Peiyi Wang[4,11,*]

[1]CAS Key Laboratory of Pathogen Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

[2]School of Medicine, Zhongda Hospital, Southeast University, NanJing 210009, China

[3]University of Chinese Academy of Sciences, Beijing 100049, China

[4]Cryo-EM Center, Department of Biology, Southern University of Science and Technology, Shenzhen 518055, China

[5]Shanghai Synchrotron Radiation Facility, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201204, China

[6]Faculty of Health Sciences, University of Macau, Macau 999078, China

[7]Cryo-EM Center, Shanxi Academy of Advanced Research and Innovation, Taiyuan 030032, China

[8]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong 999077, China

[9]College of life Science, Shanxi University, Taiyuan 03006, China

[10]These authors contributed equally

[11]Lead contact

*Correspondence: jxqi@im.ac.cn (J.Q.), gaof@im.ac.cn (G.F.G.), wangpy@sustech.edu.cn (P.W.)

https://doi.org/10.1016/j.cell.2022.01.001

SUMMARY

The coronavirus disease 2019 (COVID-19) pandemic continues worldwide with many variants arising, some of which are variants of concern (VOCs). A recent VOC, omicron (B.1.1.529), which obtains a large number of mutations in the receptor-binding domain (RBD) of the spike protein, has risen to intense scientific and public attention. Here, we studied the binding properties between the human receptor ACE2 (hACE2) and the VOC RBDs and resolved the crystal and cryoelectron microscopy structures of the omicron RBD-hACE2 complex as well as the crystal structure of the delta RBD-hACE2 complex. We found that, unlike alpha, beta, and gamma, omicron RBD binds to hACE2 at a similar affinity to that of the prototype RBD, which might be due to compensation of multiple mutations for both immune escape and transmissibility. The complex structures of omicron RBD-hACE2 and delta RBD-hACE2 reveal the structural basis of how RBD-specific mutations bind to hACE2.

INTRODUCTION

Receptor binding is a key step of virus invasion (Lu et al., 2015). Similar to severe acute respiratory syndrome coronavirus (SARS-CoV), SARS-CoV-2 uses its spike (S) protein to recognize the host receptor ACE2 (Wang et al., 2020). The C-terminal domain (CTD) of the S protein, also known as the receptor-binding domain (RBD), is responsible for ACE2 recognition and is an important determinant of host range (Wang et al., 2020; Wu et al., 2020). Additionally, the S protein is a major target for neutralizing antibodies (Liu et al., 2020; Piccoli et al., 2020). Mutations on the S protein, particularly the RBD, could lead to immune escape and undermine the efficacy of current therapies and vaccines (Garcia-Beltran et al., 2021).

SARS-CoV-2 keeps evolving into new variants due to sustained global transmission (Li et al., 2021a). Variants with evidence of increased transmissibility, severe disease, reduction in neutralization by antibodies of convalescents or vaccines, and a higher risk of eluding testing are classified as variants of concern (VOCs) (https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html). Currently, five VOCs have been announced by the World Health Organization (WHO), namely alpha, beta, gamma, delta, and the recently identified omicron. Studies have reported that VOCs, especially the beta variant, are subject to immune escape when treated with most clinically available monoclonal antibodies (Abdool Karim and de Oliveira, 2021; Ferreira et al., 2021; Li et al., 2021b; Zhou et al., 2021). Compromised protective potency of current licensed vaccines targeting previous VOCs has also been reported, but the extent of decrease in efficacy is rather mild (Abdool Karim and de Oliveira, 2021; Cao et al., 2021a; Huang et al., 2021a). The delta VOC was shown to have much higher transmissibility (Liu and

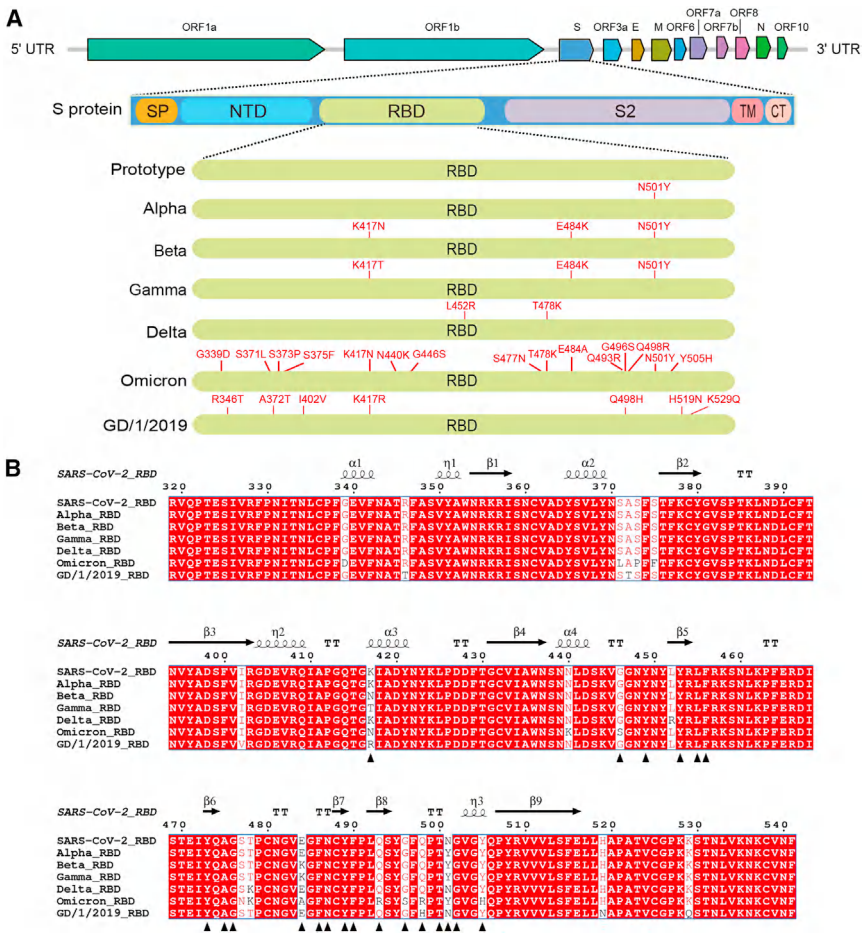


Figure 1. Mutation mapping and sequence alignment of RBDs from prototype SARS-CoV-2, VOCs, and pangolin-origin GD/1/2019

(A) Architecture of SARS-CoV-2 genome and S protein and substitution mapping of RBDs from prototype SARS-CoV-2, VOCs, and pangolin-origin GD/1/2019.

(B) Sequence alignment of abovementioned RBDs. Interacting residues of prototype SARS-CoV-2 RBD to hACE2 are labeled with black triangles. The alignment is performed by T-COFFEE and visualized by ESPript 3.0.

See also Figure S1.

Rocklöv, 2021), which could be explained by its higher replication rates and S-mediated entry (Mlcochova et al., 2021).

In contrast to other VOCs that emerged when natural immunity was predominant, omicron has emerged under the circumstances when vaccination against SARS-CoV-2 is prevalent. Genome sequencing demonstrated that omicron carries an unusually high number of mutations, especially on the S protein (23 amino acid substitutions) and its RBD (15 amino acid substitutions). Putative analyses indicate that some of the mutations are associated with enhanced viral transmission, infectivity, and immune evasions (Genovese et al., 2021; Qin et al., 2021). Recent studies showed that omicron has extensive escape of known neutralizing antibodies and sera from convalescent patients or people vaccinated with the Pfizer vaccine BNT162b2 (Cao et al., 2021b; Cele et al., 2021; Wang et al., 2022). Unveiling the molecular mechanism of receptor recognition of VOCs is essential for understanding the impact of omicron on infection, transmission, and immune escape.

In this study, we examined several VOCs, including alpha, beta, gamma, delta, and omicron, and demonstrated that omicron variant RBD binds to human ACE2 (hACE2) with comparable affinity to that of the prototype. We then determined both the crystal and cryoelectron microscopy (cryo-EM) structures of omicron RBD complexed with hACE2 and identified the role of key residues for receptor recognition. Additionally, we solved the crystal structure of a delta RBD-hACE2 complex. Our results provide important molecular information about both omicron and delta VOCs and could provide guidance for development of novel prophylaxis/therapeutics targeting the dominant omicron variant.

RESULTS

Receptor binding and efficient transduction of omicron engaged by hACE2

To demonstrate the substitution pattern of SARS-CoV-2 VOCs (alpha, beta, gamma, delta, and omicron), we mapped the substitutions of their RBDs and performed sequence alignment (Figures 1A and 1B). We also include the RBD of GD/1/2019, a pangolin-origin coronavirus closely related to SARS-CoV-2. The mapping demonstrated that omicron RBD carries an unusually high number of substitutions, among which substitution of K417, T478, E484, and N501 were previously observed in other VOCs, whereas 11 others were novel. Eight substitutions are located on the hACE2 binding residues of prototype RBD (Wang et al., 2020), some of which were previously observed on VOCs or bat-origin RaTG13 and pangolin-origin GD/1/2019 or GX/P2V/2017 (Liu et al., 2021a; Niu et al., 2021). In the phylogenetic analysis of genome, omicron is close to alpha (Figure S1A). But for RBD region, omicron is closely related to beta and gamma (Figure S1B).

We then explored the binding capacity of omicron RBD to hACE2 in comparison to that of the prototype and other variants as well as to pangolin-origin GD/1/2019. We purified RBDs from the prototype SARS-CoV-2 (the first strain of SARS-CoV-2, which was isolated from a clinical patient on January 6, 2020; GISAID: EPI_ISL_402119; Tan et al., 2020) and the variants, as well as pangolin-origin GD/1/2019, and tested their binding to hACE2-expressing BHK-21 cells with flow cytometry assay (Figure 2A). RBDs from VOCs and GD/1/2019 bound to hACE2-expressing cells to comparable levels with prototype RBD, with 73.3%–86.3% positive cells in staining experiments. We further

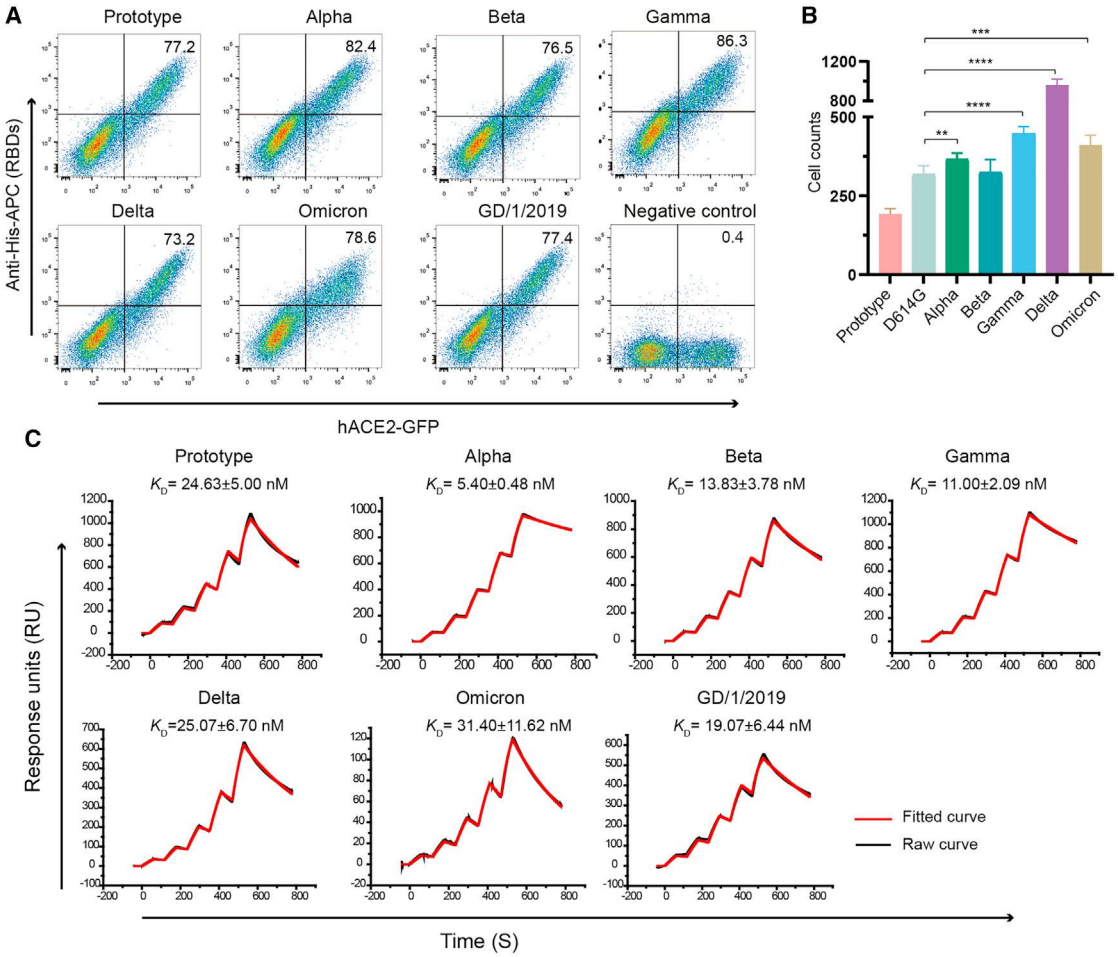


Figure 2. Binding characteristics and transduction of SARS-CoV-2 VOCs

(A) BHK-21 cells stably expressing GFP and hACE2 were incubated with His-tagged RBDs from prototype SARS-CoV-2, VOCs, and pangolin-origin GD/1/2019. Allophycocyanin (APC)-conjugated anti-His antibodies were used to detect the His-tagged protein binding to the cells. Representative results from three experiments are shown. The frequency of RBD-binding positive cells in ACE2-GFP positive cells are labeled in the upright corner.

(B) Transduction efficiency of pseudotyped prototype SARS-CoV-2, D614G variant, and VOCs. The GFP-positive cells were quantified with CQ1 confocal image cytometer (Yokogawa), and representative results from three experiments are shown. Statistical significance was analyzed with a t test comparison between VOCs and D614G.

(C) SPR characterization of RBDs from prototype SARS-CoV-2, VOCs, and pangolin-origin GD/1/2019 interacting with hACE2. Dissociation constant (K_D) indicates mean ± SD of three independent replicates. Actual and fitted curves are colored in black and red, respectively.

See also Figure S2.

measured the binding affinities of the RBDs to hACE2 with surface plasmon resonance (SPR) assay (Figure 2C; Table S1). The results showed that the prototype SARS-CoV-2 binds to hACE2 with a dissociation constant (K_D) of 24.63 nM. RBDs from alpha, beta, and gamma demonstrated enhanced affinities to various extents, ranging from 1.78- to 4.56-fold increase. On the other hand, omicron, along with delta RBD, showed no significant change in binding affinities when compared with those of the prototype RBD (Figure 2C).

To test whether omicron S protein supports virus entry with different efficiency, we tested the potential influence of the SARS-CoV-2 variants on cell infection with an assay with pseudotyped virus. Vesicular stomatitis virus (VSV)-vectored pseudotyped viruses incorporated S proteins of the prototype, the aforementioned VOCs, and D614G, the first dominant SARS-CoV-2 variant (Hou et al., 2020). These pseudotyped viruses were then used to infect Vero cells, and the GFP signals were quantified by confocal image cytometer for the infection efficiency. The results showed that all variants infected Vero cells with significantly higher efficiency (Figure 2B; Figure S2). The delta strain was shown to infect cells with the highest efficiency. The omicron strain also demonstrated an increased trend in infection efficiency, although it was milder than that of the delta strain.

Structures of hACE2 in complex with RBDs of omicron and delta variants

To further characterize the binding modes of hACE2 with omicron RBD or delta RBD, we used both the X-ray crystallography

and cryo-EM to determine the complex structures. The crystal and cryo-EM complex structures of hACE2 with omicron RBD were solved at resolutions of 3.0 Å and 3.4 Å, respectively (Table 1; Figure S3; Tables S2 and S3). The crystal structure of hACE2-delta RBD complex was also determined at a resolution of 3.35 Å (Table 1; Table S2). The electron density map of the crystal structures clearly shows the binding details of omicron RBD-hACE2 and delta RBD-hACE2 interface (Figures 3D and 3E; Figure S4). Comparatively, local features of the cryo-EM map (electrostatic potential map) of the omicron RBD-hACE2 binding interface suggest that it is much more flexible in the cryo-EM structure than it is in the crystal structure, reflecting dynamic properties of the RBD-hACE2 interaction. Thus, we used the crystal structures for further analysis.

In comparison with those of the prototype RBD-hACE2 complex, omicron RBD-hACE2 and delta RBD-hACE2 crystal structures exhibit overall similar conformation, with root-mean-square deviation (RMSD) of 0.253 Å for 746 Cα atoms and 0.494 Å for 751 Cα atoms, respectively (Figures 3A–3C; Figures S5A–S5C). When compared to the crystal structure with their own cryo-EM structure, the omicron RBD-hACE2 and delta RBD-hACE2 complexes have RMSDs of 0.944 Å and 1.018 Å, respectively.

Structural analysis further shows the accessible surface areas (ASAs) for the omicron RBD-hACE2 crystal complex structure is larger than its ASA in cryo-EM, with 1,727.2 $Å^2$ and 1,709.5 $Å^2$, respectively. A similar result is also observed in the delta RBD-hACE2 complexes, with an ASA of 1,704.0 $Å^2$ for the crystal structure and 1,656.6 $Å^2$ for the cryo-EM structure, respectively.

As previously reported (Wang et al., 2020), the binding interface of hACE2 for SARS-CoV-2 RBD is distributed over two patches. In patch 1 of omicron RBD-hACE2 complex, S19 of hACE2 forms H bonds with both A475 and N477 of RBD. Q24 of hACE2 interacts with N487 of RBD through an H bond. Y83 of hACE2 contacts Y489 and N487 of RBD by H bonds, and it also forms a π-π stacking interaction with F486 of RBD. H34 from hACE2 contacts Y453 of RBD through an H bond (Figure 3D). E35 of hACE2 forms a salt bridge with R493 from RBD (Figure 3D). Residue F486 from RBD packs against a small hydrophobic pocket in the interface formed by F28, L79, M82, and Y83 of hACE2 (Figure 3D). In patch 2 of omicron RBD-hACE2 complex, residue D38 from hACE2 forms an H bond with Y449 and a salt bridge with R498 of RBD (Figure 3D). Q42 of hACE2 also interacts with Y449 of RBD by an H bond. Y41 from hACE2 forms an H bond with T500 of RBD and also forms a π-π stacking interaction with Y501 of RBD (Figure 3D). K353 from hACE2 forms a hydrogen bond with G502 of RBD.

In patch 1 of delta RBD-hACE2 complex, S19, Q24, K31, H34, and Y83 from hACE2 contact residues of A475, N487, Q493, Y453, and Y489 from RBD through H bonds, respectively. Y83 of hACE2 forms an H bond with Y498 from RBD and a π-π stacking interaction with F486 of RBD. D30 from hACE2 binds to K417 of RBD with a salt bridge. F486 of RBD can also constitute to strong hydrophobic interactions with F28, L79, M82, and Y83 of hACE2 (Figure 3E). In patch 2, E37, D38, Y41, and K353 from hACE2 form H bonds with Y505, Y449, T500, and G496 on the RBD, respectively. Q42 of hACE2 contacts both Y449 and Q498 from RBD through H bonds (Figure 3E). It should be noted that both of the two substitutions, specifically the L452R and T478K, locate outside the RBD-hACE2 binding interface, which is consistent with the observation that similar binding affinities with ACE2 for both delta and prototype RBDs were obtained (Figure 3A; Table 1).

Next, we focused on the substitutions of RBD on the hACE2 binding interface. We labeled the residues involved in hACE2 binding on RBDs of prototype, omicron, delta, and GD/1/2019 (Figures 4A–4D). In comparison with what was seen on prototype RBD, omicron RBD possesses eight substitutions on the hACE2 recognizing interface, namely K417N, G446S, E484A, Q493R, G496S, Q498R, N501Y, and Y505H. Another substitution, S477N, implicates N477 from hACE2 and expands area of the binding interface (Figures 4E and 4F). The two substitutions observed on delta RBD, namely L452R and T478K, do not participate in hACE2 interaction (Figure 4G). Two substitutions (R417 and H498) were observed on the interface of GD/1/2019 (Figure 4H), and their impact on hACE2 binding has already been analyzed and explained in a previous study (Niu et al., 2021). As a reference, we also labeled the substituted residues in the RBD from other VOCs (Figures 4K–4M), which demonstrated that the omicron variant possesses mutations on almost all the spots where substitutions were observed in previous VOCs.

Notably, the substitutions in omicron RBD lead to drastic change of electrostatic charges (Figures 4I and 4J). Compared with other VOC RBDs and GD/1/2019 RBD, the binding surface of omicron RBD has the largest-scale positive charge region (Figure S4). Particularly, T478K, Q493R, and Q498R substitutions significantly increased positive changes, and, meanwhile, E484A decreases the negative charges. R493 and R498 of RBD might be attracted by the negative charges around E35 and D38 of hACE2, respectively. Similar results have also been observed in the previous study (Niu et al., 2021), and Q498 was mutated to H498, enhancing its binding affinity to human, mouse, or rat ACE2. On the other hand, K478 is located away from hACE2 binding surface and might only influence hACE2 binding allosterically.

Comparative analyses of hACE2 interaction with omicron RBD and prototype RBD

In order to better understand the differences of hACE2 binding with RBD variants and prototype, we investigated details of the substitutions on the RBD variant and compared their interaction patterns with those of prototype RBD (Table 1). For omicron RBD, the impact of K417N and N501Y substitutions has already been extensively analyzed by previous studies (Han et al., 2021), which showed that K417N destroyed the salt bridge with D30 of hACE2 and decreased hACE2 binding affinity, and the aromatic ring of the Y501 could make new favorable non-bonded interactions with hACE2, such as cation-π interaction with K353 of hACE2 (Figures 5A and 5G). The substitution of S477 to N477 on omicron RBD confers two new H bonds, with S19 strengthening the hACE2 RBD interaction (Figure 5B). E484 in prototype RBD forms Van der Waals' contacts with K31 from hACE2, and the substitution to A484 could exhibit mild impact on hACE2 binding (Figure 5C). Substitution of Q493 to R493 forms a salt bridge with E35 instead of the H bond (Figure 5D). Although S496 is located in the loop region and undergoes slight conformational change, the altered positioning deprived them of the H bonds with K353 (Figure 5E). Likewise, the position altering

Table 1. Amino acid residues comparison of prototype RBD, omicron RBD, and delta RBD interacting with hACE2

hACE2	Prototype RBD	Omicron RBD	Delta RBD
S19 (7/17/3)	A475 (3, 1), G476 (4)	A475 (3, 1), G476 (3), N477 (11, 2)	A475 (2), G476 (1)
Q24 (24/22/13)	A475 (4), G476 (5), N487 (15, 1)	A475 (3), G476 (5), N477 (1), N487 (13, 1)	A475 (1), G476 (2), N487 (9, 1), Y489 (1)
T27 (15/13/13)	F456 (5), Y473 (1), A475 (2), Y489 (7)	F456 (5), Y473 (1), A475 (2), Y489 (5)	F456 (7), A475 (2), Y489 (4)
F28 (7/9/8)	Y489 (7)	Y489 (9)	Y489 (8)
D30 (10/2/10)	K417 (4, 1), L455 (2), F456 (4)	L455 (1), F456 (1)	K417 (3, 1), L455 (2), F456 (5)
K31 (19/15/14)	L455 (2), F456 (5), E484 (1), Y489 (6), F490 (2), Q493 (3)	L455 (1), F456 (4), Y489 (6), R493 (4),	L455 (1), F456 (4), Y489 (4), Q493 (5, 1)
H34 (20/29/13)	Y453 (5, 1), L455 (9), Q493 (6)	Y453 (10, 1), R493 (11), S494 (8)	Y453 (8, 2), L455 (5)
E35 (8/6/0)	Q493 (8)	R493 (6, 1)	-
E37 (7/2/7)	Y505 (7)	H505 (2)	Y505 (7, 1)
D38 (15/19/11)	Y449 (9, 1), G496 (5), Q498 (1)	Y449 (7, 1), S496 (6), R498 (5, 2), Y501 (1)	Y449 (9, 1), Q498 (2)
Y41 (23/25/28)	Q498 (8), T500 (7, 1), N501 (8, 1)	R498 (3), T500 (7, 1), Y501 (15)	Q498 (13), T500 (7, 1), N501 (8)
Q42 (16/9/15)	G446 (4, 1), Y449 (4, 1), Q498 (8, 3)	Y449 (2, 1), R498 (7)	G446 (4), Y449 (2, 1), Q498 (9, 2)
L45 (4/1/2)	Q498 (3), T500 (1)	T500 (1)	Q498 (2)
L79 (2/2/0)	F486 (2)	F486 (2)	-
M82 (9/7/7)	F486 (9)	F486 (7)	F486 (7)
Y83 (20/22/18)	F486 (11), N487 (8, 1), Y489 (1)	F486 (11), N487 (10, 1), Y489 (1, 1)	F486 (8), N487 (9, 1), Y489 (1, 1)
N330 (8/8/4)	T500 (8)	T500 (8)	T500 (4)
K353 (50/57/57)	G496 (7, 1), N501 (11), G502 (4, 1), Y505 (28)	Y495 (1), S496 (2), Y501 (21), G502 (6, 1), H505 (27)	G496 (7, 1), Q498 (4), N501 (16), G502 (5, 1), Y505 (25)
G354 (11/11/10)	G502 (7), Y505 (4)	G502 (7), H505 (4)	G502 (7), Y505 (3)
D355 (9/8/9)	T500 (8, 1), G502 (1)	T500 (7), G502 (1)	T500 (8, 1), G502 (1)
R357 (3/3/3)	T500 (3)	T500 (3)	T500 (3)
R393 (1/0/1)	Y505 (1)	-	Y505 (1)
Total	288, **16**	287, **14**	247, **16**

The numbers in parentheses for prototype RBD, omicron RBD, and delta RBD residues represent the number of Van der Waals' contacts between the indicated residues with hACE2. The numbers with underline suggest numbers of potential H bonds between the pairs of residues. Van der Waals' contacts were analyzed at a cutoff of 4.5 Å and H bonds at a cutoff of 3.5 Å."-" represents that these amino acids in the hACE2 do not interact with the corresponding RBD.

caused by G446S substitution deprives S446 of the H bond with Q42. The Q498R mutant, with its longer side chain, formed a new salt bridge with D38 (Figure 5F). Interestingly, Y505H has been observed on the bat-origin RaTG13 (Liu et al., 2021a), where the substitution led to fewer Van der Waals' contacts. Mutated RaTG13 RBD carrying a H505Y substitution demonstrated strengthened interaction (Liu et al., 2021a). The H505 residue of omicron RBD also formed fewer Van der Waals' interactions with hACE2 (Figure 5H; Table 1), indicating the substitution plays an unfavorable role during hACE2 binding.

