

Supplementary Methods

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

Journal: Biochemical Genetics

Authors: 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, Yuritzia Alexandra Zaragoza, Giorgio Lagna

Affiliation: Mathematics, Engineering, and Science Achievement (MESA) and Emerging Technologies Institute (ETI), Foothill College, 12345 El Monte Road, Los Altos Hills, CA 94022, USA.

Corresponding author: Giorgio Lagna; lagnagiorgio@fhda.edu

Online Resource 1

This supplement accompanies the manuscript identified above. The frozen principal results are unchanged. Separately labeled analyses decompose the preserved and added guide sets and test conservative exclusions of additional ABE8e substrates. The original eight-case external benchmark remains distinct from the expanded follow-up benchmark. Proposed biological experiments have not been performed.

Paths beginning with `sequence_audit/` or `manuscript_audit/`, and the root `reproduce.py`, refer to the accompanying computational resource, `aku_hgd_resource_v8_3`. Its preserved computational source is the earlier release candidate `public_candidate_v4_1`. Historical packet paths beginning with `recovered_from_drive/`, `rebuild/` or `crispor_rerun/` refer to the original project archive and are not public repository paths.

Reference and variant reconstruction

The analysis preserves the historical ClinVar census and selected inventory. The explicit-coordinate reference is the GRCh38 plus-strand interval [120615000, 120695000) in `sequence_audit/inputs/rebuild/DIAG/p65/HGD_region.fa`. The shorter p60 sequence is exactly contained in it, with zero-based start 120627999 and length 54601. Its filename uses a one-based interval, chr3:120628000–120682600. V2 treated that start as zero-based while shifting its extraction formulas by one; those errors canceled numerically. They are corrected in v3 documentation and code.

Exons of ENST00000283871 are ordered in transcript direction on the minus strand. The spliced sequence is aligned to the retained 1338 nt CDS; the CDS starts at transcript offset 128, zero-based. Separate comparison to retrieved NM_000187.4 reproduced the full transcript and CDS. Coding and intronic HGVS coordinates are mapped into genomic coordinates before applying substitutions, deletions, insertions, duplications or delins. A single-variant sequence is constructed from the reference; the intended correction restores the reference allele. The historical code calls this the patient sequence, but it is not a diploid patient genome. Individual haplotypes, variants in cis and the second disease allele are not modeled and require assessment before patient-specific use. A 30 nt upstream anchor must agree in the patient and intended nicked-strand frames.

Variant inclusion is a study rule, not clinical reclassification. The eligible census includes 263 aggregate pathogenic/likely pathogenic records and eight qualifying conflicting records. Consequently, neither every database record nor every inventory variant should be described unconditionally as a confirmed pathogenic allele.

Seven legacy mappings were checked by sequence or boundary reconstruction. Six produce identical edited sequences after normalization; c.468+2T>C to c.469+2T>C is a coordinate correction at an exon boundary. The full row-level tests are in the cohort audit. VariationID 643931 has a frozen genomic HGVS but a blank transcript HGVS. A separate, reference-checked sidecar supplies NM_000187.4:c.651_652del, using [ClinGen CA915941535](#). This annotation does not admit a new design or increase historical coverage.

Sequence orientation and template repair

The historical `RT_template_seq` is the intended DNA flap in its synthesis direction. It is retained unchanged for provenance. The actual 5' to 3' RNA reverse-transcription template is its reverse complement. RNA-component sequence fields use DNA-letter notation, so T represents a U residue in the corresponding RNA.

Let n be `peg_nick_chr3_0based_base_D4plus1` and t the template length. In the plus-strand design frame the intended flap is reference $[n, n+t)$; in the minus-strand frame it is the reverse complement of reference $[n-t+1, n+1)$. The RNA RTT is the reverse complement of this flap. This convention was checked against reconstructed patient protospacers and PAMs, rather than inferred from matching one historical column alone.

The repeat-aware edited endpoint m is independently recovered by comparing patient and intended sequences downstream of the nick, identifying their common prefix and suffix. The repaired length is $\max(t_{\text{old}}, 10, m+7)$ and must not exceed the selected cap, 41 or 50 nt. Downstream homology is $t_{\text{repaired}} - m$ and must be at least 7 nt. Extending the DNA flap preserves its old prefix; the corresponding old RNA template is a suffix of the repaired RNA template. No spacer, PAM, strand or nick coordinate is altered by this extension.

The 13 nt PBS is the reverse complement of the immediately upstream sequence in the nicked-strand frame. The modeled component order is RTT then PBS. This is a defined computational PBS rule; it does not identify an experimentally optimized PBS or a complete expression construct. The public-facing fields are

flap_5to3, rtt_5to3, PBS_13_5to3, extension_RTT_PBS_5to3 and the explicitly suffixed repaired length and homology fields. Do not use the legacy sequence column as an RNA ordering sequence.

