Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mapping molecular genome-editing opportunities across Mendelian disorders using ClinVar and GenCC

Achab-Awelana Ernest

achabawelana@gmail.com

Kwame Nkrumah University of Science and Technology https://orcid.org/0000-0001-8411-7895

Research Article

Keywords: CRISPR, base editing, prime editing, disease ontology, pathogenic variants, variant ascertainment, editing windows

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11697380/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Mapping molecular genome-editing opportunities across Mendelian disorders using ClinVar and GenCC

Article category: Investigation | Original computational genetics research

Achab-Awelana Ernest[1]

[1]Thrivus University for Biomedical Science and Technology, Constellations Avenue, Lashibi, Accra GT-336-4330, Ghana

Corresponding author: Achab-Awelana Ernest | achabawelana@gmail.com

Running title: Molecular editing in Mendelian disorders

Keywords: CRISPR; base editing; prime editing; disease ontology; pathogenic variants; variant ascertainment; editing windows

Abstract

Catalog-wide estimates of genome-editing correctability may conflate nucleotide chemistry, confidence in disease attribution, target-sequence feasibility, and therapeutic effectiveness. We integrated frozen ClinVar variant classifications with Gene Curation Coalition (GenCC) assertions to analyze 6,968 high-confidence gene–disease–mode-of-inheritance relationships. Filtering 9,345,867 ClinVar source rows yielded 312,679 unique qualifying variants, of which 168,648 were sequence-resolved single-nucleotide variants (SNVs); 92,687 (54.96%) exhibited canonical adenine/cytosine base-editor reaction chemistry for hypothetical alternate-to-reference correction. Exact disease linkage, unique relationship assignment, and ClinVar review confidence of at least two stars retained 46,786 of 269,669 variant–relationship association edges. Within the stringent cohort, 2,396 relationships contained a chemistry-compatible SNV and 1,763 contained short-indel evidence; 1,427 overlapped, yielding 2,732/6,968 (39.21%) nonredundant molecular-class candidates. Variant documentation was highly unequal (Gini coefficient 0.805), with the most extensively annotated 1% of relationships contributing 28.18% of association edges. In a GRCh38-validated chromosome 21 pilot, 373 of 1,071 chemistry-compatible SNVs had an NGG PAM and an eligible modeled editing-window position; 169 lacked additional modeled same-class bystanders. A purposive translational benchmark highlighted interventions directed at compensatory or alternative loci. The results define evidence-linked molecular substrates for experimental prioritization, not the fraction of Mendelian disorders with technically deliverable or clinically validated genome-editing therapies.

Article summary

The analysis integrates public human genetic-variant classifications with curated gene–disease associations to identify molecular changes relevant to genome editing. A stringent evidence filter revealed 2,732 of 6,968 relationships with a documented base-editing chemistry or short-indel candidate. Sequence screening of chromosome 21 demonstrated substantial targeting constraints, and the number of candidate relationships was sensitive to variant ascertainment and disease-linkage rules. These results provide a reproducible starting point for prioritizing editing experiments. They do not establish that an equivalent fraction of genetic disorders has deliverable, safe, or effective genome-editing therapy.

Introduction

Programmable nucleases, deaminase-based editors and prime editors have created a continuum of genome interventions, ranging from gene disruption to precise single-base correction, small-sequence replacement, regulatory manipulation and targeted integration. The chemistry of a prospective edit is necessary to establish

whether a particular molecular tool could, in principle, reverse an allele; it is not sufficient to determine whether the corresponding disease is treatable. This distinction becomes particularly consequential in genome-wide analyses where thousands of curated disorders are represented by heterogeneous alleles, mechanisms and cellular phenotypes (Anzalone et al. 2019; Dadush et al. 2024; Gaudelli et al. 2017; Komor et al. 2016; Walton et al. 2020).

The conventional description of base-editing "correctability" begins with the six transition substitution classes and their complements. Cytosine base editors (CBEs) catalyse selected C·G-to-T·A conversions, whereas adenine base editors (ABEs) enable A·T-to-G·C conversions (Gaudelli et al. 2017; Komor et al. 2016). Whether a *pathogenic* variant can be corrected depends on the opposite, pathogenic-alternate-to-reference direction, the genomic strand, the editable position within the protospacer, the available editor architecture and local sequence context. A chemically appropriate conversion is therefore a statement about the mutation class, not an experimentally demonstrated guide–editor pair. Moreover, prime editing can program substitutions and small insertions or deletions that fall outside canonical CBE/ABE chemistry; its performance varies with sequence, pegRNA design, mismatch-repair activity and cellular context (Anzalone et al. 2019; Chen et al. 2021; Sousa et al. 2025).

The unit of aggregation matters as much as the molecular model. A single gene may be associated with multiple phenotypes and inheritance modes; a MONDO disease entity may have multiple molecular aetiologies; and individual ClinVar VariationIDs may carry multiple phenotype annotations. In large catalogs, the existence of at least one chemically addressable variant in an extensively studied locus can be mistaken for evidence that all patients with the disease are treatable. The difference between catalogued molecular opportunities and population-level therapeutic eligibility is magnified by unequal clinical sequencing, pathogenicity assessment and historical gene ascertainment (Dadush et al. 2024; Gene Curation Coalition 2022; Landrum et al. 2025).

Therapeutic editing further escapes the boundaries of direct allele reversal. Exagamglogene autotemcel (CASGEVY) edits an erythroid enhancer of *BCL11A* to augment fetal hemoglobin expression rather than correcting the causative globin alleles; it has regulatory approval for specified sickle-cell disease and transfusion-dependent β-thalassemia indications (US Food and Drug Administration 2026). Similarly, CRISPR disruption of *KLKB1* in hepatocytes can prevent hereditary-angioedema attacks without editing the typical causal *SERPING1* allele (Cohn et al. 2026). Conversely, the successful pharmacodynamic modification of a therapeutic modifier gene cannot be automatically transferred to every GenCC assertion bearing that target gene (Vafai et al. 2026). Accordingly, therapeutic targets, etiologic genes, clinical indications and variants must be represented as distinct entities.

The relevant tissue and evidence maturity introduce further non-equivalences. Ex vivo autologous hematopoietic stem-cell modification, in vivo hepatocyte-targeted lipid nanoparticles, subretinal gene editing and systemic skeletal-muscle delivery entail different anatomical, immunological and safety constraints (Cohn et al. 2026; De Caneva et al. 2019; Gillmore et al. 2021; Li et al. 2026; Musunuru et al. 2025; Ngo et al. 2026; Pierce et al. 2024; US Food and Drug Administration 2026; Vafai et al. 2026). Detection of on-target editing in a human biopsy does not itself establish durable functional benefit. The 2026 first-in-human Duchenne muscular dystrophy (*DMD*) study, for example, reported molecular evidence of muscle editing but also a treatment-related fatal serious adverse event at the higher administered dose (Li et al. 2026). An analytically useful therapeutic landscape must therefore be multidimensional rather than a single scalar "editable" label.

We designed a retrospective, reproducible analysis with three connected aims. First, to define the mutation-class landscape of validated human gene–disease–inheritance relationships using the same disease-linked denominators

and explicit confidence rules. Second, to quantify the effects of variant ascertainment, linkage stringency and reference-sequence/PAM constraints on apparent molecular opportunities. Third, to demonstrate—through structured, nonrepresentative clinical and preclinical examples—why molecular addressability, biological rationale, actionable tissue delivery, functional rescue, regulatory status and safety cannot be inferred from one another.

Our central estimand is deliberately narrow: the proportion of high-confidence GenCC relationship records for which the frozen ClinVar crosswalk contains at least one evidence-qualified nucleotide class relevant to further evaluation of genome editing. It is not the prevalence of clinically feasible, deliverable or approved genome-editing therapies.

Materials and methods

Study design, source freeze and analytical hierarchy

This was a cross-sectional secondary analysis of a locally frozen ClinVar variant-summary file and a GenCC submissions file, with additional source-linked computational and translational evaluations. The analysis/evidence cutoff was 6 November 2026. Because the upstream ClinVar snapshot release timestamp was not verified from an independent release manifest, we describe it by its uploaded-file identity and checksum rather than retrospectively assigning it to a nominal calendar-date release. Clinical evidence published after the cutoff was not introduced into the intervention catalog.

Three noninterchangeable units were retained. The molecular unit was the unique ClinVar VariationID; the disease-linked unit was the tuple (HGNC gene identifier, MONDO identifier, mode of inheritance), denoted a gene–disease–MOI relationship; and the evidence-linkage unit was an association edge joining one VariationID to one eligible relationship. A fourth translational unit, intervention–indication, was maintained separately because a treatment can target a different gene from the one establishing hereditary susceptibility. Neither the gene–disease–MOI relationship nor a MONDO identifier is equivalent to a unique patient, allele pair or therapeutic product.

The primary source file was `variant_summary.txt.gz` (457,862,838 compressed bytes; SHA-256 7c39f2d875f3cd1ccd2d568662f06a7df647dd91fc0f51a072a758b08004c72e). Two uploads were byte-identical and were not concatenated. The frozen GenCC input comprised 30,362 submitted assertions. Source files, executable scripts, unaggregated intermediate tables and audit manifests are retained in the research companion archive.

ClinVar selection and molecular normalization

ClinVar records were first restricted to GRCh38 (`Assembly=GRCh38`), followed by `OriginSimple=germline`, a germline `ClinicalSignificance` equal to Pathogenic, Likely pathogenic or Pathogenic/Likely pathogenic, and a `ReviewStatus` reflecting assertion criteria, expert-panel evaluation or a practice guideline. Specifically, low-information or clinically indeterminate records—including benign, likely benign, uncertain significance, conflicting pathogenicity and unreviewed submissions—were excluded from the primary pathogenicity screen. Surviving records were deduplicated by VariationID. The 43-field original schema and the sequential row denominators were preserved.

Literal GRCh38 VCF reference and alternate strings were used to classify single-nucleotide variants (SNVs), multiple-nucleotide substitutions, short indels, larger/complex sequence representations and nonsequence-resolvable entries. An SNV required `Type=single nucleotide variant`, single A/C/G/T nucleotides for REF

and ALT, and REF≠ALT. Accordingly, not every source record tagged as an SNV by ClinVar necessarily entered the sequence-resolved SNV denominator.

This selection uses aggregate variant-level ClinVar information. Its review stars describe consensus or expert review of the *variant classification*, not prospective validation of a specific genome editor or treatment. Because a variation-level summary can aggregate multiple disease associations, assignment of a pathogenic label was not accepted as proof of pathogenicity for every listed phenotype without condition-level RCV/SCV adjudication (Landrum et al. 2025; National Center for Biotechnology Information n.d.a).

