Zhang et al. Genome Biology (2019) 20:279
https://doi.org/10.1186/s13059-019-1897-7

Genome Biology

METHOD Open Access

PASTMUS: mapping functional elements at single amino acid resolution in human cells

Xinyi Zhang[1†], Di Yue[1†], Yinan Wang[1,2†], Yuexin Zhou[1†], Ying Liu[1†], Yeting Qiu[1], Feng Tian[1], Ying Yu[1], Zhuo Zhou[1] and Wensheng Wei[1*]

Abstract

Identification of functional elements for a protein of interest is important for achieving a mechanistic understanding. However, it remains cumbersome to assess each and every amino acid of a given protein in relevance to its functional significance. Here, we report a strategy, PArsing fragmented DNA Sequences from CRISPR Tiling MUtagenesis Screening (PASTMUS), which provides a streamlined workflow and a bioinformatics pipeline to identify critical amino acids of proteins in their native biological contexts. Using this approach, we map six proteins—three bacterial toxin receptors and three cancer drug targets, and acquire their corresponding functional maps at amino acid resolution.

Background

RNA-guided CRISPR-associated protein 9 nucleases can introduce indels (insertions or deletions) and point mutations at target genomic loci by generating DNA double-strand breaks (DSBs) and consequently activating internal repair mechanisms, especially non-homologous end-joining (NHEJ) [1, 2]. Mutagenesis, and mutations leading to a frameshift in particular, can usually abolish protein expression, making the CRISPR-Cas9 system a powerful tool for genome engineering [3, 4] and even for high-throughput functional screening [5–8]. To better understand the role of regulatory elements or protein-coding sequences, CRISPR-mediated tiling mutagenesis has been employed with relevant biological assays [9, 10].

It is of great importance for the identification of functional elements for a protein of interest to achieve a mechanistic understanding. Traditional methods mainly rely on in vitro biochemical assays, such as co-immunoprecipitation (Co-IP) combined with truncation mutagenesis [11]; however, these techniques have a low resolution, and none of them is performed in native biological contexts. Previous studies include screening of cells expressing cDNAs containing various missense mutations [12, 13], screening through generating point mutations [14, 15], screening of tiling library followed by NGS (next-generation sequencing) on enriched sgRNAs [16–20], and a recent approach named "tag-mutate-enrich" [21]. Most of these methods require the exogenous expression of cDNAs [12, 13, 21]. They are also limited by the coverage of the actual amino acids of target [12–15, 21], the types of mutation [12–15], or the resolution of the functional map [16–20]. After all, most of these methods are not designed to study mutations that are genetically recessive [12, 13, 16–21]. There is no existing method that could assess potentially all amino acids of a given protein for their functional importance, especially in the native biological contexts.

Herein, we report the development of the PArsing fragmented DNA Sequences from CRISPR Tiling MUtagenesis Screening (PASTMUS) strategy, aiming at precisely mapping functional elements and assessing the importance of each amino acid (a.a.) spanning the full length of the protein of interest.

* Correspondence: wswei@pku.edu.cn
†Xinyi Zhang, Di Yue, Yinan Wang, Yuexin Zhou and Ying Liu contributed equally to this work.
[1]Biomedical Pioneering Innovation Center, Beijing Advanced Innovation Center for Genomics, Peking-Tsinghua Center for Life Sciences, Peking University Genome Editing Research Center, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

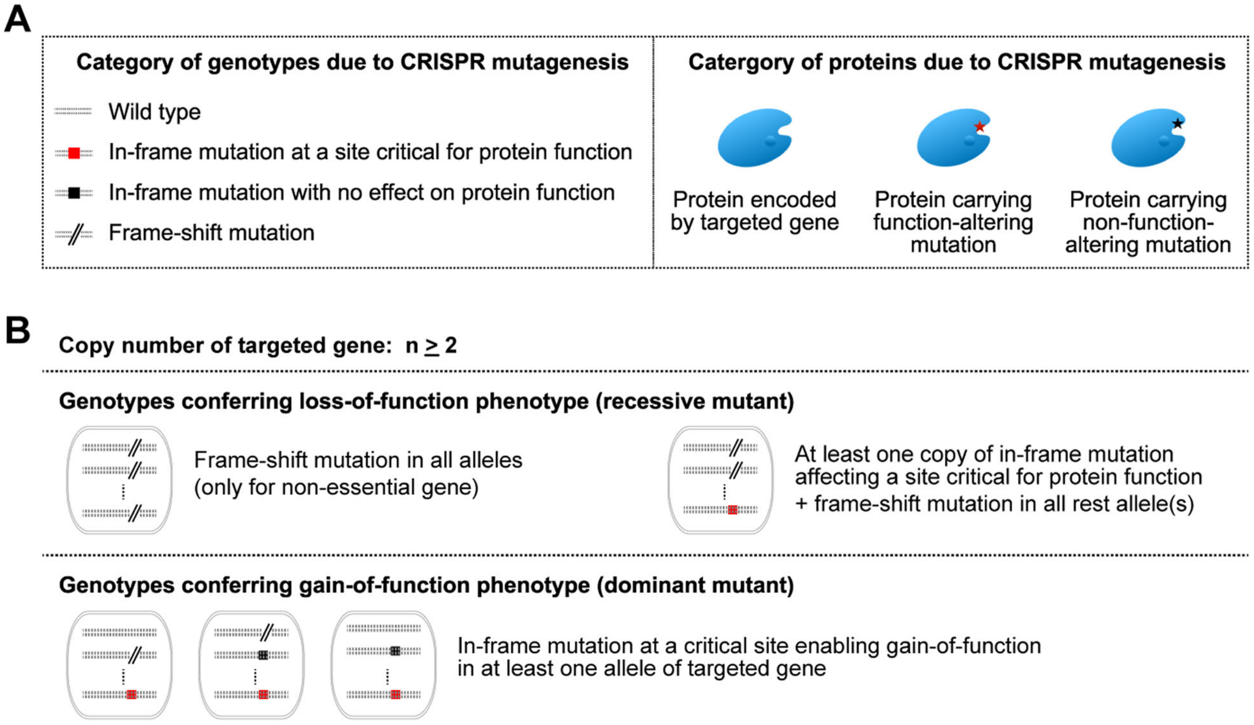
A
Category of genotypes due to CRISPR mutagenesis
Wild type
In-frame mutation at a site critical for protein function
In-frame mutation with no effect on protein function
Frame-shift mutation
Catergory of proteins due to CRISPR mutagenesis
Protein encoded by targeted gene
Protein carrying function-altering mutation
Protein carrying non-function-altering mutation
B
Copy number of targeted gene: n ≥ 2
Genotypes conferring loss-of-function phenotype (recessive mutant)
Frame-shift mutation in all alleles (only for non-essential gene)
At least one copy of in-frame mutation affecting a site critical for protein function + frame-shift mutation in all rest allele(s)
Genotypes conferring gain-of-function phenotype (dominant mutant)
In-frame mutation at a critical site enabling gain-of-function in at least one allele of targeted gene

Fig. 1 Rationale for acquiring residues critical for protein function based on phenotypic changes associated with in-frame mutations. **a** Category of genotypes or proteins due to CRISPR mutagenesis. **b** Genotypes conferring loss-of-function phenotype (for recessive mutant) and genotypes conferring gain-of-function phenotype (dominant mutant)

Results

Rationale, workflow, and bioinformatics pipeline of PASTMUS

If we would generate a library of cells containing a variety of mutations spanning the targeted gene on the genome, we could readily enrich those cells harboring proteins carrying function-altering mutations in a positive selection screening (Fig. 1a). If mutations in targeted gene are genetically recessive, cells would have complete loss of function only if (i) frameshift mutations occur in all alleles (only for non-essential genes), or (ii) in-frame mutation affecting a site critical for protein function occurs in one or more allele(s), and frameshift mutation(s) in all the rest allele(s) (Fig. 1b, Additional file 1: Figure S1). For the genetically dominant mutant, in-frame mutation at a critical site enabling gain-of-function phenotype in at least one allele of targeted gene is sufficient to confer phenotypic change (Fig. 1b, Additional file 1: Figure S1). We therefore hypothesized that if we were to apply CRISPR tiling mutagenesis and retrieve only in-frame mutations (in-frame deletions or missense mutations) that give rise to a phenotypic change of choice, we could identify critical amino acids relevant to the protein functions.

We first performed tiling mutagenesis of targeted genes using the CRISPR-spCas9 system [8, 22, 23]. To maximize the coverage density in designing sgRNAs, we included two types of protospacer-adjacent motifs (PAMs), NGG and NAG [24]. After library screening using bacterial toxins or cancer drugs, genomic DNA was extracted for the conventional PCR amplification of sgRNA barcodes, followed by NGS analysis. In addition, cDNAs obtained from reverse-transcribed RNAs of targeted genes were PCR amplified and subsequently fragmented to approximately 250 bp in length before subjected to NGS. NGS data have to be mapped with reference and applied with a series of rules to obtain frequencies of those “meaningful” mutant reads at functionally relevant sites (Fig. 2a).

To determine whether we could generate sufficient mutation variants for PASTMUS and how the sgRNA coverage (cell count per sgRNA) corresponds to mutation complexity, we performed CRISPR mutagenesis on eight sites from *CSPG4* and *HBEGF* genes (Additional file 1: Figure S2, Additional file 2: Table S1). With variable folds of sgRNA coverages, total mutations and in-frame mutation types were calculated based on NGS (at average base coverage of sequencing ~ 50,000×). Wild type loci were also sequenced following the same experimental protocol to determine the basal level of mutations due to PCR and sequencing errors. It turned out that the types of in-frame mutants, as well as the