The ten tables cover fixed, NGG, NGN, NRN and NNN selections at both caps. They contain 2364 selected-design records, including repeated loci across configurations. All selected records and all 80544 retained candidate records pass independent allele, spacer, PAM, flap and endpoint checks. The cap41 fixed set extends 63 templates; the NGG-reselected set extends 62. Every repaired template meets the modeled length and homology requirements.

Retained-pool reranking

PegRNA candidates are ranked by PAM tier, nick-to-edit distance, template length and deterministic spacer tie-breaking. The original key first reproduces the retained winners; the repaired length then replaces the length term. All ten complete current pools preserve the selected pegRNA spacer, PAM, strand and nick. The fixed cap41 pool has 7597 records across all 237 loci. V2 used an older pool of 7546 records at 234 loci; c.308_317del, c.472_496del and c.970del were unnecessarily absent from that comparison.

This is reranking and sequence verification of retained candidate pools. It is not de novo enumeration of the full ClinVar census. Nicking-guide inputs and ranking are held fixed because their parent pegRNA identities and nick positions do not change. Genome-wide count thresholds do not participate in the selection key. The repair therefore does not select candidates opportunistically to pass the screen. Candidate nomination excludes five consecutive T residues in either prime-guide spacer. This historical spacer rule is separate from four-U checks of the repaired RTT and RTT-plus-PBS components. Nicking-guide offsets use the allele on which the candidate was enumerated: the variant allele (called patient allele in historical fields) for PE3 and the intended-correction allele for predicted PE3b. The pegRNA nick is mapped into that same frame before subtraction. The 40–90 nt absolute separation rule applies to PE3 in the variant-allele frame and is not a PE3b filter. Interpret `nick_offset_BOND` together with `nick_offset_frame`; `nick_offset_BOND_intended_frame` separately reports corrected-allele separation.

Count matrices and target exclusion

The original cache contains 1538 distinct spacers. The additional cache contains 84, of which 15 overlap, producing 1607 distinct spacers. All overlapping count matrices and target-exclusion tuples agree. Each matrix has 25 suffix buckets, in the ordered Cartesian product of A,C,G,T,X, and five exact mismatch strata, 0–4.

Both aggregation programs use chromosomes 1–22, X and Y. Their effective counting scope excludes mitochondrial, alternate and unplaced sequences. Fresh addition of the 24 retained original chromosome aggregates reproduces all 192250 matrix cells; the recorded original scan manifest contains 216 files across nine batches. Additional-cache chromosome totals also cover the same 24 chromosomes. These checks do not rerun the original genome scanner or establish complete enumeration of all possible genomic alignments.

The retained site key is chromosome, spacer, zero-based 20 nt protospacer-window start and strand. The PAM lies outside that 20 nt window. One chromosome-3 hit within the inclusive reported-coordinate window 120627172–120683239 is nominated per spacer by minimizing mismatch count, position and strand. That positional hit is removed before counting. Independent reconstruction confirms the nominated tuple, including mismatch count and PAM, for all 1884 guide uses in the four fixed/NGG tables at both caps and all 625 base-guide rows. For this target-exclusion audit, the explicitly named intended-allele offset is used to reconstruct the reference-coordinate target tuple; the generic selection offset remains branch-specific. Comparing coordinates from different allele frames directly across indels would create false disagreements. These use counts are not unique spacer or experiment counts.

The principal intended-site requirement is NGG for both prime guides. Off-target models are NGG = {GG}, A3 = {GG,AG,GA}, NGN = suffixes starting G, NRN = suffixes starting A or G, and NNN = all suffixes. Actual intended PAMs must contain only A,C,G,T. The study retains no additional exact site, at most one exact one-mismatch site and at most five exact two-mismatch sites per required guide. The base arm uses the NNN count model in both compared conditions. Both prime guides must pass, and each variant is counted once in the union.

X is an unresolved observed reference character, not an IUPAC wildcard. Both inclusion and exclusion of X-containing suffix buckets are emitted. Eight nonzero buckets contribute zero counts at mismatches 0–2, two at mismatch 3 and six at mismatch 4. All 24 scenario sets and 33 principal threshold settings therefore agree between these policies. This is a measured property of the retained data, not a general argument that ambiguity is irrelevant to specificity.

Bystanders and component sensitivities

Base editing uses the retained positions 4–8 actionability definition and corrected worst-consequence bystander rule. The 625 candidate rows join one-to-one to their verdict rows; 321 satisfy actionability and the corrected bystander criterion. Wider splice-region annotations are retained under the main rule unless a more severe consequence fails it. The strict sensitivity excludes all wider splice-region bystanders, retaining 85 passing guide rows across 37 variants instead of 90 across 40.

Poly-T checks operate on the actual RTT and on RTT plus PBS. Repair of c.413G>A newly creates four U residues in the RNA template, represented as TTTTCCTCTG in DNA letters. No repaired RTT newly reaches five. The fixed cap41 inventory has nine RTTs and eighteen RTT-plus-PBS components with at least four T characters; the NGG-reselected inventory has eight and sixteen, respectively. Counts differ from the numbers passing the principal screen.