DISCUSSION

Although recombination mutations are common for coronaviruses (Su et al., 2016), the SARS-CoV-2 virus is rapidly evolving with the accumulation of multiple mutations, which will definitely worsen the situation of pandemic and threaten the public health. Both delta and omicron strains are highly contagious variants, raising the global concerns. In this study, we determined both X-ray crystallography and cryo-EM structures of omicron RBD-hACE2 complex as well as X-ray crystallography structure of delta RBD-hACE2 complex. We extensively analyzed both VOCs for hACE2-RBD binding interface and compared with that of the published delta-hACE2 cryo-EM structure (PDB ID: 7V7V).

In omicron RBD-hACE2 complex, some substituted residues decrease the binding affinity between RBD and hACE2, and others enhance the binding affinity. The K417N mutation in omicron RBD was also found in beta, gamma, and in some delta lineage RBDs. Delta (B.1.617.2) itself is not defined as containing

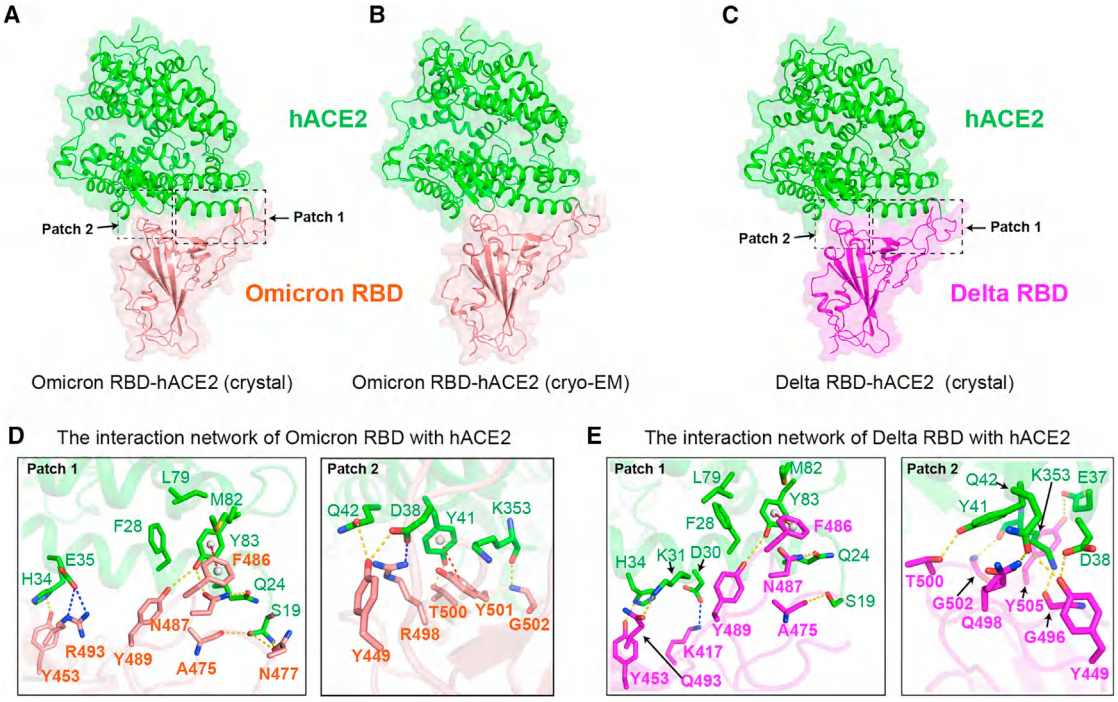


Figure 3. Overall architecture and interaction network of hACE2-omicron RBD and hACE2-delta RBD complexes

(A–C) Overall architecture of hACE2-omicron RBD crystal structure (A) and cryo-EM structure (B) and crystal structure of hACE2-delta RBD complex (C).

(D) Interaction patches in crystal structure of hACE2-omicron RBD. Side chains of interacting residues on hACE2 (green) and omicron RBD (pink) are shown as sticks and labeled appropriately. Yellow, red, and blue dashes present H bond, π-π stacking interaction, and salt bridges, respectively.

(E) Interaction patches in crystal structure of hACE2-delta RBD. Side chains of interacting residues on hACE2 (green) and omicron RBD (purple) are shown as sticks. Yellow, red, and blue dashes present H bond, π-π stacking interaction, and salt bridges, respectively.

See also Figures S3, S4, and S5.

the K417N mutation, but a sublineage (AY.2, or B.1.617.2.2) is defined as such (Baay et al., 2020; Han et al., 2021). In prototype RBD-hACE2 complex, K417 forms a salt bridge with D30 from ACE2 (Liu et al., 2021b; Wang et al., 2020). When it was mutated to N417 or T417, the absence of a salt bridge decreases the binding affinity with hACE2 (Han et al., 2021; Zhang et al., 2021b). The S477N mutation in omicron RBD has also been reported in B.1.620. S477N was shown to enhance the binding affinity of RBD with hACE2 (Singh et al., 2021). According to the omicron RBD-hACE2 complex structure, we found that N477, but not S477, forms an H bond with S19 from hACE2 to enhance the binding of RBD with hACE2. E484K mutation was reported in beta, gamma, zeta, eta, and theta variants, and E484Q mutation was reported in kappa variant (Han et al., 2021; Thye et al., 2021). As for omicron variant, E484 mutated to A484. E484 of RBD prototype forms weak contact with K31 in hACE2. When it was mutated to A484, the side chain is too short to contact with hACE2, resulting in decreased binding. Additionally, Q493K and Q498H mutations were observed in mouse-adapted strains (Huang et al., 2021b). Q498H also appears in two SARS-CoV-2-like viruses isolated from pangolins, GX/P2V/2017 and GD/1/2019 (Niu et al., 2021). Both Q493K and Q498H increase the binding affinity of SARS-CoV-2 RBD with hACE2 (Huang et al., 2021b; Niu et al., 2021). Q493Y and Q498Y were reported in the RBD of bat-origin virus RaTG13 (Liu et al., 2021a). Y493Q and Y498Q substitutions decrease the binding affinity of RaTG13 RBD with hACE2 (Liu et al., 2021a). In the omicron RBD-hACE2 complex, both Q493 and Q498 were substituted by a positive charged amino acid, arginine (R), and R493 and R498 form salt bridges with E35 and D38, respectively. N501Y is the key residue to enhance SARS-CoV-2 RBD binding to hACE2, which is also reported in alpha, beta, gamma, and theta variants (Han et al., 2021). N501 in SARS-CoV-2 RBD forms an H bond with Y41 from hACE2 and forms Van der Waals' force with K353. Multiple complex structures of SARS-CoV-2 RBD-hACE2 possessing N501Y substitution show that Y501 forms a π-π stacking interaction with Y41 and forms H bonds with K353 of hACE2 (Han et al., 2021; Nabel et al., 2021; Zhu et al., 2021). Notably, Q498R and N501Y mutations of omicron RBD were previously observed during *in vitro* evolution, which exhibited ~600-fold higher affinity with hACE2 (Zahradník et al., 2021). It was shown that Q498R is epistatic to N501Y and that R498 shifts the electrostatic feature to positive, thus attracting negative charges of R498 interacting residues. H505 is reported in the RaTG13 RBD-hACE2 complex (Liu et al., 2021a). Y505 forms van der Waals' force with E37, K353, and G354, but H505 only contacts K353 and G354 and doesn't contact E37 (Liu et al., 2021a). In our previous work, when we substituted H505 to Y505 in RaTG13 RBD, the binding affinity of RaTG13 RBD with hACE2 increased (Liu et al., 2021a). On the contrary, the Y505

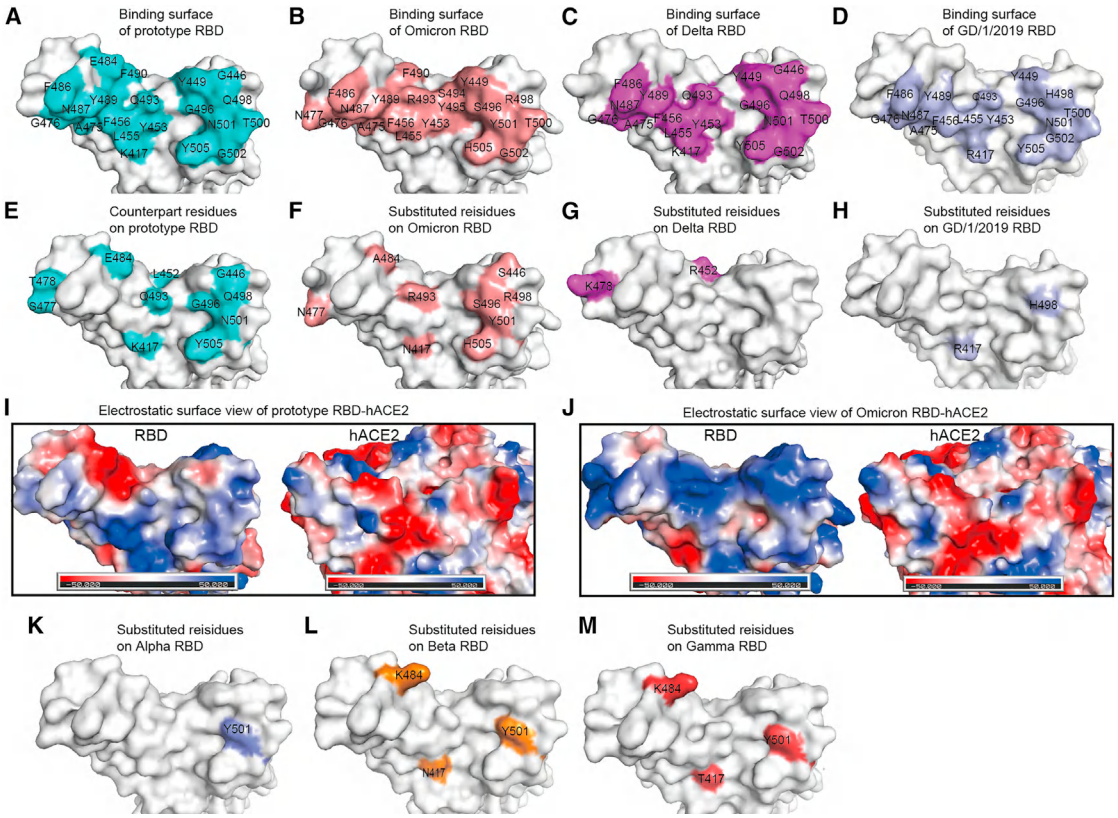


Figure 4. The RBMs and different residues on the RBDs

(A–D) The binding surfaces of hACE2 with prototype RBD (A), omicron RBD (B), delta RBD (C), and GD/1/2019 RBD (D) were labeled in cyan, salmon, magenta, and light blue, respectively.

(E and F) Nine different residues on hACE2 binding interface between omicron RBD and prototype RBD were labeled.

(I and J) Electrostatic surface view of prototype SARS-CoV-2 RBD and omicron RBD. The first panel represents the top view. The others are yielded by rotation of the former panel along a horizontal axis.

(K and L) The different residues on the RBMs between VOCs and prototype are labeled.

See also Figure S6.

in prototype RBD was substituted by H505 in omicron RBD. H505 forms fewer contacts with hACE2, which also indicates that Y505H substitution on omicron RBD could decrease the binding of omicron RBD with hACE2. In summary, K417N, G446S, E484A, G496S, and Y505H substitutions decrease the binding affinity of omicron RBD with hACE2, whereas S477N and N501Y compensatively increase the binding affinity of omicron RBD with hACE2.

Although the binding affinity of delta RBD with hACE2 is not significantly changed comparing with the prototype RBD, delta variant has higher transmissibility and a shorter incubation period, which indicates that there are some other factors affecting viral transmission. The mutations in other regions of S protein could also influence virus infection. For instance, the P681R mutation in the S protein of delta increased the level of the cleaved S2 subunit, and the level of the cleaved S2 subunit of the D614G/P681R mutation was significantly higher than that of D614G alone (Tian et al., 2021). For omicron variant, the multiple substitutions outside RBD in the S protein domain are needed to be evaluated in the future.

The emerging variants contain various important mutations in the RBD, enabling it to escape immune surveillance and thus decrease the protective potency of vaccines. Because there is only one mutation (N501Y) on the alpha variant, the efficacy of vaccine was slightly affected (Shen et al., 2021), and re-infection of alpha is found to be no greater than prototype SARS-CoV-2 virus (Graham et al., 2021). Beta variant has a great risk of re-infection because it escapes some neutralizing antibodies produced by dominant strains. The E484K substitution on RBD of beta stain is the key residue associated with a high rate of immune escape. Although there is only one amino acid (residue 417) difference in RBD between beta and gamma variants, gamma is less resistant to naturally acquired or vaccine-induced antibody responses. The transmission of delta increased a lot, but there is not much evidence for vaccine efficacy reduction. Compared with other variants, omicron possesses the largest-scale mutations in the receptor-binding motif (RBM) of RBD, in which E484 and Q493 have been observed to play important roles in immune escape. Notably, E484A, Q493K, and Q493R were repeatedly reported to emerge in immune-compromised patients or during

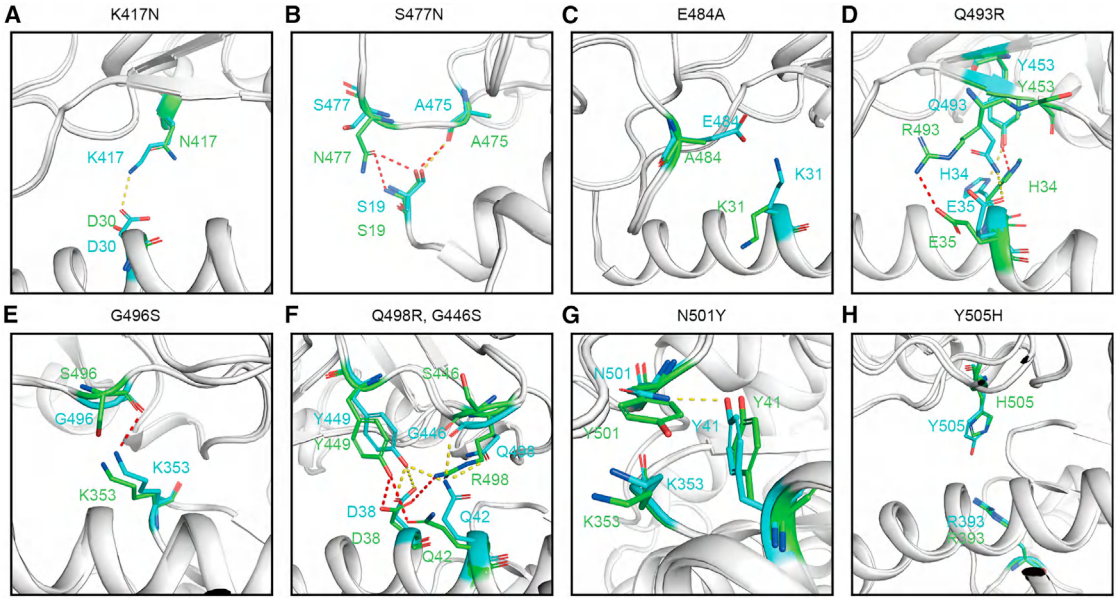


Figure 5. Structural comparison of prototype RBD-hACE2 and omicron RBD-hACE2

(A–H) The nine substitutions on the binding surface of RBD are shown. The residues in the prototype RBD are shown as cyan sticks and the residues in the omicron RBD are shown as green sticks. The yellow dashes represent H bond and salt bridges between prototype RBD and hACE2. The red dashes represent H bond and salt bridges between omicron RBD and hACE2. H-bond interactions were analyzed at a cutoff of 3.5 Å.

monoclonal antibodies treatment, which is responsible for immune escape (Choi et al., 2020; Clark et al., 2021; Focosi et al., 2021; Guigon et al., 2021; Starr et al., 2021). Interestingly, in comparison with prototype SARS-CoV-2, the neutralization capacity of sera immunized by double mRNA1273-vaccine (non-boosted) and additionally BNT162b2-boosted vaccine were reduced by about 20- and 22.7-fold respectively for omicron (Wilhelm et al., 2021).

The *in vitro* receptor binding and pseudovirus entry assays indicate that SARS-CoV-2 has a broad range of hosts (Liu et al., 2021c; Wu et al., 2020). In nature, SARS-CoV-2 has been reported to infect lots of animals including cats, dogs, lions, tigers, minks, ferrets, snow leopards, pumas, gorillas, white-tailed deer, and hippopotamuses (Gao and Wang, 2021). SARS-CoV-2 from mink can also be transmitted back to human. Several mink-related SARS-CoV-2 variants were also detected. It indicates that SARS-CoV-2 can mutate to adapt to other hosts during cross-species transmission, which inevitably facilitates the evolution of the SARS-CoV-2. Residues 493, 498, and 501 in RBD have already been identified as key sites for the host range of SARS-CoV-2. For example, a single N501Y mutation of RBD allowed it to infect mice. Q493K and Q498H mutations were also detected in mice-adaptive strains of SARS-CoV-2. Studies showed Q498H mutation appeared after only one passage, and Q493K mutation was introduced to the S protein after five passages in BALB/C mice (Huang et al., 2021b). Therefore, residues 493, 498, and 501, as well as other mutations in the RBD of omicron, could probably change the host spectrum of SARS-CoV-2 and whether omicron variant would be able to cross the species barrier needs to be evaluated in the future.

Limitations of the study

Receptor binding is the key step for SARS-CoV-2 infection. In this study, we only focused on the interaction between the RBDs of two VOCs and hACE2. However, other mutations in the spike protein outside the RBD domain could also contribute to the transmissibility. In addition, other co-receptors and co-factors could also play a role in SARS-CoV-2 infection, which should be addressed in the future studies. Moreover, pseudovirus system only represents the function of S protein, and other proteins of SARS-CoV-2 could also influence the infectivity and pathogenicity. The authentic virus infection assay needs to be further studied.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cells
- METHOD DETAILS
  - Gene Cloning
  - Protein expression and purification
  - Flow Cytometry
  - SPR Analysis

- Production and quantification of VSV pseudotyped viruses
- Pseudovirus infection assay
- Crystallization
- Data collection and structure determination
- Cryo-EM sample preparation and data collection
- Image processing
- Model building and refinement

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Flow cytometry analysis
  - Binding affinity analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2022.01.001.

ACKNOWLEDGMENTS

We are grateful to Drs G. Wong and G. Shang for revising the manuscript. We are grateful to the PMI-IMCAS Public Technology Service Center for its support on FACS assay. We thank Tong Zhao from Institute of Microbiology, Chinese Academy of Sciences, for her technical support in FACS assay. We acknowledge the staff of beamline BL02U1 and BL10U2 at the Shanghai Synchrotron Radiation Facility for assistance during data collection. We thank all staff at the Cryo-EM Center of Southern University of Science and Technology for their assistances. We thank Y.Q. Mi, H. Feng, and X.X. Bai at the Cryo-EM Center, Shanxi Academy of Advanced Research and Innovation, for their technical support on the cryo-EM. We thank Shuimu BioSciences for cryo-EM facility access and technical support during image acquisition. This work was supported by the National Key R&D Program of China (2021YFC2301401, 2020YFA0907102, 2020YFA0509202, 2021YFC0863300), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB29010202, XDB37030204), the Queensland-Chinese Academy of Sciences Collaborative Science Fund 2020 (153211KYSB20200001), the intramural special grant for SARS-CoV-2 research from the Chinese Academy of Sciences, and the Science and Technology Innovation Committee of Shenzhen Municipality (JSGG20200207155251653). S. Liu is funded by President's Excellent Postdoctoral Program of SUSTech.

AUTHOR CONTRIBUTIONS

J.Q., G.F.G., and P.W. initiated and designed the project. Specifically, J.Q. supervised and coordinated the experiments for crystals preparations, the diffraction data collection, and structure determinations. G.F.G. supervised the structural analysis, SPR and FACS assays, and pseudoviruses transductions. P.W. supervised and coordinated the samples preparations for the cryo-EM structural analysis, data collection, and cryo-EM structure determinations. Pengcheng Han, L.L., and D.Z. purified the proteins and grew the crystals with the help of B.H. and M.T. Pengcheng Han and L.L. performed the SPR analysis. D.Z. conducted the flow cytometry assay. Q.W., Pengcheng Han, and J.Q. collected the crystal structural data and solved the structures. S.L., Pu Han, and P.W. collected the cryo-EM structural data and solved the structure with the help of X.L., Q.P., L.F., and Y.G. D.L., R.Z., and X.Z. performed pseudotyped virus-related assays. Z.X., C.S., K.L., J.Q., G.F.G., and P.W. wrote and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: December 17, 2021
Revised: December 26, 2021
Accepted: December 31, 2021
Published: January 6, 2022

REFERENCES

Abdool Karim, S.S., and de Oliveira, T. (2021). New SARS-CoV-2 variants—clinical, public health, and vaccine implications. N. Engl. J. Med. *384*, 1866–1868.

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. *66*, 213–221.

Baay, M., Lina, B., Fontanet, A., Marchant, A., Saville, M., Sabot, P., Duclos, P., Vandeputte, J., and Neels, P. (2020). SARS-CoV-2: Virology, epidemiology, immunology and vaccine development. Biologicals *66*, 35–40.

Cao, Y., Yisimayi, A., Bai, Y., Huang, W., Li, X., Zhang, Z., Yuan, T., An, R., Wang, J., Xiao, T., et al. (2021a). Humoral immune response to circulating SARS-CoV-2 variants elicited by inactivated and RBD-subunit vaccines. Cell Res. *31*, 732–741.

Cao, Y.R., Wang, J., Jian, F., Xiao, T., Song, W., Yisimayi, A., Huang, W., Li, Q., Wang, P., and An, R. (2021b). B. 1.1. 529 escapes the majority of SARS-CoV-2 neutralizing antibodies of diverse epitopes. bioRxiv. https://doi.org/10.1101/2021.12.07.470392.

Cele, S., Jackson, L., Khoury, D.S., Khan, K., Moyo-Gwete, T., Tegally, H., San, J.E., Cromer, D., Scheepers, C., Amoako, D., et al.; NGS-SA; COMMIT-KZN Team (2021). SARS-CoV-2 Omicron has extensive but incomplete escape of Pfizer BNT162b2 elicited neutralization and requires ACE2 for infection. medRxiv. 2021.12.08.21267417. https://doi.org/10.1101/2021.12.08.21267417.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *66*, 12–21.

Choi, B., Choudhary, M.C., Regan, J., Sparks, J.A., Padera, R.F., Qiu, X., Solomon, I.H., Kuo, H.H., Boucau, J., Bowman, K., et al. (2020). Persistence and evolution of SARS-CoV-2 in an immunocompromised host. N. Engl. J. Med. *383*, 2291–2293.

Clark, S.A., Clark, L.E., Pan, J., Coscia, A., McKay, L.G.A., Shankar, S., Johnson, R.I., Brusic, V., Choudhary, M.C., Regan, J., et al. (2021). SARS-CoV-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms. Cell *184*, 2605–2617.e18.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501.

Ferreira, I., Datir, R., Papa, G., Kemp, S., Meng, B., Rakshit, P., Singh, S., Pandey, R., Ponnusamy, K., and Radhakrishnan, V.S. (2021). SARS-CoV-2 B. 1.617 emergence and sensitivity to vaccine-elicited antibodies. bioRxiv. https://doi.org/10.1101/2021.05.08.443253.

Focosi, D., Novazzi, F., Genoni, A., Dentali, F., Gasperina, D.D., Baj, A., and Maggi, F. (2021). Emergence of SARS-CoV-2 spike protein escape mutation Q493R after treatment for COVID-19. Emerg. Infect. Dis. *27*, 2728–2731.