High-confidence GenCC gene–disease–inheritance denominator

GenCC assertions were normalized to stable HGNC gene identifiers, MONDO disease identifiers and a harmonized inheritance category. Duplicate submissions for the same triplet were collapsed only for the relationship-level denominator; supporting and conflicting source assertions remained available for review. The strict primary cohort required at least one Strong or Definitive GenCC assertion and excluded triplets having an adverse or insufficient evidence classification under the project's prespecified conflict policy (Limited, Disputed Evidence, Refuted Evidence or No Known Disease Relationship). A Moderate-only relationship was not promoted into the high-confidence cohort (Gene Curation Coalition 2022; Gene Curation Coalition n.d.).

Inheritance groups were harmonized for descriptive analysis as autosomal recessive, autosomal dominant, X-linked, semidominant, mitochondrial and other/unknown. The categories do not encode loss-of-function, gain-of-function, haploinsufficiency, dominant-negative activity, dosage compensation or disease-specific target tissue. In particular, inheritance was never treated as an algorithmic substitute for an experimentally established molecular mechanism.

Disease-linkage crosswalk and stringent evidence policy

The frozen ClinVar and GenCC records were joined by unambiguous HGNC identifier and exact MONDO origin. A source label of `MONDO+OMIM` was retained when the matched assertion also carried the corresponding exact OMIM mapping. The set of distinct associations was represented as (VariationID, HGNC, MONDO, MOI). We did not infer phenotype equivalence from symbol similarity or free-text disease names.

The prespecified stringent linkage predicate was the conjunction of: (i) `MatchSource` in {MONDO, MONDO+OMIM}; (ii) `MultiRelationship=0`, indicating that the variant was uniquely attributable to one eligible relationship in the frozen crosswalk; and (iii) `ReviewTier` equal to ≥2-star multiple-submitter/no-conflict, expert-panel three-star or practice-guideline four-star status. The same predicate was applied to the base-chemistry and short-indel analyses, preserving a common comparison denominator.

The unique-assignment rule is conservative rather than biologically authoritative. True pleiotropic variants and clinically valid associations with imperfect MONDO mapping may be excluded. Conversely, exact ontology matching alone does not verify the causality of an asserted specific variant for the corresponding disease (Landrum et al. 2025; National Center for Biotechnology Information n.d.a). Sequential retention of association edges and relationship keys was quantified to make the evidentiary cost of stringency explicit.

Alternate-to-reference base-editor chemistry

For each sequence-resolved, pathogenic alternate allele, the hypothetical desired correction was the ALT→REF substitution. We classified this reaction as canonical ABE-compatible for A→G or T→C and CBE-compatible for

C→T or G→A, using complement-equivalent classes; all other one-base correction directions were designated outside the two canonical reaction classes. This operational chemistry assessment does not establish whether the available editor acts on the required strand at that locus or whether the target nucleotide falls within a functional deamination window (Gaudelli et al. 2017; Komor et al. 2016).

Let V_i be the set of variants meeting the stringent evidence policy for high-confidence relationship i. We defined $A_i=1$ if any $v \in V_i$ was a chemically compatible SNV, and $B_i=1$ if any $v \in V_i$ satisfied the short-indel sequence-size proxy. The molecular-route union was $U_i=A_i \vee B_i$. Across N=6,968 relationships, the corresponding estimands were $\Sigma A_i/N$, $\Sigma B_i/N$ and $\Sigma U_i/N$. Because A and B need not be disjoint, their overlap was estimated as $\Sigma(A_i B_i)/N$ and the union verified by inclusion–exclusion: $\Sigma U_i=\Sigma A_i+\Sigma B_i-\Sigma(A_i B_i)$.

These are flags for the existence of a *documented* molecular class in a gene–disease assertion. They are neither probabilities of correcting a given patient's genotype nor estimates of a clinically effective editing fraction. Canonical ABE/CBE chemistry does not include transversions, conventional nuclease disruption, prime editing or integration-based interventions.

Short-indel evidence and interpretation boundaries

A record was classified as a short-indel sequence-size proxy when both literal allele strings were composed of A/C/G/T, the absolute net difference in sequence lengths was ≤20 nucleotides and the longer allele was ≤100 nucleotides. These are operational binning thresholds inherited from the frozen Stage 1 pipeline; they are *not* empirically calibrated boundaries of an editor's performance. We counted this class independently under the same relationship and confidence filters used for ABE/CBE chemistry.

Actual prime-editor design additionally requires a suitable pegRNA, PAM or alternative targeting architecture, appropriate primer-binding-site and reverse-transcription template design, acceptable nicking geometry, cell-specific DNA-repair compatibility, by-product characterization and delivery of a comparatively large editor cargo (Anzalone et al. 2019; Chen et al. 2021). Accordingly, the short-indel class is labelled potential sequence-repair substrate, not "prime-editable variant."

Chromosome 21 sequence-context validation

For a predefined chromosome 21 pilot, candidate pathogenic/likely pathogenic SNVs were cross-checked against a GRCh38-labelled chromosome 21 FASTA by exact genomic reference-allele agreement. The alternate disease allele, rather than only the reference genome, was modeled on both strands. Twenty-nucleotide protospacers were enumerated with the target nucleotide within the one-based, PAM-distal positions 4–8 editing window.

The baseline motif was SpCas9 NGG. Under otherwise identical assumptions, recognition was broadened to SpG NGN and the preferential SpRY NRN class; NNN was analysed as a theoretical motif-permissiveness ceiling. A bystander flag was assigned where an additional susceptible base for the same edit chemistry occurred within the nominated window. We did not assign bystander clinical consequence, editing efficiency, off-target activity, by-product purity or expression of the final edited protein (Walton et al. 2020).

The PAM classes were nested motif-recognition models on the same sequence subset, not four independent intervention arms. Although SpRY has expanded recognition, the model does not presume equivalent editing activity or specificity at every NRN site. These pilot results are not extrapolated to unscanned chromosomes.

Variant ascertainment, evidence stratification and descriptive statistics

For each relationship, the number of direct disease-linked variant association edges was counted. We quantified concentration using the share of all edges contributed by the 70 most highly annotated triplets (approximately 1% of N), and the Gini coefficient for the relationship-level documentation distribution. For ascending edge counts $x_{(1)} \ldots x_{(N)}$, $\text{Gini} = 2\Sigma_i\, i \cdot x_{(i)} / (N\Sigma_i\, x_{(i)}) - (N+1)/N$. Values approaching unity indicate extreme concentration of annotation among a minority of relationships, not concentration of true disease incidence.

Separate sensitivity analyses compared all accepted ClinVar review tiers against ≥2-star and ≥3-star subsets; inherited-disease groups were reported without inferential causal tests. Earlier equal-depth rarefaction was interpreted exclusively as a demonstration that an "at least one" candidate endpoint approaches saturation as variant documentation accumulates. We did not calculate a patient-weighted treatability fraction because allele frequencies and genotype-specific eligibility denominators were not available.

The complete cohort is a filtered catalog rather than a probability sample of people or human disorders. Hence nominal binomial confidence limits and Fisher tests should not be used to suggest population inference. Counts, rates, stratified contrasts and the frozen audit trail are treated as descriptive.

Genomic compartment, gene architecture and disease biology

Mitochondrial DNA-encoded genes were separated from the nuclear chromosomal cohort, because mammalian mitochondrial DNA requires distinct RNA-independent sequence-recognition systems and organelle-access strategies (Cho et al. 2022; Cho et al. 2024; Mok et al. 2020). A mitochondrial SNV meeting the reaction-chemistry rule was not considered an SpCas9-targetable site.

For a nonrepresentative benchmark of nine genes, inclusive GRCh38 locus spans were obtained from source-verified ClinGen annotations, with locus length defined as genomic end−start+1. Genomic locus span was not substituted for coding-sequence (CDS) length, the size of an editing cargo, genomic accessibility or the number of nucleotides required for correction. Genome-wide MANE transcript/CDS and Human Phenotype Ontology annotations had not been integrated into the analysis; no missing measurements were imputed (Human Phenotype Ontology Consortium 2026; National Center for Biotechnology Information n.d.b).

Disease-mechanism and target-tissue review was conducted on selected triplets to test the annotation schema. These reviews could establish plausible functional interpretations but did not independently verify that editing an identified cell compartment could provide sufficient therapeutic rescue.

Translation and intervention-level evidence framework

The evidence catalog consisted of 14 purposively selected intervention–indication records available by the cutoff: two approved indication records corresponding to one product, nine additional human-administered intervention records and three preclinical functional-rescue examples. Source type (regulatory label, peer-reviewed human study, preclinical experiment or company disclosure), variant restriction, edited locus, causal gene, edited compartment, clinical phenotype and adverse events were retained separately.

Evidence maturity was designated L1 (in silico molecular plausibility), L2 (experimental editing demonstrated), L3 (disease-relevant preclinical functional improvement), L4 (at least one human participant administered the editing intervention) or L5 (regulatory approval for that specific indication). Levels describe evidence *maturity*, not a summed measure of benefit, safety or suitability. A fatal treatment-related event does not revert human editing

evidence to L2, but it radically changes the safety interpretation. Conversely, a biochemical response following editing a disease modifier does not verify direct correction of the causal allele.

Evidence transfer from an intervention to the GenCC triplet required alignment of the causal gene, condition identifier, inheritance or genotype context and the intervention's demonstrated indication. An exact gene-name match to the edited target, without patient-level disease causality, was insufficient. The clinical catalog was purposively selected, not a systematic registry enumeration; its small numerator was never extrapolated to all 6,968 triplets.

Reproducibility and quality control

Sequence-class, relationship-level and variant-edge tables were exported with source identifiers and retained with their scripts. The primary source hash, stage-specific manifests and independent cross-stage consistency checks were archived. For this revision we additionally reloaded the frozen 6,968-row strict molecular matrix and 269,669-edge crosswalk, checked mutually exclusive route partitions, reproduced the strict chemistry and short-indel denominators, and independently computed a filter-attrition audit and documentational Gini coefficient.

The full cohort and linkage computations are available as archived code and data tables. No new experimental data were generated; original wet-laboratory experiments, patient-level reanalysis, RNA-seq, off-target sequencing, genome-wide prime-editor enumeration and clinical effectiveness meta-analysis are outside the scope of the current work.

Results

Sequential ClinVar filtering yields 312,679 unique qualifying variants

The uploaded ClinVar table contained 9,345,867 rows across assembly representations. Restriction to GRCh38 retained 4,639,270 rows; 4,316,360 were marked germline, 338,226 had an accepted pathogenic/likely pathogenic classification, and 312,753 additionally met the review-status criterion. Deduplication removed 74 repeated eligible VariationID rows, yielding 312,679 unique screened variants (Table 1; Figure 1). The reduction from the initial source is a property of assembly duplication and our strict germline/pathogenicity/review policy rather than a count of intrinsically ineligible genetic diseases.

These denominators have different units. The first five steps are row counts; the final value represents distinct VariationIDs. The 312,679 figure should not be interpreted as unique etiologic alleles for a single disease, and certainly not as a count of affected individuals. The uploaded source contained multiple representations and an aggregate phenotype field, motivating the separate disease-linked crosswalk in subsequent analyses.

Table 1. Sequential ClinVar selection. Source observations are row counts until variation-level deduplication.