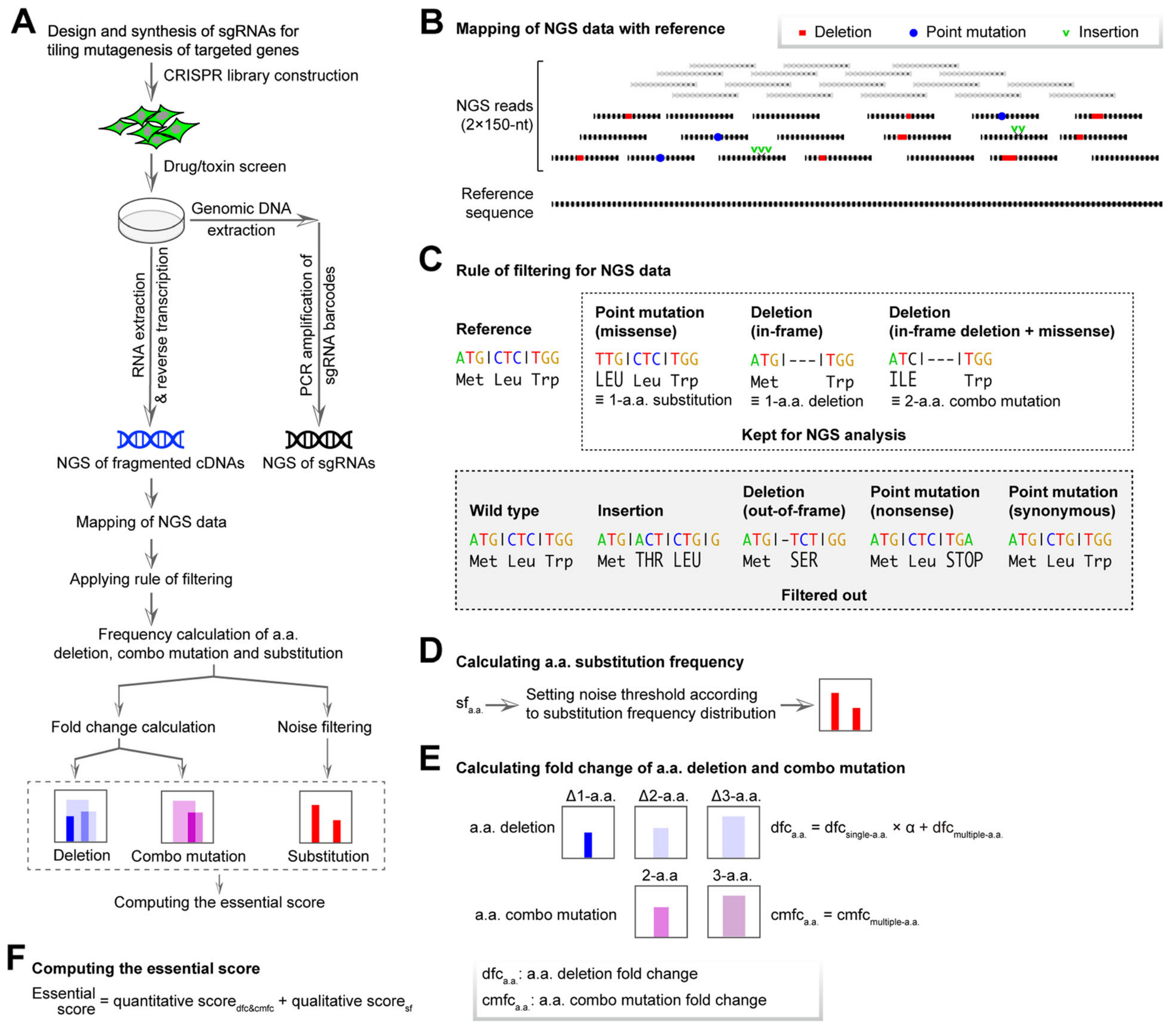
A
Design and synthesis of sgRNAs for tiling mutagenesis of targeted genes
CRISPR library construction
Drug/toxin screen
Genomic DNA extraction
RNA extraction & reverse transcription
PCR amplification of sgRNA barcodes
NGS of fragmented cDNAs
NGS of sgRNAs
Mapping of NGS data
Applying rule of filtering
Frequency calculation of a.a. deletion, combo mutation and substitution
Fold change calculation
Noise filtering
Deletion
Combo mutation
Substitution
Computing the essential score
B
Mapping of NGS data with reference
Deletion
Point mutation
Insertion
NGS reads (2×150-nt)
Reference sequence
C
Rule of filtering for NGS data
Reference
ATG|CTC|TGG
Met Leu Trp
Point mutation (missense)
TTG|CTC|TGG
LEU Leu Trp
≡ 1-a.a. substitution
Deletion (in-frame)
ATG|---|TGG
Met Trp
≡ 1-a.a. deletion
Deletion (in-frame deletion + missense)
ATC|---|TGG
ILE Trp
≡ 2-a.a. combo mutation
Kept for NGS analysis
Wild type
ATG|CTC|TGG
Met Leu Trp
Insertion
ATG|ACT|CTG|G
Met THR LEU
Deletion (out-of-frame)
ATG|-TCT|GG
Met SER
Point mutation (nonsense)
ATG|CTC|TGA
Met Leu STOP
Point mutation (synonymous)
ATG|CTG|TGG
Met Leu Trp
Filtered out
D
Calculating a.a. substitution frequency
sf a.a.
Setting noise threshold according to substitution frequency distribution
E
Calculating fold change of a.a. deletion and combo mutation
Δ1-a.a.
Δ2-a.a.
Δ3-a.a.
a.a. deletion
dfc a.a. = dfc single-a.a. × α + dfc multiple-a.a.
2-a.a
3-a.a.
a.a. combo mutation
cmfc a.a. = cmfc multiple-a.a.
dfc a.a.: a.a. deletion fold change
cmfc a.a.: a.a. combo mutation fold change
F
Computing the essential score
Essential score = quantitative score dfc&cmfc + qualitative score sf

Fig. 2 PASTMUS workflow and bioinformatics pipeline to identify critical residues of proteins. **a** PASTMUS workflow. sgRNA tiling library screening was performed, followed by amplification of sgRNA barcodes and the targeted gene cDNA for NGS. After mapping to targeted gene CDS and extracting mutation, sequencing reads with out-of-frame mutations were filtered out. The frequency of reads conferring amino acid deletion, combo mutation, and substitution were calculated for libraries before and after screening. Quantitative assessment (fold change) of a.a. deletions and combo mutations and qualitative assessment of a.a. substitutions were performed, respectively. **b–f** The essential scores of individual amino acids indicating their functional significance were calculated accordingly. The PASTMUS bioinformatics pipeline is as follows: mapping of NGS data with reference sequence (**b**), applying the filtering rules for NGS data (**c**), calculating a.a. substitution frequency and setting noise threshold for filtering (**d**), calculating fold change of a.a. deletion and a.a. combo mutation between libraries before and after screening (**e**), and computing essential scores for individual a.a. according to quantitative assessment of a.a. deletion and a.a. combo mutation, and qualitative assessment of substitution frequency (**f**)

total mutants after CRISPR mutagenesis in all eight sites, were significantly higher than the basal levels. It is also evident that the higher the sgRNA coverage, the larger the mutant varieties. It is therefore beneficial to generate a sgRNA library with as high as possible the coverage to maximize mutant complexity for screening (Additional file 1: Figure S2).

To test PASTMUS strategy in mapping functional elements of proteins, we selected three genes (*ANTXR1*, *CSPG4*, and *HBEGF*) encoding bacterial toxin receptors and three genes (*HPRT1*, *PLK1*, and *PSMB5*) encoding cancer drug targets (Additional file 3: Table S2). We chose HeLa cells to construct the CRISPR library for screening because we have determined the appropriate killing conditions in this cell line for toxins [8, 11, 25] and drugs (e.g., 6-TG targeting HPRT1 [26], BI2536 targeting PLK1 [27], and Bortezomib targeting PSMB5 [28]) (Additional file 1: Figure S3).

For the targeted genes, sgRNAs were designed in silico and synthesized on a chip as pools to construct a tiling CRISPR library covering the full lengths of the three receptor-coding genes and a library covering three drug targets (Additional file 1: Figure S3, Additional file 4: Table S3, Additional file 5: Table S4). We performed functional screens in two replicates for each of the six treatments in addition to controls without treatment. After three rounds of treatment with a toxin (PA/LFnDTA toxin, diphtheria toxin, or *Clostridium difficile* toxin B) or a drug (6-TG, BI2536, or Bortezomib), resistant cells were harvested, and genomic DNA was extracted for conventional sgRNA deciphering through NGS analysis [8, 29]. The harvested resistant cells were also subjected to total RNA isolation and reverse transcription to obtain cDNAs, which were subsequently used as templates for PCR amplification using specific primers (Additional file 5: Table S5). For genes with big sizes, such as *CSPG4*, multiple pairs of primers were used to amplify overlapping fragments to encompass their full lengths. For genes with alternative splicing, specific primer pairs were designed to ensure that all alternative transcripts were included (Additional file 1: Figure S3).

To meet the size limitation for NGS, PCR amplification of cDNA was fragmented to average 250 bp (Fig. 2a, Additional file 1: Figure S3). Since the mixtures of DNA fragments were predominantly wild type sequences, it is critical to reach enough sequencing depths to identify those small percentages of mutants. For this, we have developed a bioinformatics pipeline and applied a series of filtering procedure to process NGS data (Fig. 2). Among all types of mutations, deletion, insertion, and point mutation (Fig. 2b), we only kept those falling into one of the following two categories: missense mutation leading to amino acid substitution, and in-frame deletion leading to either a.a. deletion or a.a. combo mutation (due to the combined effect of in-frame deletion and missense mutation) (Fig. 2c). All wild type genes of targets were also sequenced following the same experimental protocol to determine the basal level of mutations.

From six gene-targeting PASTMUS libraries before screening, the levels of in-frame deletions leading to either a.a. deletions or a.a. combo mutations were significantly higher than the mock controls (Additional file 1: Figure S4); however, the levels of missense mutations leading to a.a. substitutions were indistinguishable between libraries and the mock controls (Additional file 1: Figure S4). Several recent studies [30, 31] reported that substitution frequency generated by DNA repair after CRISPR-spCas9 editing is relatively low, while the errors generated through the course of reverse transcription, PCR amplification, and NGS were predominantly point mutations rather than indels. Although we were unable to normalize missense mutation, the enrichment of such type through screening provided an affirmative answer for the functional importance of the affected amino acid (Fig. 2d).

For a.a. deletion and combo mutation types (Fig. 2c), more than 95% of amino acids encoded by the six targeted genes were covered. In particular, 98% of amino acids were covered for genes with relatively smaller sizes when counting mutations affecting up to three a.a., e.g., *HBEGF* and *HPRT1* (Additional file 1: Figure S5). To reduce potential false-positive rate, we counted only those mutations affecting ≤ 3 a.a. for further analysis. The enrichment of a.a. deletions and combo mutations could be benchmarked by fold changes with their frequencies in the original libraries before the screening. Since the affirmative role of any given amino acid could be determined by single-a.a. deletion result, we assigned a full weight to those sites and a discounted weight (based on mutation lengths) to those with multiple-a.a. deletion or combo mutation (Fig. 2e and the "Methods" section). Combining both the quantitative data for a.a. deletion or combo mutation and the qualitative data for a.a. substitution, we could compute the essential scores to obtain the importance of all amino acids in relevance to protein function (Fig. 2f).