Gao, G.F., and Wang, L. (2021). COVID-19 expands its territories from humans to animals. China CDC Wkly *3*, 855–858.

Garcia-Beltran, W.F., Lam, E.C., St Denis, K., Nitido, A.D., Garcia, Z.H., Hauser, B.M., Feldman, J., Pavlovic, M.N., Gregory, D.J., Poznansky, M.C., et al. (2021). Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell *184*, 2372–2383.e9, e2379.

Genovese, L., Zaccaria, M., Farzan, M., Johnson, W., and Momeni, B. (2021). Investigating the mutational landscape of the SARS-CoV-2 Omicron variant via ab initio quantum mechanical modeling. bioRxiv. https://doi.org/10.1101/2021.12.01.470748.

Graham, M.S., Sudre, C.H., May, A., Antonelli, M., Murray, B., Varsavsky, T., Kläser, K., Canas, L.S., Molteni, E., Modat, M., et al.; COVID-19 Genomics UK (COG-UK) Consortium (2021). Changes in symptomatology, reinfection,

and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. Lancet Public Health *6*, e335–e345.

Guigon, A., Faure, E., Lemaire, C., Chopin, M.-C., Tinez, C., Assaf, A., Lazrek, M., Hober, D., Bocket, L., Engelmann, I., and Alidjinou, E.K. (2021). Emergence of Q493R mutation in SARS-CoV-2 spike protein during bamlanivimab/etesevimab treatment and resistance to viral clearance. J. Infect. S0163-4453(21) 00435-7. https://doi.org/10.1016/j.jinf.2021.08.033.

Han, P., Su, C., Zhang, Y., Bai, C., Zheng, A., Qiao, C., Wang, Q., Niu, S., Chen, Q., Zhang, Y., et al. (2021). Molecular insights into receptor binding of recent emerging SARS-CoV-2 variants. Nat. Commun. *12*, 6103.

Hou, Y.J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K.H., 3rd, Leist, S.R., Schäfer, A., Nakajima, N., Takahashi, K., et al. (2020). SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science *370*, 1464–1468.

Huang, B., Dai, L., Wang, H., Hu, Z., Yang, X., Tan, W., and Gao, G.F. (2021a). Serum sample neutralisation of BBIBP-CorV and ZF2001 vaccines to SARS-CoV-2 501Y.V2. Lancet Microbe *2*, e285.

Huang, K., Zhang, Y., Hui, X., Zhao, Y., Gong, W., Wang, T., Zhang, S., Yang, Y., Deng, F., Zhang, Q., et al. (2021b). Q493K and Q498H substitutions in Spike promote adaptation of SARS-CoV-2 in mice. EBioMedicine *67*, 103381.

Kucukelbir, A., Sigworth, F.J., and Tagare, H.D. (2014). Quantifying the local resolution of cryo-EM density maps. Nat. Methods *11*, 63–65.

Li, J., Lai, S., Gao, G.F., and Shi, W. (2021a). The emergence, genomic diversity and global spread of SARS-CoV-2. Nature *600*, 408–418.

Li, Q., Nie, J., Wu, J., Zhang, L., Ding, R., Wang, H., Zhang, Y., Li, T., Liu, S., Zhang, M., et al. (2021b). SARS-CoV-2 501Y.V2 variants lack higher infectivity but do have immune escape. Cell *184*, 2362–2371.e9, e2369.

Liu, Y., and Rocklöv, J. (2021). The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. J. Travel Med. *28*, taab124. https://doi.org/10.1093/jtm/taab124.

Liu, L., Wang, P., Nair, M.S., Yu, J., Rapp, M., Wang, Q., Luo, Y., Chan, J.F.-W., Sahi, V., Figueroa, A., et al. (2020). Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. Nature *584*, 450–456.

Liu, K., Pan, X., Li, L., Yu, F., Zheng, A., Du, P., Han, P., Meng, Y., Zhang, Y., Wu, L., et al. (2021a). Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species. Cell *184*, 3438–3451.e10.

Liu, K., Tan, S., Niu, S., Wang, J., Wu, L., Sun, H., Zhang, Y., Pan, X., Qu, X., Du, P., et al. (2021b). Cross-species recognition of SARS-CoV-2 to bat ACE2. Proc. Natl. Acad. Sci. USA *118*, e2020216118.

Liu, Y., Hu, G., Wang, Y., Ren, W., Zhao, X., Ji, F., Zhu, Y., Feng, F., Gong, M., Ju, X., et al. (2021c). Functional and genetic analysis of viral receptor ACE2 orthologs reveals a broad potential host range of SARS-CoV-2. Proc. Natl. Acad. Sci. USA *118*, e2025373118.

Lu, G., Wang, Q., and Gao, G.F. (2015). Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends Microbiol. *23*, 468–478.

Mlcochova, P., Kemp, S.A., Dhar, M.S., Papa, G., Meng, B., Ferreira, I.A.T.M., Datir, R., Collier, D.A., Albecka, A., Singh, S., et al.; Indian SARS-CoV-2 Genomics Consortium (INSACOG); Genotype to Phenotype Japan (G2P-Japan) Consortium; CITIID-NIHR BioResource COVID-19 Collaboration (2021). SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. Nature *599*, 114–119.

Muik, A., Wallisch, A.-K., Sänger, B., Swanson, K.A., Mühl, J., Chen, W., Cai, H., Maurus, D., Sarkar, R., Türeci, Ö., et al. (2021). Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine-elicited human sera. Science *371*, 1152–1153.

Nabel, K.G., Clark, S.A., Shankar, S., Pan, J., Clark, L.E., Yang, P., Coscia, A., McKay, L.G.A., Varnum, H.H., Brusic, V., et al. (2021). Structural basis for continued antibody evasion by the SARS-CoV-2 receptor binding domain. Science, l6251.

Niu, S., Wang, J., Bai, B., Wu, L., Zheng, A., Chen, Q., Du, P., Han, P., Zhang, Y., Jia, Y., et al. (2021). Molecular basis of cross-species ACE2 interactions with SARS-CoV-2-like viruses of pangolin origin. EMBO J. *40*, e107786.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. *276*, 307–326.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Piccoli, L., Park, Y.-J., Tortorici, M.A., Czudnochowski, N., Walls, A.C., Beltramello, M., Silacci-Fregni, C., Pinto, D., Rosen, L.E., Bowen, J.E., et al. (2020). Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. Cell *183*, 1024–1042.e21, e1021.

Qin, S., Cui, M., Sun, S., Zhou, J., Du, Z., Cui, Y., and Fan, H. (2021). Genome characterization and potential risk assessment of the novel SARS-CoV-2 variant Omicron (B. 1.1. 529). Zoonoses. https://doi.org/10.15212/ZOONOSES-2021-0024.

Rohou, A., and Grigorieff, N. (2015). CTFFIND4: Fast and accurate defocus estimation from electron micrographs. J. Struct. Biol. *192*, 216–221.

Shen, X., Tang, H., McDanal, C., Wagh, K., Fischer, W., Theiler, J., Yoon, H., Li, D., Haynes, B.F., Sanders, K.O., et al. (2021). SARS-CoV-2 variant B.1.1.7 is susceptible to neutralizing antibodies elicited by ancestral spike vaccines. Cell Host Microbe *29*, 529–539.e3, e523.

Singh, A., Steinkellner, G., Köchl, K., Gruber, K., and Gruber, C.C. (2021). Serine 477 plays a crucial role in the interaction of the SARS-CoV-2 spike protein with the human receptor ACE2. Sci. Rep. *11*, 4320.

Starr, T.N., Greaney, A.J., Addetia, A., Hannon, W.W., Choudhary, M.C., Dingens, A.S., Li, J.Z., and Bloom, J.D. (2021). Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. Science *371*, 850–854.

Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., Liu, W., Bi, Y., and Gao, G.F. (2016). Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. *24*, 490–502.

Tan, W., Zhao, X., Ma, X., Wang, W., Niu, P., Xu, W., Gao, G.F., and Wu, G. (2020). A novel coronavirus genome identified in a cluster of pneumonia cases - Wuhan, China 2019-2020. China CDC Wkly *2*, 61–62.

Thye, A.Y., Law, J.W., Pusparajah, P., Letchumanan, V., Chan, K.G., and Lee, L.H. (2021). Emerging SARS-CoV-2 variants of concern (vocs): an impending global crisis. Biomedicines *9*, 1303.

Tian, D., Sun, Y., Zhou, J., and Ye, Q. (2021). The global epidemic of the SARS-CoV-2 Delta variant, key spike mutations and immune escape. Front. Immunol. *12*, 751778.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.-Y., et al. (2020). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell *181*, 894–904.e9, e899.

Wang, Y., Zhang, L., Li, Q., Liang, Z., Li, T., Liu, S., Cui, Q., Nie, J., Wu, Q., Qu, X., and Huang, W. (2022). The significant immune escape of pseudotyped SARS-CoV-2 variant Omicron. Emerg. Microbes Infect. *11*, 1–5.

Wilhelm, A., Widera, M., Grikscheit, K., Toptan, T., Schenk, B., Pallas, C., Metzler, M., Kohmer, N., Hoehl, S., Helfritz, F.A., et al. (2021). Reduced neutralization of SARS-CoV-2 omicron variant by vaccine sera and monoclonal antibodies. medRxiv. https://doi.org/10.1101/2021.12.07.21267432.

Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., et al. (2018). MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. *27*, 293–315.

Wu, L., Chen, Q., Liu, K., Wang, J., Han, P., Zhang, Y., Hu, Y., Meng, Y., Pan, X., Qiao, C., et al. (2020). Broad host range of SARS-CoV-2 and the molecular basis for SARS-CoV-2 binding to cat ACE2. Cell Discov. *6*, 68.

Zahradník, J., Marciano, S., Shemesh, M., Zoler, E., Harari, D., Chiaravalli, J., Meyer, B., Rudich, Y., Li, C., Marton, I., et al. (2021). SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. Nat. Microbiol. *6*, 1188–1198.

Zhang, Z., Zhang, Y., Liu, K., Li, Y., Lu, Q., Wang, Q., Zhang, Y., Wang, L., Liao, H., Zheng, A., et al. (2021b). The molecular basis for SARS-CoV-2 binding to dog ACE2. Nat. Commun. *12*, 4195.

Zheng, S.Q., Palovcak, E., Armache, J.-P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods *14*, 331–332.

Zhou, D., Dejnirattisai, W., Supasa, P., Liu, C., Mentzer, A.J., Ginn, H.M., Zhao, Y., Duyvesteyn, H.M.E., Tuekprakhon, A., Nutalai, R., et al. (2021). Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. Cell *184*, 2348–2361.e6, e2346.

Zhu, X., Mannar, D., Srivastava, S.S., Berezuk, A.M., Demers, J.P., Saville, J.W., Leopold, K., Li, W., Dimitrov, D.S., Tuttle, K.S., et al. (2021). Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. PLoS Biol. *19*, e3001237.

Zivanov, J., Nakane, T., Forsberg, B.O., Kimanius, D., Hagen, W.J., Lindahl, E., and Scheres, S.H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. eLife *7*, e42166.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial Strains		
Escherichia coli (E. coli) strain DH5α	TIANGEN	Cat# CB101-02
Escherichia coli (E. coli) strain BL21 (DE3)	Novagen	Cat# 69450
MAX Efficiency DH10Bac Competent *E. coli*	Invitrogen	Cat# 10361-012
Chemicals, antibody and Recombinant proteins		
PEI	Alfa	A04043896-1g
Anti-His/APC	Miltenyi Biotec	Cat# 130-119-820; RRID: AB_2751870
SARS-CoV-2 prototype RBD protein with his-tag, spike residues 319-541	This paper	accession number: EPI_ISL_402119
SARS-CoV-2 Alpha RBD protein with his-tag, spike residues 319-541	This paper	accession number: EPI_ISL_683466
SARS-CoV-2 Beta RBD protein with his-tag, spike residues 319-541	This paper	accession number: EPI_ISL_678615
SARS-CoV-2 Gamma RBD protein with his-tag, spike residues 319-541	This paper	accession number: EPI_ISL_833172
SARS-CoV-2 Delta RBD protein with his-tag, spike residues 319-541	This paper	accession number: EPI_ISL_6640916
SARS-CoV-2 Omicron RBD protein with his-tag, spike residues 319-541	This paper	accession number: EPI_ISL_410721
GD/1/2019 RBD protein with his-tag, spike residues 319-541	This paper	Accession number: EPI_ISL_410721
hACE2 protein, residues 18-615	This paper	accession number: NP_001358344
Critical Commercial Assays		
HisTrap HP 5 mL column	GE Healthcare	Cat# 17524802
HiLoad 16/600 Superdex 200 pg	GE Healthcare	Cat# 28989335
Series S Sensor Chip CM5	GE Healthcare	Cat# 29149603
Membrane concentrator	Millipore	UFC901096
Deposited Data		
Omicron RBD/hACE2 complex (Cryo-EM)	This paper	Protein Data Bank: 7WBL
Omicron RBD/hACE2 complex (crystal)	This paper	Protein Data Bank:7WBP
Delta RBD/hACE2 complex	This paper	Protein Data Bank: 7WBQ
Experimental Models: Cell Lines		
HEK293T cells	ATCC	ATCC CRL-3216
BHK-21 cells	ATCC	ATCC CCL-10
Vero	ATCC	ATCC CCL-81
Recombinant DNA		
pEGFP-N1	MiaoLingPlasmid	Cat# P0133
pEGFP-N1-hACE2	This paper	accession number: BAJ21180
pCAGGS	MiaoLingPlasmid	Cat# P0165
Software and Algorithms		
PyMOL software	Molecular Graphics System, Version 1.8 Schrö dinger	https://pymol.org/2/
BIAcore® 8K Evaluation software	GE Healthcare	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FlowJo V10	FLOWJO	https://www.flowjo.com/solutions/flowjo/downloads
Motioncor2	Zheng et al., 2017	N/A
COOT	Emsley and Cowtan, 2004	http://www.mrc-lmb.cam.ac.uk/personal/peemsley/coot/
Phenix	Adams et al., 2010	http://www.phenix-online.org/
MolProbity	Duke Biochemistry	http://molprobity.biochem.duke.edu/index.php

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Peiyi Wang (wangpy@sustech.edu.cn).

Materials Availability

All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement.

Data and Code Availability

The atomic coordinates for the Cryo-EM structure of the omicron RBD-hACE2 complex (PDB code: 7WBL), the crystal structures of the omicron RBD-hACE2 complex (PDB code: 7WBP) and Delta RBD-hACE2 complex (PDB code: 7WBQ) have been deposited in the Protein Data Bank (www.rcsb.org).

This study did not generate custom computer code.

Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cells

HEK293T cells (ATCC CRL-3216), BHK-21 cells (ATCC CCL-10), HuH7 cells and HeLa-hACE2 cells (prepared in this study) were cultured at 37°C in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS).

METHOD DETAILS

Gene Cloning

The full-length sequence of hACE2 (GenBank: NP_001358344) was cloned in pEGFP-N1 vector for flow cytometry. The extracellular domain of hACE2 (residues 1-740, GenBank: NP_001358344) fused with the Fc domain of mouse IgG (mFc) were cloned into the pCAGGS vector for protein expression. The coding sequences of RBDs from original SARS-CoV-2 (residue 319-541) and VOCs (residue 319-541) and hACE2(residues 19-615, GenBank:NP_001358344) were cloned into the pCAGGS vector. The coding sequence S proteins from original SARS-CoV-2 and VOCs (residues 1-1233) were cloned into pCAGGS vectors.

Protein expression and purification

The hACE2 fused with mFc were expressed and purified from the culture supernatants of HEK293F cells using a Protein A affinity column (GE Healthcare) and further purified by gel filtration using a SuperdexTM 200 10/300 GL (GE Healthcare). Purified proteins were stored in a buffer containing 20 mM Tris-HCl and 150 mM NaCl (pH 8.0). Proteins for SPR assay were transferred to PBST (1.8 mM KH_2PO_4, 10 mM Na2HPO4 (pH 7.4), 137 mM NaCl, 2.7 mM KCl, and 0.005% (v/v) Tween 20) buffer.

The hACE2 and RBDs from original SARS-CoV-2 and VOCs cloned in pCAGGS were expressed in HEK293F cells. Cell culture supernatants were collected, filtered with a 0.22 μm filter, purified by His-Trap HP column (GE Healthcare), and SuperdexTM 200 Increase 10/300 GL column (GE Healthcare). Purified proteins were stored in protein buffer (20 mM Tris-HCl, pH 8.0 and 150 mM NaCl).

Flow Cytometry

The plasmids containing hACE2 fused with eGFP were transfected into BHK-21 cells. A mixture containing RBDs from prototype SARS-CoV-2 (1 μg/mL) or VOCs (1 μg/mL) were incubated with the BHK-21 cells at 4°C for 1 h. Subsequently, cells were washed

with PBS thrice and stained with APC mouse anti-his secondary antibody for 1 h before being analyzed using BD FACS Canto FlowCytometer (BD Biosciences). The data of all samples were analyzed using FlowJo 7.6 (TreeStar Inc., Ashland, OR, USA)

SPR Analysis

The ACE2-mFC fusion proteins were transferred into PBST buffer (1.8 mM KH_2PO_4, 10 mM Na_2HPO_4 (pH 7.4), 137 mM NaCl, 2.7 mM KCl, and 0.05% (v/v) Tween 20) and immobilized on flow cell 2 of CM5 chip. Flow cell 1 was used as the negative control. Serially diluted RBDs from original SARS-CoV-2 and VOCs were then flowed over the chip in PBST buffer. Binding affinities were measured using a BIAcore 8K (GE Healthcare) at 25°C in the single-cycle mode. Binding kinetics were analyzed with Biacore™ Insight software (GE healthcare) using a 1:1 Langmuir binding model.

Production and quantification of VSV pseudotyped viruses

The original SARS-CoV-2, and VOC pseudoviruses were constructed with a GFP encoding replication-deficient vesicular stomatitis virus (VSV) vector backbone (VSV-ΔG-GFP) and the coding sequence of corresponding spike proteins, as previously described (Muik et al., 2021). Briefly, HEK293T cells were transfected by 30 ng of spike protein expression plasmids. The VSV-ΔG-GFP pseudovirus was added 24 h post-transfection. The inoculum was removed after incubation for 1 h at 37°C. The culture medium was then changed into DMEM supplemented with 10% FBS and 10 lg/mL of anti-VSV-G antibody (I1-Hybridoma ATCC CRL2700TM) after washing cells with PBS. The pseudoviruses were harvested 20 h post-inoculation, passed through a 0.45-lm filter (Millipore, Cat#SLHP033RB) before aliquoted, and stored at −80°C.

All pseudoviruses were treated with 0.5 U/μL BaseMuncher Endonuclease (Abcam, ab270049) for 1.5 h at 37°C to remove unpackaged RNA before quantification. Viral RNA was extracted (Bioer Technology, Cat# BYQ6.6.101711-213) and quantitated by quantitative RT–PCR (qPCR) using 7500 fast real-time PCR system (Applied Biosystems) with the primers and probe for detecting the P protein coding sequence of VSV.

Pseudovirus infection assay

The pseudovirus particles of original SARS-CoV-2 and Omicron variant were normalized to the same amount for quantitation by qRT–PCR. Then, 100 μL of each pseudovirus was added to each well of 96-well plate containing Vero cells. Uninfected Vero cells were used as control. Plates were imaged 15 h post-transfection. The numbers of fluorescent cells were determined on a CQ1 confocal image cytometer (Yokogawa). Each group contains 5 replicates. Statistical significance was determined by a two-sided unpaired Student's t test.

Crystallization

The sitting-drop method was used to obtain the Omicron RBD-hACE2 and Delta RBD-hACE2 complex crystals. In detail, purified complex proteins were concentrated to 5 and 10 mg/mL. Then, 0.8 μL protein was mixed with 0.8 μL reservoir solution. The resulting solution was sealed and equilibrated against 100 μL of reservoir solution at 18°C and 4°C. High resolution Omicron RBD-hACE2 complex crystal was grown in 0.15 M ammonium sulfate, 0.1 M Na HEPES, 20% w/v polyethylene glycol 4000, and Delta RBD-hACE2 was grown complex in 0.1 M sodium cacodylate, 25% w/v polyethylene glycol 4000.

Data collection and structure determination

The diffraction data were collected at Shanghai Synchrotron Radiation Facility (SSRF) BL02U1 (wavelength, 0.97919 Å). For data collection, the crystals were cryo-protected by briefly soaking in reservoir solution supplemented with 20% (v/v) glycerol before flash-cooling in liquid nitrogen. The dataset was processed with HKL2000 software (Otwinowski and Minor, 1997). The structure of 2 complexes were determined by the molecular replacement method using Phaser (Adams et al., 2010) with previously reported complex structure SARS-CoV-2-RBD complex with hACE2 (PDB: 6LZG). The atomic models were completed with Coot (Emsley and Cowtan, 2004) and refined with phenix.refine in Phenix (Adams et al., 2010), and the stereochemical qualities of the final models were assessed with MolProbity (Williams et al., 2018). Data collection, processing, and refinement statistics were summarized in Table S2. All structural figures were generated using Pymol software (https://pymol.org/2/).

Cryo-EM sample preparation and data collection

For the hACE2-Omicorn RBD complex, an aliquot of 3.5 μL solution (0.25 mg/mL) was applied to glow-discharged home-made graphene grids (Quantifoil Au R1.2/1.3, 300 mesh) blotted for 1.5 s with a humidity of 90% at 4°C before being plunged into liquid ethane using a Vitrobot Mark IV (Thermo Fisher). The frozen grids were loaded onto a Titan Krios transmission electron microscope (Thermo Fisher) and operated at 300 kV. The microscope is equipped with a BioQuantum energy filter and k3 direct electron detector (Gatan). A slit width of 10 eV was used for data collection. Automatic data collection was performed by using EPU software (Thermo Fisher). Movies were recorded in super-resolution counting mode at pixel size of 0.335 Å. The exposure was performed with a dose rate of 15 e^-/pixel/s and an accumulative dose of ~60 $e^-/Å^2$ for each movie, of which was fractionated into 40 sub-frames. The final defocus ranges of these two datasets were approximately −1.3~-2.3 μm.

Image processing

The drift correction of all stacks was performed with MotionCor2 (Zheng et al., 2017) to generate 2 × binned micrographs. Initial contrast transfer function (CTF) values for each micrograph were calculated with CTFFIND4.1 (Rohou and Grigorieff, 2015). Micrographs with an estimated resolution limit worse than 5 Å were discarded in the initial screening. A set of ~150,000 particles were auto-picked by Laplacian-of-Gaussian from RELION-3.1 (Zivanov et al., 2018) and then subjected to 2D classification to generate templates for auto-picking against the entire dataset. The subsequent data processing and reconstruction were performed using RELION-3.1.

For the Omicron RBD-hACE2 dataset, 1,659,778 particles were picked from 6,596 micrographs. The picked particles were then extracted and subjected to three rounds of reference-free 2D classification. A clean dataset with 863,953 particles from good 2D classes were selected to generate the initial model in RELION. Subsequently, the model was used as reference in RELION 3D classification. After the second round of 3D classification without applying symmetry, the predominant class containing a subset of 110,912 best particles shows the clear features of secondary structural elements and the high accuracy of particle alignment. These particles were subjected to 3D refinement, which yielded a reconstruction at 3.7 Å resolution. To further improve the resolution, dose-weighted images were generated by MotionCor2 with the first 2 frames and last 14 frames discarded for each stack, resulting in a reduced dose of ~35 $e^-/Å^2$. In addition, CTF refinement was performed to correct the local CTF values of each particle. Combining these parameters, a new round of 3D refinement was performed to obtain the density map at 3.4 Å resolution determined by the Fourier shell correlation (FSC) 0.143 cut-off value (Figure S3). Local resolution estimation was performed with ResMap (Kucukelbir et al., 2014).

Model building and refinement

The crystal structure of the hACE2-SARS-CoV-2 RBD complex (PDB: 6LZG) was docked into the cryo-EM density maps using CHIMERA (Pettersen et al., 2004). The model was manually corrected for local fit in COOT (Emsley et al., 2010) and the sequence register was updated based on alignment. The model was refined against the corresponding map in real space using PHENIX (Adams et al., 2010), in which the secondary structural restraints and Ramachandran restrains were applied. The stereochemical quality of each model was assessed using MolProbity (Chen et al., 2010). Statistics for model refinement and validation are shown in Table S3.

QUANTIFICATION AND STATISTICAL ANALYSIS

Flow cytometry analysis

All experiments were performed three times; one representative of each experiment is shown in Figure 2.

Binding affinity analysis

K_D values of SPR experiments were obtained with BIAcore 8K Evaluation Software (GE Healthcare), using a 1:1 binding model. The values indicate the mean ± SD of three independent experiments.

Supplemental figures

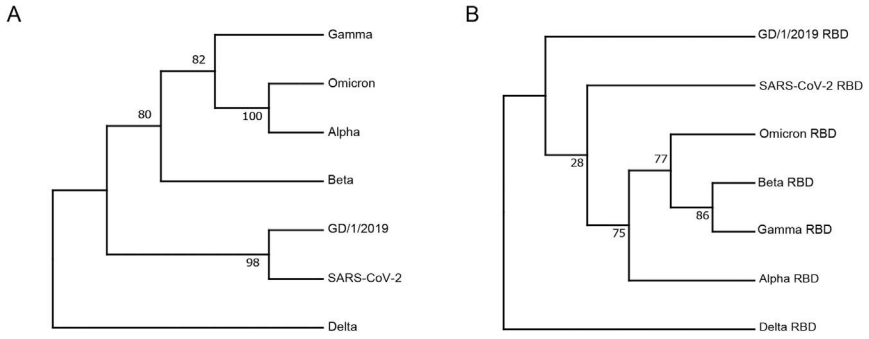


Figure S1. Phylogenetic analysis of representative SARS-CoV-2, related VOCs, and GD/1/2019, Related to Figure 1
Phylogenetic trees are depicted for nucleotide sequences of the full genome (A) and RBD (B).