The following are matched sensitivities, not alternative claims of experimentally demonstrated usability:

Rule	Fixed union	NGG-reselected union
Principal count and adopted bystander rules	110	188
Exclude wider splice-region bystanders	107	186
Exclude originally short RTTs rather than repair	94	150
Apply both preceding exclusions	91	148
Exclude repaired RNA RTTs containing TTTT	110	181
Exclude repaired RTT-plus-PBS components containing TTTT	110	180

The scaffold, stabilizing motif, linker and expression system remain unspecified. The component exclusions cannot establish whole-transcript suitability, and a four-U flag does not establish transcriptional failure. The main count comparison remains explicitly separate from those questions.

Preserved and added guide configurations

An independently written aggregation compared eleven identity and component fields for the 93 fixed canonical pairs and verified their exact preservation in the 233 reselected pairs. These fields comprise pegRNA spacer, PAM, strand and nick coordinate; nicking-guide spacer, PAM and cut coordinate; predicted

PE3/PE3b class; RNA RTT; PBS; and combined extension. Applying a common deterministic screen to preserved pairs guarantees set containment. It does not guarantee strictly positive additions; their observed magnitude depends on the added candidates and the screen.

Stratum	Available pairs	NGG screen	NGG/NAG/NGA screen
Already canonical and preserved	93	90	70
Newly canonical PE3	140	138	79

The corresponding principal retained fractions are 70/93 (75.3 per cent) and 79/140 (56.4 per cent). These conditional descriptive fractions are not causal comparisons of nicking strategies. The sole added prime variant already represented by base editing is c.879+1G>A. Prime-only retention is 70 versus 149; predicted-PE3b-only prime retention is 70 versus 70. All 33 threshold rows were independently recomputed from retained matrices and selected pairs. They are correlated sensitivity settings, not independent experimental replicates. Supplementary Table S11 and panel [followup/nested/](#) supply the pair-level and threshold-level decomposition; no genome scan was repeated.

Conservative ABE8e substrate exclusions

The frozen base-editing model keeps intended-correction eligibility and the original consequence rule at positions 4–8. Experimental characterizations of ABE8e motivate examining possible substrates beyond those positions (Zhao et al. 2024; Evanoff et al. 2026). We used two declared stress windows: positions 3–11 and 3–12, numbered from the PAM-distal end. An ABE8e guide was excluded if it contained any adenine outside positions 4–8 within the tested wider window. All 55 retained BE4 rows were held unchanged. No additional intended correction was admitted.

The script reconstructs the retained HGD transcript from the exon table and reference interval, recovers the 1338 nt coding sequence, maps each retained substitution, confirms its reference base and creates the single-variant allele. Each of the 90 principal-pass guides had one exact 20 nt spacer plus PAM match with intended target chemistry at positions 4–8. The resulting coordinates, strands, target positions and extra substrate positions are recorded for every row. These are newly executed assertions against retained inputs, not a new reference retrieval or genome scan.

Rule	Base rows	Base variants	Fixed union	Reselected union
Frozen positions 4–8 rule	90	40	110	188
Exclude ABE8e A at 3, 9, 10 or 11	64	30	100	179
Also exclude ABE8e A at 12	61	27	97	176
Omit the entire base arm	0	0	70	149

The first exclusion removes 26 of the 35 retained ABE8e rows; the second removes 29. Variants remain represented when another guide survives. These presence-based exclusions intentionally can discard benign or inactive substrates. They do not predict activity, frequency, coding or splice consequence, guide-independent effects, noncanonical editing products, or suitability of the still-unspecified full editor construct. They do not assess a wider BE4 window. The 110/188 principal result remains conditional on its declared

model; the prime-only comparison establishes that its positive prime-design increment does not depend on accepting any base candidate. Complete output sets, input hashes and mapping checks accompany Supplementary Table S12 in `panel_followup/window/`.

Reading current analytical fields

`current_views/Table_S3_current.csv` is a one-to-one projection of all 625 original base-guide rows onto the current analytical fields. `actionable_and_bystander_eligible` applies the adopted actionability/consequence rule; `NNN_mm0`, `NNN_mm1` and `NNN_mm2` are the counts used here; `principal_count_screen_pass` selects the 90 rows and 40 variants in the principal base arm. The original S3 remains available for provenance, but its `old_*`, `new_*`, `deposited_usable_main`, `bystander_ok`, `score_detail` and MIT-related columns are historical fields and must not determine the current selection. The current view introduces no new scientific result. The package data dictionary states the authoritative filtering recipe and distinguishes legacy source weights from patient counts.

External tool comparison and prior art

PrimeDesign was pinned to commit `e4fdb31c81ccd2389dbada11d81101dd8a750b18`. Before running it, the benchmark selected the first two eligible-census single-nucleotide substitutions by VariationID in each `peg-strand × PE3/PE3b` stratum, giving eight cases. Each input uses 250 bases of flank on either side and encodes patient ALT to reference REF correction. Constraints are canonical NGG, PBS13 and RTT lengths 10–41, with no silent mutations, poly-T filter or first-C filter. Native nick-distance defaults are retained and the study's PE3 distance band is evaluated downstream. The nick-output cap is raised to 1000 to remove default three-guide truncation, without changing the frozen sample or biological constraints.