We evaluated the quality of the screens based on the sgRNA fold changes between two replicates and obtained the correlation coefficients ranging from 0.71 to 0.98 (Additional file 1: Figure S6). Because all three toxin receptors are non-essential for cell viability, the sgRNAs observed after screening were uniformly distributed across their coding sequences (Fig. 3a, Additional file 1: Figures S8, S9), indicating that most of them could generate frameshift indels, resulting in the disruption of targeted protein expression. However, NGS of sgRNA-coding regions revealed little sequence-to-function information.

By applying PASTMUS with the computational pipeline, we obtained function-related amino acid maps. We purposely assigned a solid blue color to single-a.a. deletions because there is no ambiguity regarding the significance of such mutations, while we assigned blue with 15% transparency to multi-a.a. deletions (Figs. 3b and 4b, Additional file 1: Figures S8, S9, S10, S11), purple for a.a. combo mutations with higher transparent levels set for longer affected length (Figs. 3c and 4c, Additional file 1: Figures S8, S9, S10, S11), and red for the a.a. substitutions (Fig. 4d, Additional file 1: Figures S9, S10, S11).

Mapping toxin receptors

For the functional screening of *HBEGF*, which encodes a receptor for diphtheria toxin (DT), most of the resistant cells carried a.a. deletions and combo mutations in the

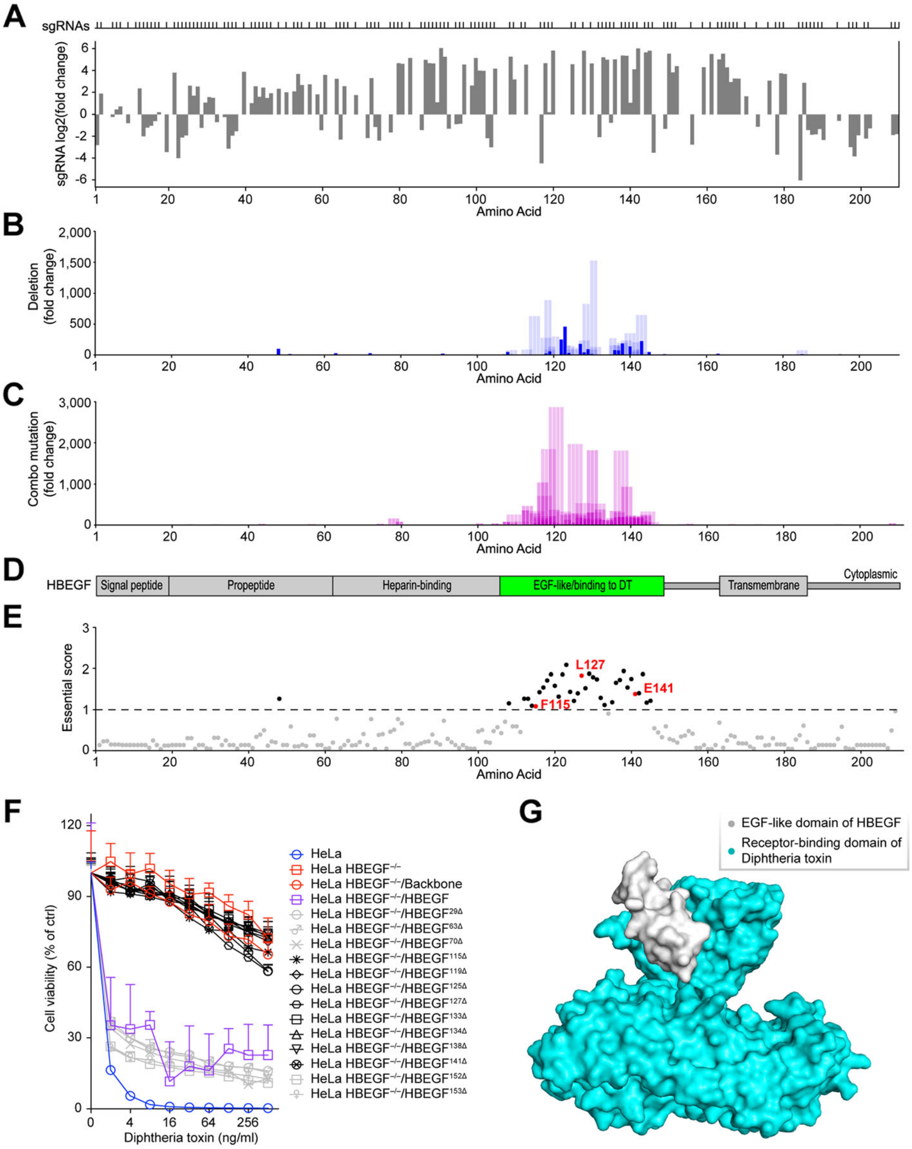

Fig. 3 Identification of HBEGF amino acids critical for diphtheria toxin (DT)-mediated cytotoxicity through PASTMUS. **a** Identification of *HBEGF*-targeting sgRNAs conferring cell resistance to DT. Distribution of sgRNAs mapped to the corresponding amino acid in HBEGF is indicated on top. **b** a.a. deletion fold change corresponding to each amino acid. The solid blue bars indicate single-a.a. deletions; bars with transparency indicate multiple-a.a. deletions. The width of the blue bar with transparency corresponds with a.a. deletion length. **c** a.a. combo mutation fold change corresponding to each a.a. Width of the bar indicates affected a.a. length. **d** Schematic diagram of HBEGF with the EGF-like domain shown in green, a known binding region for DT. **e** Essential score of each a.a. of HBEGF. Cutoff of essential score is plotted as a dashed line, with critical amino acids above the cutoff shown in black and known critical amino acids labeled in red. **f** Effects of single-a.a. deletions on the susceptibility of cells to DT. Cells were treated with different concentrations of DT, and the MTT cytotoxicity assay was performed 48 h after toxin treatment. Data are presented as the mean ± SD, *n* = 5. **g** Surface diagram of crystal structure of the complex of DT with EGF-like domain of HBEGF. The EGF-like domain of HBEGF is shown in gray, and the receptor-binding domain of DT is shown in cyan (PDB code: 1XDT)

EGF-like domain (Fig. 3b–d), a reported DT-binding site [32]. Essential scores (Fig. 3e, Additional file 6: Table S6) indicated that the amino acids with the highest scores were enriched in the EGF-like domain, further confirming the essentiality of this domain for mediating toxin binding [32]. Three amino acids, F115, L127, and E141 [32], that are known to be essential for the HBEGF-DT interactions were ranked at the top (35th, 7th, and 22nd) among all amino acids. Importantly, PASTMUS uncovered a number of novel sites in addition to these three that appeared important for receptor function (Fig. 3e). To validate these results, we expressed wild type and mutant *HBEGF* from cDNAs in HeLa *HBEGF*$^{-/-}$ cells [8] via lentiviral infection (Additional file 1: Figure S7,

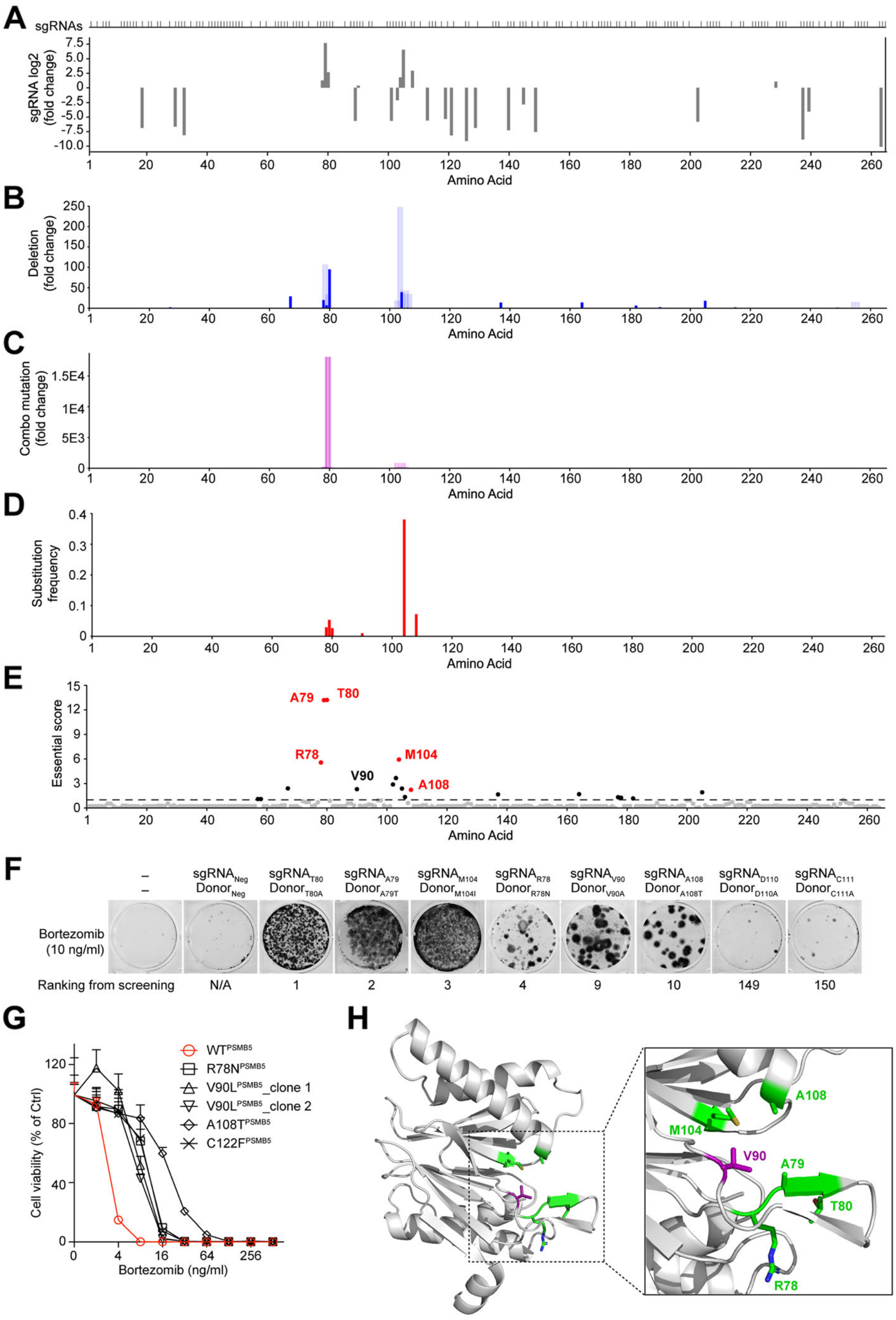

Fig. 4 Identification of PSMB5 amino acids critical for Bortezomib-mediated killing through PASTMUS. **a** Identification of *PSMB5*-targeting sgRNAs conferring cell resistance to Bortezomib. Distribution of sgRNAs mapped to the corresponding amino acid in PSMB5 is indicated on top. **b** a.a. deletion fold change corresponding to each a.a. **c** a.a. combo mutation fold change corresponding to each a.a. **d** a.a. substitution frequency corresponding to each a.a. **e** Essential score of each a.a. of PSMB5. Cutoff of essential score is plotted as a dashed line, with critical amino acids above the cutoff shown in black and known critical amino acids labeled in red. **f** MTT viability assay for indicated substitutions of PSMB5 on the susceptibility of cells to Bortezomib. **g** Effects of indicated substitutions in PSMB5 on the susceptibility of cells to Bortezomib. Data are presented as the mean ± SD, $n = 6$. **h** Crystal structure of PSMB5. Five positive amino acids and the novel critical site, V90, are labeled in green and purple, respectively (PDB code: 5LF3)