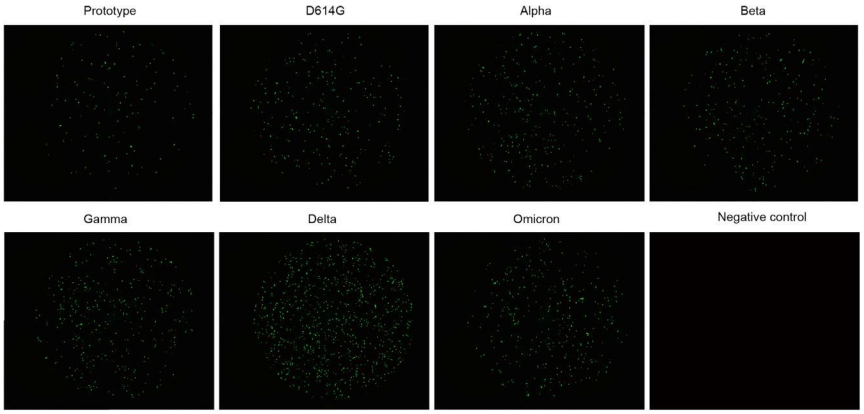


Figure S2. Entry of the pseudovirus of prototype SARS-CoV-2 and its VOCs into Vero cells, Related to Figure 2
Green fluorescent Vero cells indicate pseudovirus-transducing cells. Untransfected Vero cells were used as negative controls.

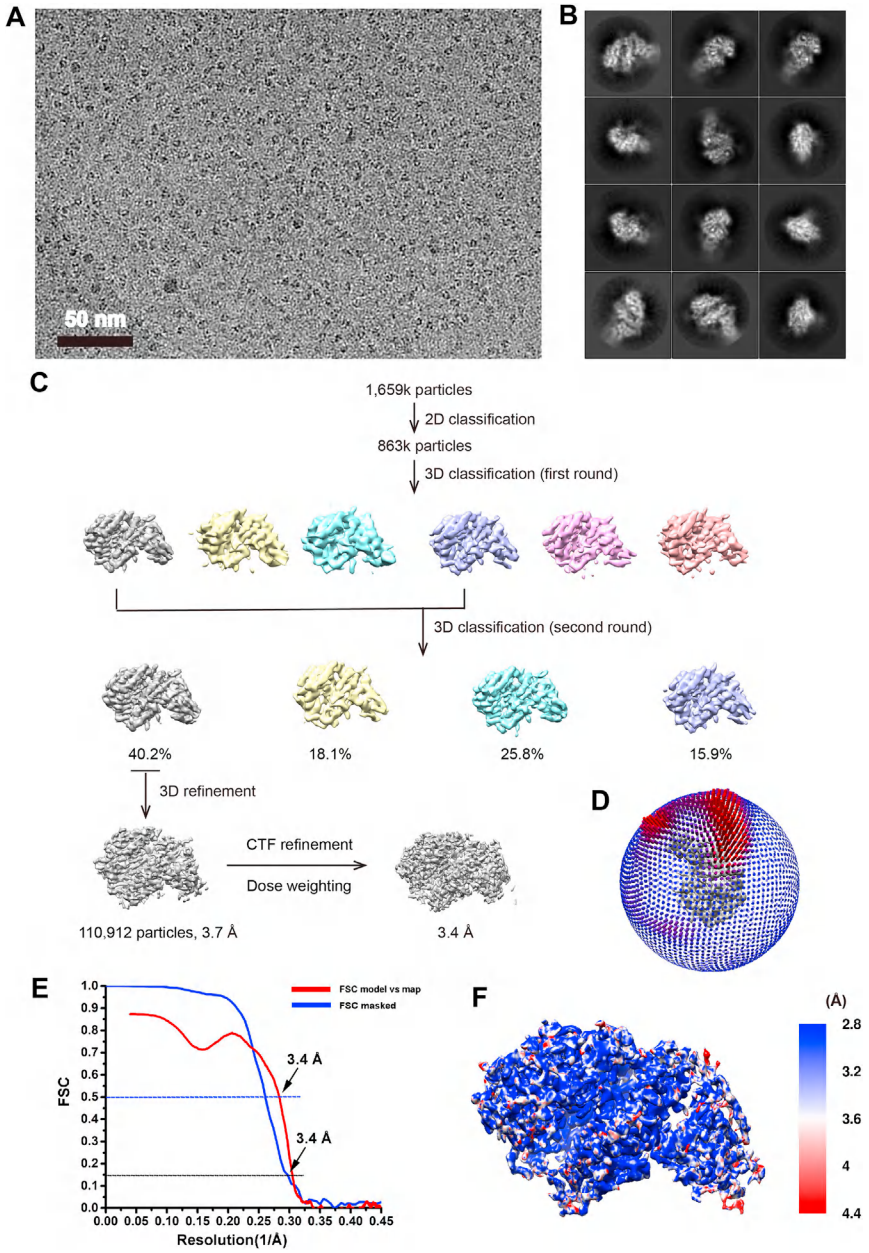


Figure S3. Cryo-EM structrual analysis of hACE2-omicron RBD, Related to Figure 3

(A) A representative cryo-EM micrograph of hACE2-Omicron RBD. Scale bar, 50 nm. (B) 2D class average images of hACE2-Omicron RBD. (C) A brief workflow of cryo-EM image processing and reconstruction. (D) Euler angle distribution of the final reconstruction. (E) The FSC curve for the reconstruction. (F) Local resolution distribution for the density map of hACE2-Omicron RBD.

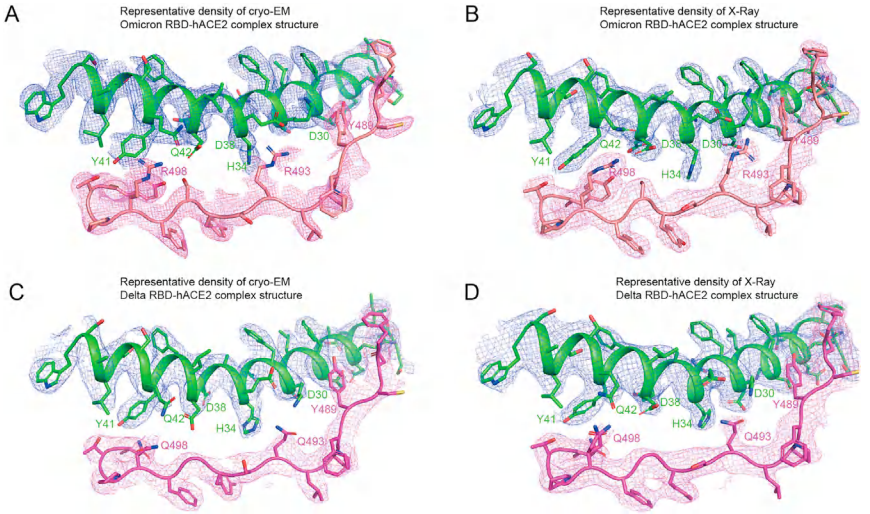


Figure S4. Representative densities and atomic models, Related to Figure 3

Representative densities and atomic models of Omicron RBD-hACE2 complex structures solved through cryo-EM (A) and X-ray (B) were shown. hACE2 was shown as green and Omicron RBD was colored by salmon. The representative densities and atomic models of Delta RBD-hACE2 complex structures determined by cryo-EM (PBD: 7V7V) (C) and X-ray (D) were shown. hACE2 was also shown as green and Delta RBD was labeled by magenta.

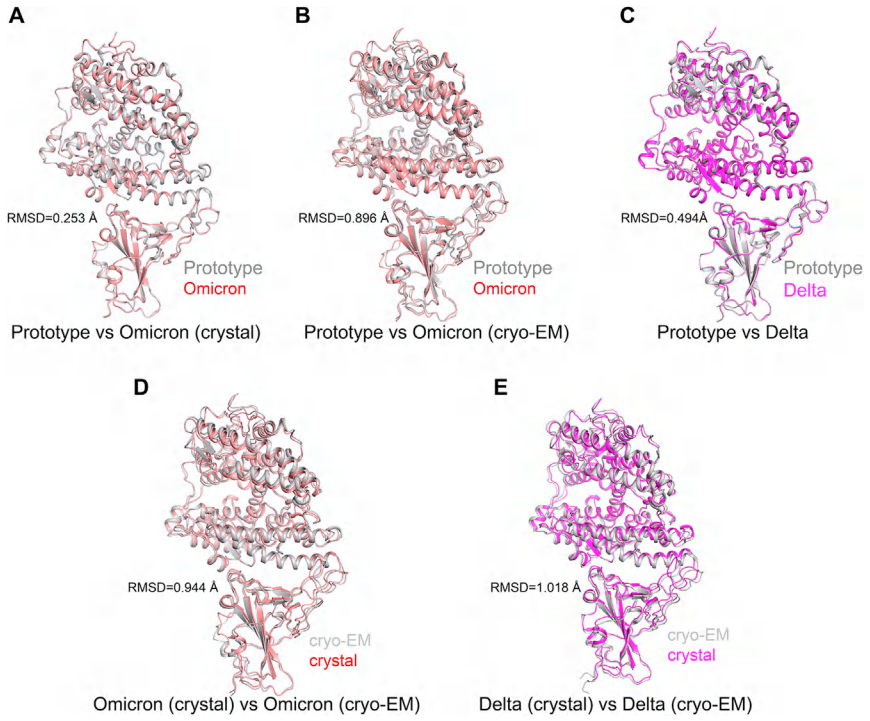


Figure S5. Overall structure comparison of hACE2 with VOC RBDs and GD/1/2019 RBD, Related to Figure 3

(A-C) The crystal structure (A) and cryo-EM structure (B) of Omicron RBD-hACE2 complex and crystal structure of Delta RBD-hACE2 complex are compared with prototype RBD-hACE2 complex and labeled by salmon. Both crystal structure and cryo-EM structure of Omicron RBD-hACE2 complex were colored by salmon. Delta RBD-hACE2 complex was labeled by magenta. Prototype RBD-hACE2 complex was labeled by white. (D and E) The structural comparations between crystal structure and cryo-EM structure of Omicron RBD-hACE2 complex (D) and Delta RBD-hACE2 complex (E). The cryo-EM of both Omicron RBD-hACE2 complex and Delta RBD-hACE2 complex were colored by white and the corresponding crystal structures were labeled by salmon and magenta, respectively.

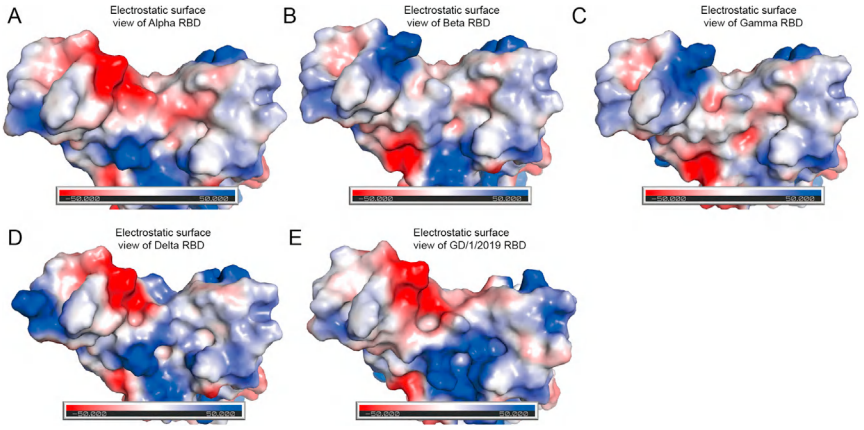


Figure S6. Electrostatic surface view, Related to Figure 4

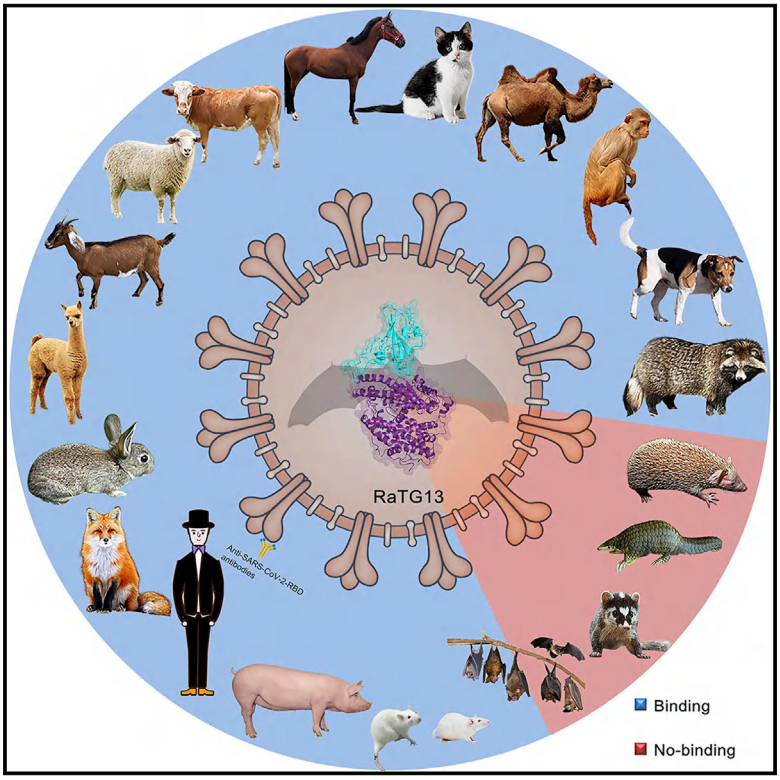
Electrostatic surface view of SARS-CoV-2 VOCs Alpha, Beta, Gamma and Delta RBDs and GD/1/2019 RBD. The first panel represents the top view. The others are yielded by rotation of the former panel along a horizontal axis.

Cell

Article

Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species

Graphical abstract



Highlights

- The complex structure of RaTG13 RBD with hACE2 was determined
- Binding of RaTG13 RBD to 24 additional ACE2 orthologs was evaluated
- Residue 501 plays a key role in determining the potential host range of RaTG13
- SARS-CoV-2 induces strong cross-protective antibodies to RaTG13 RBD

Authors

Kefang Liu, Xiaoqian Pan, Linjie Li, ..., Jianxun Qi, George F. Gao, Qihui Wang

Correspondence

jxqi@im.ac.cn (J.Q.),
gaof@im.ac.cn (G.F.G.),
wangqihui@im.ac.cn (Q.W.)

In brief

Structural and molecular analysis of the receptor binding domain of RaTG13, a coronavirus phylogenetically closely related to SARS-CoV-2, bound to the human receptor ACE2 as well as ACE2 orthologs in 24 other species provides a framework to understand its host range as well as the basis of antibody cross-reactivity between the two viruses.

Liu et al., 2021, Cell *184*, 3438–3451
June 24, 2021 © 2021 Elsevier Inc.
https://doi.org/10.1016/j.cell.2021.05.031

CellPress

CellPress

Cell

Article

Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species

Kefang Liu,[1,10] Xiaoqian Pan,[1,2,10] Linjie Li,[1,2,10] Feng Yu,[3,10] Anqi Zheng,[1] Pei Du,[1] Pengcheng Han,[1,4] Yumin Meng,[1] Yanfang Zhang,[1,5] Lili Wu,[1] Qian Chen,[1,6] Chunli Song,[6] Yunfei Jia,[1,7] Sheng Niu,[1,7] Dan Lu,[1] Chengpeng Qiao,[1] Zhihai Chen,[8] Dongli Ma,[9] Xiaopeng Ma,[9] Shuguang Tan,[1] Xin Zhao,[1] Jianxun Qi,[1,2,*] George F. Gao,[1,2,*] and Qihui Wang[1,2,11,*]

[1]CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

[2]University of the Chinese Academy of Sciences, Beijing 100049, China

[3]Shanghai Synchrotron Radiation Facility, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201204, China

[4]Department of Biomedical Engineering, Emory University, Atlanta, GA 10033, USA

[5]Laboratory of Protein Engineering and Vaccines, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

[6]Institute of Physical Science and Information, Anhui University, Hefei 230039, China

[7]College of Veterinary Medicine, Shanxi Agricultural University, Jinzhong 030801, China

[8]Center of Infectious Disease, Beijing Ditan Hospital, Capital Medical University, 100015 Beijing, China

[9]Shenzhen Children's Hospital, Shenzhen 518036, China

[10]These authors contributed equally

[11]Lead contact

*Correspondence: jxqi@im.ac.cn (J.Q.), gaof@im.ac.cn (G.F.G.), wangqihui@im.ac.cn (Q.W.)

https://doi.org/10.1016/j.cell.2021.05.031

SUMMARY

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been spreading worldwide, causing a global pandemic. Bat-origin RaTG13 is currently the most phylogenetically related virus. Here we obtained the complex structure of the RaTG13 receptor binding domain (RBD) with human ACE2 (hACE2) and evaluated binding of RaTG13 RBD to 24 additional ACE2 orthologs. By substituting residues in the RaTG13 RBD with their counterparts in the SARS-CoV-2 RBD, we found that residue 501, the major position found in variants of concern (VOCs) 501Y.V1/V2/V3, plays a key role in determining the potential host range of RaTG13. We also found that SARS-CoV-2 could induce strong cross-reactive antibodies to RaTG13 and identified a SARS-CoV-2 monoclonal antibody (mAb), CB6, that could cross-neutralize RaTG13 pseudovirus. These results elucidate the receptor binding and host adaption mechanisms of RaTG13 and emphasize the importance of continuous surveillance of coronaviruses (CoVs) carried by animal reservoirs to prevent another spillover of CoVs.

INTRODUCTION

Emerging and re-emerging pathogens threaten global public health and cause tremendous economic losses (Gao, 2018). The new coronavirus (CoV) severe acute respiratory syndrome CoV 2 (SARS-CoV-2) was detected and then isolated in early 2020 (Tan et al., 2020; The 2019-nCoV Outbreak Joint Field Epidemiology Investigation Team and Li, 2020; Wang et al., 2020a; Zhu et al., 2020). SARS-CoV-2 later spread worldwide, causing a global pandemic. As of March 5, 2021, the World Health Organization (WHO; https://covid19.who.int/) has recorded more than 114 million confirmed cases of CoV disease 2019 (COVID-19) globally and more than 2.5 million related deaths.

SARS-CoV-2 belongs to the family Coronaviridae, enveloped RNA viruses in the order Nidovirales (https://talk.ictvonline.org/taxonomy/?Redirected=true). CoVs are classified into four genera: alpha, beta, gamma, and delta CoVs. Bats are recognized as natural reservoirs of alpha and beta CoVs (Latinne et al., 2020). Increasing evidence indicates that bat CoVs (bCoVs) are the evolutionary sources of SARS-CoV (beta CoV), Middle East respiratory syndrome CoV (MERS-CoV; beta CoV), swine acute diarrhea syndrome CoV (SADS-CoV; alpha CoV), human CoV (HCoV)-OC43 (beta CoV), HCoV-NL63 (alpha CoV), and HCoV-229E (alpha CoV) (Corman et al., 2015; Gao, 2018; Huynh et al., 2012; Latinne et al., 2020; Lau et al., 2005; Li et al., 2005; Smith and Wang, 2013; Wang et al., 2018), most of which can infect humans and cause disease (Su et al., 2016). SARS-CoV and MERS-CoV have caused global outbreaks in the past two decades and confirmed cases of MERS-CoV infection are still growing in the Middle East region.

SARS-CoV-2 is closely related to several CoVs from bats (Chan et al., 2020; Liu et al., 2021; Malik et al., 2020). RaTG13,

 © 2021 Elsevier Inc.

sequenced from *Rhinolophus affinis* (intermediate horseshoe bat) in Yunnan province is the closest relative to SARS-CoV-2, with 96.2% identity of the overall genome sequence and 89.3% amino acid identity in the receptor binding domain (RBD) (Boni et al., 2020; Wu et al., 2020a; Zhou et al., 2020b). One variance of RaTG13 compared with SARS-CoV-2 is the absence of the four-residue (PRRA) insertions at the S1/S2 cleavage site on the spike (S) protein. However, such residue insertions can occur naturally in CoVs, which may contribute to RaTG13 evolving into the causative pathogen of the next potential pandemic. For example, RmYN02, another bCoV, detected in *Rhinolophus malayanus*, shares 93.3% nucleotide identity with the SARS-CoV-2 genome and contains three-amino-acid-residue (PAA) insertions at the S1/S2 cleavage site of its S protein, which is similar to that in SARS-CoV-2 (Zhou et al., 2020a). Considering the similarity between RaTG13 and SARS-CoV-2 and the devastation caused by SARS-CoV-2, RaTG13 may pose a significant threat to humans and other animals. To access this risk, the host range of RaTG13 needs to be investigated.

Receptor binding is the key step for a virus to achieve cross-species infection and transmission (Li et al., 2020b). The gain of function to interact with the receptor orthologs of another species is also a prerequisite for inter-species transmission (Su et al., 2016). Thus, evaluating the interactions between viral ligands and receptor orthologs of various animals may lead to identification of susceptible hosts. For SARS-CoV-2, angiotensin-converting enzyme 2 (ACE2) is required for human cell entry by binding to the SARS-CoV-2 RBD on the S protein (Lan et al., 2020; Shang et al., 2020; Wang et al., 2020b). We and other groups independently discovered ACE2 as the receptor and investigated the interactions between the SARS-CoV-2 RBD and ACE2 orthologs of other species. Our previous work revealed that SARS-CoV-2 might have a broad host range, including cats and dogs, which is consistent with the findings of epidemiological studies. Thus, quantifying the interactions between RaTG13 RBD and ACE2 orthologs of domestic and wild animals that are in close proximity to humans could provide clues for determining the host range of RaTG13.

Currently, the structures of SARS-CoV-2 RBD in complex with human ACE2 (hACE2), cat ACE2 (cACE2), bat ACE2 (bACE2), and pangolin ACE2 (pACE2) have been determined (Lan et al., 2020; Liu et al., 2021; Shang et al., 2020; Wang et al., 2020b; Wu et al., 2020b, 2021). These structures revealed that the SARS-CoV-2 RBD utilizes a similar binding mechanism with all four ACE2s, but it forms fewer H-bonds in the interactions with the latter three ACE2s than with hACE2, providing the molecular basis for the decreased binding affinity between the SARS-CoV-2 RBD and the three animal ACE2s. Previous work has reported that RaTG13 RBD binds to hACE2 with a much lower affinity than the SARS-CoV-2 RBD (Wrobel et al., 2020). However, the mechanism of interaction between the RaTG13 RBD and hACE2 remains unclear.

Here the complex structure of RaTG13 RBD with hACE2 was determined, and the underlying mechanism for the lower binding affinity of the RaTG13 RBD compared with that of the SARS-CoV-2 RBD to hACE2 was resolved. Moreover, the binding ability of 24 ACE2 orthologs, together with hACE2, to the RaTG13 RBD was quantified. We also evaluated the effects of six substitutions between the RaTG13 RBD and SARS-CoV-2 RBD on the interaction with hACE2 and orthologs. Specifically, 501 is most likely the key residue that determines the potential host range of RaTG13. Coincidentally, residue 501 is also the key mutation of the SARS-CoV-2 strains that have been spreading rapidly in many countries, including variant of concern (VOC) lineages 20I/501Y.V1 in the United Kingdom, 20H/501Y.V2 in South Africa, and 20J/501Y.V3 in Brazil and Japan, and also the key mutation of a mouse adaptive SARS-CoV-2 virus strain, indicating that mutation at this residue should be watched closely (Gu et al., 2020). Moreover, we also found that SARS-CoV-2 could induce a strong cross-reactive immune response to RaTG13 and identified a SARS-CoV-2 monoclonal antibody (mAb), CB6, that could cross-neutralize the RaTG13 pseudovirus. These results broaden our knowledge of the molecular mechanisms of RaTG13 RBD binding to hACE2 and the potential host adaption of RaTG13, providing evidence emphasizing the importance of continuous surveillance of CoVs carried by other potential animal reservoirs, monitoring RaTG13 and its related CoVs especially closely.

RESULTS

Complex structure of RaTG13 RBD with hACE2

Sequence alignments of the RaTG13 RBD and SARS-CoV-2 RBD were performed. Overall, the RBDs of RaTG13 and SARS-CoV-2 showed an 89.3% amino acid identity (Figure S1A). The complex structure indicates that the SARS-CoV-2 RBD forms 14 hydrogen bonds through 10 amino acids in its interaction with hACE2. Only 3 of 10 amino acids differed between the RaTG13 RBD and SARS-CoV-2 RBD (Figure S1A). However, the binding affinity of the RaTG13 RBD to hACE2 (K_D = 3.86 μM) was approximately 2-log lower than that of SARS-CoV-2 RBD to hACE2 (K_D = 21.1 nM) (Figures 1A and 1B). To investigate whether RaTG13 infects host cells by binding to hACE2, a vesicular stomatitis virus (VSV)-based pseudotyped RaTG13, which incorporates the S protein of RaTG13, was used to attempt to transduce HeLa cells expressing hACE2 (HeLa-hACE2s). A similar amount of SARS-CoV-2 pseudovirus (as determined by quantitative real-time PCR) was used as a positive control. The RaTG13 pseudovirus and SARS-CoV-2 pseudovirus were unable to transduce HeLa cells; however, both pseudoviruses were able to transduce HeLa-hACE2s, but the RaTG13 pseudovirus displayed a substantially lower transduction efficiency than the SARS-CoV-2 pseudovirus (Figure 1C).