Filtering step	Retained count	Unit
All assembly records	9,345,867	rows
GRCh38	4,639,270	rows
GRCh38 + germline	4,316,360	rows
Accepted germline P/LP	338,226	rows
Review criteria provided or higher	312,753	rows
Deduplicated VariationIDs	312,679	unique variants

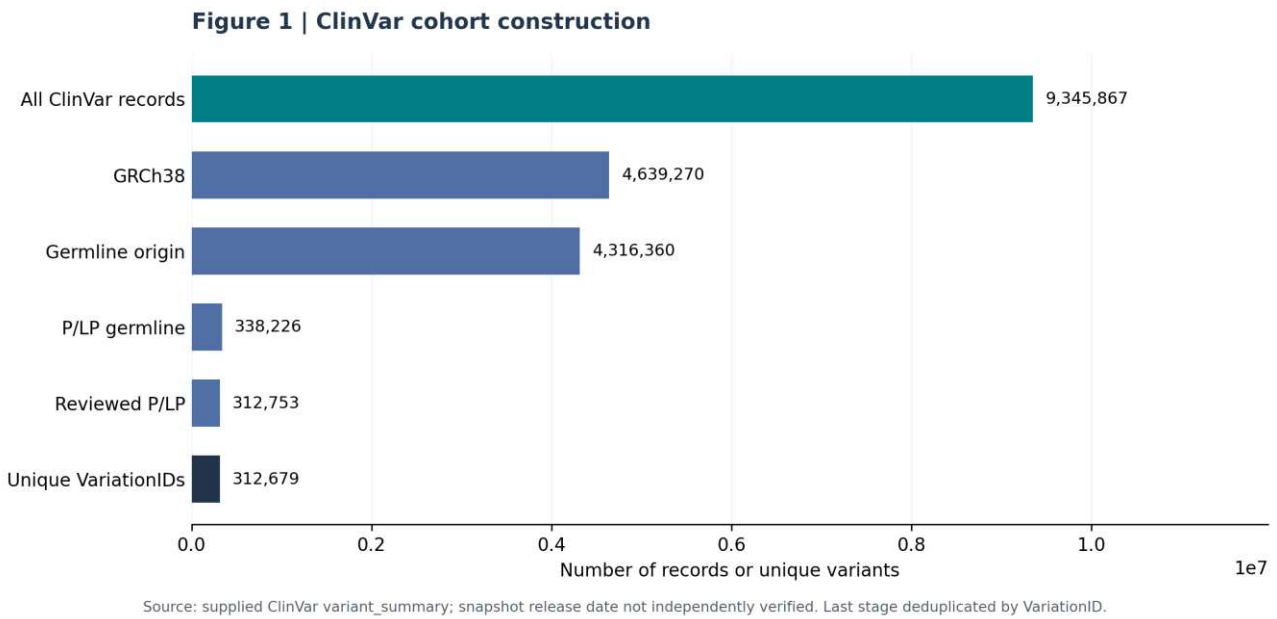


Figure 1. Flow of the source selection and deduplication filters. Original frozen ClinVar input; first stages count records, final stage counts VariationIDs.

Alt text: Flow diagram of ClinVar inclusion filters. From 9,345,867 source rows, 312,679 unique pathogenic or likely pathogenic germline variants remain after build, origin, classification, review, and deduplication filters.

Chemistry-compatible SNVs are common, but nucleotide chemistry is not a guide design

Among the 312,679 selected variants, 168,648 (53.93%) had a sequence-resolved single-nucleotide substitution; 131,830 (42.16%) fell into the prespecified short-indel size class. The remaining categories comprised 6,159 larger or complex sequence representations, 1,469 multinucleotide substitutions, 4,570 nonsequence-resolvable records and three residual sequence representations (Table 2). These mutually exclusive categories sum exactly to the unique selected-variant denominator.

Of 168,648 sequence-resolved SNVs, 92,687 (54.96%) could hypothetically be corrected by canonical ABE or CBE chemistry (Figure 2). ABE-compatible transitions accounted for 71,169/168,648 (42.20%) and CBE-compatible transitions for 21,518/168,648 (12.76%); among the 92,687 chemically compatible SNVs, 76.78% required ABE chemistry and 23.22% CBE chemistry. The remaining 75,961 (45.04%) required a different substitution mechanism. We emphasize that the two editors' representation in a catalog reflects the distribution of pathogenic reference-to-alternate calls and the direction of correction, not a head-to-head comparison of their experimental performance.

The chemistry-compatible class represents 29.64% of *all* 312,679 selected variant records because the broader denominator includes indels and structural/complex representations. The 54.96% SNV-conditional and 29.64% all-variant-conditional proportions are both correct, but they answer different questions. Neither proportion establishes that an edit can be placed within a usable window or that its phenotype can be rescued.

Table 2. Mutually exclusive sequence-representation classes among 312,679 unique variants.

Variant class	Unique variants	Share of 312,679
Sequence-resolved SNV	168,648	53.93%
Short-indel size proxy	131,830	42.16%
Larger/complex sequence representation	6,159	1.97%

Variant class	Unique variants	Share of 312,679
Multinucleotide substitution	1,469	0.47%
Alleles not sequence-resolvable	4,570	1.46%
Other sequence representation	3	<0.01%
Total	312,679	100%

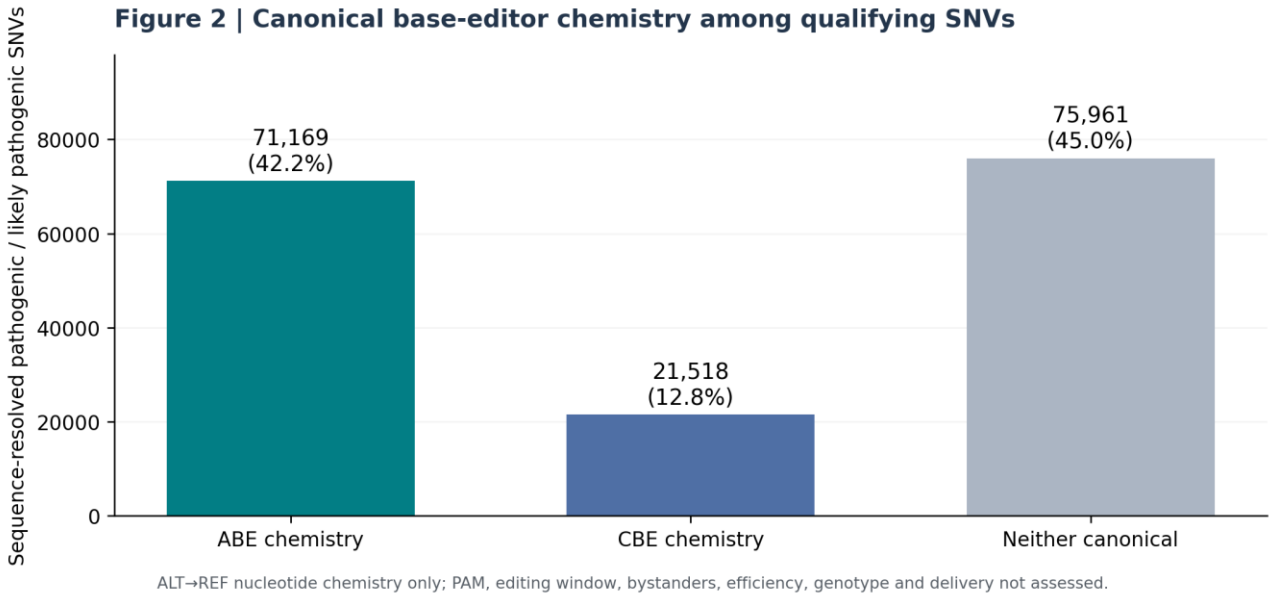


Figure 2. Canonical ALT→REF reaction-chemistry classes among 168,648 eligible SNVs. This is a nucleotide-class screen, not evidence of a functioning guide–editor design.

Alt text: Bar chart of canonical alternate-to-reference chemistry among 168,648 sequence-resolved pathogenic SNVs, distinguishing 92,687 ABE/CBE-compatible conversions from 75,961 incompatible reactions.

A conservative GenCC denominator contains 6,968 relationship triplets

The source GenCC submission inventory contained 30,362 assertions across 6,111 genes and 8,142 disease identifiers. After consolidation to exact HGNC–MONDO–MOI triplets there were 15,361 unique relationship keys. Applying the stringent Strong/Definitive, nonconflicting evidence policy retained 6,968 triplets representing 3,844 genes and 5,329 distinct MONDO identifiers.

Autosomal recessive assertions were the largest category (3,548; 50.92%), followed by autosomal dominant (2,766; 39.70%), X-linked (402; 5.77%), semidominant (183; 2.63%), mitochondrial (40; 0.57%) and other or unknown (29; 0.42%). Under broad linkage, 5,475/6,968 (78.57%) triplets had at least one ClinVar variant, 5,242 (75.23%) had a linked SNV, and 4,242 (60.88%) had an eligible short-indel record. The remaining 1,493 triplets had no directly linked variant, an observation best understood as a combination of catalog coverage and ontology-linkage limitations.

Inherited-disease strata showed unequal proportions with a stringent molecular-class signal (Table 3; Figure 3). Autosomal recessive relationships had 1,587/3,548 (44.73%) either-class flags compared with 967/2,766 (34.96%) autosomal dominant relationships. The corresponding descriptive difference was 9.77 percentage points. This is not evidence that recessive disorders are intrinsically easier to treat: the proportions are confounded by diagnostic testing, disease-gene ascertainment, variant class distributions, severity and phenotype-specific submission practices. The semidominant category's low rate (6/183) particularly illustrates that classification and annotation composition can dominate apparent editability.

Table 3. Inheritance-stratified direct variant coverage and stringent molecular-route flags. Mitochondrial records require a separate editor domain.