Additional file 5: Table S7). We verified five top-ranking sites (G119, K125, I133, C134, and Y138), three known positive sites (F115, L127, and E141), and five other sites below the threshold (L29, D63, D70, N152, and R153). HeLa *HBEGF*$^{-/-}$ cells appeared to be completely resistant to DT, and wild type *HBEGF* expression recovered cell sensitivity to the toxin. The expression of all mutant forms of HBEGF with a single-a.a. deletion in one of these five top-ranking sites (G119, K125, I133, C134, and Y138) or known positive sites (F115, L127, and E141) failed to rescue the sensitivity of cells to DT, while mutant HBEGF with single-a.a. deletion in one of five other sites (L29, D63, D70, N152, and R153) could make the rescue like wild type (Fig. 3f). Crystal structure of the complex of DT toxin with EGF-like domain of HBEGF illustrated the interaction between toxin and its receptor-binding site (Fig. 3g, Additional file 7: Video S1). Notably, the fact that very few amino acids outside of the DT-binding domain of HBEGF were screened out indicated a low false-positive rate of PASTMUS.

For the anthrax toxin receptor, ANTXR1, all resistant cells carried a variety of a.a. deletions and combo mutations across the entire coding region, including known sites corresponding to PA binding (Additional file 1: Figure S8). In addition to the known PA-binding sites [33] and transmembrane domain, a number of novel amino acids showing variable levels of importance were identified (Additional file 1: Figure S8). The most important amino acids for the function of ANTXR1 in mediating anthrax toxicity were determined by computing essential scores, including one known site, H57 [33] (Additional file 1: Figure S8).

For CSPG4, the receptor of *Clostridium difficile* toxin B (TcdB) [11], critical amino acids were mainly located in the first and last two CSPG repeats (Additional file 1: Figure S9). The first CSPG repeat is a known TcdB-binding site [11], while the last two repeats represented novel findings. Importantly, unlike the above two cases involving HBEGF and ANTXR1, in which the most informative data came from a.a. deletions and combo mutations, a.a. substitutions affecting Q780 in CSPG4 were highly enriched (in red, Additional file 1: Figure S9), suggesting a critical role of this residue in mediating TcdB toxicity.

Mapping cancer drug targets

Regarding the three genes encoding cancer drug targets, *HPRT1* is a non-essential gene, while both *PLK1* and *PSMB5* are essential for cell viability [34]. For HPRT1, 6-TG screening of the library showed that most of the sgRNAs were enriched and evenly distributed (Additional file 1: Figure S10), similar to those in bacterial toxin screens (Fig. 3a, Additional file 1: Figures S8, S9). The significance of each amino acid throughout the protein was indiscernible from sgRNA sequencing analysis (Additional file 1: Figure S10). The PASTMUS approach revealed a great number of sites that appeared important for HPRT1 in mediating cell sensitivity to 6-TG (Additional file 1: Figure S10). These findings were consistent with the known structure of tetrameric HPRT1 [26] (Additional file 1: Figure S10).

For the essential targets, PLK1 and PSMB5, sgRNA sequencing provided the approximate locations of certain critical amino acids where sgRNAs generated in-frame mutations (Fig. 4a, Additional file 1: Figure S11). Because sgRNA enrichment provided indirect evidence with low resolution, we reasoned that PASTMUS strategy would provide a more precise and comprehensive map with enhanced details. Indeed, more amino acids that appeared critical for protein function were identified with high accuracy in both PSMB5 and PLK1 (Fig. 4b–d, Additional file 1: Figure S11). Notably, top essential amino acids were identified mainly from a.a. substitutions, complemented by a variable number of a.a. deletions and combo mutations (Fig. 4d, e, Additional file 1: Figure S11). We again identified both known critical sites in PSMB5 for its interaction with Bortezomib (R78, A79, T80, M104, and A108) [35–37] and novel essential residues such as V90 (Fig. 4d, e). Similarly, we identified the residues C67 and R136, which are known to be critical for the BI2536-PLK1 interaction [37, 38], as well as a novel essential residue F183 (Additional file 1: Figure S11).

Because a.a. substitution was the predominant mutant type conferring drug resistance for both PSMB5 and PLK1, we decided to employ the ssODN-mediated method [39] to generate specific substitutions, instead of a.a. deletions, for validation. We selected eight sites (R78, A79, T80, V90, M104, A108, D110, and C111) in PSMB5, among which D110 and C111 were bottom-ranked and served as controls. The mutant types from the screening results or previous reports were preferentially chosen for substitution in the validation experiments. For the remainder, alanine was used for substitution (Additional file 5: Table S8). The transfected cells with donors containing one of six substitutions (R78N, A79T, T80A, V90A, M104I, and A108T) produced a variable number of Bortezomib-resistant colonies (Fig. 4f). In comparison, D110A and C111A failed to produce Bortezomib resistance, demonstrating that our method of validation was reliable (Fig. 4f). Interestingly, the C111 site has previously been reported to be important for PSMB5 in SW1573 and CEM cells [36, 40], in contrast to our screening and validation results (Fig. 4f). This discrepancy suggests that either the role of this amino acid varies with biological context, or the mutation leading to the correct type of a.a. substitution was missing in the original library. To verify the

Bortezomib-resistant cells, we sequenced the genomic region of the target loci and confirmed that all six sites contained the expected substitutions (Additional file 1: Figure S12, Additional file 5: Table S9). To further verify our results, we isolated single clones (Additional file 1: Figure S13) and performed the cell viability assay. We demonstrated that the following substitutions conferred Bortezomib resistance: R78N, V90L, and A108T (Fig. 4g). Among these substitutions, T80 and A108 have been previously reported to be involved in the direct binding of PSMB5 to Bortezomib [35–37], and substitutions of R78, A79, and M104 have been reported to disrupt the structures of the drug-binding sites and consequently confer Bortezomib resistance [14, 37, 41]. For the novel site, V90, we confirmed that V90L conferred drug resistance with two independent clones (Fig. 4g). Crystal structure of PSMB5 showed that V90 together with five known critical amino acids, R78, A79, T80, V90, M104, and A108, were all located in the pocket of PSMB5 that interacts with Bortezomib (Fig. 4h).

For PLK1, it has been reported that R136 and C67 are critical amino acids for BI2536 and F183 is structurally important for PLK1 binding to BI2536 [2, 37, 38]. A substitution in each of these three sites was confirmed to confer BI2536 resistance (Additional file 1: Figure S11).

PASTMUS reveals substitution patterns for critical residues

Since each amino acid has 19 kinds of non-synonymous substitutions, we hypothesized that different substitutions might have distinct effects. We retrieved missense mutation data of top hits from PSMB5 and PLK1 and performed substitution pattern analysis. Indeed, there was a clear pattern preference of substitution for these amino acids to confer cell resistance to drugs (Fig. 5a). In the case of PSMB5 at the site M104, PASTMUS identified three substitution variants (M104V, M104I, and M104N) that conferred Bortezomib resistance (Fig. 5a, b). To determine whether these were the only 3 resistance-conferring substitutions and how powerful the PASTMUS strategy is in generating substitution variety, we expressed wild type and 19 kinds of PSMB5 M104 mutants in HeLa cells via lentiviral infection (Additional file 5: Table S10). Besides these three, several other substitution variants of M104 also conferred drug resistance. With a compatible exogenous expression of all substitution variants (Fig. 5c), M104V appeared much more resistant than many other variants, especially to high dosage of Bortezomib (Fig. 5d). Interestingly, PASTMUS identified that V90 had a preference in glutamate (E) that conferred Bortezomib resistance (Fig. 5a). We expressed wild type and PSMB5 V90E mutants in HeLa cells (Additional file 5: Table S10). With a comparable expression level of wild type (Additional file 1: Figure S14), V90E conferred significant resistance to Bortezomib (Additional file 1: Figure S14). Owing to the high possibility that multiple substitution variants at critical sites could confer drug resistance, we did not suffer from a high level of false discovery rate for PASTMUS, even though we could not generate all substitution types for a given site in the original tiling mutagenesis library.

Sequencing depth in PASTMUS

Because in-frame deletions and missense mutations generated by tiling CRISPR mutagenesis comprised only small percentages of all sequenced fragments, we would like to determine the proper sequencing depth to detect these rare events. To this end, we performed subsampling to the reads of *HBEGF* and *PSMB5* genes at different sequencing coverage of libraries before and after screening (Additional file 1: Figures S15, S16). For original libraries before the screening, sequencing depth of 1.5E7× and 1E7× appeared sufficient for *HBEGF* and *PSMB5*, respectively (Additional file 1: Figure S15, S16). The difference between these two genes is likely because *PSMB5* is an essential gene, and its out-of-frame mutants cause cell death. After the positive screening, however, the required sequencing depth became 1E6×, much lower than the original library before screening (Additional file 1: Figure S15, S16). Hence, we would recommend a sequencing depth of 1.5E7× and 1E7× for original libraries before screening targeting non-essential and essential gene, respectively, and 1E6× for libraries post-screening.

Structural superposition of protein functional maps

To compare the functional maps with their corresponding protein structures, we highlighted those critical residues identified from PASTMUS on the surface diagram of three cancer drug targets (PSMB5, PLK1, and HPRT1) (Fig. 6a, d, g). For PSMB5, functional mapping results at the linear sequence format revealed little information regarding the spatial correlations of those critical sites and the drug (Additional file 1: Figure S4); however, this became self-explanatory as most of those identified amino acids were located in the pocket embracing Bortezomib (Fig. 6a, b, Additional file 8: Video S2) in the 3D structure. Comparing the wild type PSMB5 and its M104 mutants, the slight changes of amino acid side chains for M104I and M104V (Fig. 6c) were likely responsible for weakened interaction of PSMB5 with Bortezomib, thereby resulting in drug resistance. M104V appeared to have much shorter side chain than M104 and M104I, and thus conferred Bortezomib resistance at a much higher dosage (Fig. 5d).