To analyze the molecular mechanism behind this difference, the complex protein of RaTG13 RBD with hACE2 was prepared by co-infection of insect cells with baculoviruses expressing the RaTG13 RBD and hACE2 (Figure S1B). The stable complex was further purified and crystallized. Finally, we obtained diffraction data and determined the complex structure at a resolution of 3.1 Å (Table 1). There are two RBD and hACE2 complex molecules in one asymmetric unit. The electron densities of residues S19–A614 of hACE2 and residues T333–P527 of RaTG13 RBD were observed clearly. Five N-glycans linked to N53, N90, N103, N322, and N546 on hACE2 and one N-glycan linked to N343 of RaTG13 RBD were identified (Figure 1D).

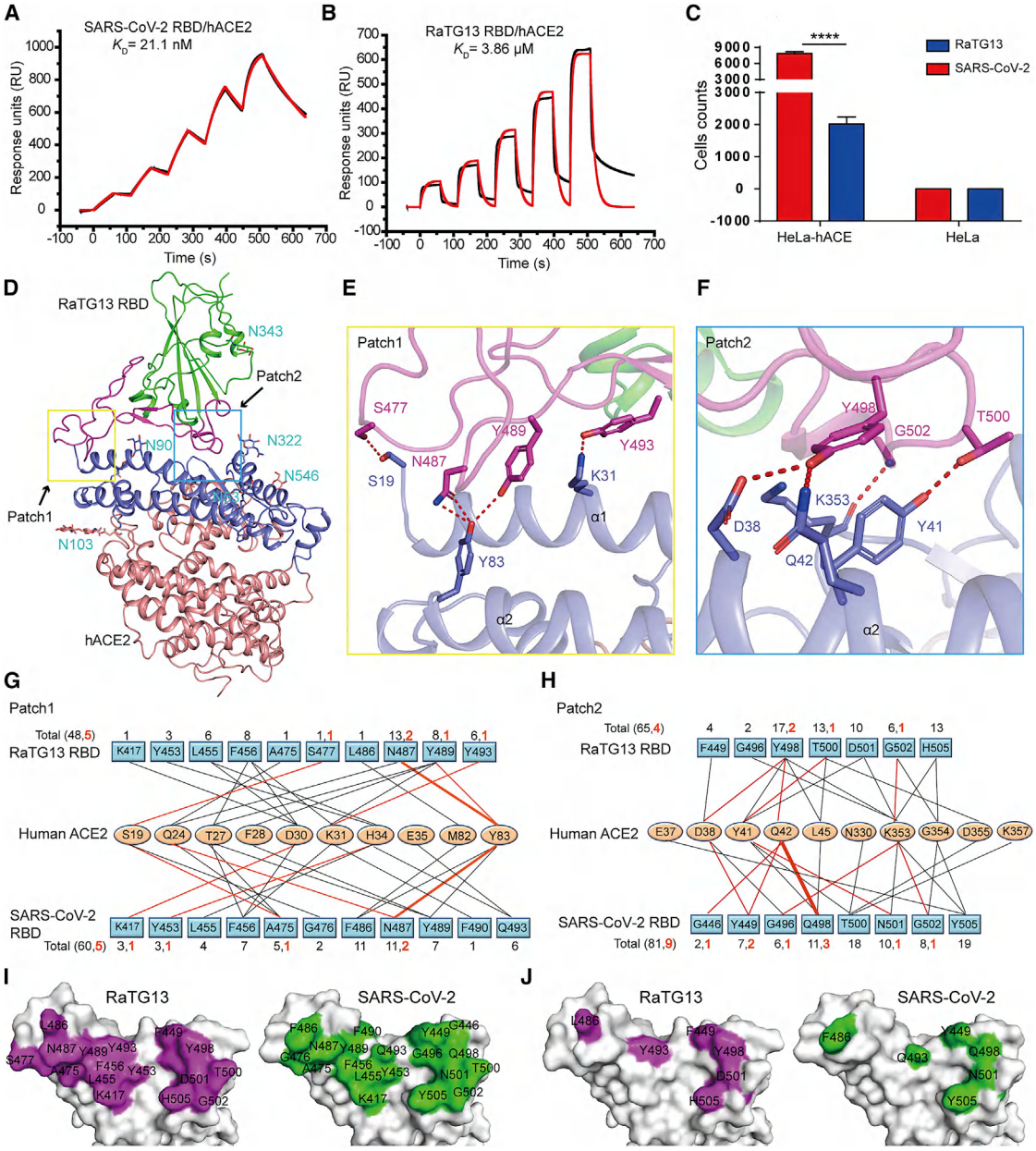


Figure 1. Structural basis of binding between the RaTG13 RBD and hACE2

(A and B) SPR characterization of the SARS-CoV-2 RBD (A) and RaTG13 RBD (B) interacting with hACE2.

(C) Entry of the SARS-CoV-2 and RaTG13 pseudovirus into HeLa cells expressing the hACE2 (HeLa-hACE2s). Data represent the mean ± SD of six replicates. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$ (two-tailed unpaired t test).

(D) Overall structure of the RaTG13 RBD in complex with hACE2. Boxes indicate the patches of the RaTG13 RBD/hACE2 complex.

(E and F) The hydrogen bond networks of patch 1 (E) and patch 2 (F). The complex structure is shown as a cartoon, and residues taking part in hydrogen bond formation are shown as sticks.

(G and H) Residues involved in the interaction of hACE2 with the RaTG13 RBD or SARS-CoV-2 RBD are listed and connected by solid lines. Black lines indicate vdw contacts, and red lines represent an H-bond or salt bridge.

(I) The binding surface of hACE2 with the RaTG13 RBD (left) or SARS-CoV-2 RBD (right).

(J) Six different residues between RaTG13 RBD and SARS-CoV-2 RBD among RaTG13 RBD binding sites are labeled.

See also Figure S1.

Table 1. Crystallographic data collection and refinement statistics

	SARS2-CoV-2 RBD/dACE2
Data collection	
Space group	$P2_1$
Cell dimensions	
a, *b*, *c* (Å)	81.28, 122.88, 110.74
α, β, γ (°)	90.00, 95.07, 90.00
Resolution (Å)	50.00–3.10 (3.21–3.10)
Unique reflections	39622 (3906)
Completeness (%)	99.7 (99.5)
R_{merge}	0.106 (1.310)
$I/\sigma I$	13.9 (1.1)
$CC_{1/2}$	0.985(0.685)
Redundancy	6.8 (6.9)
Refinement	
Resolution (Å)	48.94–3.10
No. of reflections	39,578
R_{work}/R_{free}	0.2260/0.2500
No. of atoms	
Protein	12,914
Ligand/ion	2
Water	0
B-factors	
Protein	118.8
Ligand/ion	219.8
Water	
RMSD	
Bond lengths (Å)	0.002
Bond angles (°)	0.478
Ramachandran statistics (%)	
Favored	96.44
Allowed	3.56
Disallowed	0.00

Values in parentheses are for highest-resolution shell.

Overall, the complex structure of the RaTG13 RBD with hACE2 is extremely similar to that of the SARS-CoV-2 RBD with hACE2 (Lan et al., 2020; Shang et al., 2020; Wang et al., 2020b; Yan et al., 2020), with a root-mean-square deviation (RMSD) of 1.012 Å. Similar to other beta CoV RBDs, RaTG13 RBD consists of a core subdomain and an external subdomain (Han et al., 2017). The external subdomain is dominated by a disulfide bond-stabilized flexible loop that connects two small β strands (Figure 1D). The binding interface of hACE2 with RaTG13 RBD is distributed into two patches. Patch 1 is located on the N-terminal α1 and α2 helixes, and Patch 2 is on the conformational surface consisting of the residues from α1 together with a β-hairpin constituted by β3 and β4. In patch 1, residues S19 and K31 from α1 and Y83 from α2 form a hydrogen bond network with residues S477, Y493, N487, and Y489 of the RaTG13 RBD (Figures 1E and 1G; Table 2). In patch 2, residues D38, Q42, and Y41 from α2 and K353 from the loop linking β3 and β4 form a hydrogen bond network with residues Y498, T500, and G502 (Figures 1F and 1H; Table 2).

Comparison of the complex structures of the SARS-CoV-2 RBD with hACE2 and the RaTG13 RBD with hACE2

The total contacts of the RaTG13 RBD with hACE2 (113 contacts) are significantly less than those of SARS-CoV-2 with the same receptor (141 contacts). Specifically, there are 48 contacts between the RaTG13 RBD and hACE2 in patch 1, which is fewer than those in the SARS-CoV-2 RBD with the hACE2 complex (60 contacts) (Figure 1G). Similarly, in patch 2, fewer contacts are formed between the RaTG13 RBD and hACE2 than between the SARS-CoV-2 RBD and hACE2 (65 versus 81) (Figure 1H). Moreover, there are substantially fewer hydrogen bonds in patch 2 of the RaTG13 RBD with hACE2 than of the SARS-CoV-2 RBD with the hACE2 complex (4 versus 9) (Figure 1H).

The buried surface area of the RaTG13 RBD bound to hACE2 is 1604 $Å^2$, which is smaller than that of the SARS-CoV-2 RBD with the hACE2 complex (1,773 $Å^2$). Specifically, residues G446 and F490 of the SARS-CoV-2 RBD contributed to contact with hACE2, whereas the corresponding residues of the RaTG13 RBD (G446 and Y490) did not (Figure 1I). On the binding surface with hACE2, the RaTG13 RBD displays six differences compared with the SARS-CoV-2 RBD: Y449F, F486L, Q493Y, Q498Y, N501D, and Y505H (Figure 1J).

The binding capacity of RaTG13 RBD to ACE2 orthologs from 24 different species

To evaluate the potential host range of RaTG13, we analyzed the sequence of hACE2 and of 24 other animals belonging to nine orders, including primates (human and monkey), Lagomorpha (rabbit), Rodentia (mouse and rat), Pholidota (Malayan pangolin), Carnivora (cat, civet, fox, dog, and raccoon dog), Perissodactyla (horse), Artiodactyla (pig, wild Bactrian camel, alpaca, bovine, goat, and sheep), Chiroptera (intermediate horseshoe bat, least horseshoe bat, little brown bat, fulvous fruit bat, greater horseshoe bat, and big-eared horseshoe bat), and Afrotheria (lesser hedgehog tenrec) (Figure S1C). The 16 key residues in hACE2 responsible for the interaction with the RaTG13 RBD were highlighted and compared with the 24 ACE2 orthologs. Sequence alignment showed a range of 0–9 residue substitutions in the ACE2 orthologs. Overall, residues distributed in patch 1 are more diverse than residues in patch 2 (Figure S1C). Specifically, residues F28 and D355 are fully conserved among all ACE2 orthologs compared in this study. Residues S19, Y41, Q42, L45, and G354 of hACE2 were substituted in no more than three ACE2 orthologs among the 24 species (Figure S1C).

We subsequently tested the binding of ACE2 orthologs with the RaTG13 RBD or SARS-CoV-2 RBD using eGFP-fused ACE2 orthologs expressed on the cell surface via flow cytometry. The MERS-CoV RBD was the negative control. The RaTG13 RBD interacted with cells expressing primate (human and monkey), Lagomorpha (rabbit), Carnivora (cat, fox, dog, and raccoon dog), Perissodactyla (horse), and Artiodactyla (pig, wild Bactrian camel, bovine, goat, and sheep) ACE2s (Figures 2A and S2). The SARS-CoV-2 RBD also interacted with

Table 2. Comparison of RaTG13 RBD and SARS-CoV-2 RBD binding to hACE2

RaTG13 RBD/SARS-CoV-2 RBD	hACE2 (RaTG13 RBD)	hACE2 (SARS-CoV-2 RBD)
K417	D30 (1)	D30 (3,**1**)
G446		Q42 (2,**1**)
F449/Y449	D38 (4)	D38 (5,**1**), Q42 (2,**1**)
Y453	H34 (3)	H34 (3,**1**)
L455	H34 (5), D30 (1)	H34 (4)
F456	T27 (5), D30 (2), K31 (1)	T27 (5), D30 (1), K31 (1)
A475	Q24 (1)	S19 (3,**1**), Q24 (1), T27 (1)
G476		S19 (2)
S477	S19 (1,**1**)	
L486/F486	M82 (1)	M82 (4), Y83 (7)
N487	Q24 (8), Y83 (5,**2**)	Q24 (7,**1**), Y83 (4,**1**)
Y489	T27 (2), F28 (3), K31(2), Y83 (1,**1**)	T27 (2), F28 (4), Y83 (1)
F490		K31 (1)
Y493/Q493	K31 (6,**1**)	H34 (3), E35 (3)
G496	K353 (2)	D38 (1), K353 (5,**1**)
Y498/Q498	Y41 (6), Q42 (3,**1**), D38 (5,**1**), K353 (2), L45 (1)	Y41(5), Q42(5,**3**), L45(1)
T500	Y41 (7,**1**), L45 (1), D355 (5)	Y41 (6), N330 (3), D355 (6), R357 (3)
D501/N501	Y41 (5), K353 (5)	Y41 (5,**1**), K353 (5)
G502	K353 (3,**1**), 354 (3)	K353 (3,**1**), G354 (5)
H505/Y505	K353 (11), G354 (2)	E37 (3), K353 (14), G354 (2)
Total	113, 9	141, 14

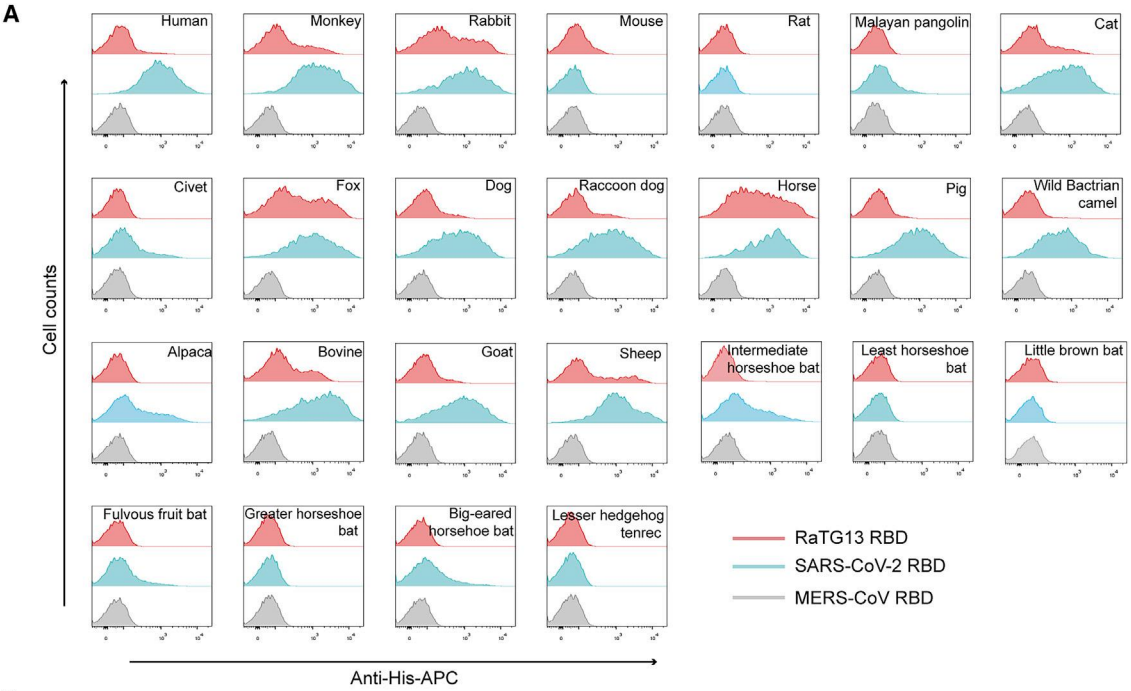
The numbers without underline in parentheses of hACE2 (RaTG13 RBD) and hACE2 (SARS-CoV-2 RBD) residues represent the number of vdw contacts between the indicated hACE2 residues with the RaTG13 RBD or SARS-CoV-2 RBD. Underlined numbers with bold format suggest numbers of potential H-bonds between the pairs of residues. vdw contact was analyzed at a cutoff of 4 Å and H-bonds at a cutoff of 3.3 Å.

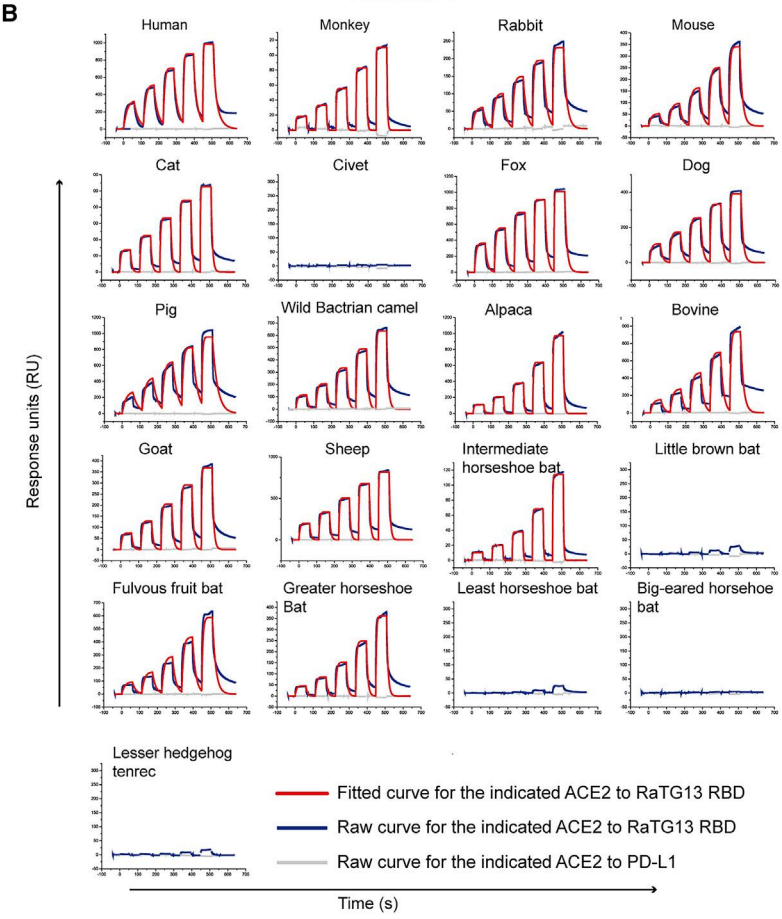
the abovementioned ACE2s, albeit with a higher fluorescence shift. Additionally, the SARS-CoV-2 RBD also interacted with Malayan pangolin, alpaca, intermediate horseshoe bat, and big-eared horseshoe bat ACE2s, as reported previously, whereas the RaTG13 RBD does not (Figures 2A and S2). Another differing binding feature was that the RaTG13 RBD, but not the SARS-CoV-2 RBD, showed positive binding to mouse ACE2 (Figures 2A and S2). No detectable interactions were observed between either RBDs with rat, least horseshoe bat, little brown bat, fulvous fruit bat, greater horseshoe bat, or lesser hedgehog tenrec ACE2 orthologs (Figures 2A and S2).

To clarify the interactions between the RaTG13 RBD and ACE2 orthologs from different animals, we quantified their binding affinities via surface plasmon resonance (SPR) assay. The RaTG13 RBD was capable of binding to human, monkey, rabbit, cat, fox, dog, raccoon dog, horse, pig, bovine, wild Bactrian camel, goat, sheep, and mouse ACE2, with equilibrium dissociation constant (K_D) values of 0.5–7.14 μM, but were unable to interact with Malayan pangolin, civet, little brown bat, least horseshoe bat, big-eared horseshoe bat, or lesser hedgehog tenrec ACE2s (Figure 2B). Notably, the RaTG13 RBD displayed an ~5-fold stronger affinity to horse ACE2 than to hACE2 (0.5 ± 0.12 μM versus 2.7 ± 1.27 μM) (Figure 2B). We performed a BLAST search to identify the 10 ACE2 orthologs that are most phylogenetically related to horse ACE2 and observed that most residues interacting with the RaTG13 RBD are conserved (Figure S3). Three of these ten ACE2 orthologs, wild Bactrian camel, cat, and alpaca ACE2, were included here, all of which showed binding to the RaTG13 RBD (Figure 2B). Although flow cytometry results showed no interactions between RaTG13 RBD and ACE2 orthologs of rat, alpaca, intermediate horseshoe bat, fulvous fruit bat, or greater horseshoe bat (Figures 2A and S2), SPR results showed RaTG13 can bind to these ACE2s, with K_D values of 11.42–38.63 μM) (Figures 2A and 2B).

Distinctive binding sites between the RaTG13 RBD and SARS-CoV-2 RBD

To evaluate the roles of six substitutions between the SARS-CoV-2 RBD and RaTG13 RBD in binding to the human receptor, we mutated the residues in the RaTG13 RBD to those of their SARS-CoV-2 RBD counterparts. Thus, we prepared six RaTG13 RBD mutants containing F449Y, L486F, Y493Q, Y498Q, D501N, and H505Y, respectively. Another mutant, called the 6 mutant, where all six RaTG13 residues were exchanged for their SARS-CoV-2 RBD counterparts, was also prepared. We evaluated their ability to interact with hACE2. Compared with the wild-type RaTG13 RBD, the F449Y, L486F, D501N, and H505Y mutants showed stronger interactions with hACE2, with binding affinity increased by about ~3.5-, ~3.1-, ~8.0-, and ~4.0-fold respectively, whereas the Y493Q mutation substantially decreased the binding affinity to hACE2, and the Y498Q mutation abolished the interaction with hACE2 (Figures 3A, 3B, and S4). More importantly, the binding affinity of the 6 mutant to human, mouse, horse, and intermediate horseshoe bat





Species	Binding affinity (μM)
Human	2.7 ± 1.27
Monkey	2.18 ± 1.32
Rabbit	1.81 ± 0.03
Mouse	4.85 ± 0.72
Rat	11.42 ± 3.68
Malayan pangolin	--
Cat	3.65 ± 0.22
Civet	--
Fox	1.2 ± 0.15
Dog	4.36 ± 1.16
Raccoon dog	2.24 ± 0.53
Horse	0.5 ± 0.12
Pig	4.03 ± 1.26
Wild bactrian camel	7.14 ± 0.75
Alpaca	38.63 ± 3.5
Bovine	1.32 ± 0.73
Goat	6.15 ± 4.38
Sheep	3.85 ± 0.52
Intermediate horseshoe bat	26.9 ± 9.07
Little brown bat	--
Fulvous fruit bat	11.39 ± 5.26
Greater horseshoe bat	21.9 ± 6.41
Least horseshoe bat	--
Big-eared horseshoe bat	--
Lesser hedgehog tenrec	--

(legend on next page)

ACE2 orthologs showed similar binding characteristics to those of the SARS-CoV-2 RBD (Figure 3B). The pseudovirus infection assay also indicated that the RaTG13 pseudovirus integrating the S protein with the 6 mutant increased transduction to HeLa-hACE2 cells, although this was still significantly weaker than the SARS-CoV-2 pseudotyped virus (Figures 3C and 3D).

Structural analysis revealed the basis of why the RaTG13 RBD has a lower binding affinity with hACE2 compared with that of the SARS-CoV-2 RBD. Some differing residues on the RaTG13 RBD form fewer contacts than their SARS-CoV-2 RBD counterparts, including F449, L486, D501, and H505. Specifically, RaTG13 RBD F449 contacts the D38 of hACE2 through van der Waals (vdw) force (4 contacts), whereas Y449 of the SARS-CoV-2 RBD forms stronger interactions with two potential hydrogen bonds (with D38 and Q42 of hACE2) and seven vdw contacts (Figure 4A). Moreover, SARS-CoV-2 RBD F486 closely packs against the F28, L79, M82, and Y83 residues of hACE2 and forms a small patch of hydrophobic interactions at the interface, and N501 forms a hydrogen bond with Y41 of hACE2. In contrast, RaTG13 RBD L486 interacts with M82 through only one vdw contact (Figure 4B), and D501 contacts Y41 through vdw contacts (Figure 4C). H505 of the RaTG13 RBD and Y505 of SARS-CoV-2 are in contact with K353 and G354, but SARS-CoV-2 Y505 also binds to an extra E37 on hACE2 (Figure 4D).

Some other substituted residues on RaTG13 RBD form weaker interactions compared with their SARS-CoV-2 RBD counterparts. RaTG13 RBD Y493 interacts with hACE2 K31 through a hydrogen bond, whereas SARS-CoV-2 RBD Q493 only interacts with hACE2 H34 and E35 through vdw interactions because of different conformations (Figure 4E). Additionally, although RaTG13 RBD Y498 and SARS-CoV-2 RBD Q498 are able to form two potential hydrogen bonds with hACE2 D38 and Q42, RaTG13 RBD Y498 is 0.1 Å closer to hACE2 Q42 compared with SARS-CoV-2 RBD Q498. Thus, substitution from Y498 to Q498 potentially weakened the interactions between RaTG13 RBD Y498 and hACE2, explaining the inability of the RaTG13 RBD Y498Q mutant to interact with the human receptor (Figure 4F).