Inheritance	All triplets	Any linked variant	Strict ABE/CBE	Strict short indel	Either class
Autosomal dominant	2,766	2,092	868	482	967 (35.0%)
Autosomal recessive	3,548	2,866	1,366	1,185	1,587 (44.7%)
Mitochondrial	40	36	19	6	20 (50.0%)
Other/unknown	29	27	0	0	0 (0.0%)
Semidominant	183	147	5	5	6 (3.3%)
X-linked	402	307	138	85	152 (37.8%)

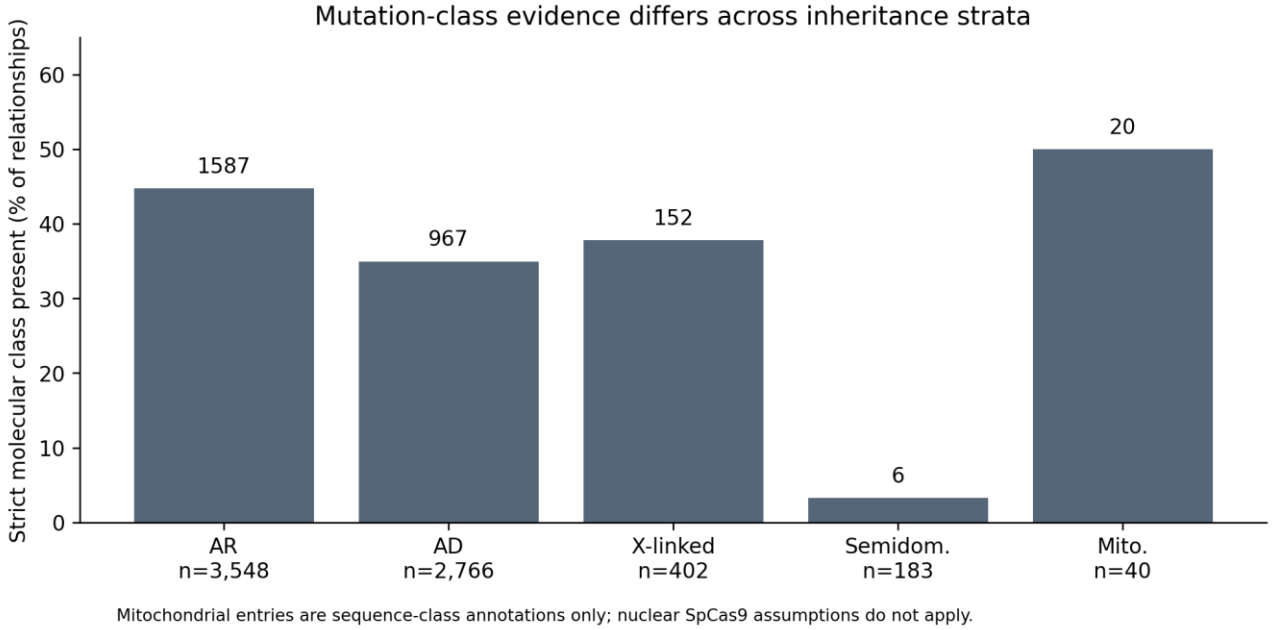


Figure 3. Descriptive difference in strictly linked molecular-class evidence by inheritance. Rates are fractions of relationship records, not probabilities of disease treatability.

Alt text: Grouped chart comparing the proportion of high-confidence gene–disease–inheritance relationships with strict molecular-class evidence across inheritance groups; category sizes vary and are not estimates of treatability.

Exact disease linkage and confidence rules materially restrict the molecular evidence

The direct GenCC–ClinVar crosswalk contained 269,669 variant–relationship association edges, derived from 192,050 distinct ClinVar VariationIDs and representing 5,475 relationship keys. Exact MONDO-origin linkage retained 260,561 edges and 5,396 keys. Requiring that each variant map to only one eligible relationship reduced the set to 118,637 edges and 4,013 keys. Imposing the additional ≥2-star ClinVar requirement further reduced the evidence to 46,786 edges and 2,917 triplets (Table 4; Figure 4). Thus the strict definition retained only 17.35% of the original association edges.

The reduction must not be described as the removal of "false" variants. It represents an intentionally conservative approach to ambiguous disease mapping and conflicting or low-confidence annotation. Excluded associations may be biologically valid, including legitimately pleiotropic variants. Conversely, retained associations are not individually certified pathogenic for a condition without ClinVar assertion-specific review. The contrast makes transparent how strongly the molecular opportunity estimate depends on evidence attribution.

This attrition also distinguishes an underlying biological question from a data-quality question. The presence of a disease-linked variant under permissive mapping is conceptually different from the presence of a confidently

attributable variant relevant to a precise gene–disease–MOI assertion. Because the latter is the appropriate starting point for genotype-specific therapeutic prioritization, stringent results are emphasized for the principal molecular screen and permissive counts are retained as sensitivity context.

Table 4. Sequential attrition of distinct variant–relationship edges under the prespecified strict linkage rules.

Evidence condition	Edges	Triplets	Unique VariationIDs
All disease-linked edges	269,669	5,475	192,050
Exact MONDO-origin matches	260,561	5,396	190,982
Unique relationship assignment	118,637	4,013	118,637
Plus ClinVar review >=2 stars	46,786	2,917	46,786

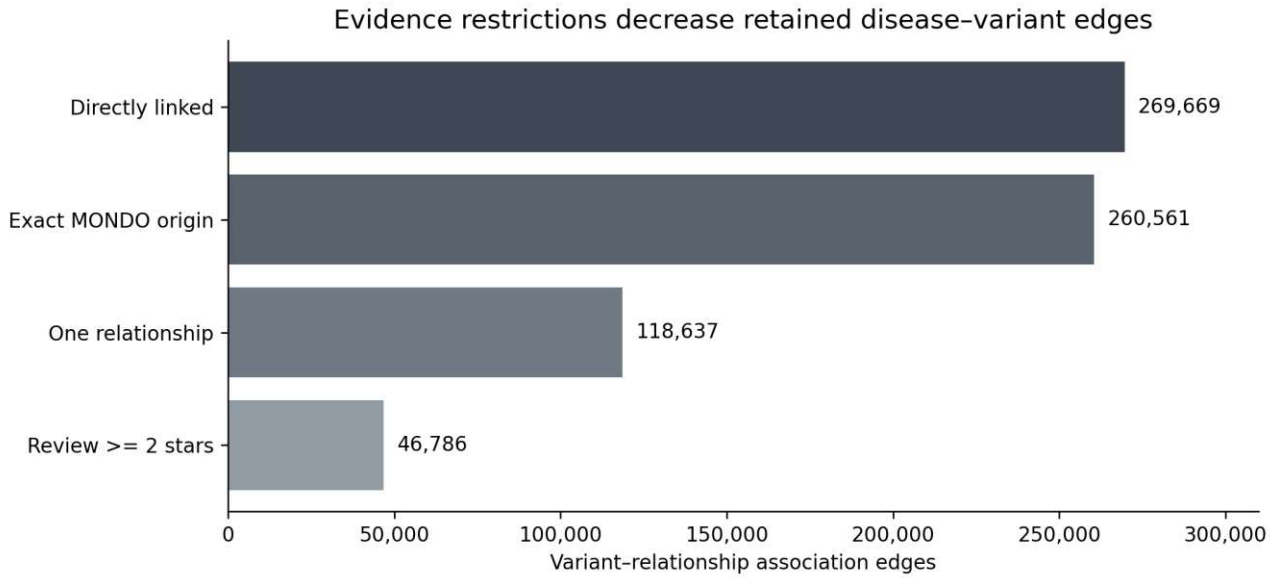


Figure 4. Sequential evidence filters; bars show association edges. Discarded edges are not necessarily biologically incorrect; stringent mapping lowers ambiguity at the expense of sensitivity.

Alt text: Sequential attrition chart showing the reduction of 269,669 variant–disease association edges to 46,786 with exact identifiers, unique assignment, and stricter ClinVar review.

Mutation-class union identifies 2,732 candidate relationship records, not 2,732 therapies

Within the same high-confidence 6,968-triplet denominator, 2,396 relationships (34.39%) had at least one stringent, canonical ABE/CBE-chemistry-compatible SNV, and 1,763 (25.30%) had a stringent short-indel record. These sets were markedly overlapping: 1,427 triplets (20.48% of all triplets) contained both types. Inclusion–exclusion therefore yielded 2,732 nonredundant triplets, 39.21% of the full cohort (Table 5; Figure 5).

The mutually exclusive partition was 969 triplets (13.91%) with chemistry-only evidence, 1,427 (20.48%) with both classes, 336 (4.82%) with short-indel evidence but no stringent chemistry-compatible SNV, and 4,236 (60.79%) with neither class. The 336 short-indel-only triplets increase the number of identified molecular evidence opportunities by 14.02% relative to the 2,396 strict base-chemistry triplets, but this incremental gain must not be reported as a prime-editing-success rate.

Conditioning only on the 5,475 triplets with any directly linked ClinVar variant raises the observed either-class ratio to 49.90%. Conditioning instead on the 2,917 triplets with at least one strict edge raises it further to 93.66%. The striking dependence on denominator demonstrates a core methodological risk: restricting analyses to well-characterized, confidently matched disease assertions can create an apparent near-saturation of addressable *variant*

classes without establishing broad clinical eligibility. We retain the full 6,968 denominator for the primary rate and disclose both conditional sensitivities.

Table 5. Mutually exclusive stringent molecular-evidence classes (N=6,968 gene–disease–MOI triplets).

Class	Triplets	% of all triplets	Interpretation
Chemistry-compatible SNV only	969	13.91	Reaction class; no guide proof
Both SNV chemistry and short indel	1,427	20.48	Overlapping molecular classes
Short-indel size class only	336	4.82	Repair-substrate triage, not prime design
Neither stringent class	4,236	60.79	No qualifying evidence in these two classes
Either stringent class	2,732	39.21	Nonredundant union; not treatable disorders

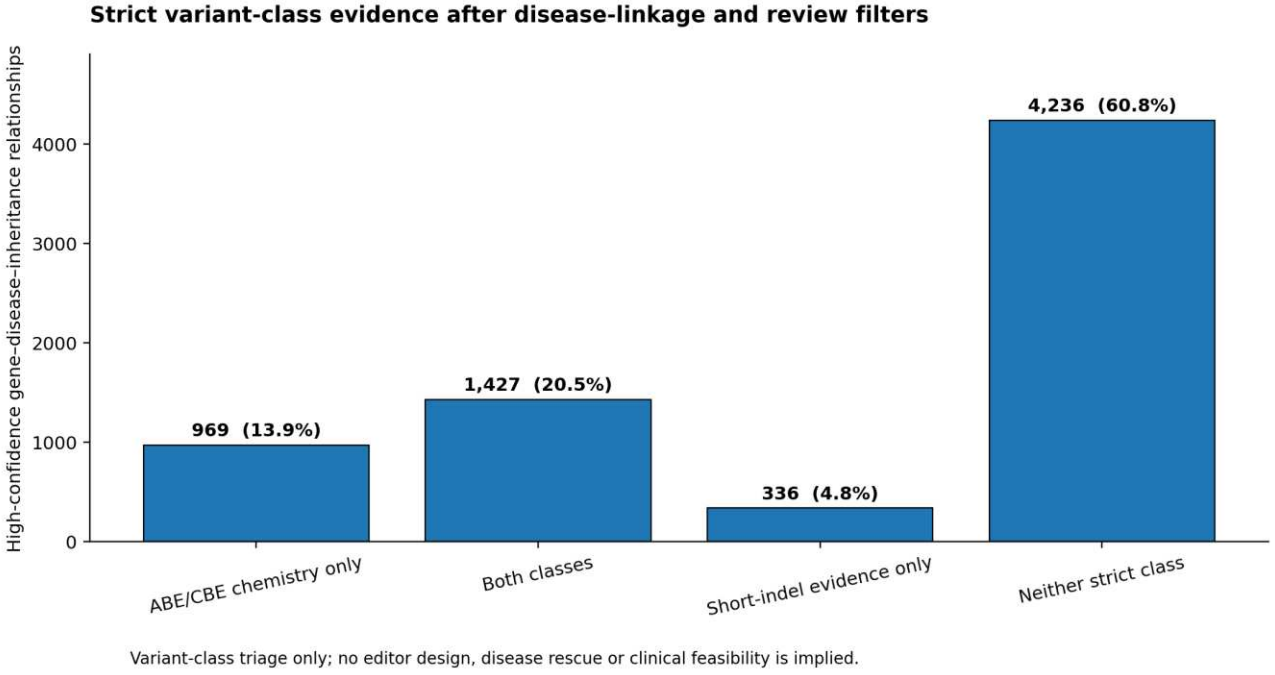


Figure 5. Overlap of stringent canonical ABE/CBE chemistry and short-indel evidence among 6,968 triplets. Class membership does not establish editing efficacy.