Similarly, for PLK1, most of the identified residues (L59, C67, R136, and N181-L184) are located within the

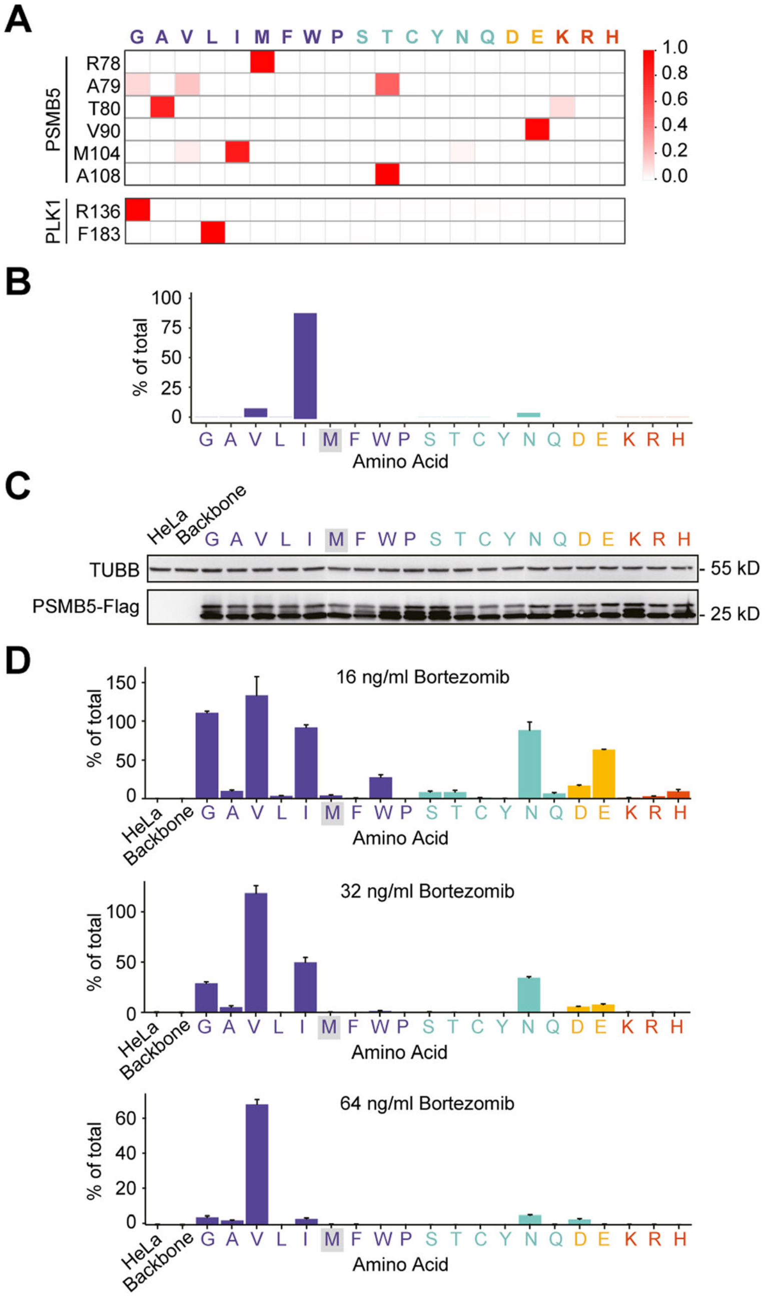

Fig. 5 Substitution pattern analysis of top hits from PSMB5 and PLK1. **a** Heat maps showing the substitution pattern of top amino acids of PSMB5 and PLK1. The 20 amino acids are classified into 4 groups with different colors according to their side-chain properties: non-polar (purple), polar (aqua), acidic (yellow), and basic (red). **b** PSMB5 M104 substitution pattern enriched in PASTMUS. **c** The expression of all PSMB5 M104 substitution variants. **d** Effects of M104 substitution variants on Bortezomib-mediated cell cytotoxicity with indicated dosages. Data are presented as the mean ± SD, *n* = 3

pocket that binds to BI2536 (Fig. 6d, e, Additional file 9: Video S3). F183, a structurally important site [42], showed a direct interaction with BI2536 through π-π stacking between aromatic rings (Fig. 6f). F183L identified from PASTMUS might disrupt the π-π stacking, leading to drug resistance (Fig. 5a, Additional file 1: Figure S11). Notably, several critical amino acids identified are located outside of the binding pocket (Fig. 6d), suggesting that these types of mutations could remotely alter binding pocket. This is particularly important because such amino acids are hardly predictable from the crystal structure.

For HPRT1, a transferase catalyzing the conversion of guanine to guanine monophosphate and hypoxanthine

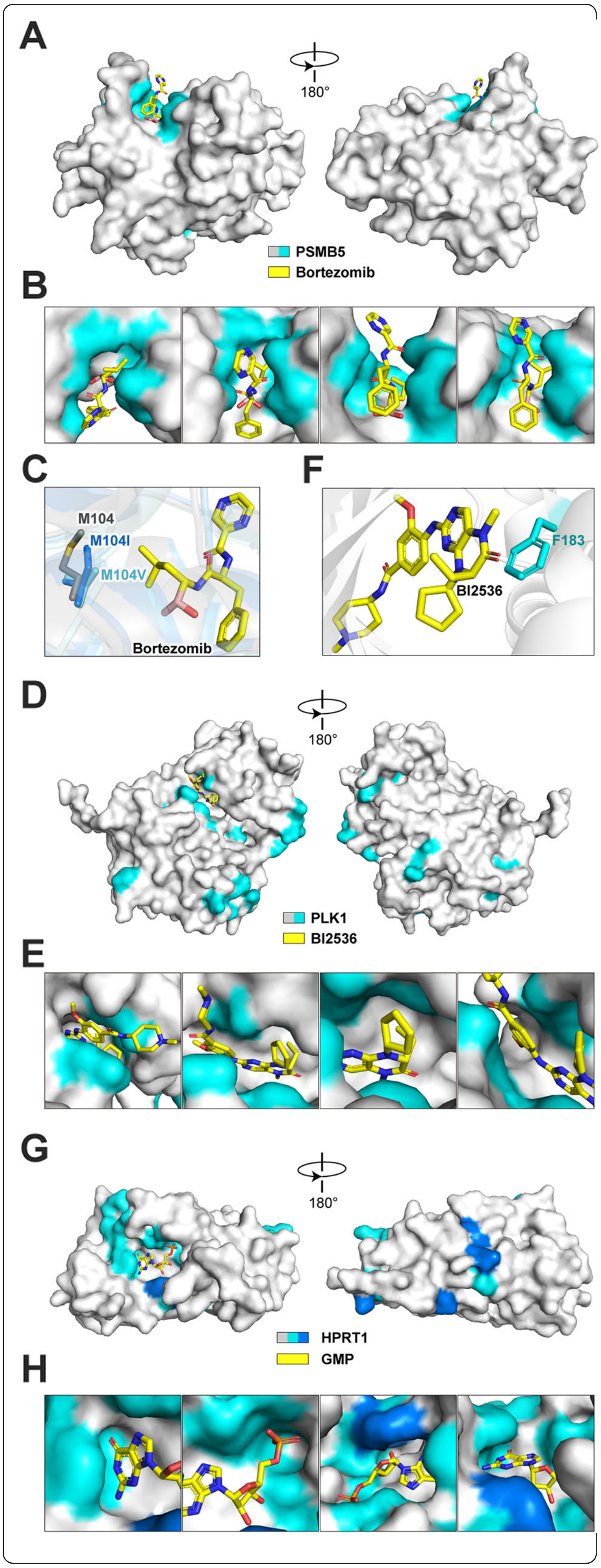

Fig. 6 Functional maps of PSMB5, PLK1, and HPRT1 at structure level. **a** Surface diagram of PSMB5 with Bortezomib. PSMB5 is shown in gray, and critical sites enriched in PASTMUS are shown in cyan. Bortezomib is shown in yellow (PDB code: 5LF3). **b** Zoomed-in view shown the interface of PSMB5 and Bortezomib. **c** Structural comparison of PSMB5 M104, M104I, and M104V interacts with Bortezomib (PDB code: 2F16, 4QVN, and 4QVQ). **d** Surface diagram of PLK1 with BI2536. PLK1 is shown in gray, and critical sites are shown in cyan. BI2536 is shown in yellow (PDB code: 2RKU). **e** Zoomed-in view shown the interface of PLK1 and BI2536. **f** Interaction of PLK1 F183 and BI2536. **g** Surface diagram of HPRT1 with bound GMP. HPRT1 is shown in gray, and critical sites are shown in blue (critical sites known for dimer or tetramer interaction are colored dark blue) (PDB code: 1HMP). **h** Zoomed-in view shown the interface of HPRT1 and GMP

to inosine monophosphate, its crystal structure bound with guanine monophosphate (GMP) showed that the identified amino acids were mostly housed within the pocket embracing GMP (Fig. 6g, h, Additional file 10: Video S4). This was consistent with the finding that 6-thioguanine (6-TG) acts as a purine analog to inhibit HPRT1 [26]. Moreover, several amino acids shown in dark blue were known sites essential for HPRT1 dimer or tetramer interaction. Mutations of these types of amino acids might disrupt HPRT1 tetramer formation, resulting in the loss of protein function and consequently the 6-TG resistance.

Discussion

Although previous studies using tiling mutagenesis have been reported to identify critical residues of proteins of interest [16–44], PASTMUS strategy is different from all of them. Our method enabled the identification of functionally important sites of the protein of interest at its native biological context and could work for both dominant and recessive mutations, regardless of the target gene size. The use of truncation mutagenesis to identify potential functional domains is often laborious. It is also technically difficult, if not impossible, to assess the significance of every amino acid spanning the full length of the protein of interest. Gill and colleagues recently described a method for mapping functionally relevant mutations in a protein of interest in bacteria or yeast; however, this method relies heavily on the homologous recombination rate, preventing its effective application in higher eukaryotes [45]. Moreover, PASTMUS allows multiple genes to be scanned simultaneously to identify functional elements in their corresponding proteins.

PASTMUS could delineate a functional map of a protein. Importantly, PASTMUS reveals critical amino acids that are located in both “reasonable” and “unreasonable-appearing” sites based on protein’s structural data. Those residues out of the catalytic domain or drug-binding pocket may provide valuable information for in-depth mechanisms of function of the protein. In addition, when co-crystallization data are missing, PASTMUS might be helpful to determine the precise binding sites of small chemical compounds, or even help to calibrate protein structures.