Key amino acids of RaTG13 and SARS-CoV-2 RBDs for binding to different ACE2s and the role of residue 501

To further evaluate changes in binding affinity because of the six RaTG13 mutants, we chose the ACE2 orthologs from horse, mouse, and intermediate horseshoe bat for in-depth analysis. Horse ACE2 was chosen because it binds to the RaTG13 RBD with a higher affinity than hACE2, whereas mouse ACE2 was chosen for its distinct binding feature to the RaTG13 RBD and SARS-CoV-2 RBD. As the origin of RaTG13, intermediate horseshoe bat ACE2 was also included in this analysis.

Our results showed a broad spectrum of binding abilities to these ACE2 orthologs among the RaTG13 RBD mutants. Specifically, the H505Y mutant increased the binding affinity with the three chosen ACE2s by ~4.3-fold (mouse ACE2), ~4.6-fold (horse ACE2), and ~7.3-fold (intermediate horseshoe bat), respectively (Table S3). The D501N mutant displayed increased binding affinities to the ACE2s of human and intermediate horseshoe bat by 1 and 2 orders of magnitude, respectively, and decreased binding affinity with mouse ACE2 by ~14-fold (Table S3). The L486F mutant also displayed decreased interaction with mouse ACE2 (~1.8-fold) but favored the interaction with hACE2 (Table S3). Unlike D501N and L486F, which increased adaptation to the human ACE2 and decreased affinity for the mouse ACE2, F449Y showed benefit to the association to hACE2 (~2.7-fold) and mouse ACE2 (~2.5-fold) but abolished binding to its natural host ACE2 (Table S3; Figures 3A, 3B, and S4). The Y493Q mutant could not interact with ACE2s from mouse or intermediate horseshoe bat and displayed substantial decreased interaction with human ACE2 (~3.7-fold) but favored association with horse ACE2 (2.1-fold increase) (Figures 3A, 3B, and S4). The Y498Q mutant maintained binding to horse ACE2 with substantially decreased binding affinity (~13.6-fold), whereas the other four mutants (F449Y, L486F, Y493Q, and D501N) had little effect on the interaction with the horse receptor, with the K_D fluctuating within a ~3-fold range (Figures 3A, 3B, and S4).

Differential amino acids on ACE2 orthologs may also have a significant effect on the interaction with viral RBDs. The complex structure revealed that the Y41 and K353 residues of hACE2 mediate the contacts with RaTG13 RBD D501. However, Y41 is relatively conserved among the 25 ACE2s analyzed, suggesting that it may not play an important role in determining binding affinity. In contrast, K353 is present in the majority of the ACE2 orthologs, but not in mouse and rat ACE2, where it has been substituted with H353. Thus, considering that K353 of hACE2 is also a key residue for binding with the SARS-CoV-2 RBD and that SARS-CoV-2 is unable to infect mice and rats, we introduced K353H into hACE2 to investigate the relevance for the ability of the SARS-CoV-2 RBD to bind to mouse and rat ACE2s. Through flow cytometry analysis, we found that the SARS-CoV-2 RBD (containing N501) is more likely to bind to cells expressing hACE2 (containing K353) compared with cells expressing the hACE2 K353H mutant. In contrast, RaTG13 RBD (containing D501) binds preferentially to the K353H mutant of hACE2 (Figure 4G). Introducing D501N to RaTG13 RBD shifted the binding preference of the parental RBD from H353 to K353 (Figure 4G). These results indicate that RaTG13 RBD D501 and SARS-CoV-2 RBD N501 favor binding with H353 and K353, respectively, suggesting that RaTG13 RBD D501

Figure 2. Binding between ACE2 orthologs and the SARS-CoV-2 RBD or RaTG13 RBD

(A) Flow cytometry analysis of binding between 25 ACE2 orthologs and the SARS-CoV-2 RBD or RaTG13 RBD. HEK293T cells expressing EGFP-fused ACE2s are stained with the indicated His-tagged proteins (RaTG13 RBD, SARS-CoV-2 RBD, and MERS-CoV RBD). An anti-His/Allophycocyanin (APC) antibody is used to detect His-tagged proteins. The MERS-CoV RBD is used as a negative control.

(B) SPR characterizations of the binding between 25 ACE2 orthologs and the RaTG13 RBD. ACE2s with a mouse Fc (mFc) tag are immobilized on a CM5 chip. SPR characterizations of the binding affinity between the RaTG13 RBD and each ACE2 ortholog are shown. PD-L1 is used as a negative control. Raw and fitted curves are displayed as blue and red lines, respectively. Data represent the mean ± SD of three independent experiments.

See also Figures S2 and S3 and Tables S1 and S2.

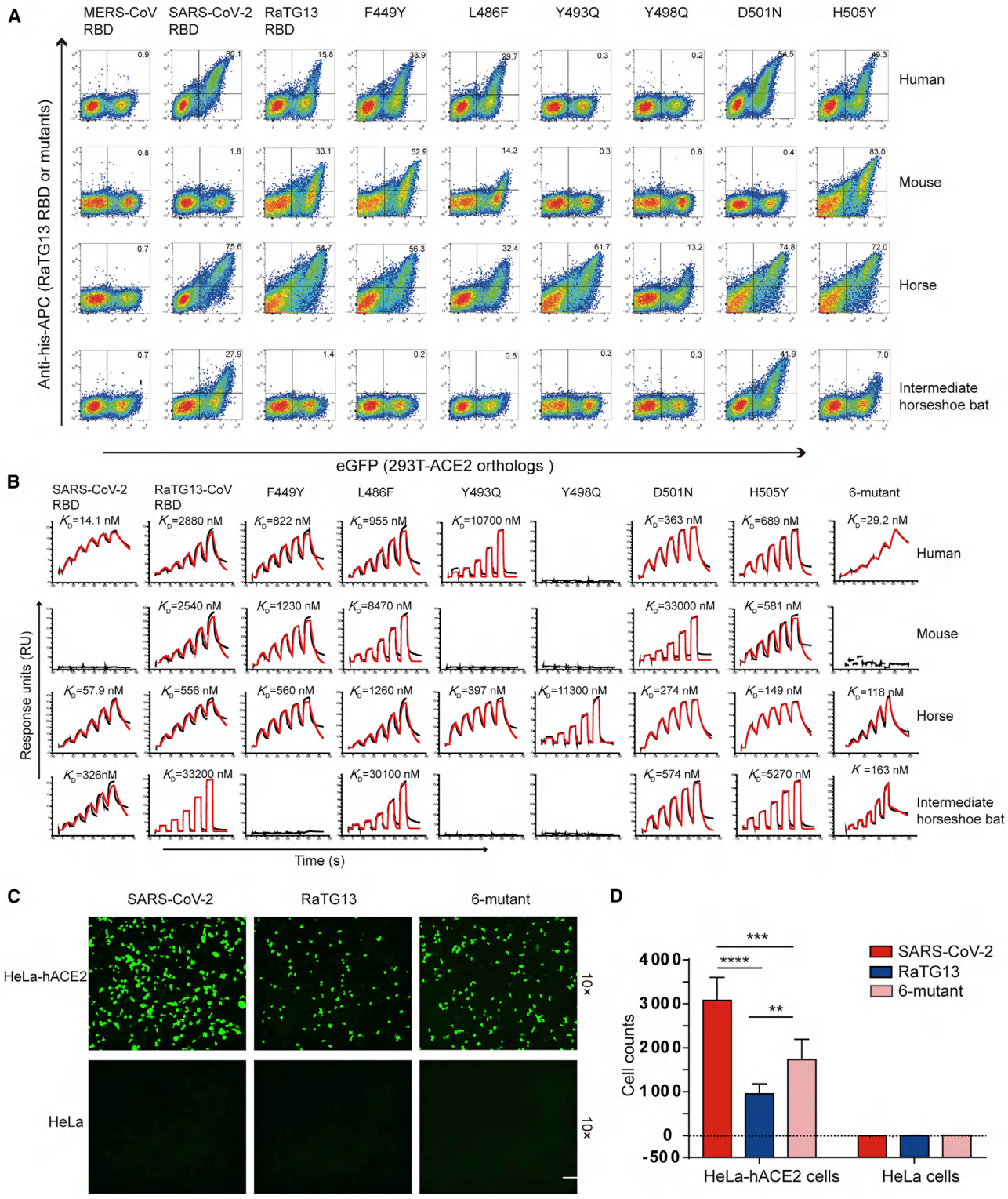


Figure 3. Mutational analysis of the key residues in RaTG13 involved in interaction with hACE2

(A) Flow cytometry analysis of binding between the six RaTG13 RBD mutants and ACE2s from human, mouse, horse, or intermediate horseshoe bat. SARS-CoV-2 RBD and MERS-CoV RBD are used as control.

(B) SPR analysis of the binding affinity between wild-type/mutated RaTG13 RBD with ACE2s from human, mouse, horse, or intermediate horseshoe bat, respectively. Raw and fitted curves are displayed as blue and red lines, respectively.

(legend continued on next page)

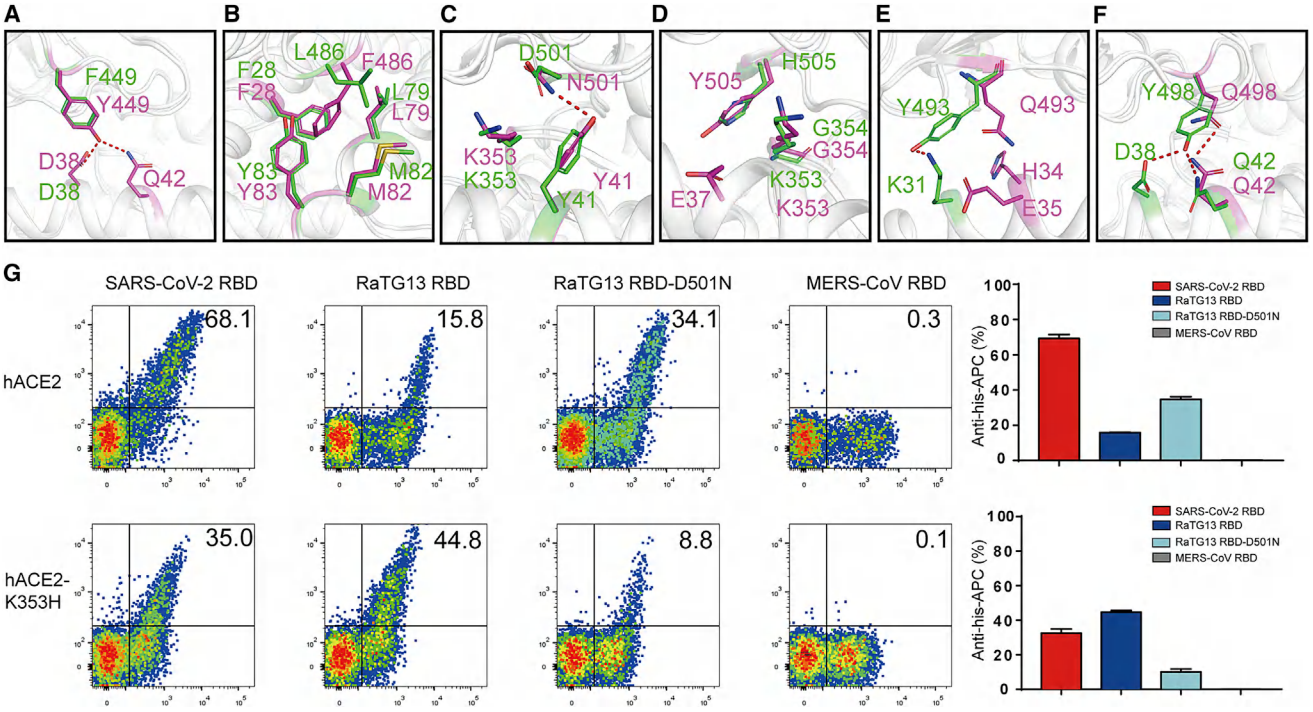


Figure 4. Structural and functional analysis of the role of each key residue between ACE2 orthologs and the RaTG13 RBD or SARS-CoV-2 RBD

(A–F) Structural comparison of the binding between hACE2 and the RaTG13 RBD or SARS-CoV-2 RBD. Substituted residues on the RaG13 RBD and SARS-CoV-2 RBD are shown in green and purple, respectively. Key residues are shown as sticks with corresponding colors.

(G) Flow cytometric analysis of the effect of D501 and N501 of the RaTG13 RBD and K353 and H353 of hACE2. Bar graphs indicate the ratio of anti-His-positive cells in HEK293T cells with hACE2 expression.

plays a key role in the ability to bind to mouse and rat ACE2s and providing an explanation for the inability of SARS-CoV-2 to interact with mouse and rat ACE2s.

The cross-reactive immune response of SARS-CoV-2 to RaTG13

To assess the cross-reactive immune response of SARS-CoV-2 to RaTG13, we tested binding of the RaTG13 RBD with six COVID-19 convalescent sera by enzyme-linked immunosorbent assay (ELISA). The SARS-CoV-2 antibodies in all six convalescent donors' sera could cross-recognize the RaTG13 RBD with an endpoint titer of 10^3–10^4, which was slightly lower than that of the SARS-CoV-2 RBD (Figures 5A and 5B). In contrast, sera from three healthy donors showed no binding to the RaTG13 RBD or SARS-CoV-2 RBD (Figures 5A and 5B). We then evaluated the neutralization activity of three COVID-19 convalescent sera against SARS-CoV-2 and RaTG13 pseudoviruses. Interestingly, although all three sera showed effective neutralization against the RaTG13 pseudovirus, only one completely neutralized the SARS-CoV-2 pseudovirus, whereas the other two only partially neutralized the SARS-CoV-2 pseudovirus at the highest concentration (Figure 5C).

To assess the ability of SARS-CoV-2 RBD mAbs to cross-react with the RaTG13 RBD, we tested the binding affinity of six SARS-CoV-2 RBD mAbs (REGN10933, REGN10987, C110, H4, B38, CB6, and S2H14) with the RaTG13 RBD (Barnes et al., 2020; Hansen et al., 2020; Piccoli et al., 2020; Shi et al., 2020; Wu et al., 2020c) to the RaTG13 RBD. Our data showed that CB6 binds to the RaTG13 RBD with the highest binding affinity (39.1 nM) among these six mAbs (Figure 5D). The binding affinities of REGN10933, REGN10987, C110, and B38 with the RaTG13 RBD were decreased 100- to 10,000-fold compared with their corresponding affinities for the SARS-CoV-2 RBD (Figure 5D). H4 and S2H14, which bind to the SARS-CoV-2 RBD with high affinities (1.76 nM and 4.12 nM, respectively), lost the ability to interact with the RaTG13 RBD (Figure 5D). Flow cytometry analysis indicates that, in the presence of the CB6 antibody, the numbers of events in which hACE2-expressing cells bind to SARS-CoV-2 or RaTG13 were decreased substantially in a CB6 concentration-dependent manner (Figures 5E and S5).

(C) Entry of the pseudovirus of SARS-CoV-2, RaTG13, and 6-mutant RaTG13 into HeLa-hACE2s. Green fluorescent HeLa-hACE2s indicate pseudovirus-transducing-cells. Untransfected HeLa cells were used as negative controls. The scale bar indicates 150 μm.

(D) Statistics for transduction of the SARS-CoV-2, RaTG13, and 6-mutant RaTG13 pseudoviruses. Data represent the results of six replicates. All data are presented as mean ± SD. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$ (two-tailed unpaired t test).

See also Figure S4 and Table S3.

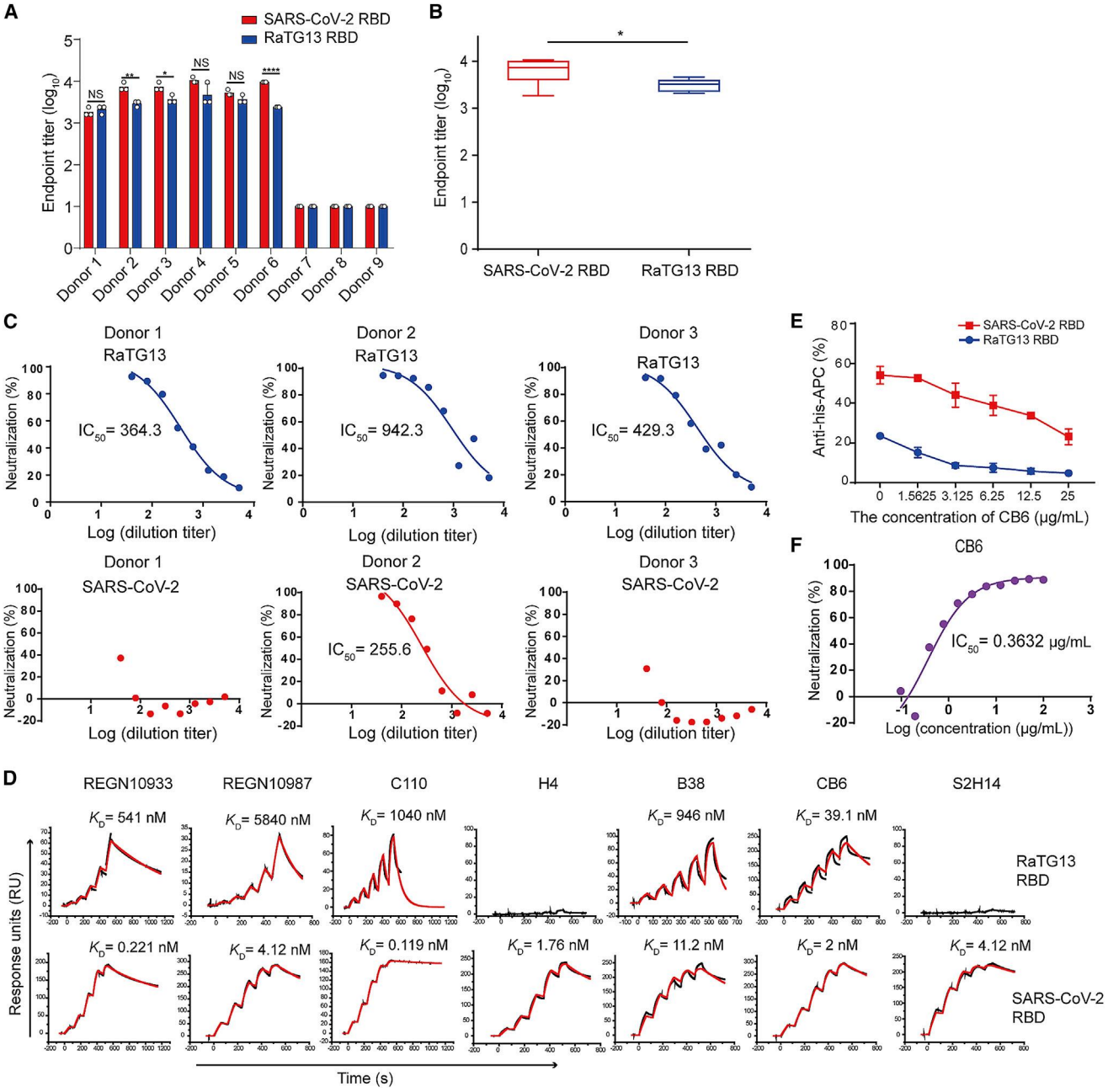


Figure 5. The cross-reactive immune response of SARS-CoV-2 to RaTG13

(A) Enzyme-linked immunosorbent assay (ELISA) measurement of the titers of SARS-CoV-2 RBD- and RaTG13 RBD-specific immunoglobulin G (IgG) in serum samples.

(B) Statistic of the titers of SARS-CoV-2 RBD- and RaTG13 RBD-specific IgG in six convalescent individuals' serum samples. Data are presented as mean ± SD. *$p < 0.05$ (two-tailed unpaired t test).

(C) Neutralization of the RaTG13 or SARS-CoV-2 pseudovirus by antisera from three convalescent individuals.

(D) SPR characterization of the binding affinities of the SARS-CoV-2 RBD or RaTG13 RBD with the indicated antibodies.

(E) The ratio of SARS-CoV-2 RBD- and RaTG13 RBD-binding cells among Baby Hamster Syrian Kidney cells with hACE2 expression (BHK-hACE2 cells) in the presence of CB6. Data represent the results of three replicates. Data are presented as mean ± SD.

(F) Neutralization of the RaTG13 or SARS-CoV-2 pseudovirus by the CB6 antibody.

See also Figure S5 and Table S4.

Finally, by applying CB6 to the RaTG13 pseudovirus, we found that CB6 could effectively neutralize the RaTG13 pseudovirus, with a half maximal inhibitory concentration (IC_{50}) of 0.3632 μg/mL (Figure 5F).

DISCUSSION

Increasing studies indicate that SARS-CoV-2 shares a common ancestor with bCoVs, and bats are suspected to be the reservoir host. So far, RaTG13, sequenced from bats, is the virus most closely related to SARS-CoV-2 (Zhou et al., 2020b). Quantification of the association between RaTG13 and its human receptor and those of wild and domestic animals is critical for evaluating the potential inter-species transmission of RaTG13 to control and prevent related infectious diseases, especially in humans (Su et al., 2016).

The binding affinity between the RaTG13 RBD and hACE2 is approximately 70-fold lower than that between the SARS-CoV-2 RBD and hACE2 because of the substantially reduced area of the buried surface and decreased network of hydrogen bonds in the RaTG13 RBD-hACE2 complex structure. In addition, substituting six residues on the RaTG13 RBD (F449Y, L486F, Y493Q, Y498Q, D501N, and H505Y) with their counterparts on the SARS-CoV-2 RBD rescued the binding affinity of RaTG13 with hACE2. Such observation is consistent with a recent report that identified the importance of these RaTG13 mutations for binding to hACE2 (Conceicao et al., 2020). Here we provide substantially more data indicating the effect of these mutations on binding with ACE2s from mouse, horse, and intermediate horseshoe bat.

The cryoelectron microscopy (cryo-EM) structure of the RaTG13 S protein was reported recently, and through structural alignments of the RaTG13 S protein and SARS-CoV-2 S protein, RaTG13 RBD Y493 was speculated to confer a potential steric clash to hACE2 (Wrobel et al., 2020). However, the structure of the RaTG13 RBD in complex with hACE2 reported here indicates that RaTG13 RBD Y493 forms a hydrogen bond with hACE2 K31.

Most of the 24 ACE2 orthologs have weaker binding affinity for the RaTG13 RBD than the SARS-CoV-2 RBD, with ACE2s from mouse and rat as two exceptions. Through analysis of mutations, we identified D501 on RaTG13 as the key residue to bind to ACE2s of mouse and rat, which also explains why SARS-CoV-2 (N501) is unable to interact with mouse and rat ACE2s. N501Y was identified in a mouse adaptive SARS-CoV-2 virus strain that enhances the interaction between SARS-CoV-2 RBD and mouse ACE2 (Gu et al., 2020). SARS-CoV-2 carrying the N501Y mutation has been spreading widely in many countries with different mutations, including VOCs 20I/501Y.V1 (United Kingdom), 20H/501Y.V2 (South Africa), and 20H/501Y.V3 (Brazil/Japan), indicating that this mutation is the "hotspot" for host adaptation. It is believed that human infection with novel CoVs is related to a "hot jump" of the viruses from reservoir (bat)-intermediate-host. Our work here with bat-origin RaTG13 further confirms that residue 501 is the key amino acid position in the S protein for bat-intermediate-host adaptation, which implicates that RaTG13 has the potential for interspecies transmission.

Among the ACE2 orthologs from humans and 24 animals, horse ACE2 has the highest binding affinity with the RaTG13 RBD even after introduction of the Y498Q substitution on the RaTG13 RBD, which abolishes binding with ACE2s from human, mouse, and intermediate horseshoe bat. Our previous study indicates that horse ACE2 also possesses high binding affinity for the SARS-CoV-2 RBD (Wu et al., 2020b). However, so far, no evidence indicating susceptibility of horse ACE2 to SARS-CoV-2 has been reported. Other host factors may evolve, and further studies are needed. Because non-SARS-CoV-2 beta CoVs are widespread in horses (Haake et al., 2020), which could potentially result in recombination between CoVs, horses and other species with similar ACE2s, especially ones in close proximity to bats, should be monitored closely to prevent inter-species CoV transmission, which could ultimately result in human infection.

Interestingly, our data, together with previous reports, indicate that RaTG13 possesses different binding spectra to ACE2 orthologs compared with those of SARS-CoV-2, suggesting that RaTG13 has undergone diverse evolutionary routes compared with SARS-CoV-2. For example, although civets and pangolins have been proposed to play roles in transmission of SARS-CoV and SARS-CoV-2, respectively (Guan et al., 2003; Lam et al., 2020; Xiao et al., 2020), our data showed that the ACE2 orthologs of these species do not interact with the RaTG13 RBD. Our previous work showed that Rodentia (mouse and rat) ACE2s were unable to bind to the SARS-CoV-2 RBD (Wu et al., 2020b), whereas the data presented here indicate that they can bind to the RaTG13 RBD.