All eight cases match `peg spacer/PAM/strand/length`, corrected RNA RTT, PBS13, combined extension, nick spacer/PAM, PE3/PE3b classification and nick offset after coordinate-frame alignment. The original v2 template strings match only after reverse complementation. This is a bounded sequence and annotation comparison on canonical-PAM substitutions. It does not validate indel performance, other PAM models, the full inventory, ranking superiority, off-target enumeration or biological activity. The exact sample, benchmark code, commands and initial/truncation-corrected outputs are retained in `manuscript_audit/`, together with the upstream license and pinned provenance. The upstream PrimeDesign program is a separately retrieved dependency; it is not redistributed in this package.

Expanded diagnostic benchmark

The follow-up protocol, sample, allele strings, preparation code and pinned dependency hashes were frozen before tool execution. The primary sample contains 19 cases: 15 selected pairs (14 length-changing variants and the substitution with the longest selected RTT, 40 nt) plus the four retained cap41 cases with no admissible NGG pegRNA. It covers every represented length-change class by peg strand: deletion and duplication on both strands, insertion on both strands and delins on the minus strand. The source has no plus-strand delins; plus-strand insertion has one row. Selected boundary examples include repaired RTT10, longest length-changing RTT30 and overall RTT40. No selected RTT41 design exists in the source. The original eight-SNV benchmark is unchanged.

The primary run used the same unmodified PrimeDesign commit, canonical NGG, PBS13, RTT10–41, a nick-output cap of 1000, native nick-distance defaults and no silent-mutation, poly-T or first-C filters. It emitted 2974 component rows across settings; these are not independent experimental observations. Thirteen of 15 selected configurations had exact matched pegRNA rows. In all 13 matched cases, RNA RTT, PBS, extension, nick spacer/PAM, predicted class, branch-specific offset and uniquely reconstructed intended-allele offset agreed. The two non-emitted selected pegRNAs were `c.184_187dup` and `c.357dup`.

For those two duplications, the frozen prefix-first representation of the changed span gives the selected minus-strand guide a negative distance to the explicitly annotated edit in PrimeDesign. A separately frozen post hoc diagnostic used suffix-first repeat-copy annotation. Before execution, equality of the complete variant sequence and complete intended sequence was asserted against the primary input. No allele, selected guide, RTT limit or upstream setting changed. Both selected configurations were then emitted and matched all compared components and distances. This explains representation-sensitive emission; it does not convert the primary 13/15 result into a prespecified 15/15 success. Both freezes, annotations, raw outputs and nonmatches are retained.

An initial comparison used the generic branch-offset field as though it always described the intended allele. Eight length-changing cases differed by the corresponding edit-length shift. Historical source inspection and independent target-string geometry established the correct fields, without changing stored sequence or coordinate values. In c.664_674dup, for example, PE3 separation is 47 nt in the variant allele and 36 nt after correction; the former satisfies the adopted selection band. All 13 primary matched cases agree when the correct field and orientation are used. The initial comparator and discrepancies remain in the annex alongside the adjudication.

No-admissible comparator	Emitted peg component rows	Reason no emitted peg passes checked study rules
c.308_317del	9	Rows with enough downstream homology have a spacer with eight consecutive T residues
c.319C>T	8	The only row meeting downstream homology has the same excluded T-rich spacer
c.449C>T	0	No pegRNA is emitted for the frozen inputs and settings
c.472_496del	7	All emitted templates have fewer than seven downstream homology bases

The post-output checks apply the retained 30 nt invariant upstream anchor, RTT10–41, at least seven downstream homology bases and the five-T spacer exclusion. No emitted peg survives, so no such peg can yield a fully admissible pair. This bounded result does not establish exhaustive design impossibility, full nick-ranking equivalence or biological performance. The four cases are a separate availability group, not four additional component-agreement successes. Supplementary Table S13 and `panel_followup/expanded_benchmark/` retain every case, component comparison, diagnostic and input/output hash.

Existing HGD components

Morris component files independently contain 77 HGD pegRNA rows across 26 IDs, 124 PE3 nick rows across 16 IDs and 11 PE3b nick rows across six IDs. The PE3b identifier suffix -b must be normalized before joining. Across the component files, the union is 28 HGD identifiers; the component counts do not enumerate complete paired designs. The public resource retains the extraction script, source identifiers and frozen hashes. The original local archive preserves the raw source; the public deposit omits that archive and its HGD row extracts and supplies retrieval instructions. These data establish prior HGD and PE3b design coverage.