Alt text: Four-category comparison for 6,968 GenCC relationships: base chemistry only 969; short-indel evidence only 336; both 1,427; neither 4,236.

Review-tier sensitivity and heavy-tailed clinical variant ascertainment

The relation between review confidence and apparent chemical addressability was not monotonic. Across the full screened cohort, 92,687 of 168,648 SNVs (54.96%) met canonical ABE/CBE chemistry. Restricting to ≥2-star clinical review yielded 41,913/65,935 (63.57%), whereas the expert-panel/practice-guideline subset yielded 3,488/6,579 (53.02%; Table 6). Within the main cohort, the one-star SNV stratum contained 50,774/102,713 chemistry-compatible calls (49.43%), compared with 38,425/59,356 (64.74%) for exactly two-star calls. The pattern reflects shifts in which genes and variant classes receive higher levels of clinical interpretation; it cannot be explained as a simple monotonic relationship between biological editability and review rigor.

Variant documentation was substantially unequal across disease assertions. The number of directly linked association edges per triplet ranged from zero to 4,435, with a median of seven. The top 70 triplets, approximately

1% of the cohort, contributed 28.18% of all 269,669 edges. A newly recalculated Gini coefficient of 0.8047 quantified the extreme concentration (Figure 6). These counts were independently reproduced from the complete 6,968-row frozen matrix and precisely reconciled to the edge-level crosswalk.

As the number of catalogued alleles for a disease increases, the chance of encountering at least one variant with a common chemistry rises even if the disease's actual therapeutic barriers do not change. Previously archived equal-depth rarefaction analyses demonstrated this ascertainment ceiling effect; the present concentration analysis explains the mechanism by showing how uneven documentation loads the relationship-level "any variant" statistic. For that reason, we did not interpret the relationship-level flags as a clinical burden-adjusted estimate of eligible patients or conditions.

Table 6. Sensitivity of canonical ABE/CBE reaction chemistry to ClinVar review-confidence threshold.

Nested review threshold	All variants	Sequence-resolved SNVs	ABE/CBE SNVs	Among SNVs
≥1 star	312,679	168,648	92,687	54.96%
≥2 stars	103,659	65,935	41,913	63.57%
≥3 stars	12,593	6,579	3,488	53.02%

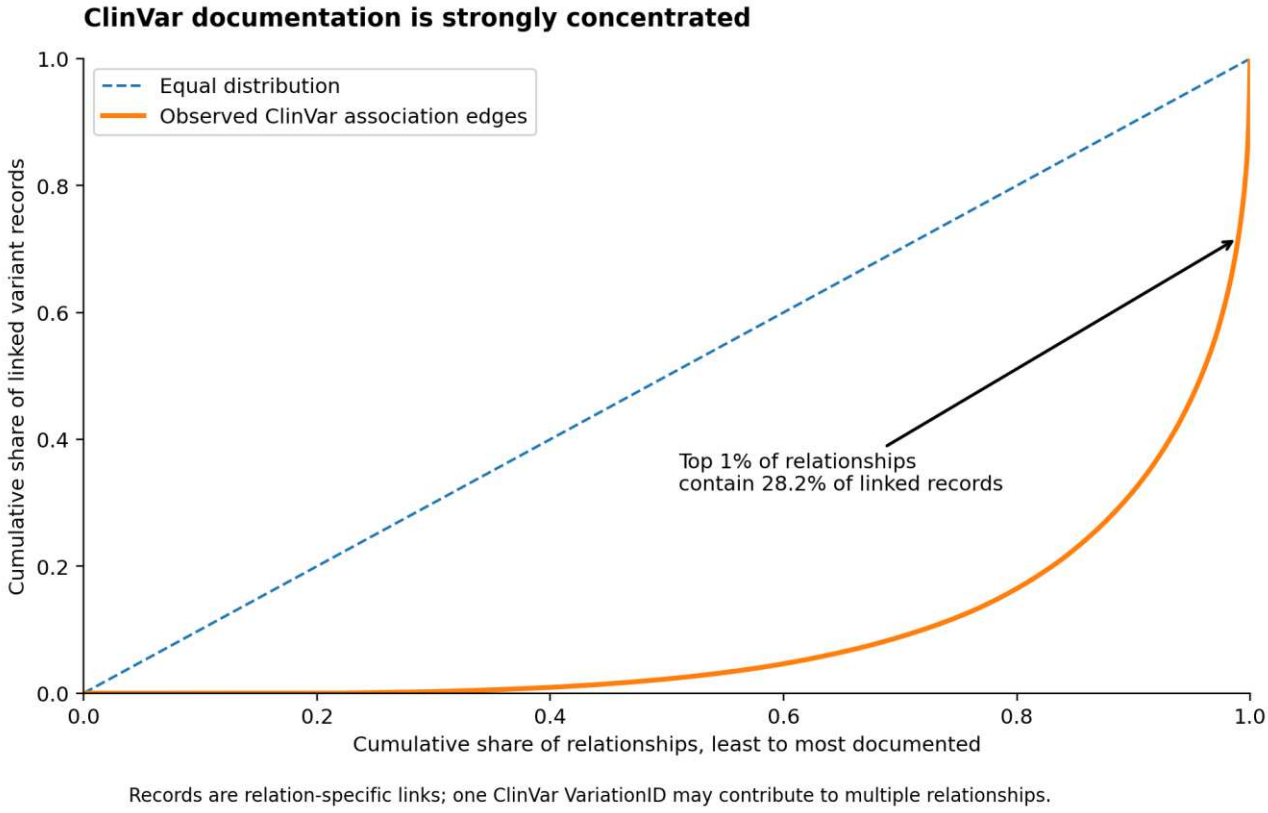


Figure 6. Lorenz distribution for disease–variant association edges in the frozen GenCC cohort. The 70 most annotated triplets contribute 28.18% of total edges (Gini=0.805).

Alt text: Lorenz curve showing marked concentration of variant documentation across high-confidence relationships. The Gini coefficient is 0.805 and the top 1% of relationships contribute 28.18% of edges.

The chromosome 21 pilot reveals a substantial gap between chemistry and local guide access

Chromosome 21 contained 1,847 qualifying pathogenic/likely pathogenic SNVs in the sequence-context pilot, of which 1,071 had canonical ALT→REF reaction chemistry. Exact GRCh38 reference validation was incorporated before guide enumeration. Under the baseline 20-nucleotide protospacer and fixed positions 4–8 window, 373 of

the 1,071 (34.83%) had an NGG-associated guide candidate (Table 7; Figure 7). Of these 373, only 169 (45.31%) had a candidate free of *additional same-class editable nucleotides within the assumed window*. The corresponding chemistry-to-nominally-bystander-free reduction was 169/1,071 (15.78%).

Expanding PAM recognition from NGG to NGN increased guide-recognized variants to 865 (80.77% of the chemistry set) and nominally bystander-free candidates to 421; SpRY preferred NRN yielded 1,022 (95.42%) and 567, respectively. The synthetic NNN ceiling reached all 1,071 chemistry-compatible variants, but only 759 lacked another modeled same-class editable base in at least one candidate. These are set-theoretic motif results, not empirically measured editing windows or expected patient response probabilities.

A more permissive PAM creates additional opportunities to *nominate* guides, but does not erase constraints from allelic sequence, window position, bystander consequence, chromatin accessibility, protein function and off-target risk. The 1,022 SpRY preferred-NRN recognized variants should not be translated into a "95% editable" fraction because editor activity varies by locus, PAM and substrate (Walton et al. 2020). The chromosome 21 pilot is also not a random sample of the genome: local sequence composition and the disease-gene mix can differ from other chromosomes.

Table 7. Reference-validated chromosome 21 PAM/editing-window pilot (1,071 canonical chemistry-compatible SNVs).

Recognition model	Any model guide	% of chemistry set	No additional modeled same-class bystander	% of chemistry set
SpCas9 NGG	373	34.83%	169	15.78%
SpG NGN	865	80.77%	421	39.31%
SpRY preferred NRN	1,022	95.42%	567	52.94%
NNN theoretical ceiling	1,071	100%	759	70.87%

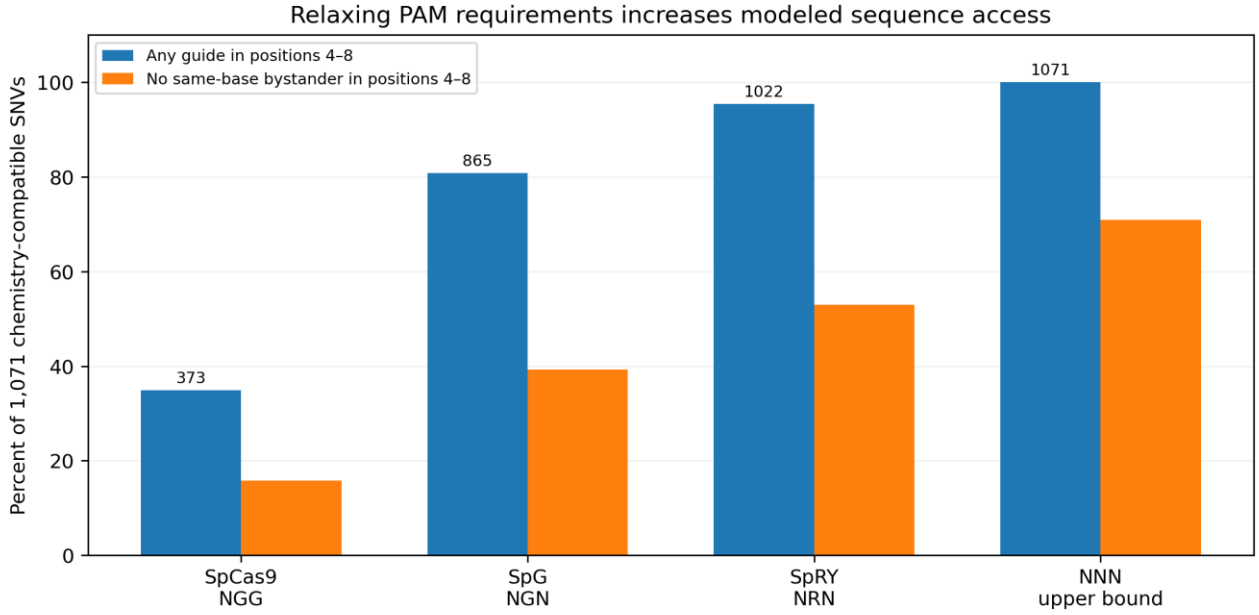


Figure 7. Nested computational guide-recognition models on chromosome 21. Editor kinetics, activity and off-target/bystander clinical effects were not assessed.