The intermediate horseshoe bat is the host from which RaTG13 was originally sequenced, but its ACE2 only binds weakly to the RaTG13 RBD. Interestingly, a similar phenomenon has been observed in other CoVs, including SARS-CoV (Wu et al., 2020b), SARS-related CoVs (Guo et al., 2020), and MERS-CoV (Yuan et al., 2020), possibly because of the unique immune system of bats. Specifically, some pivotal receptors for natural killer (NK) cell activation are not expressed in bat NK cells, and the signaling pathway of type I interferons (IFNs) is suppressed by reduction of the stimulator of IFN genes (STING) pathway in bats (Pavlovich et al., 2018; Xie et al., 2018; Zhang et al., 2013). Besides the suppressed innate immune response, bats also have a special adaptive immune system, exemplified by the unusual peptide presentation features brought by specific 3-amino-acid insertion in major histocompatibility complex (MHC) class I (Lu et al., 2019). In addition, polymorphisms of *Rhinolophus sinicus* ACE2 confer varied susceptibility to SARS-related CoV infection (Guo et al., 2020). The polymorphism of ACE2 may also exist in intermediate horseshoe bats, which show different binding capacities to the RaTG13 RBD.

RaTG13 has a broad host range, which suggests that acquisition of other adaptive mutations can further broaden its host range. Thus, RaTG13 and related CoVs can possibly jump the species barrier to humans and have pandemic potential when they acquire adaptive mutations required for human infection. Evaluating the cross-reactive immune response of SARS-CoV-2 to RaTG13 revealed that serum samples from SARS-CoV-2 convalescent donors could more efficiently cross-neutralize the RaTG13 pseudovirus than the SARS-CoV-2 pseudovirus,

probably because of the lower binding affinity between hACE2 and the RaTG13 RBD than the SARS-CoV-2 RBD. Thus, lower concentrations of neutralizing antibodies are required to block binding of the receptor to the RaTG13 RBD than to the SARS-CoV-2 RBD; therefore, fewer neutralizing antibodies are needed to neutralize the RaTG13 pseudovirus than the SARS-CoV-2 pseudovirus. Moreover, our studies using SARS-CoV-2 RBD monoclonal antibodies show that many of these antibodies interact with RaTG13 with a 10^3-fold reduced binding affinity. One exception with high binding affinity with RaTG13 is CB6, a SARS-CoV-2 mAb identified previously by our group that could bind to RaTG13 with high affinity and neutralize RaTG13 pseudovirus infection of HeLa-hACE2s.

Limitations of the study

Our approach for analyzing the host range of RaTG13 by studying the interaction between the RaTG13 RBD and ACE2 is limited because we only examine the binding affinity for the ACE2 receptor. RaTG13 may potentially infect cells through other receptors besides ACE2. Further studies are required to identify whether RaTG13 utilizes other receptors for cell entry. Cell entry of SARS-CoV-2 also requires S protein priming by cellular proteases, such as transmembrane protease serine 2 (TMPRSS2) and furin (Hoffmann et al., 2020), which should be addressed in the future studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cells
  - Sera Samples
- METHOD DETAILS
  - Gene Cloning
  - Protein Expression and Purification
  - Flow Cytometry Assay
  - SPR Analysis
  - Enzyme-linked Immunosorbent Assay
  - Production and Quantification of Pseudoviruses
  - Pseudovirus Infection Assay
  - Neutralization Assay
  - Crystallization
  - Data Collection and Structure Determination
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Binding Studies
  - Flow Cytometry Analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2021.05.031.

ACKNOWLEDGMENTS

We are grateful to William Ningda Gao for revising the manuscript and Hillary Yida Zhang for constructing the RaTG13 6-mutant plasmid. We are grateful to the Pathogenic Microbiology and Immunology Public Technology Service Center for its support regarding the flow cytometry assay. We acknowledge the staff of beamline BL19U1 at the Shanghai Synchrotron Radiation Facility for assistance during data collection. We also thank Y. Chen, B. Zhou, and Z. Yang from the Institute of Biophysics, Chinese Academy of Sciences, for technical support with the SPR assay. We acknowledge Z. Shi from Wuhan Institute of Virology, Chinese Academy of Sciences, for providing the ACE2 sequence of *Rhinolophus affinis*. This work was supported by the Ministry of Science and Technology of the People's Republic of China; the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB29010202, XDB29040203, and XDB37030204); an intramural special grant for SARS-CoV-2 research from the Chinese Academy of Sciences, National Natural Science Foundation of China (81922044), and the National Science and Technology Major Project (2018ZX10101004-001 and 2018ZX09711003-002-001). Q.W. is supported by the Youth Innovation Promotion Association of the CAS (2018119). G.F.G is supported by the Yanqi Lake Meeting organized by the academic divisions of the CAS.

AUTHOR CONTRIBUTIONS

J.Q., G.F.G., and Q.W. initiated and designed the project. K.L. and X.P. purified the proteins and grew the crystals with help from Y.M., Y.Z., L.W., and C.S. X.P. and L.L. performed the SPR analysis with help from Y.J., S.N., and C.Q. K.L., X.P., and L.L. conducted the flow cytometry assay with help from A.Z., Q.C., and L.W. F.Y. and J.Q. collected the structural data and solved the structures. K.L., F.Y., S.T., and J.Q. analyzed the data and prepared the figures with help from D.M. and X.M. A.Z., P.H., and X.Z. performed pseudovirus-related assays. D.L. and Z.C. collected the serum samples. K.L., P.D., Q.W., and G.F.G. wrote and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 7, 2021
Revised: March 8, 2021
Accepted: May 19, 2021
Published: May 24, 2021

REFERENCES

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. *66*, 213–221.

Barnes, C.O., Jette, C.A., Abernathy, M.E., Dam, K.A., Esswein, S.R., Gristick, H.B., Malyutin, A.G., Sharaf, N.G., Huey-Tubman, K.E., Lee, Y.E., et al. (2020). SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. Nature *588*, 682–687.

Boni, M.F., Lemey, P., Jiang, X., Lam, T.T., Perry, B.W., Castoe, T.A., Rambaut, A., and Robertson, D.L. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. *5*, 1408–1417.

Chan, J.F., Kok, K.H., Zhu, Z., Chu, H., To, K.K., Yuan, S., and Yuen, K.Y. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg. Microbes Infect. *9*, 221–236.

Conceicao, C., Thakur, N., Human, S., Kelly, J.T., Logan, L., Bialy, D., Bhat, S., Stevenson-Leggett, P., Zagrajek, A.K., Hollinghurst, P., et al. (2020). The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins. PLoS Biol. *18*, e3001016.

Corman, V.M., Baldwin, H.J., Tateno, A.F., Zerbinati, R.M., Annan, A., Owusu, M., Nkrumah, E.E., Maganga, G.D., Oppong, S., Adu-Sarkodie, Y., et al. (2015). Evidence for an ancestral association of human coronavirus 229E with bats. J. Virol. *89*, 11858–11870.

Dai, L., Zheng, T., Xu, K., Han, Y., Xu, L., Huang, E., An, Y., Cheng, Y., Li, S., Liu, M., et al. (2020). A universal design of betacoronavirus vaccines against COVID-19, MERS, and SARS. Cell *182*, 722–733.e11.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Gao, G.F. (2018). From "A"IV to "Z"IKV: attacks from emerging and re-emerging pathogens. Cell *172*, 1157–1159.

Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y.Q., Wang, Y., Teng, Y., Zhao, Z., Cui, Y., et al. (2020). Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science *369*, 1603–1607.

Guan, Y., Zheng, B.J., He, Y.Q., Liu, X.L., Zhuang, Z.X., Cheung, C.L., Luo, S.W., Li, P.H., Zhang, L.J., Guan, Y.J., et al. (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science *302*, 276–278.

Guo, H., Hu, B.J., Yang, X.L., Zeng, L.P., Li, B., Ouyang, S., and Shi, Z.L. (2020). Evolutionary arms race between virus and host drives genetic diversity in bat severe acute respiratory syndrome-related coronavirus spike genes. J. Virol. *94*, e00902–e00920.

Haake, C., Cook, S., Pusterla, N., and Murphy, B. (2020). Coronavirus infections in companion animals: virology, epidemiology, clinical and pathologic features. Viruses *12*, 1023.

Han, X., Qi, J., Song, H., Wang, Q., Zhang, Y., Wu, Y., Lu, G., Yuen, K.Y., Shi, Y., and Gao, G.F. (2017). Structure of the S1 subunit C-terminal domain from bat-derived coronavirus HKU5 spike protein. Virology *507*, 101–109.

Hansen, J., Baum, A., Pascal, K.E., Russo, V., Giordano, S., Wloga, E., Fulton, B.O., Yan, Y., Koon, K., Patel, K., et al. (2020). Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. Science *369*, 1010–1014.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A., et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell *181*, 271–280.e8.

Huynh, J., Li, S., Yount, B., Smith, A., Sturges, L., Olsen, J.C., Nagel, J., Johnson, J.B., Agnihothram, S., Gates, J.E., et al. (2012). Evidence supporting a zoonotic origin of human coronavirus strain NL63. J. Virol. *86*, 12816–12825.

Lam, T.T., Jia, N., Zhang, Y.W., Shum, M.H., Jiang, J.F., Zhu, H.C., Tong, Y.G., Shi, Y.X., Ni, X.B., Liao, Y.S., et al. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature *583*, 282–285.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., and Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature *581*, 215–220.

Latinne, A., Hu, B., Olival, K.J., Zhu, G., Zhang, L., Li, H., Chmura, A.A., Field, H.E., Zambrana-Torrelio, C., Epstein, J.H., et al. (2020). Origin and cross-species transmission of bat coronaviruses in China. Nat. Commun. *11*, 4235.

Lau, S.K., Woo, P.C., Li, K.S., Huang, Y., Tsoi, H.W., Wong, B.H., Wong, S.S., Leung, S.Y., Chan, K.H., and Yuen, K.Y. (2005). Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc. Natl. Acad. Sci. USA *102*, 14040–14045.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., Zhang, H., et al. (2005). Bats are natural reservoirs of SARS-like coronaviruses. Science *310*, 676–679.

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., et al. (2020a). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell *182*, 1284–1294.e9.

Li, Y., Wang, H., Tang, X., Fang, S., Ma, D., Du, C., Wang, Y., Pan, H., Yao, W., Zhang, R., et al. (2020b). SARS-CoV-2 and three related coronaviruses utilize multiple ACE2 orthologs and are potently blocked by an improved ACE2-Ig. J. Virol. *94*, e01283, 20.

Liu, K., Tan, S., Niu, S., Wang, J., Wu, L., Sun, H., Zhang, Y., Pan, X., Qu, X., Du, P., et al. (2021). Cross-species recognition of SARS-CoV-2 to bat ACE2. Proc. Natl. Acad. Sci. USA *118*, e2020216118.

Lu, G., Hu, Y., Wang, Q., Qi, J., Gao, F., Li, Y., Zhang, Y., Zhang, W., Yuan, Y., Bao, J., et al. (2013). Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. Nature *500*, 227–231.

Lu, D., Liu, K., Zhang, D., Yue, C., Lu, Q., Cheng, H., Wang, L., Chai, Y., Qi, J., Wang, L.F., et al. (2019). Peptide presentation by bat MHC class I provides new insight into the antiviral immunity of bats. PLoS Biol. *17*, e3000436.

Malik, Y.S., Sircar, S., Bhat, S., Sharun, K., Dhama, K., Dadar, M., Tiwari, R., and Chaicumpa, W. (2020). Emerging novel coronavirus (2019-nCoV)-current scenario, evolutionary perspective based on genome analysis and recent developments. Vet. Q. *40*, 68–76.

Muik, A., Wallisch, A.K., Sänger, B., Swanson, K.A., Mühl, J., Chen, W., Cai, H., Maurus, D., Sarkar, R., Türeci, Ö., et al. (2021). Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine-elicited human sera. Science *371*, 1152–1153.

Otwinowski, Z., and Minor, W. (1997). [20] Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. *276*, 307–326.

Pavlovich, S.S., Lovett, S.P., Koroleva, G., Guito, J.C., Arnold, C.E., Nagle, E.R., Kulcsar, K., Lee, A., Thibaud-Nissen, F., Hume, A.J., et al. (2018). The egyptian rousette genome reveals unexpected features of bat antiviral immunity. Cell *173*, 1098–1110.e18.

Piccoli, L., Park, Y.J., Tortorici, M.A., Czudnochowski, N., Walls, A.C., Beltramello, M., Silacci-Fregni, C., Pinto, D., Rosen, L.E., Bowen, J.E., et al. (2020). Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. Cell *183*, 1024–1042.e21.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., and Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. Nature *581*, 221–224.

Shi, R., Shan, C., Duan, X., Chen, Z., Liu, P., Song, J., Song, T., Bi, X., Han, C., Wu, L., et al. (2020). A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. Nature *584*, 120–124.

Smith, I., and Wang, L.F. (2013). Bats and their virome: an important source of emerging viruses capable of infecting humans. Curr. Opin. Virol. *3*, 84–91.

Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., Liu, W., Bi, Y., and Gao, G.F. (2016). Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. *24*, 490–502.

Tan, S., Zhang, H., Chai, Y., Song, H., Tong, Z., Wang, Q., Qi, J., Wong, G., Zhu, X., Liu, W.J., et al. (2017). An unexpected N-terminal loop in PD-1 dominates binding by nivolumab. Nat. Commun. *8*, 14369.

Tan, W., Zhao, X., Ma, X., Wang, W., Niu, P., Xu, W., Gao, G., and Wu, G. (2020). A novel coronavirus genome identified in a cluster of pneumonia cases-Wuhan, China 2019-2020. China CDC Weekly *2*, 61–62.

The 2019-nCoV Outbreak Joint Field Epidemiology Investigation Team, and Li, Q. (2020). Notes from the Field: An Outbreak of NCIP (2019-nCoV) Infection in China — Wuhan, Hubei Province, 2019−2020. China CDC Weekly *2*, 79–80.

Wang, L., Su, S., Bi, Y., Wong, G., and Gao, G.F. (2018). Bat-origin coronaviruses expand their host range to pigs. Trends Microbiol. *26*, 466–470.

Wang, C., Horby, P.W., Hayden, F.G., and Gao, G.F. (2020a). A novel coronavirus outbreak of global health concern. Lancet *395*, 470–473.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.Y., et al. (2020b). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell *181*, 894–904.e9.

Wrobel, A.G., Benton, D.J., Xu, P., Roustan, C., Martin, S.R., Rosenthal, P.B., Skehel, J.J., and Gamblin, S.J. (2020). SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat. Struct. Mol. Biol. *27*, 763–767.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020a). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269.

Wu, L., Chen, Q., Liu, K., Wang, J., Han, P., Zhang, Y., Hu, Y., Meng, Y., Pan, X., Qiao, C., et al. (2020b). Broad host range of SARS-CoV-2 and the molecular basis for SARS-CoV-2 binding to cat ACE2. Cell Discov. *6*, 68.

Wu, Y., Wang, F., Shen, C., Peng, W., Li, D., Zhao, C., Li, Z., Li, S., Bi, Y., Yang, Y., et al. (2020c). A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. Science *368*, 1274–1278.

Wu, L., Su, J., Niu, S., Chen, Q., Zhang, Y., Yan, J., Shi, Y., Qi, J., Gao, G., and Wang, Q. (2021). Molecular basis of pangolin ACE2 engaged by COVID-19 virus. Chin. Sci. Bull. *66*, 73–84.

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J.J., Li, N., Guo, Y., Li, X., Shen, X., et al. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature *583*, 286–289.

Xie, J., Li, Y., Shen, X., Goh, G., Zhu, Y., Cui, J., Wang, L.F., Shi, Z.L., and Zhou, P. (2018). Dampened STING-dependent interferon activation in bats. Cell Host Microbe *23*, 297–301.e4.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science *367*, 1444–1448.

Yuan, Y., Qi, J., Peng, R., Li, C., Lu, G., Yan, J., Wang, Q., and Gao, G.F. (2020). Molecular basis of binding between Middle East respiratory syndrome coronavirus and CD26 from seven bat species. J. Virol. *94*, e01387, e01319.

Zhang, G., Cowled, C., Shi, Z., Huang, Z., Bishop-Lilly, K.A., Fang, X., Wynne, J.W., Xiong, Z., Baker, M.L., Zhao, W., et al. (2013). Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science *339*, 456–460.

Zheng, S.Q., Palovcak, E., Armache, J.P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods *14*, 331–332.

Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E.C., et al. (2020a). A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. Curr. Biol. *30*, 2196–2203.e3.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al. (2020b). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al.; China Novel Coronavirus Investigating and Research Team (2020). A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. *382*, 727–733.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial strains		
Escherichia coli (E. coli) strain DH5α	TIANGEN	Cat# CB101-02
Escherichia coli (E. coli) strain BL21 (DE3)	Novagen	Cat# 69450
MAX Efficiency DH10Bac Competent *E. coli*	Invitrogen	Cat# 10361-012
Chemicals, antibodies, and recombinant proteins		
PEI	Alfa	A04043896-1g
Anti-His/APC	Miltenyi Biotec	Cat# 130-119-820; RRID: AB_2751870
SARS-CoV-2 RBD protein with his-tag, spike residues 319-541, accession number: EPI_ISL_402119	This paper	N/A
Goat anti-human IgG-HRP	Thermo Fisher	Cat# 31410
SARS-CoV-2 RBD-F449Y	This paper	N/A
SARS-CoV-2 RBD-L486F	This paper	N/A
SARS-CoV-2 RBD-Q493Y	This paper	N/A
SARS-CoV-2 RBD-Q498Y	This paper	N/A
SARS-CoV-2 RBD-D501N	This paper	N/A
SARS-CoV-2 RBD-H505Y	This paper	N/A
SARS-CoV RBD protein with his-tag, spike residues 306-527, accession number: NC_004718	This paper	N/A
cACE2 protein, residues 18-740, accession number: NP_001034545.1	This paper	N/A
Critical commercial assays		
HisTrap HP 5 mL column	GE Healthcare	Cat# 17524802
HiLoad 16/600 Superdex 200 pg	GE Healthcare	Cat# 28989335
Series S Sensor Chip CM5	GE Healthcare	Cat# 29149603
Sensor Chip Protein A	GE Healthcare	Cat# 29127556
Membrane concentrator	Millipore	UFC901096
Deposited data		
RaTG13 RBD/hACE2 complex	This paper	Protein Data Bank: 7DRV
Experimental models: Cell lines		
Sf9 Cells, SFM Adapted	Invitrogen	Cat# 11496015
High Five cells	Invitrogen	Cat# B85502
HEK293T cells	ATCC	ATCC CRL-3216
Recombinant DNA		
pEGFP-N1	MiaoLingPlasmid	Cat# P0133
pEGFP-N1-hACE2, accession number: BAJ21180	This paper	N/A
pEGFP-N1-monkey ACE2, accession number: A0A2K5X283	This paper	N/A
pEGFP-N1-rabbit ACE2, accession number: G1TEF4	This paper	N/A
pEGFP-N1-mouse ACE2, accession number: Q8R0I0	This paper	N/A
pEGFP-N1-rat ACE2, accession number: Q5EGZ1	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
pEGFP-N1-malayan pangolin ACE2, accession number: XP_017505746	This paper	N/A
pEGFP-N1-catACE2, accession number: Q56H28	This paper	N/A
pEGFP-N1-civet ACE2, accession number: Q56NL1.1	This paper	N/A
pEGFP-N1-fox ACE2, accession number: XP_025842512.1	This paper	N/A
pEGFP-N1-dog ACE2, accession number: J9P7Y2	This paper	N/A
pEGFP-N1-raccoon dog ACE2, accession number: ABW16956.1	This paper	N/A
pEGFP-N1-horse ACE2, accession number: F6V9L3	This paper	N/A
pEGFP-N1-pig ACE2, accession number: A0A220QT48	This paper	N/A
pEGFP-N1-wild Bactrian camel ACE2, accession number: XP_006194263.1	This paper	N/A
pEGFP-N1-alpaca ACE2, accession number: XP_006212709.1	This paper	N/A
pEGFP-N1-bovine ACE2, accession number: Q58DD0	This paper	N/A
pEGFP-N1-goat ACE2, accession number: XP_005701129.2	This paper	N/A
pEGFP-N1-sheep ACE2, accession number: W5PSB6	This paper	N/A
pEGFP-N1-intermediate horseshoe bat ACE2, accession number: W5PSB6	This paper	N/A
pEGFP-N1-Least horseshoe bat ACE2, accession number: W5PSB6	This paper	N/A
pEGFP-N1-little brown bat ACE2, accession number: G1PXH7	This paper	N/A
pEGFP-N1-fulvous fruit bat ACE2, accession number: D8WU01	This paper	N/A
pEGFP-N1-greater horseshoe bat ACE2, accession number: B6ZGN7	This paper	N/A
pEGFP-N1-big-eared horseshoe bat ACE2, accession number: XP_007538670.1	This paper	N/A
pEGFP-N1-lesser hedgehog tenrec ACE2, accession number: XP_004710002.1	This paper	N/A
pFastbac1	Invitrogen	10360014
pFastbac-SARS-CoV-2 RBD -His, residues 319-541, accession number: EPI_ISL_402119	This paper	N/A
pFastbac- SARS-CoV RBD -His, residues 306-527, accession number: NC_004718	This paper	N/A
pCAGGS	MiaoLingPlasmid	Cat# P0165
pET21a	Novagen	Cat# 69740
pET21a-hACE2, residues 18-740, accession number: NP_001034545.1	This paper	N/A
Software and algorithms		
PyMOL software	Molecular Graphics System, Version 1.8 Schrödinger	https://pymol.org/2/
BIAcore 8K Evaluation software	GE Healthcare	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FlowJo V10	FLOWJO	https://www.flowjo.com/solutions/flowjo/downloads
Motioncor2	Zheng et al., 2017	N/A
COOT	Emsley and Cowtan, 2004	https://www2.mrc-lmb.cam.ac.uk/personal/peemsley/coot/
Phenix	Adams et al., 2010	http://www.phenix-online.org/
MolProbity		N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Qihui Wang (wangqihui@im.ac.cn).

Materials availability

All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement.

The number of replicates carried out for each experiment is described in the figure/table legends.

Data and code availability

The atomic coordinates for the crystal structure of the RaTG13 RBD and hACE2 complex have been deposited in the Protein Data Bank (https://www.rcsb.org/) (PDB: 7DRV).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cells

HEK293T cells (ATCC CRL-3216), HeLa cells (ATCC CCL-2), BHK-21 cells (ATCC CCL-10), HeLa-hACE2 cells (prepared in this study), and BHK-hACE2 (prepared in this study) were cultured at 37°C in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS).

Sera Samples

The information of convalescent and healthy donors is included in the Table S5.

METHOD DETAILS

Gene Cloning

The full-length coding sequence of the 25 ACE2 orthologs was synthesized and cloned into the pEGFP-N1 vector for flow cytometry (Table S1). The extracellular domain of these 25 ACE2 orthologs fused with the Fc domain of mouse IgG (mFc) were cloned into the pCAGGS vector for protein expression. The coding sequences of RaTG13 RBD (residues 319-541, GenBank: QHR63300.2), SARS-CoV-2 RBD (residues 319-541, GISAID: EPI_ISL_402119), MERS-CoV RBD (residues 367-606, GenBank: JX869059) and hACE2 (residues 19-615, NCBI Reference Sequence: NP_001358344.1) were cloned into pFastBac vectors (Dai et al., 2020; Lu et al., 2013). The wild-type RaTG13 RBD (residues 319-541, GenBank: QHR63300.2) and mutated RaTG13 RBDs (F449Y, L486F, Q493Y, Q498Y, D501N and H505Y) were cloned in pCAGGS vectors. The coding sequence of RaTG13 S (residues 1-1233) and SARS-CoV-2 S (residues 1-1255) were cloned into pCAGGS vectors. The variable region of MAbs, REGN10933 (PDB: 6XDG), REGN10987 (PDB: 6XDG), C110 (PDB: 7K8P), and S2H14 (PDB: 4JX3) fused with the constant region of IgG1 were cloned into pCAGGS vectors. The extracellular domain of PD-L1 was constructed as previously described.

Protein Expression and Purification

The 25 ACE2 orthologs fused with mFc and 7 MAbs were expressed and purified from the culture supernatants of HEK293F cells using a Protein A affinity column (GE Healthcare) and further purified by gel filtration using a Superdex^{TM} 200 10/300 GL (GE Healthcare). Purified proteins were stored in a buffer containing 20 mM Tris-HCl and 150 mM NaCl (pH 8.0). Proteins for SPR assay were transferred to PBST (1.8 mM KH_2PO_4, 10 mM Na_2HPO_4 (pH 7.4), 137 mM NaCl, 2.7 mM KCl, and 0.005% (v/v) Tween 20) buffer.

The recombinant baculovirus-expressed proteins were expressed and purified as previously described (Tan et al., 2017). Briefly, the RaTG13 RBD, SARS-CoV-2 RBD, MERS-CoV RBD, and hACE2 proteins were expressed by infecting Hi5 cells with recombinant baculovirus containing the corresponding coding sequences. The RaTG13 RBD/hACE2 complex protein was expressed by co-infecting Hi5 cells with recombinant baculovirus containing the coding sequence of RaTG13 RBD and hACE2. The wild-type RaTG13 RBD (residues 319-541, GenBank: QHR63300.2) and mutated RaTG13 RBDs (F449Y, L486F, Q493Y, Q498Y, D501N and H505Y) cloned in pCAGGS were expressed in HEK293F cells. Cell culture supernatants were collected, filtered with a 0.22 μm filter, purified by His-Trap HP column (GE Healthcare), and further purified by Superdex™ 200 Increase 10/300 GL column (GE Healthcare). Purified proteins were stored in protein buffer (20 mM Tris-HCl, pH 8.0 and 150 mM NaCl).