Although PASTMUS is capable of generating abundant mutations on almost all amino acids across the target protein, the function-altering mutation does not necessarily indicate that the affected site is directly relevant to protein function. For non-essential genes, at least two types of mutations could be identified from PASTMUS. The first is the mutation on a site that is critical for protein function. The second type is the mutation on a site that is critical to maintain the overall protein conformation or structure. For instance, we identified many hits that were located within the transmembrane domain of ANTXR1 (Additional file 1: Figure S8), a region that is important for the presence of receptor on the cell surface, not necessarily directly involved in toxin endocytosis.

For gain-of-function mutations, the “hitchhiking effect” could be a possible source of false positive (Additional file 1: Figure S1). Among the results of six gene-targeting PASTMUS screening, we did not find out those kinds of false-positive sites. For adjacent amino acids that appear as hits in our PSMB5 screening, we could verify the importance of R78, A79, and T80. Thus, this “hitchhiking effect” would not be a severe problem in PASTMUS strategy. This is likely because that the frequency of any “meaningful” in-frame mutation is extremely low, which makes the case of two or more different in-frame mutation variants in the same cell a very rare incidence.

Conclusions

We report a high-throughput strategy, PASTMUS, that provides a streamlined workflow and a bioinformatics pipeline for identifying critical elements of proteins in their native biological contexts. We mapped six proteins and acquired corresponding comprehensive functional maps at a single amino acid resolution; these maps contained both known domains or sites and novel amino acids that are critical for drug or toxin sensitivity. This method revealed comprehensive and precise single amino acid substitution patterns for critical residues. Because both a.a. deletions and combo mutations could be determined and quantified in the original libraries before the screening, PASTMUS could be readily applied in negative screening. Moreover, PASTMUS strategy is also suited for acquiring functional maps of regulatory elements, such as non-coding RNA, promoters, and enhancers.

Methods

Cells and reagents

Stably Cas9-expressing HeLa cells [8] and HEK293T cells were cultured in Dulbecco’s modified Eagle’s medium (DMEM, Corning) containing 10% fetal bovine serum (FBS, CellMax) under 5% CO_2 at 37 °C. All cells were checked for the absence of mycoplasma contamination. STR analysis was used for cell line authentication.

Plasmid construction

The sgRNA vector (pLenti-sgRNA-GFP) was cloned by replacing the U6 promoter in pLL3.7 (Addgene) with the human U6 promoter, ccdB cassette, and sgRNA scaffold. The Cas9 expression vector (pLenti-OC-IRES-BSD) has been previously reported [8]. pcDNA-HBEGF and pcDNA-PSMB5 were cloned by replacing the KRAB-dCas9 element of pHR-SFFV-KRAB-dCas9-P2A-mCherry (Addgene) with the human HBEGF or PSMB5 coding sequence and 3×FLAG. Vectors expressing cDNA of HBEGF with single-a.a. deletions were constructed via PCR-based site-directed mutagenesis (PfuUltraII Fusion HS DNA Polymerase, STRATAGENE). The primers used for these purposes are listed in Additional file 5: Table S7. Vectors expressing cDNAs of PSMB5 M104 and V90 substitutions were constructed via PCR-based site-directed mutagenesis (PrimeSTAR HS DNA Polymerase, Takara). Primers used for the construction of M104 and V90 substitution mutants are listed in Additional file 5: Table S10.

sgRNA library design

The hg19 CDS sequences of targeted genes were downloaded from UCSC genome browser (https://genome.ucsc.edu/), and all potential sgRNAs with NAG or NGG PAM sequence were designed using a homemade script to build the library.

Construction of the CRISPR/Cas9 sgRNA library

Two tiling libraries were constructed to include 1236 and 3712 sgRNAs targeting 3 drug-associated proteins and 3 toxin receptors, respectively (Additional file 4: Table S3). Array-based oligos encoding sgRNAs were synthesized and amplified via PCR with corresponding primers (Additional file 5: Table S4) that included the BsmBI recognition site at the 5′ end. The amplified DNA products were ligated into the vector using the Golden Gate method. The ligation mixture was then transformed into Trans1-T1 competent cells (Transgen) to generate the plasmid library [23, 29, 46]. The sgRNA plasmid library was subsequently transfected into HEK293T cells, together with two viral packaging plasmids, pVSVG and pR8.74 (Addgene), using the X-tremeGENE HP DNA transfection reagent (Roche). HeLa cells were then infected with a low MOI (~ 0.3) of lentivirus, and EGFP$^+$ cells were collected 48 h after infection via FACS.

Genome preparation and sequencing

Genomic DNAs of single site mutagenesis of HBEGF and CSPG4 were extracted after culturing for 14 days using the DNeasy Blood and Tissue kit (Qiagen). The targeted regions were amplified via 24 cycles of PCR (NEBNext Ultra II Q5 Master Mix). For each targeted region, 10 different forward primers and 10 different reverse primers were used to increase the diversity of the NGS library, each of which contained 1–10 additional nucleotides [47] (Additional file 2: Table S1). So, the potential bias of Illumina Sequencing that may affect screening results could be minimized [47]. PCR products from each library were purified using the DNA Clean & Concentrator-5 kit (Zymo Research Corporation) and indexed with different adaptors (NEB #7335, #7500) for NGS analysis.

Library screening

For TcdB screening, four 150-mm dishes were plated with 3.5×10^6 cells each as one experimental replicate. For each round of screening, cells were treated with an appropriate concentration: 70 ng/ml for the first round and 100 ng/ml for the second and third rounds. The details of DT and PA/LFnDTA screenings were the same as described in our previous report [8]. For 6-TG screening, six 150-nm dishes were plated with 3×10^6 cells for each experimental replicate. Two hundred fifty nanograms per milliliter of 6-TG was used in the first and second rounds, and 300 ng/ml 6-TG was used in the third round of screening. For Bortezomib screening, seven 150-nm dishes were plated with 2×10^6 cells for each experimental replicate. For each round of screening, cells were treated with variable doses of Bortezomib as follows: 10 ng/ml for the first round, 16 ng/ml for the second round, 28 ng/ml for the third round, and 40 ng/ml for the fourth round. For BI2536 screening, two 150-nm dishes were plated with 3.5×10^6 cells for each experimental replicate. For each round of screening, cells were treated with 4 ng/ml of BI2536 for the first round, 5 ng/ml for the second round, and 6 ng/ml for the third round.

The resistant cells from each screening were collected for genomic DNA and total RNA extraction, followed by reverse transcription. The sgRNA-coding regions and cDNAs of the targeted genes obtained through PCR amplification were then subjected to NGS analysis.

Identification of sgRNA sequences

Genomic DNAs were extracted from library cells (cell number corresponding to 1000× sgRNA coverage) using DNeasy Blood and Tissue kit (Qiagen). sgRNA regions were amplified via 26 cycles of PCR using primers [5–8] annealing to the flanking sequences of the sgRNAs. PCR products from each replicate were purified with DNA Clean & Concentrator-5 (Zymo Research Corporation), indexed with different adaptors (NEB #7370, #7335, #7500), and analyzed via NGS.

cDNA preparation and sequencing

Total RNAs were extracted from library cells using RNAprep Pure Cell/Bacteria Kit (TIANGEN), and cDNAs were synthesized using Quantscript RT Kit (TIANGEN). A two-step method was employed to construct libraries for NGS. The first step consisted of PCR amplification of the cDNA (26 cycles; PrimeSTAR HS DNA Polymerase, Takara). Primers used for different genes are listed in Additional file 5: Table S5. The coding sequence of *CSPG4* was approximately 6.9 kb in length, and three amplification reactions were employed to obtain overlapping fragments (~ 50 bp) encompassing its full length. After purified, cDNAs from each gene were sheared to ~ 250 bp using the Covaris S2 system (Additional file 1: Figure S3). The sheared products were purified and concentrated using DNA Clean & Concentrator-5 kit (Zymo Research Corporation) and indexed with different adaptors (NEB #7370, #7335, #7500) for NGS analysis.

Evaluation of mutation variety generated by CRISPR mutagenesis

Sequencing reads were trimmed, and the remaining reads were filtered to remove those with base quality below 30 before subjected to mapping with the reference sequences of targeted genes using Bowtie2 2.3.4.3 and sorted using SAMtools 1.9. Because different samples had variable volumes of sequencing data, mapped reads were down sampled (by sambamba-0.6.9) to approximately 100,000 reads for further analysis. Mutation types of both sgRNA libraries and wild type controls (without sgRNA) were calculated (using R package "CrispRVariants") covering ± 20 nt of estimated Cas9 cutting sites (3-bp upstream of PAM). Mutations with read counts less than 5 were removed from the analysis.

Computational methods for the identification of critical amino acids

Sequencing reads were trimmed, filtered, and mapped as described above. We only considered those reads containing in-frame mutations leading to either a.a. deletion, combo mutation, or a.a. substitution. Here, mutations with read count < 9 were also removed.

For fragments leading to a.a. deletions, we computed the deletion frequency ($\mathrm{Freq}^{\mathrm{del}}$) for each deletion type. For a deletion type x, we computed $\mathrm{Freq}^{\mathrm{del}-x}$ as follows:

$$\text{Freq}^{\text{del\_}x} = \frac{\text{number of reads with deletions of del\_}x}{\text{total number of reads covered region del\_}x}$$

For fragments containing a.a. combo mutations, their frequencies ($\text{Freq}^{\text{combo}}$) were calculated. For a combo mutation type *y*, we calculated $\text{Freq}^{\text{combo}-y}$ as follows:

$$\text{Freq}^{\text{combo\_}y} = \frac{\text{number of reads with combo mutations of combo\_}y}{\text{total number of reads covered region combo\_}y}$$

Then, fold change of a.a. deletion (*dfc*) and a.a. combo mutation (*cmfc*) were calculated, respectively. For a deletion type *x* and combo mutation type *y*, dfc_x and $cmfc\_y$ were calculated as follows:

$$dfc\_x = \frac{\text{Freq}^{\text{del\_}x} \text{ after screening}}{\text{Freq}^{\text{del\_}x} \text{ before screening}}$$

$$cmfc\_y = \frac{\text{Freq}^{\text{combo\_}y} \text{ after screening}}{\text{Freq}^{\text{combo\_}y} \text{ before screening}}$$

To estimate the significance of individual amino acid, length of affected a.a. and its position (*dfc* and *cmfc*) were taken into calculation. Fold change of single-a.a. deletion was assigned a weight (*w*). Fold changes of multiple-a.a. deletion and combo mutation (*dfc* and *cmfc*) were divided by squared of affected length, and the value was assigned to each affected amino acid.