Relation to existing design resources

Resource	Scope and evidence	Relation to this study
PrimeDesign / PrimeVar (Hsu et al. 2021)	Paired prime-guide design, configurable templates, a ClinVar design database and selected cellular tests	Existing paired-guide resource; the pinned tool is used here for component comparison, not a ranking-superiority test
Morris et al. (2021)	Genome-wide computational prime-editing designs across PAM/template choices, including HGD and PE3b components	The retained 28 HGD component identifiers are not complete pairs and cannot be compared directly with the current variant count
EditABLE (Maxim et al. 2026)	Integrates several editing approaches, including PrimeDesign; disease-focused application and selected in vitro validation	Earlier multimodal integration; its disease and input units differ from this HGD census
CRISPR-GPT (Qu et al. 2026)	Language-model support for editing workflows, with task evaluations and cellular demonstrations	Prior AI-assisted editing work; not used or evaluated in this project
Present HGD resource	Within-inventory canonical-PAM pair comparison, constant base arm, retained-count screens, sequence checks and sensitivity analyses	Computational representation of 110 versus 188 variants under shared rules; no biological experiments or measured educational outcomes

The table compares published scope, not relative accuracy or current live-interface functionality. Database snapshots, variant eligibility, assemblies, correction directions, PAM/RTT settings and observational units differ. No percentage or identifier count should be interpreted as a direct coverage improvement over another resource. CRISPR-GPT is cited as prior AI-assisted editing work; it was not used or benchmarked in this project.

Published series and conditional arithmetic

Source-specific comparisons are primary. Vilboux's 79 probands define 158 nominal disease-chromosome positions under diploid accounting. The table has 157 variant detections, including three probands with three detections. For each, all three two-variant choices are enumerated, producing 27 joint configurations. Four unidentified disease chromosomes remain in the denominator. Differences are calculated within the same configuration before bounds are taken.

Direct extraction confirms the Slovak table's 12 rows summing to 104 and the Italian table's 27 rows summing to 68, with exact agreement to the retained weights after normalization. Their cumulative ascertainment and observational units differ. The Slovak count exceeds two times its stated 50 families; this remains a source interpretation limit. No selected Slovak–Italian overlap was demonstrated, but independence and complete deduplication were not established.

Pending full-text adjudication, the Turkish series is excluded from the revised pooled sensitivity. Its abstract does support seven unrelated families and three homozygous R225H families, so calling that genotype information wholly unsourced would be wrong. The remaining nominal weights sum to 346, with fixed representation 187–190, reselected representation 284–286 and paired gain 96–97. These are heterogeneous-source conditional counts. They are neither confidence intervals nor estimates of population coverage or patient eligibility. All 27 paired worlds and additional source-omission sensitivities are supplied.

Prospective assessment of the three inventory gaps

The gap annex is a separate, unadopted analysis of c.448del, c.435-1G>C and c.651_652del. It does not alter the frozen census, selected inventory or principal comparison. Transcript positions were reconstructed from the retained exon table and reference sequence. For c.651_652del, equality of the complete edited local reference confirms the repeat-shifted genomic alias. Each variant uses 400 reference bases on either side. Candidate enumeration applies the implemented P125/P172 local rules with a minimum RTT length of 10, caps of 41 and 50, and NGG, NGN, NRN or NNN intended-site recognition. The older P97 reach additions are not included.

For each of the 24 variant/model/cap configurations, candidates are ranked by PAM tier, edit distance, RTT length and spacer. The first-ranked pegRNA is paired using the retained PE3b preference and PE3 distance rule; all enumerated pegRNAs and each selected pair are supplied. Thus the screening result for a selected pair is not an exhaustive screen of every alternative. Cached counts are joined by spacer only when the nominated target tuple also matches the prospective sequence context. P171 supplies three additional cached spacers after verification of 19 overlapping count matrices and target tuples against the other caches. P171 raw chromosome records were not freshly reaggregated in this annex.

The outputs distinguish no pegRNA, an available pair, a known failing guide and an unevaluable guide. They report both A3 and the recognition model's own count set; those sets are not necessarily nested. Two missing nicking-guide caches are listed, but their paired pegRNA already fails the tested limit, so scanning those nicking guides alone cannot change the rejection. No new genome scan was performed. The complete script, 24 configuration summaries, 20 selected pairs and all local pegRNA candidates are retained in `sequence_audit/gap_assessment/`; Supplementary Table S10 contains the selected-pair sidecar.

Reproduction and limitations

The package's root `reproduce.py` executes the portable audits from bundled inputs. Python 3.10 or later and the standard library suffice for the computational tables; optional figure regeneration requires matplotlib and NumPy. Core outputs can be regenerated from an extracted package without the original project, network access, an API key or a paid service. An isolated run of the distributed package reproduced all 26 core scientific tables against frozen hashes. Four optional verified third-party inputs enable the full 35-comparison workflow; that larger profile was not newly rerun during this packaging revision. The input and output manifests identify the exact materials checked.

The computational workflow reconstructs sequences, reranks retained pools, recombines retained count evidence, extracts public source data and compares a bounded external-tool sample. It does not claim a fresh genome scan, complete construct design, validation in cells or animals, clinical benefit, measured student learning or a total economic project cost of \$1800. The \$1800 is the reported nine-month Claude Max subscription expenditure only.