Alt text: Chart of chromosome 21 PAM/window compatibility for 1,071 chemistry-compatible variants: 373 NGG, 865 NGN, 1,022 preferred NRN, and 1,071 theoretical unrestricted NNN candidates; nominal bystander-free counts are reported separately.

Mitochondrial gene assertions form a separate molecular and delivery domain

Of the 6,968 high-confidence relationship records, 6,928 involved nuclear genes and 40 involved mitochondrial DNA-encoded genes (Figure 8). The stringent ABE/CBE chemistry count partitioned exactly into 2,377 nuclear and 19 mitochondrial relationships. Twenty mitochondrial relationships had at least one of the two molecular-evidence classes, but the 19 chemistry flags cannot be carried into the conventional SpCas9/guide-PAM denominator. Treating a mitochondrial A→G or C→T correction as though it were performed by a nuclear Cas9–deaminase fusion would conflate distinct molecular delivery and targeting architectures.

RNA-independent DdCBE/TALED platforms have demonstrated mitochondrial base editing in experimental systems, but their context requirements, editing breadth, mitochondrial import, off-target effects and heteroplasmy-shifting effects differ from nuclear editors (Cho et al. 2022; Cho et al. 2024; Mok et al. 2020). A complete mitochondrial therapeutic landscape must additionally model tissue-specific mutant-load thresholds, mitochondrial DNA copy number, organelle access and persistence of edited genomes. Therefore, mitochondrial records are retained in the molecular-class inventory but excluded from claims about conventional nuclear PAM/window design.

Table 8. Genomic-compartment sensitivity for the stringent molecular-evidence screen.

Compartment	GenCC triplets	Strict ABE/CBE chemistry	Strict short indel	Either strict class
Nuclear genes	6,928	2,377	1,757	2,712
Mitochondrial-encoded genes	40	19	6	20
Combined reporting denominator	6,968	2,396	1,763	2,732

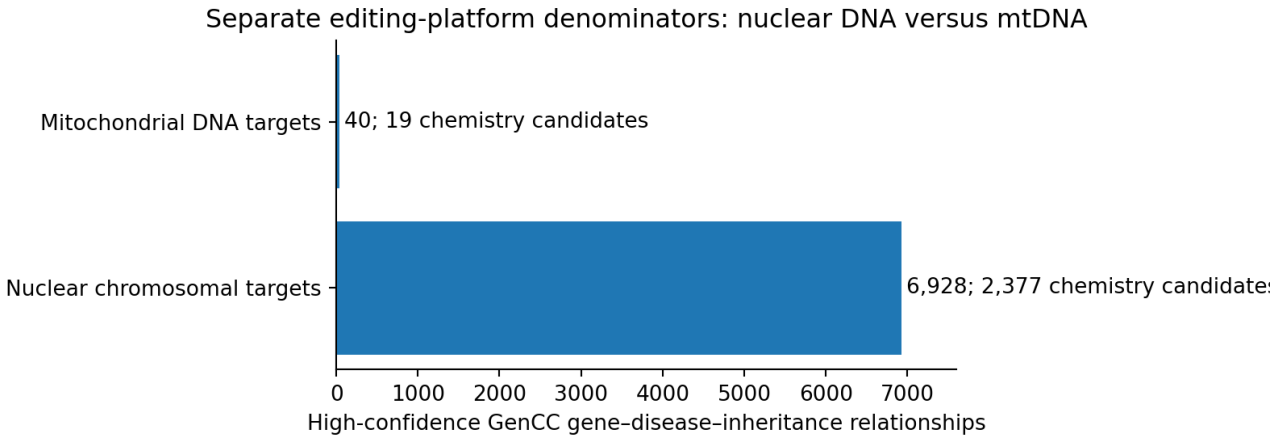


Figure 8. Nuclear and mitochondrial GenCC denominator separation. Mitochondrial chemical compatibility is not evidence of a conventional nuclear Cas9 design.

Alt text: Two-domain nuclear versus mitochondrial comparison showing 6,928 nuclear and 40 mitochondrial GenCC relationships, with 2,377 and 19 respective strict base-chemistry flags. Mitochondrial conversions require distinct editor technologies.

Gene-locus span demonstrates why target architecture and replacement payload are distinct

A source-verified pilot across nine selected genes illustrated the breadth of genomic architecture, from the 1,608-bp *HBB* locus to the 2,220,167-bp *DMD* locus—a 1,380.7-fold difference in inclusive GRCh38 span (Table 9; Figure 9). This heterogeneity may influence available regulatory elements, intron/exon architecture and the complexity of certain gene-integration interventions. It does not, by itself, quantify the difficulty of an allele-specific base edit,

which may require a localized recognition sequence independent of total gene span. Selected inclusive spans were derived from ClinGen gene-record annotations (Clinical Genome Resource n.d.).

Importantly, *BCL11A* is the edited *regulatory target* in CASGEVY, while the causal hemoglobinopathy genes are different. Its entire 103-kb locus span is not a therapeutic editing cargo; the strategy targets an erythroid-specific enhancer. Likewise, *DMD*'s large genomic span does not imply that the entire 2.22-Mb locus must be packaged into a vector for exon-specific editing, and the feasibility of alternative dystrophin-restoration approaches must be assessed in their respective architectures. A genome-wide transcript-level analysis was not possible because the fixed MANE/GENCODE resources were not materialized; no genome-wide CDS-size estimates are reported.

Table 9. Source-verified inclusive GRCh38 genomic locus spans in a nonrepresentative therapeutic benchmark.

Gene	Inclusive locus span	Biological context
HBB	1,608 bp	β-globin causal disease locus
TTR	6,945 bp	Therapeutic silencing in amyloidosis
PCSK9	25,305 bp	Hepatic modifier-gene disruption
CEP290	93,073 bp	Variant-restricted retinal splice intervention
BCL11A	103,405 bp	Erythroid enhancer therapeutic target
PAH	121,553 bp	Enzyme restoration substrate
CFTR	188,641 bp	Airway epithelial variant correction
CPS1	201,423 bp	Liver-directed personalized editing
DMD	2,220,167 bp	Genotype-dependent dystrophin restoration

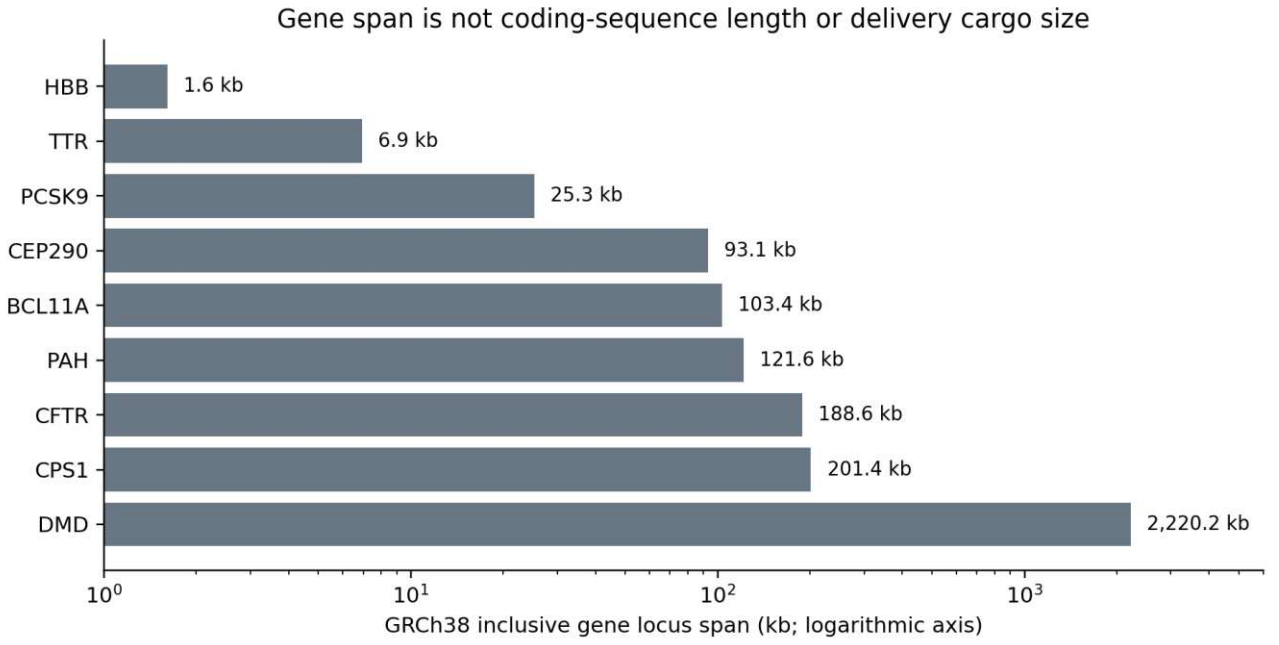


Figure 9. Log-scaled gene-locus comparison across nine selected genes. Genomic span must not be interpreted as transcript length, CDS size or editing-delivery cargo.

Alt text: Log-scale locus-length chart for nine genes, ranging from the 1,608-base HBB locus to the 2,220,167-base DMD locus; gene span is not editor payload length.

Selected clinical evidence decouples the edited target from the hereditary causal gene

The source-linked translational portfolio contained 14 purposively selected intervention–indication entries, including two separate approved indications for the single ex vivo BCL11A-enhancer editing product CASGEVY, nine additional human-administered editing interventions and three disease-relevant preclinical functional-rescue

examples. Ten intervention records could be transferred conservatively to an exact source GenCC relationship under the intervention-specific causality and inheritance policy; this count describes coverage of the deliberately selected program catalog, not the number of all genetic diseases with clinical genome-editing evidence.

Several therapeutic examples are mechanistically informative (Table 10). FDA approval of CASGEVY for recurrent-vaso-occlusive-crisis sickle-cell disease and transfusion-dependent β-thalassemia in patients aged two years and older, as of the July 2026 supplemental approval, rests on ex vivo disruption of a cell-type-selective *BCL11A* enhancer and consequent fetal hemoglobin induction (US Food and Drug Administration 2026). Neither intervention directly corrects the inherited *HBB* or β-globin regulatory/coding variants. This is a decisive counterexample to treating direct disease-linked ABE/CBE chemical correction as a necessary condition of therapeutic success.