That is,

$$dfc_{\text{a.a.}_i} = w \times dfc \text{ of single a.a.}_i + \sum\nolimits_j \frac{dfc_j \text{ affect a.a.}_i}{(\text{length of } dfc_j)^2}$$

$$cmfc_{\text{a.a.}_i} = \sum\nolimits_k \frac{cmfc_k \text{ affect a.a.}_i}{(\text{length of } cmfc_k)^2}$$

Where, a. a.$_i$ is the *i*th amino acid along with the targeted protein, dfc_j is the fold change of *j*th a.a. deletion which affect a. a.$_i$, and $cmfc_k$ is the fold change of *k*th a.a. combo mutation which affect a. a.$_i$.

For fragments containing a.a. substitutions, we computed the substitution ratio of amino acid *i* ($\text{Freq}^{\text{sub}}_{\text{Aa.a.}_i}$) as follows:

$$\text{Freq}^{\text{sub}}_{\text{a.a.}_i} = \frac{\text{number of reads with subsitutions of a.a.}_i}{\text{total number of reads covered a.a.}_i}$$

Because we could not quantify a.a. substitution frequency before the screening (Additional file 1: Figure S4), we estimated the effects of a.a. substitutions qualitatively by setting a cutoff frequency as follows:

$$\text{Cut–off}^{\text{sub}} = \text{mean of } \log 10\, \text{Freq}^{\text{sub}} + 3 \times \text{standard deviation of } \log 10\, \text{Freq}^{\text{sub}}$$

We gave a qualitative score to amino acid substitution frequency (sf_score) as follows:

$$sf\_score_{\text{a.a.}_i} = \begin{cases} 2, \text{Freq}^{\text{sub}}_{\text{a.a.}_i} > \text{Cutoff}^{\text{sub}} \\ 0, \text{Freq}^{\text{sub}}_{\text{a.a.}_i} \leq \text{Cutoff}^{\text{sub}} \end{cases}$$

Finally, we estimated the functional importance of each amino acid in a semi-quantitative way by assigning essential scores:

$$\text{quantitative\_effect}_{\text{a.a.}_i} = \text{normalization of } \log\left(dfc_{\text{a.a.}_i} + cmfc_{\text{a.a.}_i}\right)$$

$$\text{Score}_{\text{a.a.}_i} = -\log(p)\text{of quantitative\_effect}_{\text{a.a.}_i} + \text{sf\_score}_{\text{a.a.}_i}$$

Validation of the screening results

For the validation of critical substitutions of PSMB5 and PLK1, sgRNAs were designed near the mutation sites, and each 119-nt ssODN donor encoded one amino acid substitution for a validated residue. All sgRNAs and ssODN donor sequences are listed in Additional file 5: Table S8. HeLa cells were transfected with 1 μg of sgRNA and 2 μg of the ssODN donor in six-well plates. Fourteen days post-transfection, 1.5×10^5 cells were seeded in six-well plates 24 h before drug selection. Cells were treated with corresponding drugs at the proper dosages for 72 h: Bortezomib (10 ng/ml) and BI2536 (10 ng/ml). Genomic DNAs of drug-resistant cells were extracted using TIANamp Genomic DNA Kit (TIANGEN). The mutated loci were amplified using TransTaq DNA Polymerase High Fidelity (Transgen) and purified using a Universal DNA Purification Kit (TIANGEN). Primers are listed in Additional file 5: Table S9. PCR fragments were cloned into pEASY-T5 Zero Cloning Kit (Transgen) for sequencing.

Western blotting

Cell lysates were resolved on 10% SDS/PAGE gels (Bio-rad) for electrophoresis and transferred to PVDF membrane (Millipore) by Trans-blot Turbo transfer system (Bio-rad). After blocking with 5% non-fat milk at 37 °C for 1 h, probed with anti-FLAG antibody (MBL) and anti-β-tubulin antibody (CWBIO) overnight at 4 °C, the membrane was incubated with goat anti-mouse IgG-HPR secondary antibody (Jackson Immunoresearch) 1 h at room temperature. Clarity Western ECL Substrate Kit (Bio-rad) and Chemi-doc system (Bio-rad) were used to detect protein bands.

Cytotoxicity assay

Cells were seeded in 96-well plates 24 h before drug or toxin treatment (5000 cells for diphtheria toxin and 3000 cells for Bortezomib), and different concentrations of Bortezomib or DT were added. Cells were incubated at 37 °C for 48 h (DT) or 72 h (Bortezomib) before the addition of 1 mg/ml of MTT (3-[4,5-dimethylthiazol-2-

yl]-2,5-diphenyltetrazolium bromide) [25, 48]. Spectrophotometer readings at 570 nm were collected using BioTek Cytation5 (BioTek Instruments).

Structure analysis

Structures of the complex of diphtheria toxin with EGF-like domain of HBEGF (PDB code: 1XDT) [49], PSMB5 with Bortezomib (PDB code: 5LF3, 2F16, 4QVN, and 4QVQ) [50–52], PLK1 with BI2536 (PDB code: 2RKU) [42], and HPRT1 bound with GMP (PDB code: 1HMP) [53] were downloaded from the Protein Data Bank. The structures were analyzed using the PyMOL Molecular Graphics System, version 2.0, Schrödinger, LLC (https://pymol.org/2/).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1897-7.

Additional file 1: Figure S1. Rationale of acquiring residues critical for protein function based on phenotypic changes associated with in-frame mutations (copy number of targeted gene: $n = 2$). **Figure S2.** CRISPR mediated single site mutagenesis of *HBEGF* and *CSPG4*. **Figure S3.** Experimental conditions for PASTMUS screening. **Figure S4.** Distribution of amino acid (a.a.) deletion (a), combo mutation (b) and a.a. substitution (c) frequency in the original libraries (before screening) and the mock controls (wild type without sgRNA targeting). **Figure S5.** Relation between mutation affected length and a.a. coverage in libraries before screening. **Figure S6.** Scatter plot of sgRNA fold changes after screening on a log scale between two replicates. **Figure S7.** Expression of each HBEGF mutant for validation. **Figure S8.** Identification of ANTXR1 amino acids critical for PA/LFnDTA mediated cytotoxicity through PASTMUS. **Figure S9.** Identification of CSPG4 amino acids critical for TcdB mediated cytotoxicity through PASTMUS. **Figure S10.** Identification of HPRT1 amino acids critical for 6-TG mediated killing through PASTMUS. **Figure S11.** Identification of PLK1 amino acids critical for BI2536 mediated killing through PASTMUS. **Figure S12.** Sequencing chromatogram of mutated sites in PSMB5 locus from cells with or without ssODN donor transfection. **Figure S13.** DNA sequencing analysis of mutated alleles in the human PSMB5 locus from Bortezomib-resistant cell clones. **Figure S14.** Effects of PSMB5 V90E substitution variant on Bortezomib mediated killing. **Figure S15.** Correlation of NGS depth and data quality for PASTMUS screening of HBEGF. **Figure S16.** Correlation of NGS depth and data quality for PASTMUS of PSMB5.

Additional file 2: Table S1. Information of CRISPR mediated single site mutagenesis of *HBEGF* and *CSPG4*.

Additional file 3: Table S2. Information of six genes used in PASTMUS.

Additional file 4: Table S3. sgRNA sequences used for PASTMUS.

Additional file 5: Table S4. Primers used for sgRNA oligos amplification. **Table S5.** Primers used for cDNA amplification. **Table S7.** Primers used to generate different deletion mutants for *HBEGF*. **Table S8.** sgRNAs and ssODNs used for *PSMB5* and *PLK1* mutants validation. **Table S9.** Primers used for *PSMB5* genome amplification. **Table S10.** Primers used to generate PSMB5 M104 and V90 mutants. **Table S11.** Summary of candidate validations.

Additional file 6: Table S6. Essential scores of amino acids from six proteins through PASTMUS.

Additional file 7: Video S1. Crystal structure of the complex of DT toxin with EGF-like domain of HBEGF.

Additional file 8: Video S2. Crystal structure of PSMB5 with Bortezomib.

Additional file 9: Video S3. Crystal structure of PLK1 with BI2536.

Additional file 10: Video S4. Crystal structure of HPRT1 with bound GMP.

Acknowledgements

We acknowledge the staff of the BIOPIC High-throughput Sequencing Center (Peking University) and Genetron Health (Beijing, China) for their assistance with the NGS analysis and the flow cytometry core at National Center for Protein Sciences (Peking University), particularly Liying Du, Hongxia Lv, Xuefang Zhang, and Fei Wang for help with Fluorescence-Activated Cell Sorting. We thank High-performance Computing Platform of Peking University for the process of data analysis. We thank Ning Gao (Peking University) and Nieng Yan (Princeton University) for their advice on structure analysis. We thank lab members and Pengfei Yuan (EdiGene, Inc.) for helpful discussions.

Peer review information

Yixin Yao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

WW conceived and supervised the project. XZ, DY, YZ, YQ, FT, and YY performed the experiments. YW and YL performed the bioinformatics analysis. XZ, DY, YW, YL, YZ, ZZ, and WW wrote the manuscript. All authors read and approved the final manuscript.

Funding

This project was supported by funds from the National Science Foundation of China (NSFC31430025), the Beijing Municipal Science & Technology Commission (Z181100001318009), the Beijing Advanced Innovation Center for Genomics at Peking University, the Peking-Tsinghua Center for Life Sciences (WW), the National Science Foundation of China (31870893), the National Major Science & Technology Project for Control and Prevention of Major Infectious Diseases in China (2018ZX10301401) (ZZ), and the National Science Foundation of China (31801097) (YL).

Availability of data and materials

The accession number for the raw sequencing data reported in this paper is NCBI Sequence Read Archive (SRA): SRP230665 (PRJNA590617) [54]. Source code written by R for PASTMUS is available at https://bitbucket.org/WeiLab/pastmus [55] and a demo of the computational pipeline at https://figshare.com/articles/PASTMUS_mapping_functional_elements_at_single_amino_acid_resolution_in_human_cells/10435370.

Ethics approval and consent to participate

Not applicable for this study.

Consent for publication

Not applicable for this study.

Competing interests

A patent has been filed relating to the data presented, which would not limit the academic use of the data and the method. WW is a founder and scientific advisor for EdiGene. The other authors declare that they have no competing interests.

Author details

[1]Biomedical Pioneering Innovation Center, Beijing Advanced Innovation Center for Genomics, Peking-Tsinghua Center for Life Sciences, Peking University Genome Editing Research Center, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China. [2]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China.

Received: 18 September 2019 Accepted: 22 November 2019
Published online: 16 December 2019

References

1. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–21.