Cohort verification and public tools

Research developed across successive academic quarters within Foothill College's MESA and ETI programs. Undergraduate researchers joined at different stages and worked with a PhD advisor on shared verification tasks. Retained packet records support the documented student-led work. They are records of submitted checks and corrections, rather than an exhaustive account of every contribution or a quantitative measure of learning. Individual finding totals are not used here to characterize the cohort or determine authorship.

The tools below had distinct roles. A web service and a local implementation of public software are identified separately. The historical packet checks and the later frozen eight-case benchmark are also distinct; neither implies that every listed tool tested every inventory entry.

Tool or resource	Documented use	Scope of retained evidence
UCSC hg38 sequence API	Reference alleles, flanks, strand and coordinate checks	Sequence retrieval and reference comparisons documented in prime-editing packets
pegFinder web tool	Comparison of candidate prime-editing guide components	Retained three-way comparison and packet review records; Chow et al. (2021)
PrimeDesign command-line implementation	Design-component and nicking-guide comparisons	Historical packet comparisons plus the separately frozen eight-case benchmark; Hsu et al. (2021)
CRISPOR web tool	Guide-level off-target output and exact-spacer comparisons	Retained findings tables and rerun summaries; Haeussler et al. (2016)
ViennaRNA RNAfold and seqfold	Exploratory RNA-folding summaries	Partial summary records; these do not support complete expressed-guide structural validation

Historical packet records identify the following evidence locations within the retained project archive. These pointers describe the source of the cohort verification record. The public computational resource excludes identifiable student worksheets and private correspondence.

- `recovered_from_drive/01_prime_editing/Package4_ThreeWay_v2/tab3_Method_and_reproducibility.md`: UCSC, pegFinder and local PrimeDesign methods, with associated comparison tables.
- `rebuild/DIAG/p75/P75_RAJ_BATCHB_PACKET_REVIEW.md`: retained packet comparison review.
- `recovered_from_drive/01_prime_editing/Package2_tab2_PE3_OffTarget_Findings.md`: CRISPOR findings and companion comparison records.
- `crispor_rerun/RESULTS_2026-08-21.md` and `crispor_rerun/crispor_results_2026-08-21.csv`: retained CRISPOR rerun summaries.
- `recovered_from_drive/01_prime_editing/fold_versions/FOLD_A_tab2_Verification.csv`: partial folding-verification summaries, including a ViennaRNA 2.7.2 version label.

Some retained packets concern other target loci; their numerical results are not transferred into the HGD analysis. Folding summaries do not replace missing raw structures, base-pair probability matrices or full-construct sequences. Cas-OFFinder is not claimed as an independently verified completed cohort check because the surviving execution evidence does not support that claim. These distinctions preserve the scope of what the retained records establish.

Public tool references and access points include the UCSC API documentation (<https://genome.ucsc.edu/goldenPath/help/api.html>), pegFinder (Chow et al. 2021; <https://doi.org/10.1038/s41551-020-00622-8>), PrimeDesign (Hsu et al. 2021; <https://doi.org/10.1038/s41467-021-21337-7>), CRISPOR (Haeussler et al. 2016; <https://doi.org/10.1186/s13059-016-1012-2>) and ViennaRNA (<https://www.tbi.univie.ac.at/RNA/>). Tool availability and defaults may change; retained versions, inputs and outputs define the historical checks.

AI provenance and resource accounting

A retrospective recovery on 19 September 2026 examined the retained project history and local Codex session metadata. The historical model log, assembled on 1 September 2026, covers 50 analysis units and originally records one self-reported model identifier. The project git history contains 190 commits dated 21 August–11 September 2026, of which 82 carry AI model-name attributions: 26 name “Claude Fable 5” and

56 name “Claude Opus 5”. Seventy-seven commits carry session identifiers, representing four distinct session labels. The first-addition commits for 12 of the 50 historical units and 84 later units carry the “Claude Opus 5” attribution. The remaining 38 historical units have no recovered model label. A first-addition commit establishes that files entered version control with that attribution; it does not establish which model created every file or executed each analysis. Commit messages alone also do not establish that their trailers were generated automatically or that a session actually switched models. The original log is retained alongside the recovered records rather than overwritten.

For OpenAI Codex, a minimized extract contains 100 retained turn-context records from 11 project-associated sessions, dated 11–19 September 2026 UTC; all record the exact model label `gpt-6-astr`. These are local client metadata, not vendor response records. Six additional project-associated session files contained no recoverable turn-context model entry and support no model assignment. Identifiers recovered for these recorded turns are not extrapolated to earlier work or to undocumented executions. The public extract includes identifiers and timestamps, but no conversation text, private correspondence or personal filesystem paths.

Dates are reported by evidence type. Git author timestamps date the retained commit metadata; client timestamps date turn-context records; two retained benchmark process records on 14 September 2026 separately report start and end times. File modification times and dates of last deposited artifacts do not establish model execution dates. Neither a subscription name nor an instruction requesting a model setting establishes actual model use. The oldest surviving file date does not establish the project’s start date. Vendor account-side records were not examined, and no vendor-issued response identifier was recovered in this review. Recovery did not repeat any scientific computation. The corresponding author confirmed review of AI-assisted content and approval of the current scientific revisions by all authors.