The 2025 personalized *CPS1* intervention instead represents direct, genotype-specific editing in vivo. Two infusions of an individualized lipid-nanoparticle base editor in a child with severe carbamoyl-phosphate synthetase 1 deficiency were followed by increased protein tolerance and reduced nitrogen-scavenger exposure over early follow-up; the original report explicitly cautioned that longer-term assessment was required (Musunuru et al. 2025). The ability to manufacture one highly specific treatment does not imply that every *CPS1* variant has a compatible editor or that the disease is generically curable.

The 2026 randomized HAELO trial of in vivo *KLKB1* disruption in hereditary angioedema with C1 inhibitor deficiency provides stronger controlled clinical efficacy evidence for an *indirect modifier-target* strategy. Eighty patients were randomized (52 active, 28 placebo); the reported monthly attack rate between weeks 5 and 28 was 0.26 versus 2.10, an adjusted relative difference of −87% (95% confidence interval −93% to −78%) (Cohn et al. 2026). Yet this treatment does not require direct correction of the typical causal *SERPING1* locus; furthermore, the report does not authorize transferring that response to a specific MONDO–MOI genotype record without source-level etiologic verification.

The VERVE-102 phase 1 trial illustrates the reciprocal attribution problem: its edited therapeutic target, *PCSK9*, is the same as a gene that can *also* cause familial hypercholesterolaemia, but enrollment included individuals with heterozygous familial hypercholesterolaemia or premature coronary disease without an etiologic requirement that each carry a PCSK9-causal variant. Thirty-five participants were treated and showed dose-dependent *PCSK9* and LDL reductions (Vafai et al. 2026). We therefore withheld this program from a direct genotype-specific *PCSK9* disease assertion despite compelling evidence that the target can be modified in vivo.

Early human editing evidence also exists for retinal *CEP290* disease, which is genotype- and anatomical-site restricted (Pierce et al. 2024). For *DMD*, biopsy-based editing evidence from a 2026 first-in-human study was accompanied by a treatment-related fatal serious adverse event at the higher dose, characterized by complement activation and systemic inflammatory response (Li et al. 2026). Evidentiary L4 classification correctly describes human administration; it must never be read as a declaration of acceptable safety or durable clinical utility.

Preclinical evidence further clarifies intervention diversity. Optimized prime editing repaired *CFTR* F508del in patient-derived airway epithelial cells and restored CFTR-channel activity to more than 50% of wild-type values in the reported experiments, providing functional support for an allele-specific airway strategy but not proof of successful clinical delivery (Sousa et al. 2025). Separately, an AAV/SaCas9 experiment inserted a therapeutic *UGT1A1* sequence into the mouse albumin locus and produced disease-relevant rescue in a Crigler–Najjar model;

the edited genomic locus (Alb) differed from the disease-causing gene (Ugt1a1) and the expressed cargo (De Caneva et al. 2019). These examples are not interchangeable with untargeted AAV gene supplementation.

Table 10. Mechanistically representative intervention–indication evidence. Selected examples, not an exhaustive clinical trial census.

Intervention / disease	Causal gene → edited target	Strategy and compartment	Evidence at cutoff	Main attribution or safety constraint
CASGEVY / SCD and TDT	*HBB*/globin disorder → *BCL11A* enhancer	Ex vivo erythroid HSC enhancer editing	L5; 2 approved indications	Indirect modifier; myeloablation/engraftment risks
Personalized *CPS1*	*CPS1* → *CPS1*	LNP ABE; hepatocytes	L4, N-of-1	Allele-specific; limited follow-up
NTLA-2001 / ATTR	*TTR* → *TTR*	LNP nuclease disruption; hepatocytes	L4	Mutant and wild-type *TTR* lowering; longitudinal safety
VERVE-102 / hypercholesterolemia	Heterogeneous etiology → *PCSK9*	LNP ABE; hepatocytes	L4 phase 1	Do not transfer to all PCSK9-causal FH patients
EDIT-101 / retinal degeneration	*CEP290* → *CEP290*	Subretinal Cas9 splice correction	L4	IVS26 variant restriction; residual cell viability
Lonvoguran / HAE	Usually *SERPING1* → *KLKB1*	In vivo hepatocyte nuclease disruption	L4; randomized phase 3	Modifier target; phenotype vs causal-genotype distinction
HG302 / *DMD*	*DMD* → *DMD*	Systemic AAV/hfCas12Max; skeletal muscle	L4	Treatment-related fatal serious adverse event
Optimized *CFTR* F508del prime editing	*CFTR* → *CFTR*	Prime editing; airway cell models	L3 experimental rescue	Culture proof, not in vivo airway delivery
SaCas9-targeted *UGT1A1* insertion	*UGT1A1* → *ALB* locus	Targeted cDNA integration; mouse liver	L3 preclinical rescue	Edited locus differs from gene and cargo

Mechanism review confirms that inheritance and gene names are insufficient to assign treatment routes

A stratified 240-triplet probability sample was previously constructed from the 6,959 relationships without conservative clinical-program transfer. Twenty-four deliberately selected triplets received initial mechanism/tissue reviews across earlier project phases; the remaining 216 were unreviewed. Critically, the 24 were not a completed probabilistic review and no independent dual-review adjudication had been performed. They were used exclusively to test whether our annotation schema could capture biological heterogeneity, not to estimate the prevalence of a tractable mechanism across all disorders.

Mechanistically distinct exemplars included activating RET-associated multiple endocrine neoplasia, *PTEN* haploinsufficiency/loss-of-function disorders, recessive *AP4S1* neuronal trafficking deficiency, *GABRA1* epilepsy with variant-dependent gain- and loss-of-function effects, EDAR-related ectodermal phenotypes with inheritance-dependent signalling consequences, and mitochondrial *MT-CO3* disease for which heteroplasmy and organelle-specific tools are central considerations. In each case, the affected organ is not necessarily equivalent to the therapeutically sufficient edited cell population, and inheritance alone does not specify whether knockout, allele replacement, enhancer tuning or targeted integration is appropriate. The differential biological interpretation of EDAR-associated inheritance is consistent with clinical genetics reference material (GeneReviews n.d.).

The five-level evidence framework is illustrated in Table 11 and Figure 10. Evidence maturity must be combined with separate assessments of genotype specificity, disease mechanism, therapeutic cell access, functional rescue and adverse events. These dimensions explain why a uniform “editable disease” classification is not identifiable from our current source combination.

Table 11. Interpretation of evidence-maturity categories: each level is orthogonal to clinical safety and benefit.

Level	Minimum supporting observation	What it does NOT establish
L1	Molecular class/sequence model supports a candidate	Editor activity, tissue delivery or benefit
L2	On-target editing demonstrated experimentally	Disease-specific functional rescue
L3	Functional improvement in a disease-relevant preclinical model	Human pharmacology, dosing or long-term safety
L4	Intervention administered to a human participant	Efficacy, acceptable risk or regulatory approval
L5	Regulatory approval for a stated indication	Transferability to all alleles, diseases or settings

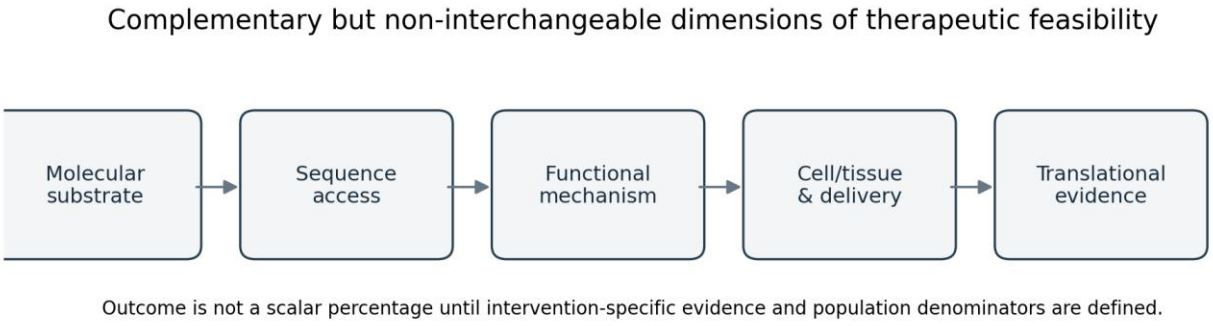


Figure 10. Operational separation of complementary feasibility dimensions. Molecule-level screening is one prerequisite and cannot substitute for phenotype, delivery and clinical evidence.

Alt text: Conceptual framework depicting five separate assessment domains: nucleotide chemistry, disease mechanism, target-cell access, clinical evidence maturity, and safety, rather than one binary editability score.

Discussion

Defining the estimand resolves an important ambiguity in the genome-editing literature

The central quantitative result of this study is a disease-linked molecular inventory, not an estimate of the clinical reach of genome-editing therapies. Using a fixed high-confidence GenCC denominator, 2,732 of 6,968 relationship records contained at least one strictly attributable pathogenic variant in a canonical ABE/CBE chemistry class or a prespecified short-indel category. The result has a clear interpretation as evidence-linked molecular substrate availability under stated database conditions. It becomes misleading if converted directly into "39.21% of monogenic disorders can be edited today."

The distinction is not semantic. A relationship-level binary indicator discards information about how many patients carry a specific allele, whether the variant causes disease by loss or acquisition of function, the fraction of cells requiring correction, the fraction of tissues that can actually be reached, off-target consequences and clinical durability. A disease with a thousand unique catalogued variants requires a variant- or mechanism-specific intervention denominator. Conversely, a therapy designed around one validated pathogenic allele may make a genuine difference to a patient with that genotype without implying all disease alleles are addressable.

The analysis also demonstrates why no single molecular editor provides a universal measure of a disease's potential tractability. ABE/CBE chemistry provides an efficient screening vocabulary for selected transitions (Dadush et al. 2024; Gaudelli et al. 2017; Komor et al. 2016), but intentionally excludes therapies based on gene knockout, enhancer disruption, transversion repair, prime editing and site-specific integration (Anzalone et al. 2019; Chen et al. 2021; De Caneva et al. 2019; US Food and Drug Administration 2026). Therapeutic mechanisms are not nested solely within direct variant reversal. An accurate landscape is therefore a matrix indexed by etiology, genotype, target, tissue, editing modality and evidence maturity, not a scalar percentage of annotated diseases.

Disease identifier and review-stringency filters are scientific design choices, not administrative details

Only 46,786 of 269,669 direct disease–variant association edges met all three stringent source-identification, unique-assignment and review-tier constraints. The attrition is large enough to materially change the observed opportunity landscape. Such dependence on crosswalk policy is especially important when clinical variant summaries aggregate phenotype labels. If the same pathogenic SNV is connected to multiple disease names, a permissive algorithm can create an apparently broad range of disorder-specific targets even when condition-specific assertions have not been reconciled.

Our conservative policy trades sensitivity for interpretability. The retained set is less likely to contain cross-disease assignments generated by loosely overlapping ontologies, but it may systematically exclude genuine pleiotropy, allelic heterogeneity and clinically valid gene–disease associations with mixed evidence histories. Consequently, both the strict and the permissive analyses should be interpreted as alternative evidence-defined estimands rather than a simple hierarchy of accuracy. For a therapeutic-development decision, condition-specific ClinVar RCV/SCV records, original segregation/functional evidence and current molecular diagnosis are required (Landrum et al. 2025; National Center for Biotechnology Information n.d.a).