2. Burkard ME, Santamaria A, Jallepalli PV. Enabling and disabling polo-like kinase 1 inhibition through chemical genetics. ACS Chem Biol. 2012;7:978–81.
3. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339:819–23.
4. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via Cas9. Science. 2013;339: 823–6.
5. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343:84–7.
6. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343:80–4.
7. Koike-Yusa H, Li Y, Tan EP, Velasco-Herrera Mdel C, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol. 2014;32:267–73.
8. Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, Wei W. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature. 2014;509:487–91.
9. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Saturation editing of genomic regions by multiplex homology-directed repair. Nature. 2014;513: 120–3.
10. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, Chen DD, Schupp PG, Vinjamur DS, Garcia SP, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. 2015;527:192–7.
11. Yuan P, Zhang H, Cai C, Zhu S, Zhou Y, Yang X, He R, Li C, Guo S, Li S, et al. Chondroitin sulfate proteoglycan 4 functions as the cellular receptor for Clostridium difficile toxin B. Cell Res. 2015;25:157–68.
12. Brenan L, Andreev A, Cohen O, Pantel S, Kamburov A, Cacchiarelli D, Persky NS, Zhu C, Bagul M, Goetz EM, et al. Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. Cell Rep. 2016;17: 1171–83.
13. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N, et al. Prospective functional classification of all possible missense variants in PPARG. Nat Genet. 2016;48: 1570–5.
14. Hess GT, Fresard L, Han K, Lee CH, Li A, Cimprich KA, Montgomery SB, Bassik MC. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat Methods. 2016;13:1036–42.
15. Ma Y, Zhang J, Yin W, Zhang Z, Song Y, Chang X. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat Methods. 2016;13:1029–35.
16. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, Syed T, Emons BJ, Gifford DK, Sherwood RI. High-throughput mapping of regulatory DNA. Nat Biotechnol. 2016;34:167–74.
17. Ipsaro JJ, Shen C, Arai E, Xu Y, Kinney JB, Joshua-Tor L, Vakoc CR, Shi J. Rapid generation of drug-resistance alleles at endogenous loci using CRISPR-Cas9 indel mutagenesis. PLoS One. 2017;12:e0172177.
18. Donovan KF, Hegde M, Sullender M, Vaimberg EW, Johannessen CM, Root DE, Doench JG. Creation of novel protein variants with CRISPR/Cas9-mediated mutagenesis: turning a screening by-product into a discovery tool. PLoS One. 2017;12:e0170445.
19. Neggers JE, Kwanten B, Dierckx T, Noguchi H, Voet A, Bral L, Minner K, Massant B, Kint N, Delforge M, et al. Target identification of small molecules using large-scale CRISPR-Cas mutagenesis scanning of essential genes. Nat Commun. 2018;9:502.
20. He W, Zhang L, Villarreal OD, Fu R, Bedford E, Dou J, Patel AY, Bedford MT, Shi X, Chen T, et al. De novo identification of essential protein domains from CRISPR-Cas9 tiling-sgRNA knockout screens. Nat Commun. 2019;10:4541.
21. Pettitt SJ, Krastev DB, Brandsma I, Drean A, Song F, Aleksandrov R, Harrell MI, Menon M, Brough R, Campbell J, et al. Genome-wide and high-density CRISPR-Cas9 screens identify point mutations in PARP1 causing PARP inhibitor resistance. Nat Commun. 2018;9:1849.
22. Chang N, Sun C, Gao L, Zhu D, Xu X, Zhu X, Xiong JW, Xi JJ. Genome editing with RNA-guided Cas9 nuclease in zebrafish embryos. Cell Res. 2013; 23:465–72.
23. Peng J, Zhou Y, Zhu S, Wei W. High-throughput screens in mammalian cells using the CRISPR-Cas9 system. FEBS J. 2015;282:2089–96.
24. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31:827–32.
25. Qian L, Cai C, Yuan P, Jeong SY, Yang X, Dealmeida V, Ernst J, Costa M, Cohen SN, Wei W. Bidirectional effect of Wnt signaling antagonist DKK1 on the modulation of anthrax toxin uptake. Sci China Life Sci. 2014;57: 469–81.
26. Duan J, Nilsson L, Lambert B. Structural and functional analysis of mutations at the human hypoxanthine phosphoribosyl transferase (HPRT1) locus. Hum Mutat. 2004;23:599–611.
27. Steegmaier M, Hoffmann M, Baum A, Lenart P, Petronczki M, Krssak M, Gurtler U, Garin-Chesa P, Lieb S, Quant J, et al. BI 2536, a potent and selective inhibitor of polo-like kinase 1, inhibits tumor growth in vivo. Curr Biol. 2007;17:316–22.
28. Chen D, Frezza M, Schmitt S, Kanwar J, Dou QP. Bortezomib as the first proteasome inhibitor anticancer drug: current status and future perspectives. Curr Cancer Drug Targets. 2011;11:239–53.
29. Zhu S, Li W, Liu J, Chen CH, Liao Q, Xu P, Xu H, Xiao T, Cao Z, Peng J, et al. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. Nat Biotechnol. 2016;34:1279–86.
30. Allen F, Crepaldi L, Alsinet C, Strong AJ, Kleshchevnikov V, De Angeli P, Palenikova P, Khodak A, Kiselev V, Kosicki M, et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat Biotechnol. 2019:37:64–72.
31. Shen MW, Arbab M, Hsu JY, Worstell D, Culbertson SJ, Krabbe O, Cassa CA, Liu DR, Gifford DK, Sherwood RI. Predictable and precise template-free CRISPR editing of pathogenic variants. Nature. 2018;563:646–51.
32. Mitamura T, Umata T, Nakano F, Shishido Y, Toyoda T, Itai A, Kimura H, Mekada E. Structure-function analysis of the diphtheria toxin receptor toxin binding site by site-directed mutagenesis. J Biol Chem. 1997;272: 27084–90.
33. Fu S, Tong X, Cai C, Zhao Y, Wu Y, Li Y, Xu J, Zhang XC, Xu L, Chen W, Rao Z. The structure of tumor endothelial marker 8 (TEM8) extracellular domain and implications for its receptor function for recognizing anthrax toxin. PLoS One. 2010;5:e11203.
34. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell. 2015;163:1515–26.
35. Lu S, Wang J. The resistance mechanisms of proteasome inhibitor bortezomib. Biomark Res. 2013;1:13.
36. Franke NE, Niewerth D, Assaraf YG, van Meerloo J, Vojtekova K, van Zantwijk CH, Zweegman S, Chan ET, Kirk CJ, Geerke DP, et al. Impaired bortezomib binding to mutant beta5 subunit of the proteasome is the underlying basis for bortezomib resistance in leukemia cells. Leukemia. 2012;26:757–68.
37. Wacker SA, Houghtaling BR, Elemento O, Kapoor TM. Using transcriptome sequencing to identify mechanisms of drug action and resistance. Nat Chem Biol. 2012;8:235–7.
38. Murugan RN, Park JE, Kim EH, Shin SY, Cheong C, Lee KS, Bang JK. Plk1-targeted small molecule inhibitors: molecular basis for their potency and specificity. Mol Cells. 2011;32:209–20.
39. Richardson CD, Ray GJ, DeWitt MA, Curie GL, Corn JE. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat Biotechnol. 2016;34:339–44.
40. de Wilt LH, Jansen G, Assaraf YG, van Meerloo J, Cloos J, Schimmer AD, Chan ET, Kirk CJ, Peters GJ, Kruyt FA. Proteasome-based mechanisms of intrinsic and acquired bortezomib resistance in non-small cell lung cancer. Biochem Pharmacol. 2012;83:207–17.
41. Suzuki E, Demo S, Deu E, Keats J, Arastu-Kapur S, Bergsagel PL, Bennett MK, Kirk CJ. Molecular mechanisms of bortezomib resistant adenocarcinoma cells. PLoS One. 2011;6:e27996.
42. Kothe M, Kohls D, Low S, Coli R, Rennie GR, Feru F, Kuhn C, Ding YH. Selectivity-determining residues in Plk1. Chem Biol Drug Des. 2007;70:540–6.
43. Gasperini M, Findlay GM, McKenna A, Milbank JH, Lee C, Zhang MD, Cusanovich DA, Shendure J. CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. Am J Hum Genet. 2017;101:192–205.
44. Shi J, Wang E, Milazzo JP, Wang Z, Kinney JB, Vakoc CR. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol. 2015;33:661–7.
45. Garst AD, Bassalo MC, Pines G, Lynch SA, Halweg-Edwards AL, Liu R, Liang L, Wang Z, Zeitoun R, Alexander WG, Gill RT. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nat Biotechnol. 2017;35:48–55.

46. Zhu S, Zhou Y, Wei W. Genome-wide CRISPR/Cas9 screening for high-throughput functional genomics in human cells. Methods Mol Biol. 2017; 1656:175–81.
47. Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD, Sanjana NE, Zhang F. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc. 2017;12:828–63.
48. Wei W, Lu Q, Chaudry GJ, Leppla SH, Cohen SN. The LDL receptor-related protein LRP6 mediates internalization and lethality of anthrax toxin. Cell. 2006;124:1141–54.
49. Louie GV, Yang W, Bowman ME, Choe S. Crystal structure of the complex of diphtheria toxin with an extracellular fragment of its receptor. Mol Cell. 1997;1:67–78.
50. Schrader J, Henneberg F, Mata RA, Tittmann K, Schneider TR, Stark H, Bourenkov G, Chari A. The inhibition mechanism of human 20S proteasomes enables next-generation inhibitor design. Science. 2016;353: 594–8.
51. Groll M, Berkers CR, Ploegh HL, Ovaa H. Crystal structure of the boronic acid-based proteasome inhibitor bortezomib in complex with the yeast 20S proteasome. Structure. 2006;14:451–6.
52. Huber EM, Heinemeyer W, Groll M. Bortezomib-resistant mutant proteasomes: structural and biochemical evaluation with carfilzomib and ONX 0914. Structure. 2015;23:407–17.
53. Eads JC, Scapin G, Xu Y, Grubmeyer C, Sacchettini JC. The crystal structure of human hypoxanthine-guanine phosphoribosyltransferase with bound GMP. Cell. 1994;78:325–34.
54. Zhang X, Yue D, Wang Y, Zhou Y, Liu Y, Qiu Y, Tian F, Yu Y, Zhou Z, Wei W. PASTMUS: mapping functional elements at single amino acid resolution in human cells. Seq Read Arch. 2019; Dataset. https://www.ncbi.nlm.nih.gov/sra/PRJNA590617.
55. Zhang X, Yue D, Wang Y, Zhou Y, Liu Y, Qiu Y, Tian F, Yu Y, Zhou Z, Wei W. PASTMUS: mapping functional elements at single amino acid resolution in human cells. Bitbucket Softw. https://bitbucket.org/WeiLab/pastmus.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC