

Guide reselection expands canonical PAM candidates in an HGD editing resource for alkaptonuria

William Zane Duncan

Foothill College

Annie de Paula Bento

Foothill College

Laura Nadia Cabrera

Foothill College

Quoc Huy Hoang

Foothill College

Maddison Lila Kahn

Foothill College

Zhuolang Li

Foothill College

Raj Mathur

Foothill College

Malaika Theresa Mekkattuparamban

Foothill College

Cristian Emanuel Vasquez

Foothill College

Yuritzia Alexandra Zaragoza

Foothill College

Giorgio Lagna

lagnagiorgio@fhda.edu

Foothill College

Research Article

Keywords: alkaptonuria, HGD, prime editing, base editing, guide selection, AI-assisted research

Posted Date: September 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11097253/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Rare monogenic diseases offer defined molecular targets, but sequence-level editing options require explicit selection and verification. Using artificial intelligence assisted analysis and public tools within a community-college research cohort, we assembled an editing resource for the homogentisate 1,2-dioxygenase gene (HGD) in alkaptonuria. The inventory contained 278 variants and represented 268 of 271 eligible records in a defined ClinVar census. We compared filtering fixed prime-editing designs with re-selection requiring canonical NGG protospacer-adjacent motifs at both guide targets. Re-selection preserved 93 compatible configurations and added 140, all using a second nick without predicted edited-allele discrimination (PE3), while retaining 92 predicted allele-discriminating configurations (PE3b). Under a shared genomic count screen including NGG, NAG and NGA sites, 70 preserved and 79 added configurations remained. With an unchanged base-editing arm, combined representation increased from 110 to 188 variants, or 106 to 180 of 271 eligible ClinVar records. Preserving fixed configurations makes nonnegative gain structural; a 33-setting threshold grid quantified additional representation, with gains of 2–121. Conservative base-window exclusions and a prime-only control retained the increase. Sequence reconstruction and independent PrimeDesign comparisons checked component conventions. The result identifies nicking-guide selection as a source of additional canonical candidates and motivates planned tests of editing activity, product purity and restoration of HGD function.

Introduction

Each rare disease affects a small population, yet rare diseases collectively impose a substantial health burden (Nguengang Wakap et al. 2020). Within this diverse group, monogenic disorders connect disease biology to a defined molecular defect. Base and prime editing create opportunities to act on that defect, but identifying a disease-causing variant is only the beginning of an experimental programme. Investigators must determine which correction strategies are available, how candidate sequences were selected and what evidence supports taking them into biological tests.

Advances in artificial intelligence (AI) tools offer a way to engage more researchers in this preparatory work. Domain-specific systems such as CRISPR-GPT already assist gene-editing workflows (Qu et al. 2026). Assistance with scripting, data organization and explanation can make public genomic resources more approachable, provided that outputs are checked against reference sequences and established tools. We explored this opportunity within Foothill College, a California community college, through its Mathematics, Engineering, and Science Achievement (MESA) program and Emerging Technologies Institute (ETI). Across successive academic quarters, undergraduate researchers joined the cohort at different stages and worked with a PhD advisor to verify AI-assisted analyses. The project asked whether this setting could produce an inspectable, disease-focused research resource with a clear path to experimental follow-up.

Alkaptonuria provides a strong biological focus for such a feasibility study. It is a rare autosomal recessive systemic metabolic disorder caused by deficiency of homogentisate 1,2-dioxygenase, encoded

by *HGD*. Accumulation of homogentisic acid (HGA) leads to ochronosis and progressive connective-tissue disease (Phornphutkul et al. 2002). Its defined enzyme defect and measurable metabolite connect sequence correction to functional endpoints. Mouse studies establish the importance of hepatic *HGD* activity in controlling circulating HGA, giving a rationale for liver-directed approaches (Hughes et al. 2019). Nitisinone reduces HGA production, with clinical benefit demonstrated in SONIA 2 (Ranganath et al. 2020) and an FDA indication for reducing urinary HGA in adults with alkaptonuria (FDA 2025). Restoration of *HGD* activity is a complementary research objective. The early clinical use of patient-specific liver-directed base editing in another rare metabolic disorder illustrates the broader translational opportunity, without establishing efficacy for alkaptonuria (Musunuru et al. 2025).

Base editors install specified nucleotide conversions (Komor et al. 2016; Gaudelli et al. 2017), while prime editing can encode substitutions and small insertions or deletions (Anzalone et al. 2019). Both depend on guide placement, and the protospacer-adjacent motif (PAM) recognized by the Cas domain helps define the available targets. Engineered Cas9 variants expand this range and also expand the genomic sites relevant to specificity assessment (Walton et al. 2020). Established design resources already address pathogenic variants, including *HGD* (Hsu et al. 2021; Morris et al. 2021), and disease-focused and multimodal tools support further candidate analysis (Kaukonen et al. 2022; Maxim et al. 2026). Our contribution is a within-*HGD* comparison of fixed and reselected paired guides under shared screening rules, with a defined census and reproducible component records. The earlier Morris resource already contains *HGD* prime-editing and PE3b components; its 28 *HGD* identifiers are not directly comparable with our variant count because the source snapshots and units differ. Supplementary Methods compares the scope of these resources without treating them as an accuracy ranking.

The central comparison asks how much candidate availability depends on selection itself. Filtering an existing design by a PAM requirement can discard a locus even when another compatible design is available. We therefore compared fixed-design filtering with re-selection constrained to canonical NGG PAMs, assessed how off-target PAM inclusion changes retention, and held the base-editing arm constant. This connects the resource's scientific purpose to its research setting: a concrete design question that can be investigated through public data, computational analysis and documented verification.

Materials and methods

Variant census and candidate design

The analytical census was an *HGD* ClinVar query retrieved on 1 September 2026, cross-checked against the 31 August 2026 variant-summary release and deduplicated by VariationID (Landrum et al. 2018). Reference annotation used GRCh38, NM_000187.4 and MANE Select transcript ENST00000283871. Eligible sequence changes were precisely delimited, confined to *HGD* and no larger than 50 bp. We included aggregate pathogenic or likely pathogenic records and separately tracked conflicting records with at least one criteria-based pathogenic or likely pathogenic submission, no criteria-based benign or

likely benign submission, and gnomAD allele frequency below 0.4 per cent. This inclusion rule did not reclassify clinical variants. A crosswalk related census records to the selected inventory; exclusions and legacy-coordinate adjudications are supplied in Supplementary Methods.