The files under `provenance/ai/` in Online Resource 2 provide the historical log, unit-to-commit mapping, minimized commit and Codex metadata extracts, evidence definitions and the adjudication of recovery claims. Student verification with the public tools described above is documented separately from AI assistance. No comparison of AI systems or measurement of educational outcomes was performed.

The Mathematics, Engineering, and Science Achievement (MESA) program and Emerging Technologies Institute (ETI) at Foothill College provided funding and support for this project. MESA and ETI provided nine months of Claude Max subscription support, totaling \$1,800. This amount records direct subscription expenditure only; internship payments and other program and institutional support are not included. A full accounting of project costs is not reported here. The study did not perform a cost-effectiveness comparison or measure educational outcomes.

Proposed biological validation programme

This section describes future work proposed in response to the computational findings. It is not a methods record for experiments already performed, and no animal facility, collaborator, model, timetable or experimental approval is asserted. The proposed sequence is component finalization, focused cellular testing, functional confirmation and then consideration of a suitable in vivo model.

Choosing an informative initial panel

The first panel should answer whether the additional canonical configurations produce useful correction and whether their product profiles differ from the alternatives. It should be small enough to complete with the available experimental capacity. Final panel size and variant identities should be fixed before generating assay results, with the full selection and exclusion record retained.

Candidate class	Scientific reason for inclusion	Practical selection criteria
Predicted PE3b retained in both analyses	Establish activity of a configuration preserved by the selection comparison	Canonical intended PAMs, principal-rule retention, unambiguous variant and component sequence
Newly available canonical PE3	Test the principal expansion in candidate availability	Principal-rule retention and a sequence-compatible experimental comparison; retain the original broader-PAM alternative where a meaningful comparison is feasible
Base-editing candidate	Evaluate a complementary correction approach	Modeled intended conversion with no excluded coding or canonical-splice bystander, plus an assay that can resolve intended and bystander products

Within each class, prefer variants with clear disease evidence, reliable reference mapping and an accessible model that can test the intended correction. Recurrence in a documented source series can support relevance, but heterogeneous pooled weights must not be treated as population prevalence. Favorable component flags improve practical readiness; they do not predict editing activity. Complete expressed-guide checks are required even when the RTT and PBS lack a four-U flag, because the spacer, scaffold and junctions contribute additional sequence context.

The design question should remain distinct from an editor comparison. Where the original broader-PAM PE3b and recovered canonical PE3 require different editor constructs, a direct experiment changes both recognition architecture and nicking strategy. Such a result is a comparison of whole configurations and cannot by itself isolate the causal effect of PE3 versus PE3b. Where technically appropriate, additional matched configurations can help separate these effects. Experimental collaborators should finalize this comparison before prerecording the panel and analysis plan.

Finalizing components before experimental work

For each selected variant, record the reference assembly and transcript, exact variant and intended alleles, both guide spacers and PAMs, guide orientation, corrected RNA RTT, PBS and full expressed-guide sequence. Select and document the editor construct, scaffold, linker, any stabilizing motif and expression/delivery format. Confirm that their targeting assumptions match the intended sites and that the experimental material is sequence-compatible with the human design. Resolve motif and whole-transcript expression flags in the final construct record. A successfully reconstructed spacer and template is a starting specification, not a complete experimental reagent.

A prospective analysis plan should define the intended molecular and functional endpoints, comparison groups, independent biological replication, sequencing analysis and criteria for repeating or stopping an experiment. Quantitative acceptance thresholds should be agreed with the experimental team before outcomes are known; no activity threshold is inferred from the present count screen.

Cellular correction and functional testing

The immediate biological question is whether the selected configuration installs the intended correction with a useful product profile. A suitable cellular model should permit verification of the target allele and measurement of intended correction, indels and other unintended local products, including base-editing bystanders where relevant. Appropriate unedited/untreated and comparison controls should distinguish the

editing effect from delivery and assay background. The assay should report independent biological replicates and uncertainty at the correct experimental unit; sequencing reads are not independent biological replicates.

Functional follow-up should ask whether correction restores HGD activity and improves HGA handling in a model capable of expressing the relevant pathway. An edit fraction alone is not evidence of metabolic rescue, and a convenient sequence-editing assay may need a separate functional model. Advance candidates on the combined evidence of reproducible correction, product quality and functional improvement. A failed candidate should remain in the result record to inform subsequent selection.

Specificity work should begin from the actual selected editor and guide sequences, integrating locus-specific follow-up with an empirical method appropriate to that editor. Homology counts nominate sites for investigation; they neither establish absence of unintended editing nor quantify its frequency. The choice and scope of empirical testing should be documented with its detection limits and should address the relevant on-target and off-target products (Kim et al. 2020; Kwon et al. 2022; Liang et al. 2023).

Considering an in vivo model

Progression to an in vivo model should follow reproducible cellular correction and functional evidence, together with a suitable facility, collaborators and applicable institutional review. An Hgd-deficient animal does not necessarily carry the human variant or the human guide-binding sequence. Model selection must therefore establish whether the candidate corrects the intended humanized allele, an appropriately homologous target, or a separately designed model allele. Those alternatives test different questions and should be named explicitly.

The proposed initial in vivo goals are hepatic molecular correction, HGD activity, circulating HGA, urinary HGA as a complementary measure, persistence of the response and tolerability of the chosen delivery approach. These endpoints connect the intervention to the known hepatic contribution to HGA control (Hughes et al. 2019). Long-term ochronotic and musculoskeletal outcomes would require a separate duration and study design. No fraction of corrected hepatocytes required for benefit is inferred from residual Hgd messenger-RNA measurements in the published animal work.

Additional census and selection details

The retained census and inventory use separate units. An explicit VariationID-to-HGVS crosswalk distinguishes the 665 queried records, 302 disease-associated records, 271 eligible records and 278 inventory variants. Later database updates are separate deltas and must not silently replace the analytical snapshot. Inclusion of conflicting records requires a criteria-based pathogenic or likely pathogenic submission, no criteria-based benign or likely benign submission and gnomAD allele frequency below 0.4 per cent. Eligible events must be precisely delimited, confined to HGD and at most 50 bp. Excluded records remain in the exclusion register.

The original 237 fixed paired designs comprise 230 predicted PE3b and seven PE3 configurations. Among the 144 without NGG at both intended sites, 140 fail at the nicking guide alone, three at the pegRNA alone and one at both. At cap41, the fixed canonical subset is 92 PE3b plus one PE3; re-selection retains all 93 and adds 140 PE3 configurations. Of the recovered loci, 134 originally used broader-PAM PE3b and six already used PE3. The cap50 reselected set contains 235 paired configurations: 93 PE3b and 142 PE3.

Base-editing models use ABE8e or BE4 substitution chemistry and SpRY targeting assumptions, with positions 4–8 measured from the PAM-distal end. Coding and splice annotations are considered in parallel; the more severe determines disposition. Canonical splice donor or acceptor changes, missense, stop-gained and stop-lost fail the adopted rule. The wider splice-region band extends three nucleotides into the exon and covers intronic positions 3–8. That positional annotation alone is retained in the principal rule and excluded

in the strict sensitivity. A low-tier positional label cannot mask a coding consequence. The 321 rows passing actionability and this bystander rule represent 108 variants before count screening; the principal NNN screen retains 90 rows across 40 variants.

Among Vilboux's 79 probands, 72 have two detected variants, four have one and three have three. The three possible two-causal-variant choices for each of the last three generate the 27 paired configurations described above. These are conditional causal assignments, not experimentally established phase. The historical six-source nominal-360 pooling is retained only in the audit record; the current supplementary sensitivity excludes Turkey and uses the stated nominal-346 heterogeneous weights.

Supplementary table index

The CSV files supplied within Online Resource 2 (ESM_2.zip) preserve the quantitative release accompanying the comparison. No values were changed for the narrative revision.

Table	Contents	Unit or interpretation
S1	Variant census and inventory crosswalk	Database record and design-variant membership
S2	Prime-editing candidate components	Selected configuration records
S3	Base-editing candidates	Guide rows with variant and consequence annotations
S4a	PAM and template-cap scenarios	Matched screening outcomes
S4b	Threshold grid	Thirty-three exploratory threshold settings
S4c	Reference ambiguity sensitivities	Inclusion or exclusion of X-containing PAM buckets
S5a	Published-series results	Source-specific observational units
S5b	Conditional pooled arithmetic	Heterogeneous weights, not independent patient counts
S5c	Paired causal-assignment scenarios	Twenty-seven matched configurations
S5d	Legacy alias checks	Reference and edited-sequence agreement
S6	Candidate-pool controls	Retained-pool reconstruction and ranking checks
S7	Component and bystander sensitivities	Matched variant-set restrictions
S8	PrimeDesign comparison	Eight frozen single-nucleotide cases
S9	Existing Morris HGD resource	Component-file counts, not complete pairs
S10	Exploratory inventory-gap candidates	Separate prospective selected-pair sidecar
S11	Preserved and added guide decomposition	Pair identities, screening strata and matched threshold sets
S12	Conservative ABE8e substrate exclusions	Row mapping, substrate flags and variant-set sensitivity
S13	Expanded external benchmark	Fifteen selected cases, four availability cases and separate post hoc repeat diagnostic

References

- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- FDA (2025) Harliku nitisinone prescribing information. Revised June 2025. https://www.accessdata.fda.gov/drugsatfda_docs/label/2025/209449s018lbl.pdf
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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:eaa04774. <https://doi.org/10.1126/sciadv.aao4774>

- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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)
- 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>
- 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>
- 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>
- 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>
- 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>

- 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
- 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>