Flow Cytometry Assay

Plasmids containing the 25 ACE2 orthologs fused with eGFP were transfected into HEK293T cells. Cells were harvested after 24 h post-transfection, washed thrice, and incubated with 100 μL 30 μg/mL test proteins (SARS-CoV-2 RBD, RaTG13 RBD and MERS RBD with histidine tag) at 37°C for 30 min. Subsequently, the cells were washed thrice with PBS to remove redundant protein and incubated with anti-His/APC antibodies (1:500, Miltenyi Biotec) at 37°C for 30 minutes. Finally, cells were washed for thrice before being analyzed using a BD FACS Calibur Flow Cytometer (BD Biosciences).

The plasmid containing recombinant hACE2 or mutated ACE2s was stably transfected into BHK-21 cells (BHK-hACE2). A mixture containing SARS-CoV-2 RBD (10 μg/mL), RaTG13 RBD (10 μg/mL) and CB6 antibody were pre-incubated at 4°C for 30 min, and then incubated with the BHK-hACE2 cells at 4°C for 30 min. Subsequently, cells were washed with PBS thrice and stained with APC mouse anti-his secondary antibody for 30 min before being analyzed using BD FACS Canto Flow Cytometer (BD Biosciences). The data of all samples were analyzed using FlowJo 7.6 (TreeStar Inc., Ashland, OR, USA).

SPR Analysis

The 25 ACE2-mFC fusion proteins were transferred into HBST buffer (20 mM HEPES (pH 7.4), 150 mM NaCl, and 0.005% (v/v) Tween 20) and immobilized on the CM 5 chip (Table S2). Then, serially diluted wild-type or mutated RaTG13 RBD proteins flowed over the chip in HBST buffer. PD-L1 protein was used as negative control. Binding affinities were measured using a BIAcore 8K (GE Healthcare) at 25°C in the single-cycle mode. Binding kinetics were analyzed with Biacore™ Insight software (GE healthcare) using a 1:1 Langmuir binding model.

The seven recombinant antibodies (2 μg/mL) were first captured on flow cell 2 of the protein A sensor chip (GE Healthcare) at more than 500 response units (Table S4). Flow cell 1 was used as the negative control. Then, serially diluted SARS-CoV-2 RBD or RaTG13 RBD proteins flowed over the chip in PBST buffer. Response Units (RU) were measured with a BIAcore 8K (GE Healthcare) at 25°C in single-cycle mode. The antibodies were regenerated with 10 mM glycine-HCl (pH 1.5). The equilibrium dissociation constants (K_D) of each pair of interactions were calculated using BIAcore® 8K Evaluation Software (GE Healthcare) by fitting to a 1:1 Langmuir binding model.

Enzyme-linked Immunosorbent Assay

The SARS-CoV-2 and RaTG13 RBD proteins were first diluted to 2 μg/mL with 0.05 M carbonate-bicarbonate buffer (pH 9.6) and then coated onto 96-well ELISA plates (Corning, USA) by overnight incubation at 4°C. Plates were blocked in 5% skim milk in PBST and incubated with serially diluted serum samples. Each serum sample was prepared in triplicate. Subsequently, plates were incubated with goat anti-human IgG-HRP antibody (Thermo Fisher, USA) and developed with 3,3′,5,5′-Tetramethylbenzidine (TMB) substrate. The reactions were stopped with 2 M sulfuric acid. Absorbance at 450 nm was measured using a microplate reader (PerkinElmer, USA). The endpoint titer was defined as the highest reciprocal dilution of serum to give an absorbance greater than 2.1-fold that of the background values. Antibody titers below the limit of detection were determined to be half the limit of detection.

Production and Quantification of Pseudoviruses

The RaTG13 and SARS-CoV-2 pseudoviruses were constructed with a GFP-encoding replication-deficient vesicular stomatitis virus (VSV) vector backbone (VSV-ΔG-GFP) and the coding sequence of corresponding spike proteins, as previously described (Li et al., 2020a; Muik et al., 2021). HEK293T cells were transfected with 30 μg of the spike protein expression plasmids. The VSV-ΔG-GFP pseudovirus was added 24h post-transfection. The inoculum was removed after incubation for 1 h at 37°C. The culture medium were then changed into DMEM supplemented with 10% FBS and 10 μg/mL of anti-VSV-G antibody (I1Hybridoma ATCC® CRL2700) after washing the cells with PBS. The pseudoviruses were harvested 20 h post-inoculation, passed through a 0.45 μm filter (Millipore, Cat#SLHP033RB) before aliquoted and stored at −80°C.

All pseudoviruses were treated with 0.5U/μL BaseMuncher endonuclease (Abcam, ab270049) for 1.5 h at 37°C to remove unpackaged RNA before quantification. Viral RNA was extracted (Bioer Technology, Cat# BYQ6.6.101711-213) and quantified by quantitative RT-PCR (qPCR) using 7500 fast Real-Time PCR System (Applied Biosystems) with the primers and probe for detecting the P protein coding sequence of VSV.

Pseudovirus Infection Assay

The pseudovirus particles of SARS-CoV-2, wild-type RaTG13 and mutated RaTG13 were normalized to the same amount for quantitation by qRT-PCR. Then, 100 μL of each pseudovirus was added to each well of the 96 well plate containing HeLa-hACE2 cells.

Untransfected HeLa cells were used as controls. The plates were imaged 15 h post-transfection. The imaging and analysis of fluorescent cells was determined using a CQ1 confocal image cytometer (Yokogawa, Japan). Each group contains 6 replicates.

Neutralization Assay

Neutralization assay was performed as previously described (Dai et al., 2020). Briefly, 50 μL of serially diluted human sera or monoclonal antibody were incubated with 50 μL of pseudoviruses at 37°C for 1 h before being added onto pre-plated HeLa-hACE2 cells. The transducing units (TU) numbers were calculated on a CQ1 confocal image cytometer (Yokogawa).

Crystallization

The sitting-drop method was used to obtain high-resolution RaTG13 RBD/hACE2 complex crystals. In detail, the purified complex proteins were concentrated to 10 mg/mL. Then, 0.8 μL protein was mixed with 0.8 μL reservoir solution. The resulting solution was sealed and equilibrated against 100 μL of the reservoir solution at 18°C. High-resolution RaTG13 RBD/hACE2 complex crystals were grown in 0.1 M succinic acid (pH 7.0), 0.1 M bicine pH 8.5, and 30% v/v polyethylene glycol monomethyl ether 550.

Data Collection and Structure Determination

Reservoir solution supplemented with 20% (v/v) glycerol was prepared as anti-freezing buffer for freezing crystals. Crystals were picked up from the groove by using a mini loop and soaked in anti-freezing buffer for a few seconds. Then, crystals were picked up and frozen by soaking in liquid nitrogen. Diffraction data were collected at the Shanghai Synchrotron Radiation Facility (SSRF) BL19U. The dataset was processed with HKL2000 software as previously described (Otwinowski and Minor, 1997). The structure of the RaTG13 RBD/hACE2 complex was determined by the molecular replacement method using Phaser (Adams et al., 2010) with a previously reported complex structure of the SARS-CoV-2-RBD complex with human ACE2 (PDB: 6LZG). The atomic models were completed using Coot (Emsley and Cowtan, 2004) and refined with phenix.refine in Phenix (Adams et al., 2010), and the stereochemical qualities of the final models were assessed using MolProbity . Data collection, processing, and refinement statistics are summarized in Table 1. All structural figures were generated using Pymol software (https://pymol.org/2/).

QUANTIFICATION AND STATISTICAL ANALYSIS

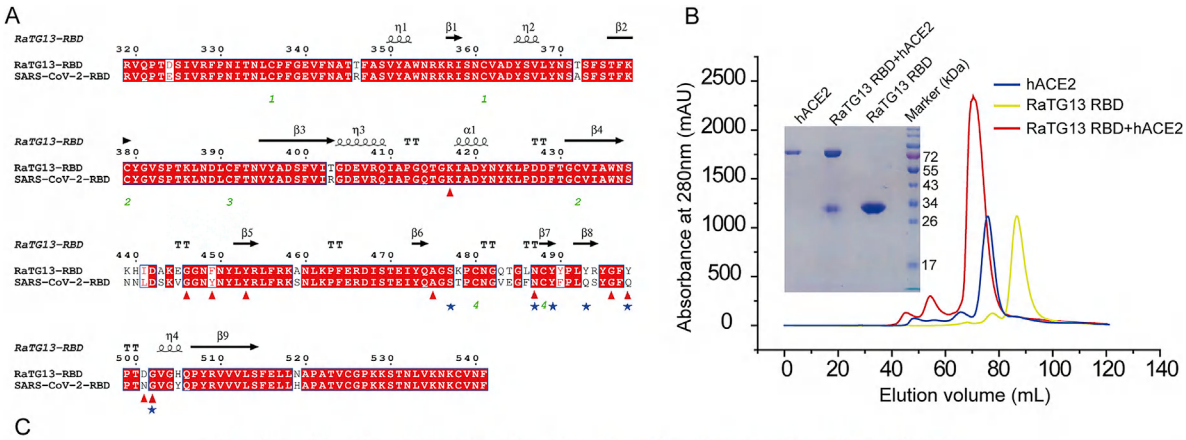
Binding Studies

K_D values for SPR experiments were obtained with BIAcore 8K Evaluation Software (GE Healthcare), using a 1:1 binding model. The values shown are the mean ± SD of three replicates.

Flow Cytometry Analysis

All experiments were performed in triplicate; one representative result was shown in Figures 2, 3, and 4.

Supplemental figures



C

	P1	P1	P1	P1	P1	P1	P1	P2	P2	P2	P2	P1	P1	P2	P2	P2		
	19	24	27	28	30	31	34	38	41	42	45	82	83	353	354	355	Substitutions	
Human	S	Q	T	F	D	K	H	D	Y	Q	L	M	Y	K	G	D		Primates
Macaca	S	Q	T	F	D	K	H	D	Y	Q	L	M	Y	K	G	D	0	
Rabbit	S	L	T	F	E	K	Q	D	Y	Q	L	T	Y	K	G	D	4	Lagomorpha
Mouse	S	N	T	F	N	N	Q	D	Y	Q	L	S	F	H	G	D	7	Rodentia
Rat	S	K	S	F	N	K	Q	D	Y	Q	L	N	F	H	G	D	7	
Malayan pangolin	S	E	T	F	E	K	S	E	Y	Q	L	N	Y	K	H	D	6	Pholidota
Cat	S	L	T	F	E	K	H	E	Y	Q	L	T	Y	K	G	D	4	Carnivora
Civet	S	L	T	F	E	T	Y	E	Y	Q	V	T	Y	K	G	D	7	
Fox	S	L	T	F	E	K	Y	E	Y	Q	L	T	Y	K	G	D	5	
Dog	S	L	T	F	E	K	Y	E	Y	Q	L	T	Y	K	G	D	5	
Raccoon dog	S	L	T	F	E	K	Y	E	Y	Q	L	T	Y	R	G	D	6	
Horse	S	L	T	F	E	K	S	E	H	Q	L	T	Y	K	G	D	6	Perissodactyla
Pig	F	L	T	F	E	K	L	D	Y	Q	L	T	Y	K	G	D	5	Artiodactyla
Wild bactrian camel	S	L	T	F	E	E	H	D	Y	Q	L	T	Y	K	G	D	4	
Alpaca	S	L	T	F	K	E	H	D	Y	Q	L	I	Y	K	G	D	4	
Bovine	S	Q	T	F	E	K	H	D	Y	Q	L	T	Y	K	G	D	2	
Goat	S	Q	T	F	E	K	H	D	Y	Q	L	T	Y	K	G	D	2	
Sheep	S	Q	T	F	E	K	H	D	Y	Q	L	T	Y	K	G	D	2	
Little brown bat	S	K	I	F	E	N	S	D	H	E	L	T	Y	K	G	D	8	Chiroptera
Intermediate horseshoe bat	S	R	I	F	D	N	R	E	Y	Q	L	N	Y	K	G	D	6	
Fulvous fruit bat	S	L	T	F	E	K	T	D	Y	Q	L	T	Y	K	G	D	4	
Greater horseshoe bat	S	L	K	F	D	D	S	N	H	Q	L	N	F	K	G	D	8	
Least horseshoe bat	S	K	K	F	N	D	S	D	Y	Q	L	N	Y	K	G	D	6	
Big-eared horseshoe bat	S	E	K	F	D	K	S	D	Y	E	L	N	Y	K	G	D	5	
Lesser hedgehog tenrec	S	Q	S	F	T	T	N	N	Y	Q	L	K	F	L	N	D	9	Afrotheria

Figure S1. Structure-based sequence alignment and protein expression, related to Figure 1

(A) Structure-based sequence alignment of SARS-CoV-2 RBD and RaTG13 RBD. Coils indicate α helices, and black arrows indicate β strands. Conserved residues are highlighted in red. Residues of SARS-CoV-2 RBD or RaTG13 RBD contact with hACE2 through hydrogen bonds are marked with red triangle or blue star, respectively. Sequence alignment is generated with ClustalX and ESPript. (B) Gel filtration profiles of hACE2 (blue), RaTG13 RBD (yellow) and the RaTG13 RBD/hACE2 complex (red) were analyzed and displayed. The separation profiles of each pooled samples on SDS-PAGE are shown in reducing conditions (+DTT). (C) Characteristics of the RaTG13 RBD-binding residues of ACE2s. The conserved residues among 25 ACE2 orthologs are shown as black letters and the residue substitutions in the ACE2 of 24 species comparing with human ACE2 are shown as red letters.

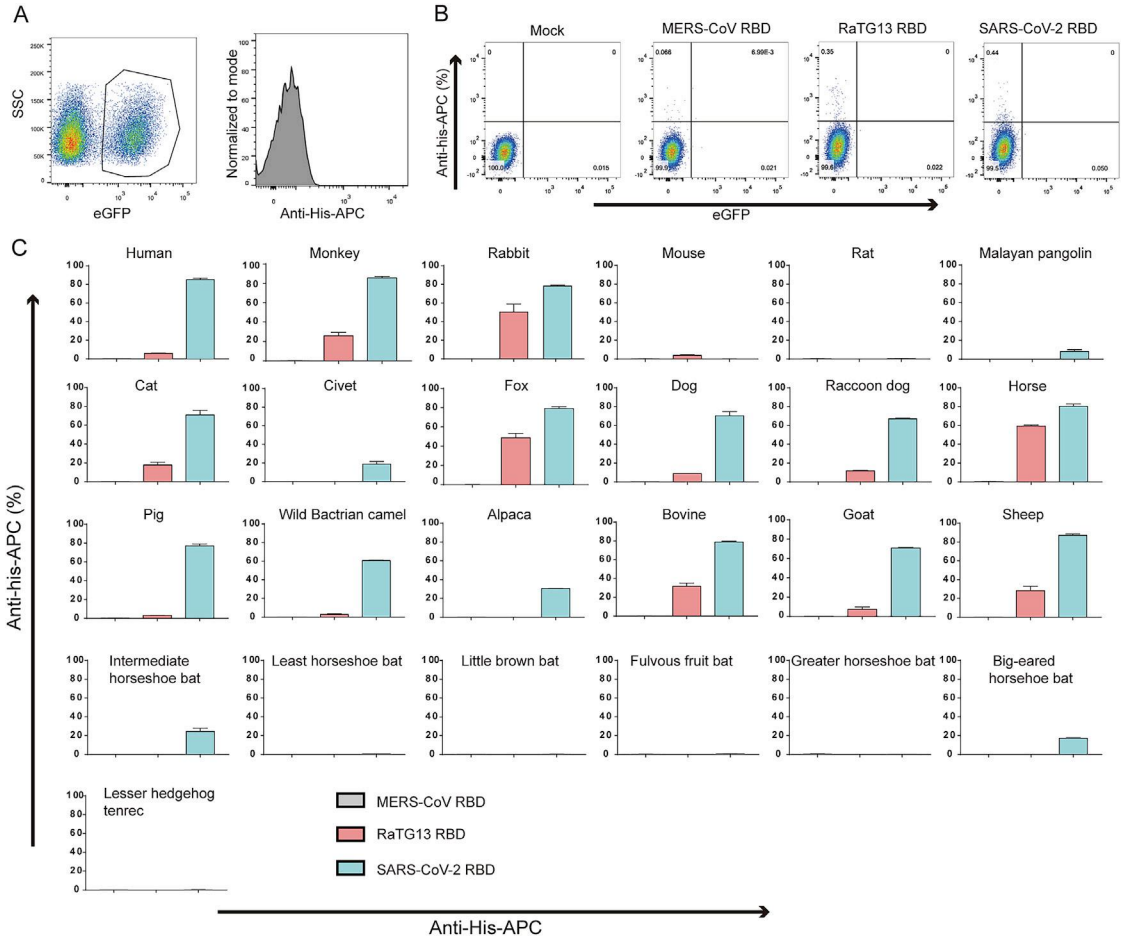


Figure S2. Gating strategy and statistics for flow cytometric analysis of the binding between ACE2s and the SARS-CoV-2 RBD or RaTG13 RBD, related to Figure 2

(A) Gating strategy for flow cytometric analysis of the binding between ACE2s and SARS-CoV-2 RBD or SARS-CoV RBD. eGFP-positive HEK293T cells are gated first, followed by analysis of anti-his-APC positive cells. (B) Dot plot of untransfected HEK293T cells stained by MERS-CoV RBD, RaTG13 RBD and SARS-CoV-2 RBD proteins. (C) Frequency of SARS-CoV-2 RBD or RaTG13 RBD positive HEK293T cells expressing hACE2. MERS-CoV RBD is used as negative control. Data represent the results of three replicates, and error bars show the SD of each measurement.

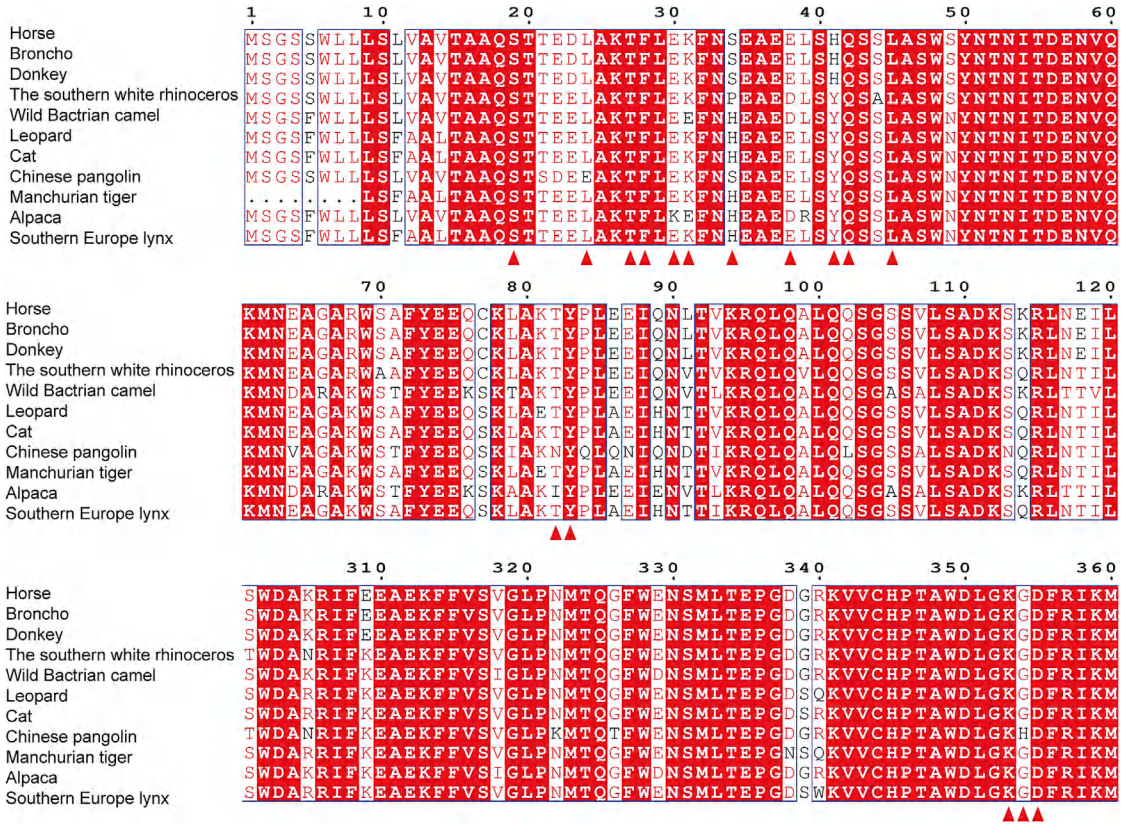


Figure S3. Structure-based sequence alignment of horse ACE2 and 10 ACE2 orthologs most closely related to horse ACE2, related to Figure 2

Coils indicate α helices, and black arrows indicate β strands. Conserved residues are highlighted in red. Residues highlighted in blue boxes are highly (80%) conserved, with consensus amino acids in red. Residues of hACE2 marked with blue star indicate key binding residues with RaTG13 RBD. Sequence alignment is generated with ClustalX, and ESPript.

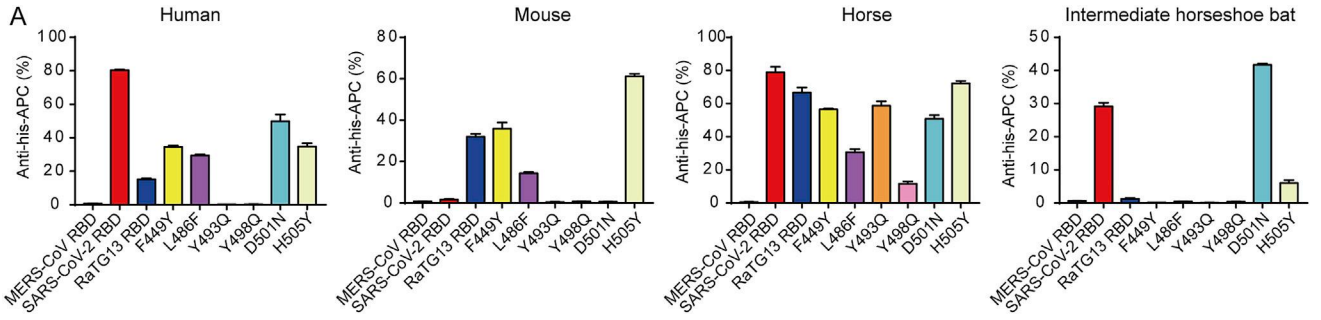


Figure S4. Percentage of ACE2-positive cells among total EGFP-positive cells, related to Figure 3

The percentage of SARS-CoV-2 RBD and RaTG13 RBD positive cells in HEK293T cells with different ACE2 othologs. Data represent the results of three replicates, and error bars show the SD of each measurement.

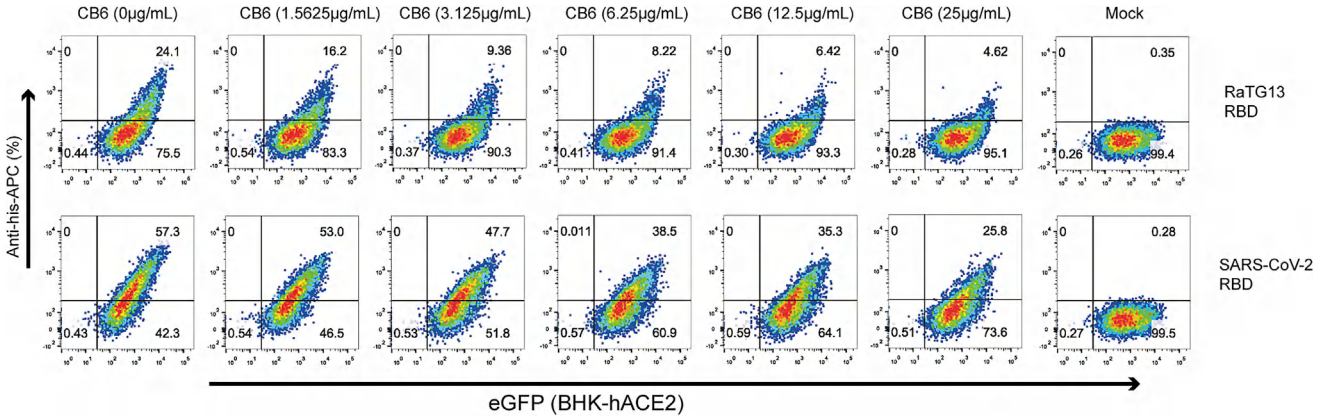


Figure S5. Blocking of the binding between HeLa-hACE2 cells and the SARS-CoV-2 RBD or RaTG13 RBD by CB6, related to Figure 5

CB6 was serially diluted by 2 folds, followed by incubation with SARS-CoV-2 RBD or RaTG13 RBD. Blocking efficacies were analyzed through staining BHK-hACE2 cells. Cells stained with SARS-CoV-2 RBD or RaTG13 RBD proteins was used as a positive control. Unstained HeLa-hACE2 cells were used as negative controls.