Prime-editing selection used reconstructed variant and intended-correction alleles at 237 variant-specific loci. Each modeled allele contains the nominated variant in a reference background; individual haplotypes and compatibility with the second allele were not modeled. Prime-editing guide RNA (pegRNA) ranking prioritized PAM tier, nick-to-edit distance, template length and deterministic sequence tie-breaking. The second guide nicks the opposite strand. Selection preferred predicted PE3b configurations, intended to discriminate the corrected allele, before PE3 alternatives without that predicted discrimination and with an absolute nick separation of 40–90 nt measured on the modeled variant allele (Anzalone et al. 2019; Doman et al. 2022). Candidate nomination excluded five consecutive T residues in either prime-guide spacer, separately from the later four-U template sensitivity. These rankings did not use genome-wide off-target counts. The reproducible analysis starts from retained candidate pools and selected nicking configurations; it reranks pegRNA candidates without replaying the complete historical nick-guide enumeration. We compared retained fixed configurations with re-selection requiring NGG at both intended targets, using a 41 nt template cap and a separate 50 nt sensitivity.

Independent reconstruction distinguished the intended DNA flap from its reverse-complement RNA reverse-transcription template (RTT), reported 5' to 3' in DNA-letter notation. Final components require an RTT of at least 10 nt and at least 7 nt of downstream homology. We derived a 13 nt primer-binding sequence (PBS) and the modeled RTT-then-PBS extension. All ten retained fixed/PAM-model and template-cap selections were repaired and their candidate pools reranked. Full scaffold, linker and expression-construct specifications remain to be selected before experimental use; the release supplies spacer and template components with derived PBS sequences.

The base-editing arm modeled ABE8e or BE4 chemistry with SpRY targeting assumptions and protospacer positions 4–8 numbered from the PAM-distal end (Komor et al. 2017; Richter et al. 2020; Walton et al. 2020). For potential bystanders, the most severe coding or splice consequence determined retention. Missense, stop-gained, stop-lost and canonical splice-dinucleotide changes failed the adopted rule. A wider splice-region annotation alone was retained and examined separately under a stricter sensitivity. These rules describe possible sequence consequences within the adopted positions 4–8 window, rather than their frequencies. ABE8e can act outside that window (Zhao et al. 2024; Evanoff et al. 2026). We therefore added conservative substrate-presence exclusions for ABE8e guides containing adenines at positions 3, 9, 10 or 11, and a second check including position 12, after reconstructing each retained guide on its variant allele. These checks leave BE4 and intended-correction eligibility unchanged and do not classify the extra substrates as harmful or predict their activity.

Matched screening and verification

Specificity screening reused retained genomic Hamming-distance counts across chromosomes 1–22, X and Y, with one nominated occurrence per spacer in the *HGD* window excluded positionally. The principal rule permitted no additional exact match, at most one site with exactly one mismatch and at most five with exactly two mismatches; both required prime guides had to pass. Additional-perfect-match exclusion follows a common uniqueness principle (Perez et al. 2017; Schmidt et al. 2025). The particular nonzero allowances are internal screening conventions without experimental calibration for these editor configurations. Counts were not weighted by mismatch position or relative PAM activity, and omitted bulges and population variation (Haeussler et al. 2016).

The principal comparison required NGG at both intended prime-editing targets and included NGG, NAG and NGA in the off-target count model. The base arm used the NNN model with identical bystander rules in both conditions. Each variant was counted once in the union. In this inventory, re-selection preserved all already-canonical fixed pairs. Their screened sets are therefore nested under a common deterministic rule; the sensitivity analyses measure how many additional configurations remain, rather than provide independent tests of algorithmic superiority. Sensitivities varied PAM inclusion, template cap, wider splice-region exclusion, RNA-component four-U flags and 33 combinations of one- and two-mismatch limits. Detailed reference scope, target-exclusion checks and ambiguity handling are reported in Supplementary Methods.

Verification reconstructed all 80,544 retained candidate records and 2,364 selected records across ten configurations; these totals include recurring designs. Independent implementations compared variant sets across 24 scenarios and the 33-setting threshold grid. An external PrimeDesign benchmark used eight frozen canonical-PAM single-nucleotide cases spanning pegRNA strand and predicted PE3/PE3b class, with matched sequence constraints and a pinned software commit (Hsu et al. 2021). A separately frozen follow-up examined 15 selected configurations, including 14 length-changing variants, and four cases without an admissible NGG pegRNA under the retained rules. Two primary nonmatches prompted a separately recorded equivalent-annotation diagnostic. Nick-distance comparisons distinguished the variant-allele PE3 selection frame from the intended-allele PE3b frame and the separately reported corrected-allele distance. The analyses are retrospective and exploratory. Verification reused retained count evidence and did not repeat genome-wide scans.

Public tools and cohort verification

Student-led verification included checking reference alleles and coordinates against UCSC hg38 sequences, comparing designs with pegFinder and the open-source PrimeDesign command-line implementation, and examining guide-level outputs from CRISPOR (Chow et al. 2021; Hsu et al. 2021; Haeussler et al. 2016). Retained packet records document submitted comparisons, discrepancies and subsequent corrections across the cohort. Supplementary Methods lists the tools, the evidence retained for each and the scope of additional exploratory folding checks.

Claude (Anthropic) assisted with scripting, preparation of verification tasks and analysis; OpenAI Codex assisted with source inspection, independent computational checks, figures and manuscript drafting and revision. Retained commit attributions name “Claude Fable 5” and “Claude Opus 5”; recovered Codex client records name gpt-6-astra. These records identify recorded labels, not independently authenticated models for every execution. Historical gaps and the evidence supporting each identifier are documented in Supplementary Methods and Online Resource 2. AI output was treated as material requiring verification against reference sequences and public tools. Writing assistance extended beyond copy editing, and the authors take responsibility for the final manuscript. The study did not compare AI systems or measure the educational effect of participation.

Relation to published variant series

Retained variants were mapped to six published *HGD* series, preserving their different ascertainment schemes and observational units (Uyguner et al. 2003; Vilboux et al. 2009; Zatkova et al. 2012; Nemethova et al. 2016; Khalil et al. 2021; Abdelkhalek et al. 2023). Source-specific results are primary. For ambiguous multivariant probands, differences were calculated within matched causal-assignment scenarios. Conditional pooled arithmetic, with the Turkish series excluded pending full-text adjudication, is confined to Supplementary Table S5 and Supplementary Methods; it does not estimate population coverage or patient eligibility.

Results

A resource spanning a defined HGD census

The ClinVar query contained 665 records, of which 302 met the disease-associated inclusion rule: 294 aggregate pathogenic or likely pathogenic records and eight qualifying conflicting records. Sequence-change eligibility retained 271 records, including 263 pathogenic or likely pathogenic records and all eight qualifying conflicting records. The selected inventory contained 278 distinct variants: 237 prime-editing loci and 125 base-editing variants, with 84 shared. It represented 268 of 271 eligible census records (98.9 per cent), alongside ten inventory variants outside that census (Fig. 1; Table 1).

Table 1
Composition and scope of the resource

Quantity	Count	Unit and interpretation
ClinVar query	665	Distinct VariationIDs
Disease-associated census	302	294 pathogenic/likely pathogenic plus 8 qualifying conflicting records
Eligible sequence-change census	271	263 pathogenic/likely pathogenic plus 8 qualifying conflicting records
Eligible records in selected inventory	268	Intersection with the selected inventory; 3 gaps
Selected inventory	278	Distinct HGVS variants; includes 10 outside the eligible census
Prime-editing inventory	237	Selected paired-guide designs at distinct loci
Base-editing inventory	625	Candidate rows across 125 variants
Base rows after actionability and bystander rules	321	Candidate rows across 108 variants, before the principal count screen
Base rows after NNN count screening	90	Candidate rows across 40 variants
Counts refer to the fixed release analysed here. The ClinVar census and design inventory are different sets and their counts are not interchangeable.		

The three gaps were c.448del, c.435-1G > C and c.651_652del. They remain explicitly identified in the resource. A bounded follow-up found no NGG pegRNA for the first two under the tested caps and rules; the first-ranked pair for c.651_652del failed the principal count screen. The latter also had a historical candidate, so its inventory absence is not evidence of design impossibility. The separate exploratory gap analysis is provided in Supplementary Table S10 and does not change the frozen comparison.

Re-selection expands canonical targeting through a change in nicking strategy

Only 93 of the 237 fixed prime-editing designs had NGG at both intended targets. Of the 144 that did not, 140 failed this requirement at the nicking guide alone. Re-selection under the NGG requirement produced paired configurations at 233 loci (Fig. 2). All 93 already compatible configurations were preserved, including 92 predicted PE3b designs, and compatibility was recovered at 140 additional loci.

At every recovered locus, the pegRNA spacer stayed the same and the nicking-guide spacer changed. All 140 recovered configurations were PE3; 134 had previously used a non-NGG-compatible predicted PE3b configuration. For example, c.879 + 1G > A retained its GGG-PAM pegRNA and corrected template, while its TGA-PAM predicted PE3b nicking guide was replaced by a TGG-PAM PE3 guide. This paired

alternative passes the principal count screen; it does not establish editing activity at that locus. The reselected set therefore contained 92 predicted PE3b and 141 PE3 designs. Canonical targeting became available at more loci, with an explicit change in the nicking strategy to be examined experimentally. The 50 nt cap yielded 235 paired NGG configurations, reported in the supplement.

The gain persists when specificity screening includes additional PAMs

An NGG-only off-target model retained 90 fixed and 228 reselected prime-editing designs. Including NAG and NGA reduced those totals to 70 and 149 under the same count limits (Fig. 2; Table 2). The 149 comprised the same 70 predicted PE3b configurations retained from the fixed set plus 79 PE3 configurations. The principal screen retained 70 of 93 preserved pairs and 79 of 140 added pairs. Broadening the off-target PAM set removed 20 preserved and 59 added pairs from the NGG-only results. These fractions describe selected strata, not relative biological performance. Restricting prime designs to predicted PE3b leaves 70 in each condition (Supplementary Table S11).

Table 2
Matched comparison with canonical intended-site PAMs

Measure	Fixed designs	NGG re-selection
Prime-editing loci considered	237	237
Designs with NGG at both intended sites	93	233
Predicted PE3b configurations among these	92	92
Predicted PE3 configurations among these	1	141
Prime designs passing NGG-only off-target counts	90	228
Prime designs passing NGG/NAG/NGA off-target counts	70	149
Predicted PE3b among designs passing NGG/NAG/NGA	70	70
Predicted PE3 among designs passing NGG/NAG/NGA	0	79
Base-editing variants retained at NNN	40	40
Combined variants with broader prime off-target search	110	188
Eligible ClinVar records in that combined set	106	180
The template cap is 41 nt. Count limits are zero additional exact matches, one one-mismatch site and five two-mismatch sites per required guide. Base-editing screening and bystander rules are identical between columns. The combined union deducts overlap between the two editing arms. Specificity screening does not establish biological safety or editing activity.		

The constant base-editing arm retained 90 candidate rows across 40 variants. Combining this arm with the principal prime-editing sets increased the union from 110 to 188 variants, adding 78 and removing none. One newly retained prime-editing locus, c.879 + 1G > A, was already represented by base editing, accounting for the difference between 79 additional prime designs and 78 additional variants. Within the eligible census, representation increased from 106 of 271 records to 180 of 271; 176 of the latter carried aggregate pathogenic or likely pathogenic classifications.

Sequence checks and sensitivity analyses support the comparison

Re-selection represented more variants in all 33 examined threshold combinations, with gains ranging from 2 to 121. With the one-mismatch limit fixed at one, raising the two-mismatch limit from zero to five increased the fixed union from 8 to 110 variants and the reselected union from 12 to 188 (Fig. 3). The grid thus quantifies strictly positive additions whose magnitude depends on the screening rule. The absence of losses follows from preservation of the fixed canonical pairs, rather than from 33 independent demonstrations of superiority.

Sequence repair extended 63 of 237 fixed and 62 of 233 NGG-reselected templates at the 41 nt cap. Repaired components met the length and homology rules, and reranking preserved pegRNA winners in all ten retained selection tables. All 24 scenario variant sets and all 33 threshold comparisons were unchanged. Excluding repaired RNA templates with four-U runs retained unions of 110 and 181; extending that exclusion to RTT plus PBS retained 110 and 180. Additional splice-region and short-template restrictions also preserved the direction of the comparison (Supplementary Methods). The new ABE8e substrate-presence exclusions left 64 guide rows across 30 base-editing variants for positions 3–11, and 61 rows across 27 variants when position 12 was included. Combined sets were 100 versus 179 and 97 versus 176, respectively. Omitting the entire base arm retained the prime-only comparison of 70 versus 149. These conservative exclusions can discard benign or inactive substrates and do not certify a broader editing window (Supplementary Table S12).

In the eight-case PrimeDesign benchmark, all sampled pegRNA spacers, PAMs, strands, corrected RNA templates, PBS sequences and extensions agreed. Nicking-guide components, predicted PE3/PE3b classes and coordinate-adjusted nick offsets also agreed in all eight cases. This bounded agreement supports the sequence conventions used for the sampled substitutions and complements the broader internal reconstruction checks. The expanded primary benchmark recovered exact selected components in 13 of 15 configurations. Two duplication cases, c.184_187dup and c.357dup, were not emitted under the frozen edit annotation; a separate post hoc diagnostic using equivalent repeat-copy annotations recovered both, with identical complete variant and intended sequences. All 13 primary matched configurations agreed on template, PBS, extension, nicking components, predicted class and correctly interpreted distances. No emitted pegRNA in the four additional availability cases met all checked historical constraints (Supplementary Table S13). These results identify annotation-sensitive tool behavior without changing the primary 13/15 denominator or the original eight-case result.

Discussion

This feasibility study connects a rare-disease question to a computational resource that can guide experimental choices. Its main finding is actionable at the design stage: selecting alternatives within a targeting constraint can retain substantially more variants than filtering an already selected inventory. For *HGD*, NGG-constrained re-selection increased the combined set from 110 to 188 variants under the principal count rules. Because the compatible fixed pairs are preserved, no-loss under common screening is structural. The informative result is that 79 of 140 added canonical PE3 configurations pass the principal screen, with one already represented in the constant base arm. Threshold and window sensitivities quantify how those additions depend on the assumptions; sequence repair preserved the compared variant sets. The resource makes these relationships visible at the level of individual variants and paired guides.

The additional designs also reveal what should be tested next. Most fixed configurations lost NGG compatibility because of their preferred nicking guide, and all 140 recovered configurations used PE3. At 134 loci, this replaced a broader-PAM predicted PE3b strategy. PE3b is intended to favor nicking of the edited allele; PE3 and PE3b can differ in editing efficiency and unwanted products (Anzalone et al. 2019; Doman et al. 2022). The added canonical configurations therefore offer testable alternatives, rather than evidence that one nicking strategy is uniformly preferable. Screening choices matter independently: including NAG and NGA reduced retained designs while preserving the benefit of re-selection.

Alkaptonuria offers a coherent route from these sequence-level questions to disease biology. Correction can be evaluated through *HGD* activity and *HGA* metabolism, and the importance of hepatic enzyme activity provides a basis for considering liver-directed validation (Hughes et al. 2019). The resource complements genome-wide design databases by bringing a defined *HGD* census, paired alternatives, sequence provenance and sensitivity analyses into one disease-focused framework (Hsu et al. 2021; Morris et al. 2021). Mapping to published series adds context for selecting variants for study, while the heterogeneous source records remain unsuitable for estimating how many patients could benefit.

From computational feasibility to biological validation

We propose beginning with a small, informative panel: canonical-PAM predicted PE3b candidates retained in both analyses, newly available PE3 alternatives that test the principal selection trade-off, and base-editing candidates with favorable modeled bystander profiles. Within these classes, prioritization should combine unambiguous sequence specifications, relevance to documented disease variants and the practical availability of an appropriate assay. This approach concentrates initial effort on whether the added design options produce useful correction and whether the alternative strategies differ in product quality. It also keeps the first experiments manageable while preserving their connection to the main scientific finding.

The first stage would measure intended correction and unwanted sequence products in cells, followed by *HGD* activity and *HGA* handling in a suitable functional model. Candidates with reproducible

correction and functional improvement would then be considered for in vivo assessment of hepatic correction, HGD function, HGA, persistence and tolerability. Complete reagent specifications, species and target-sequence compatibility, and the experimental comparison must be established first. Where editor recognition architecture and nicking strategy both change, experiments compare whole configurations. The model, facilities and timetable remain to be specified; Supplementary Methods gives the proposed stages and decision criteria.

A research and educational contribution

The project developed through a continuing community-college cohort, supported by MESA and ETI, in which students joined at different points and contributed to a shared research question. AI assistance supported computational work, while public sequence and design tools gave students concrete ways to check outputs and report disagreements. Packet records preserve this verification activity across successive quarters. The resulting resource provides a practical connection between undergraduate research and a medical problem with consequences for patients and families. Its educational value here lies in the opportunity to participate in that work; learning gains, the causal contribution of AI assistance and clinical impact were not measured.

Several boundaries define the next stage. The genomic screen is an uncalibrated sequence-count model, not a measurement of editor activity or biological safety; empirical specificity studies show why homology searches and editing outcomes must be evaluated separately (Kim et al. 2020; Kwon et al. 2022; Liang et al. 2023). The screen also omits population variation and guide-independent editor effects (Doman et al. 2020; Cancellieri et al. 2023). The current release provides components rather than complete experimental constructs. The base-editing labels are conditional on modeled windows and consequences; actual construct-specific bystanders, individual sequence backgrounds and both disease alleles require further assessment. Planned biological follow-up will assess editing efficiency, product purity and restoration of HGD function. The full technical limitations and historical provenance are retained in Supplementary Methods so that investigators can evaluate and extend the resource.

Conclusion

AI-assisted analysis and cohort-based verification at a community college produced an inspectable computational resource for alkaptonuria research. The principal result is that guide re-selection expands canonical-PAM candidate availability, with a measurable change in nicking strategy and strong dependence on screening assumptions. The resource establishes a basis for choosing a focused experimental panel and testing whether these computational alternatives can restore HGD function.

Declarations

Funding. The Mathematics, Engineering, and Science Achievement (MESA) program and Emerging Technologies Institute (ETI) at Foothill College provided funding and support for this project.

Author contributions. William Zane Duncan, Annie de Paula Bento, Laura Nadia Cabrera, Quoc Huy Hoang, Maddison Lila Kahn, Zhuolang Li, Raj Mathur, Malaika Theresa Mekkattuparamban, Cristian Emanuel Vasquez and Yuritzia Alexandra Zaragoza contributed to the cohort's validation and verification of AI-assisted outputs using public resources and tools; retained packet records support the documented work. Giorgio Lagna supervised the research and serves as corresponding author. All authors reviewed and approved the manuscript.

Competing interests. The authors declare no competing interests.

Ethics. This study analysed public variant records, published genotype descriptions and computational research outputs. It reports no newly performed cell or animal experiments and no assessment of student learning outcomes. No additional ethics approval was required for this public-data and computational study.

Use of artificial intelligence. Claude assisted with research scripting, preparation of verification tasks, design reconstruction and internal analysis. OpenAI Codex assisted with source inspection, independent computational checks, figure preparation and manuscript drafting and revision. Writing assistance extended beyond copy editing. Public-tool comparisons, computational checks and retained packet records support verification of the reported work. Historical model-log limitations are detailed in Supplementary Methods. The corresponding author has reviewed the AI-assisted content and confirms responsibility for its accuracy. Responsibility for the final manuscript rests with its authors.

Data and code availability. The computational resource, including code, retained inputs, corrected component sequences, candidate-pool controls, benchmark records and figure source data, is supplied as Online Resource 2. The accompanying Zenodo resource has the archival identifier 10.5281/zenodo.22737227. Original project code is licensed under MIT; original project data, figures and documentation are licensed under CC BY 4.0, with third-party terms retained. Source material excluded from redistribution is identified with retrieval instructions.

Acknowledgement

We acknowledge funding and program support from MESA and ETI at Foothill College. We thank Sophia Kim and Zachary Cembellin for their support and leadership, and Maryam Nasibi, Mireya Ayala-Valdez and Johara Ahmed for helping establish the project during its early phase. We also thank Megan J. Agajanian and Konstantin Kalaitzidis.

References

1. Abdelkhalek ZS, Mahmoud IG, Omair H, Abdulhay M, Elmonem MA (2023) Homogentisate 1,2-dioxygenase (HGD) gene variants in young Egyptian patients with alkaptonuria. *Sci Rep* 13:14374. <https://doi.org/10.1038/s41598-023-41200-7>

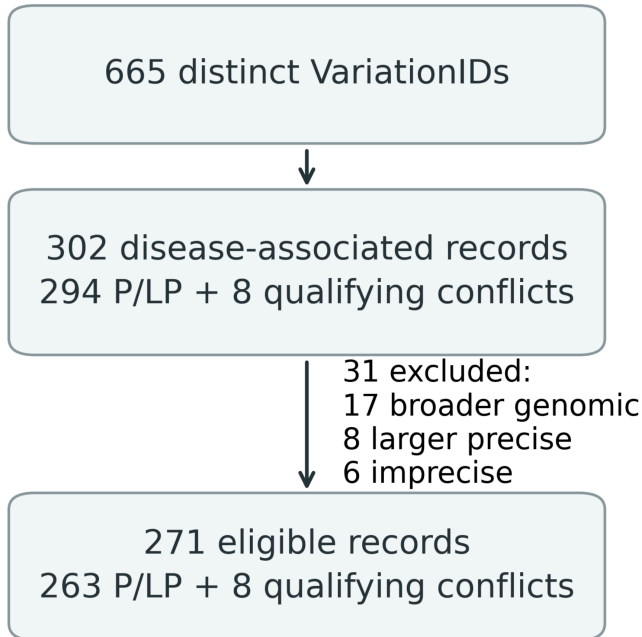
2. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM et al (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576:149–157.
<https://doi.org/10.1038/s41586-019-1711-4>
3. Cancellieri S, Zeng J, Lin LY, Tognon M, Nguyen MA, Lin J et al (2023) Human genetic diversity alters off-target outcomes of therapeutic gene editing. *Nat Genet* 55:34–43.
<https://doi.org/10.1038/s41588-022-01257-y>
4. Chow RD, Chen JS, Shen J, Chen S (2021) A web tool for the design of prime-editing guide RNAs. *Nat Biomed Eng* 5:190–194. <https://doi.org/10.1038/s41551-020-00622-8>
5. Doman JL, Raguram A, Newby GA, Liu DR (2020) Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. *Nat Biotechnol* 38:620–628.
<https://doi.org/10.1038/s41587-020-0414-6>
6. Doman JL, Sousa AA, Randolph PB, Chen PJ, Liu DR (2022) Designing and executing prime editing experiments in mammalian cells. *Nat Protoc* 17:2431–2468. <https://doi.org/10.1038/s41596-022-00724-4>
7. Evanoff M, Korpai S, Krill ZD, Cowan QT, Komor AC (2026) Precise, minimally evolved adenine base editors generated through mutation reversion analysis. *Nat Biotechnol*. Published online 18 March 2026. <https://doi.org/10.1038/s41587-026-03045-z>
8. FDA (2025) Harliku nitisinone prescribing information. Revised June 2025.
https://www.accessdata.fda.gov/drugsatfda_docs/label/2025/209449s018lbl.pdf
9. Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, Liu DR (2017) Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* 551:464–471.
<https://doi.org/10.1038/nature24644>
10. Haeussler M, Schöning K, Eckert H, Eschstruth A, Mianné J, Renaud JB et al (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* 17:148. <https://doi.org/10.1186/s13059-016-1012-2>
11. Hsu JY, Grünwald J, Szalay R, Shih J, Anzalone AV, Lam KC et al (2021) PrimeDesign software for rapid and simplified design of prime editing guide RNAs. *Nat Commun* 12:1034.
<https://doi.org/10.1038/s41467-021-21337-7>
12. Hughes JH, Liu K, Plagge A, Wilson PJM, Sutherland H, Norman BP et al (2019) Conditional targeting in mice reveals that hepatic homogentisate 1,2-dioxygenase activity is essential in reducing circulating homogentisic acid and for effective therapy in the genetic disease alkaptonuria. *Hum Mol Genet* 28:3928–3939. <https://doi.org/10.1093/hmg/ddz234>
13. Kaukonen M, McClements ME, MacLaren RE (2022) CRISPR DNA base editing strategies for treating retinitis pigmentosa caused by mutations in rhodopsin. *Genes (Basel)* 13:1327.
<https://doi.org/10.3390/genes13081327>
14. Khalil R, Ali D, Mwafi N, Alsaraireh A, Obeidat L, Albsoul E et al (2021) Variant analysis of alkaptonuria families with significant founder effect in Jordan. *Biomed Res Int* 2021:1515641.
<https://doi.org/10.1155/2021/1515641>

15. Kim DY, Moon SB, Ko JH, Kim YS, Kim D (2020) Unbiased investigation of specificities of prime editing systems in human cells. *Nucleic Acids Res* 48:10576–10589. <https://doi.org/10.1093/nar/gkaa764>
16. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533:420–424. <https://doi.org/10.1038/nature17946>
17. Komor AC, Zhao KT, Packer MS, Gaudelli NM, Waterbury AL, Koblan LW et al (2017) Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci Adv* 3:eaao4774. <https://doi.org/10.1126/sciadv.aao4774>
18. Kwon J, Kim M, Bae S, Jo A, Kim Y, Lee JK (2022) TAPE-seq is a cell-based method for predicting genome-wide off-target effects of prime editor. *Nat Commun* 13:7975. <https://doi.org/10.1038/s41467-022-35743-y>
19. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S et al (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 46:D1062–D1067. <https://doi.org/10.1093/nar/gkx1153>
20. Liang SQ, Liu P, Ponnienselvan K, Suresh S, Chen Z, Kramme C et al (2023) Genome-wide profiling of prime editor off-target sites in vitro and in vivo using PE-tag. *Nat Methods* 20:898–907. <https://doi.org/10.1038/s41592-023-01859-2>
21. Maxim DS, Sostena J, Johnson NS, Wu DW, Charu V, Carter JN, Anand S, Church GM, Bhalla V (2026) Designing genome editing experiments with EditABLE. *Genome Biol* 27:214. <https://doi.org/10.1186/s13059-026-04095-x>
22. Morris JA, Rahman JA, Guo X, Sanjana NE (2021) Automated design of CRISPR prime editors for 56,000 human pathogenic variants. *iScience* 24:103380. <https://doi.org/10.1016/j.isci.2021.103380>
23. Musunuru K, Grandinette SA, Wang X, Hudson TR, Briseno K, Berry AM et al (2025) Patient-specific in vivo gene editing to treat a rare genetic disease. *N Engl J Med* 392:2235–2243. <https://doi.org/10.1056/NEJMoa2504747>
24. Nemethova M, Radvanszky J, Kadasi L, Ascher DB, Pires DE, Blundell TL et al (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. *Eur J Hum Genet* 24:66–72. <https://doi.org/10.1038/ejhg.2015.60>
25. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, Rath A (2020) Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 28:165–173. <https://doi.org/10.1038/s41431-019-0508-0>
26. Perez AR, Pritykin Y, Vidigal JA, Chhangawala S, Zamparo L, Leslie CS, Ventura A (2017) GuideScan software for improved single and paired CRISPR guide RNA design. *Nat Biotechnol* 35:347–349. <https://doi.org/10.1038/nbt.3804>
27. Phornphutkul C, Introne WJ, Perry MB, Bernardini I, Murphey MD, Fitzpatrick DL et al (2002) Natural history of alkaptonuria. *N Engl J Med* 347:2111–2121. <https://doi.org/10.1056/nejmoa021736>

28. Qu Y, Huang K, Yin M, Zhan K, Liu D, Yin D et al (2026) CRISPR-GPT for agentic automation of gene-editing experiments. *Nat Biomed Eng* 10:245–258. <https://doi.org/10.1038/s41551-025-01463-z>
29. Ranganath LR, Psarelli EE, Arnoux JB, Braconi D, Briggs M, Bröijersén A et al (2020) Efficacy and safety of once-daily nitisinone for patients with alkaptonuria (SONIA 2): an international, multicentre, open-label, randomised controlled trial. *Lancet Diabetes Endocrinol* 8:762–772. [https://doi.org/10.1016/s2213-8587\(20\)30228-x](https://doi.org/10.1016/s2213-8587(20)30228-x)
30. Richter MF, Zhao KT, Eton E, Lapinaite A, Newby GA, Thuronyi BW et al (2020) Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat Biotechnol* 38:883–891. <https://doi.org/10.1038/s41587-020-0453-z>
31. Schmidt H, Zhang M, Chakarov D, Bansal V, Mourelatos H, Sánchez-Rivera FJ et al (2025) Genome-wide CRISPR guide RNA design and specificity analysis with GuideScan2. *Genome Biol* 26:41. <https://doi.org/10.1186/s13059-025-03488-8>
32. Uyguner O, Goicoechea de Jorge E, Cefle A, Baykal T, Kayserili H, Cefle K et al (2003) Molecular analyses of the HGO gene mutations in Turkish alkaptonuria patients suggest that the R58fs mutation originated from central Asia and was spread throughout Europe and Anatolia by human migrations. *J Inherit Metab Dis* 26:17–23. <https://doi.org/10.1023/a:1024063126954>
33. Vilboux T, Kayser M, Introne W, Suwannarat P, Bernardini I, Fischer R et al (2009) Mutation spectrum of homogentisic acid oxidase (HGD) in alkaptonuria. *Hum Mutat* 30:1611–1619. <https://doi.org/10.1002/humu.21120>
34. Walton RT, Christie KA, Whittaker MN, Kleinstiver BP (2020) Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* 368:290–296. <https://doi.org/10.1126/science.aba8853>
35. Zatkova A, Sedlackova T, Radvansky J, Polakova H, Nemethova M, Aquaron R et al (2012) Identification of 11 novel homogentisate 1,2 dioxygenase variants in alkaptonuria patients and establishment of a novel LOVD-based HGD mutation database. *JIMD Rep* 4:55–65. https://doi.org/10.1007/8904_2011_68
36. Zhao N, Zhou J, Tao T, Wang Q, Tang J, Li D et al (2024) Evolved cytidine and adenine base editors with high precision and minimized off-target activity by a continuous directed evolution system in mammalian cells. *Nat Commun* 15:8140. <https://doi.org/10.1038/s41467-024-52483-3>

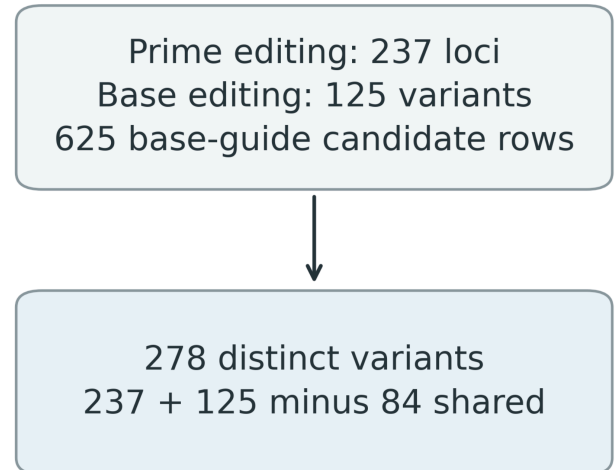
Figures

a ClinVar census



P/LP: pathogenic or likely pathogenic.
Counts describe different units; they are not additive.

b Selected inventory



c Census / inventory mapping

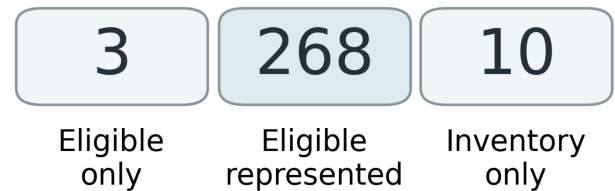
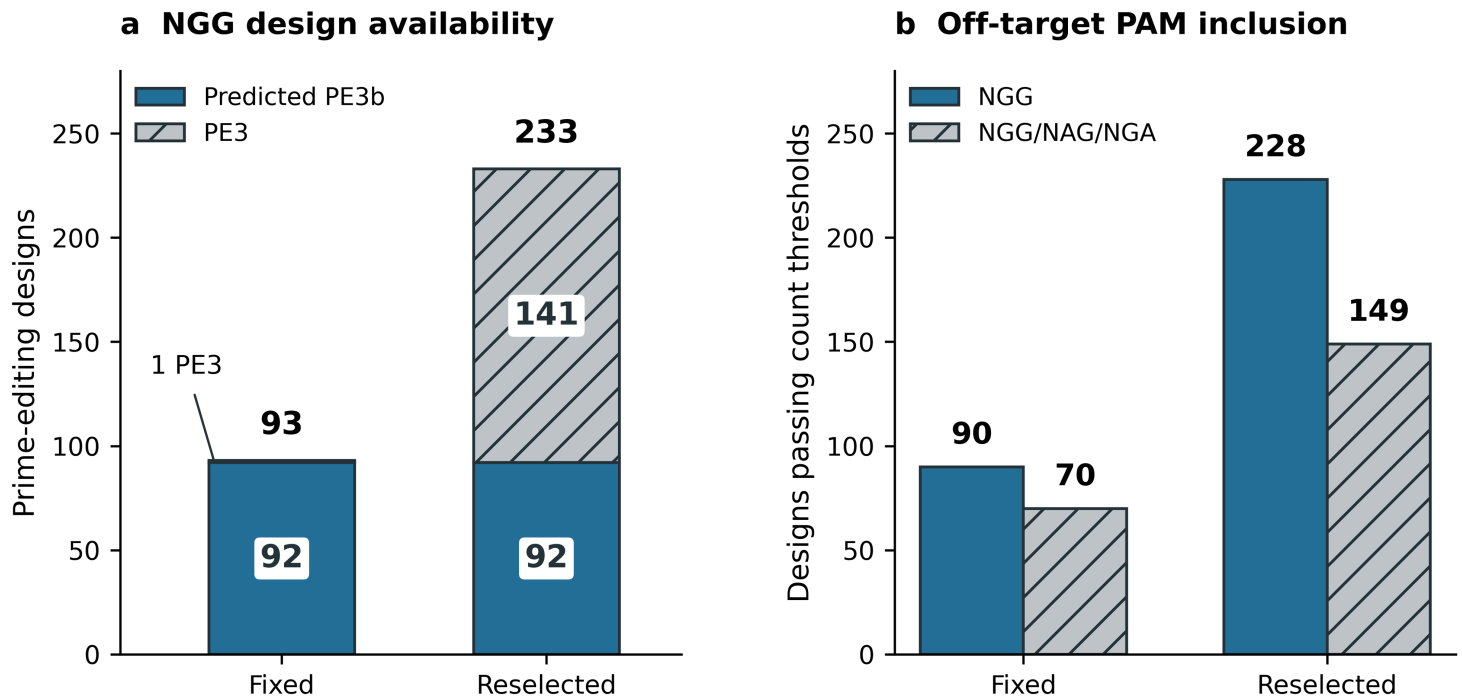


Figure 1

Scope of the HGD resource

Relationship between the ClinVar query, the eligible disease-associated census and the selected design inventory. The inventory comprises 278 distinct variants represented by 237 prime-editing loci and 125 base-editing variants, with 84 shared variants. It intersects 268 of 271 eligible ClinVar records. The three missing census records (VariationIDs 2112733, 2850560 and 643931) and ten additional inventory variants are reported separately. Candidate nomination does not imply passage of subsequent screening or complete reagent specifications.

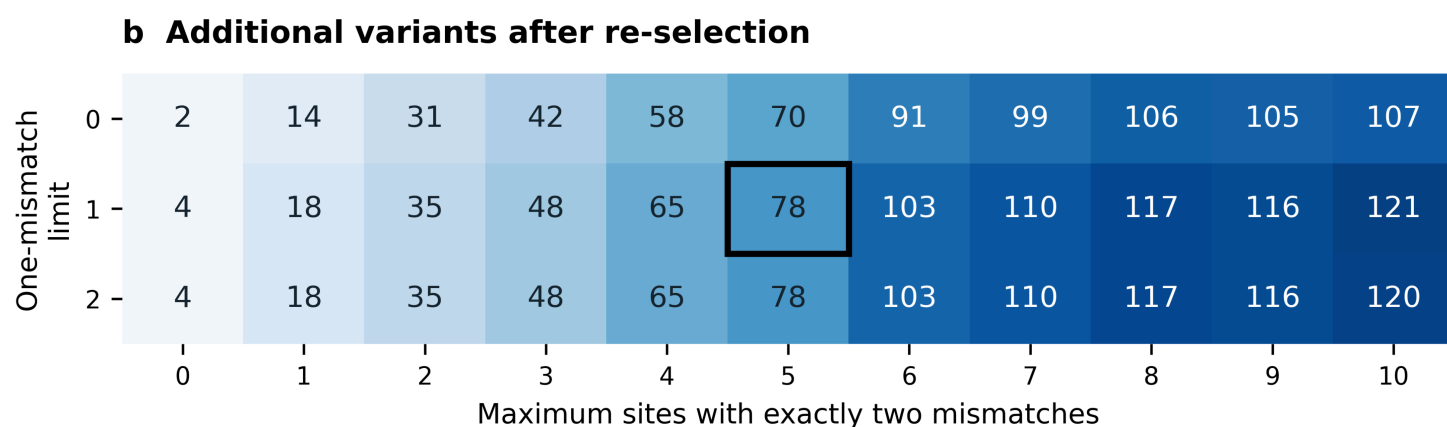
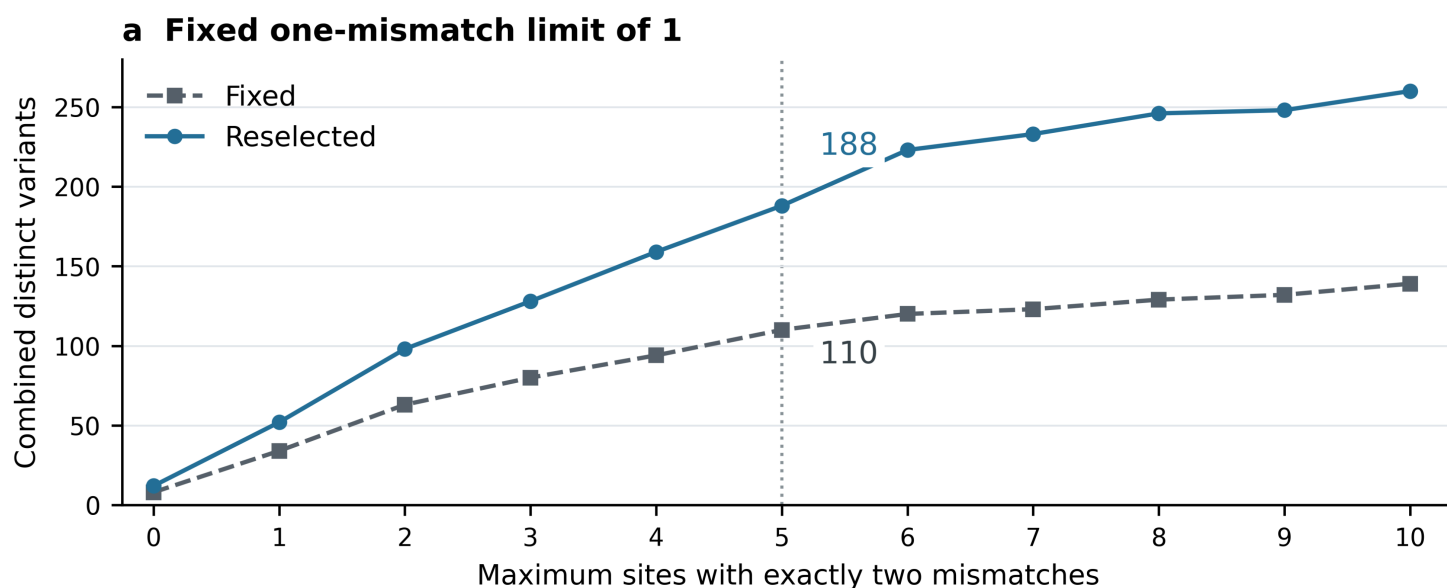


Both intended-site PAMs: NGG | Opportunity set: 237 loci | Template cap: 41 nt

Figure 2

Guide selection and specificity screening

(a) NGG-compatible prime-editing designs in the fixed and reselected sets, separated into predicted PE3b and PE3 configurations. (b) Designs passing the reference mismatch-count limits when the off-target search includes NGG alone or NGG, NAG and NGA. Both intended-site PAMs must be NGG in every comparison. The 237-locus opportunity set and 41 nt template cap are shared; repaired templates are used in both conditions; four loci have no reselected NGG design at this cap. Counts describe candidate designs, not editing efficiencies. Base editing is omitted from these panels to isolate the prime-editing comparison.



Reference cell outlined. No additional exact matches permitted.
 Prime off-target PAMs: NGG/NAG/NGA; constant base arm: NNN.

Figure 3

Threshold dependence of combined variant retention

(a) Numbers of represented variants under the fixed and reselected NGG designs across two-mismatch limits, holding the one-mismatch limit at one. (b) Additional variants represented after re-selection across the full threshold grid. The additional exact-match limit is zero throughout. Prime-editing off-target counts include NGG, NAG and NGA; the base-editing arm uses NNN and the same corrected bystander rule in both conditions. Each variant is counted once in the union. The marked reference setting permits one one-mismatch site and five two-mismatch sites. The fixed canonical configurations are preserved in the reselected set, so common filtering guarantees no-loss; the grid shows the number of additional represented variants. These are deterministic counts, not independent replicates; no sampling error bars are implied.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1.pdf](#)
- [ESM2.zip](#)
- [Graphicalabstract.pdf](#)