The particularly high conditional molecular-class fraction among records already possessing a stringent linked variant (93.66%) should be viewed as an ascertainment-conditioned proportion. The strict review criteria preferentially preserve well-annotated variants and the selected two molecular classes encompass a large share of these known alleles. This observation must not be inverted to assert that nearly all inherited disorders have a feasible treatment path. We therefore recommend publishing the full-cohort and conditionally restricted denominators side by side.

Annotation inequality produces an "at least one" saturation problem

The Gini coefficient of 0.805 and the observation that approximately 1% of gene–disease triplets contributed 28.18% of direct variant edges demonstrate a pronounced long-tailed distribution of clinical variant documentation. This inequality is consistent with historical concentration of genetic research in a subset of severe, highly penetrant, frequently tested or clinically actionable genes. It also reflects differential laboratory submission practices and the varying sizes and mutational target spaces of genes.

A binary screen asking whether any of the n documented variants falls into a common nucleotide class becomes progressively easier to satisfy as n increases. Under a stylized independent Bernoulli model with compatible-class probability p, the chance of at least one compatible observation is $1-(1-p)^n$. Although real variants are correlated and the biological/ascertainment assumptions of such a model are violated, the elementary relation illustrates the mathematical direction of bias. An "at least one" statistic is therefore partly a surrogate for documentation depth unless the variant universe is standardized or appropriately conditioned.

Prospective studies should report mutation-class composition with variant count or fixed-depth sensitivity, genotype prevalence and clinically causal variant assertions. The observed relation between inheritance and candidate flags is also susceptible to this bias. Higher strict-class rates among autosomal-recessive than autosomal-dominant GenCC triplets do not establish greater inherent amenability of recessive disorders; an appropriate biological comparison would require similar variant ascertainment, tissue biology, repair thresholds and disease mechanism.

Expanded PAM recognition narrows but does not remove the sequence-to-function gap

The chromosome 21 analysis provides an instructive intermediate level of realism between a chemical conversion and a functional editor. Under conventional NGG, only 373 of 1,071 chemical candidates were nominated within a fixed positions 4–8 editing window, and fewer than half of those retained a candidate without additional same-class editable bases within that same assumed window. Broader PAM recognition markedly expanded nomination, illustrating the practical importance of engineered Cas targeting domains (Walton et al. 2020). Yet the transition from 34.83% NGG-recognized candidates to 95.42% SpRY-preferred-NRN candidates is a change in *motif space*, not a direct empirical estimate of success.

Bystander presence is likewise a qualitative flag requiring variant- and transcript-level adjudication. A neighboring change may be synonymous, pathogenic, splice-disruptive, or therapeutically beneficial; a nominally bystander-free window may still produce RNA off-target or genomic off-target events. The actual product distribution depends on protein engineering, guide expression, strand exposure, DNA repair and cellular state. The fixed 4–8 window used here is a reproducible scenario rather than a universal physical property of every base editor.

Prime editing adds further design dimensions. For a short insertion or deletion, the same nominal change may be realized with different pegRNA template designs, primer-binding-site lengths, nicking strategies and mismatch-repair contexts. Chen and colleagues demonstrated that manipulating mismatch repair and editor architecture can substantially change editing yield and product purity across targets (Chen et al. 2021). It would therefore be inaccurate to label the 336 new indel-only relationships “prime-editable” without explicit prime-editor design, cellular validation and disease-relevant functional rescue. Our indel class is best used to prioritize additional sequence analysis, not to claim therapeutic coverage.

Molecular pathogenesis and gene dosage determine what constitutes a therapeutically correct edit

Even perfect editing of a chosen nucleotide does not guarantee restoration of physiology. For recessive loss-of-function disorders, the intervention may need to restore sufficient enzyme or structural-protein function in a critical cell fraction, and a small amount of cellular correction can sometimes be disproportionately useful if corrected cells enjoy selective advantage. For haploinsufficiency, dosage restoration and cell-type expression are central. For dominant-negative or activating alleles, adding wild-type expression or ablating both alleles may fail or worsen disease. In many genotype-dependent channelopathies and developmental syndromes, a universal gene-level knockout decision is biologically indefensible.

Causal genes and edited targets must be distinguished explicitly. The *BCL11A* erythroid-enhancer strategy exploits a compensatory haematologic program, and its tissue-selective regulatory target matters: indiscriminate disruption of the entire *BCL11A* gene would not reproduce the benefit–risk profile of enhancer editing. Similarly, *KLKB1* disruption reduces the abundance of a downstream mediator of hereditary-angioedema attacks rather than correcting C1 inhibitor deficiency itself (Cohn et al. 2026; US Food and Drug Administration 2026). These

examples imply that a catalog restricted to ALT→REF correction will systematically undercount plausible modifier-based approaches, yet conversely a successful modifier edit cannot be attributed to every inherited disorder associated with that modifier gene.

Genomic locus length also has to be interpreted in relation to intervention architecture. Although the source-verified *DMD* locus is orders of magnitude larger than *HBB*, an exon-specific nuclease approach is not equivalent to full-length gene replacement. Similarly, the requirement to express a large protein may constrain gene addition but does not, by itself, prevent local base editing. Gene length, CDS length, regulatory structure, edit template size, protein turnover and delivery cargo must be maintained as separate quantitative fields. A unified atlas that conflates them will misclassify therapies.

Tissue reach, clinical evidence and safety create bottlenecks orthogonal to DNA chemistry

Tissue access is not a passive implementation detail. Ex vivo editing of autologous hematopoietic cells enables controlled exposure to editing reagents, clone/product-quality assessment and re-infusion, though transplantation and myeloablative conditioning carry substantial risks (US Food and Drug Administration 2026). Hepatocyte-targeted lipid nanoparticles have enabled clinically important in vivo nuclease and base-editor delivery (Cohn et al. 2026; Gillmore et al. 2021; Musunuru et al. 2025; Vafai et al. 2026). In contrast, achieving sufficiently distributed, durable and safe genome editing across skeletal muscle, brain or mature retina introduces distinct delivery limitations and potentially irreversible disease-timing effects (Li et al. 2026; Ngo et al. 2026; Pierce et al. 2024).

The quality of clinical evidence also differs substantially across editing programs. A personalized, first-in-human case report can demonstrate manufacturing and molecular feasibility without proving durable efficacy (Musunuru et al. 2025). A randomized phase 3 trial with a defined clinical endpoint provides a different evidence class for a particular intervention and indication (Cohn et al. 2026). The fatal event reported in the 2026 *DMD* trial demonstrates why evidence of human on-target editing must not be conflated with a favourable risk–benefit balance (Li et al. 2026). All three are human-administered, but the external validity, patient populations and safety conclusions differ radically.

A disease-level claim should be formulated at the combination of (i) genotype or validated compensatory mechanism, (ii) edited target, (iii) relevant cell compartment, (iv) dose or editing fraction required for benefit, (v) delivery mode, (vi) functional evidence, (vii) human evidence and (viii) safety. This is a more demanding claim than a sequence-compatible guide and may remain uncertain even in conditions with early clinical interventions.

Nuclear and mitochondrial editing must not share an unqualified recognition denominator

Mitochondrial variants illustrate a categorical rather than merely quantitative model mismatch. Human mtDNA is housed in a separate organellar environment and lacks the same straightforward guide-RNA import path available to nuclear Cas9 editing. DdCBE and engineered TALED approaches provide RNA-independent methods for selected mitochondrial substitutions, but require their own sequence-context models, delivery systems and off-target assessments (Cho et al. 2022; Cho et al. 2024; Mok et al. 2020). The 19 mitochondrial GenCC triplets labelled chemically compatible in our initial SNV screen do not inherit the NGG-window results of chromosome 21.

Moreover, mitochondrial genetic disorders are shaped by heteroplasmy, maternal transmission, tissue-specific thresholds and mtDNA copy-number dynamics. A mutation-specific editor might reduce or shift the mutant genome fraction without homogenizing all mitochondria, and a proportion of corrected molecules need not map

linearly onto physiological recovery. Future therapeutic mapping must use mitochondrial-specific baseline and edited heteroplasmy, organellar distribution, target-cell biology and developmental timing, rather than applying a nuclear SNV correction score to mitochondrial disease.

Strengths, limitations and a defined path to clinical-grade mapping

This analysis has several strengths. It employs a prespecified high-confidence disease denominator; preserves source checksums and edge-level associations; separates variation, gene–disease–MOI and intervention–indication objects; uses identical strict evidence predicates for the two molecular classes; reproduces an independent molecular-route union by inclusion–exclusion; quantifies linkage attrition and heavy-tailed ascertainment; and tests the implications of explicit genome sequence/PAM constraints in a reference-validated subset. It also demonstrates the value of conservative causal-gene versus modifier-gene attribution using primary clinical evidence.

The limitations are material and govern all conclusions. ClinVar is not a population-representative mutation census, and source-level P/LP classifications can aggregate multiple phenotypes. GenCC relationship counts are not unique-disease or patient denominators. Strict mapping excludes valid ambiguous relationships and does not replace RCV/SCV re-adjudication. The molecular indel criterion is a size proxy, not a designed prime-editing system. Sequence-specific editor modeling was restricted to chromosome 21 and used fixed window/PAM assumptions. Genome-wide MANE transcript structure, systematically mapped HPO phenotypes, chromatin accessibility, gene dosage and all relevant off-target consequences were not incorporated.

The 14 clinical/preclinical intervention–indication records were selected to exercise diverse mechanisms; they are not an exhaustive global clinical trial registry or a probability sample. Only 24 of a designed 240-case mechanism-review sample received single-pass, nonindependent source review; the 216 remaining sample cases were not adjudicated. These facts preclude a prevalence estimate of currently treatable disorders. We did not conduct experimental editing, measure delivery efficiency, determine tissue rescue thresholds, perform a patient-genotype prevalence analysis or estimate adverse-event incidence across programs. It would be methodologically inappropriate to substitute a confidence interval around the 39.21% catalog proportion for any of these missing quantities.

A clinical-grade extension should proceed in a defined order: exact RCV/SCV phenotype-specific evidence and normalized transcript accession; systematic variant- and MOI-specific molecular mechanism adjudication; nuclear and mitochondrial sequence-specific editor/pegRNA design with bystander and off-target prediction; experimentally demonstrated functional rescue or matched patient-cell assays; therapeutic cell and delivery evidence; and independent clinical trial/regulatory adjudication with safety and follow-up. The 240-case review protocol should be completed by independent masked reviewers before any probability-weighted disease-level extrapolation. Disease-specific annotation, not the addition of further generic editor motifs, is the main remaining information bottleneck.

Finally, the evidence cutoff is a scientific property of the study: trial readouts, regulatory labels and editor architectures can evolve rapidly. The preserved dataset should be released with versioned source and code identifiers, enabling reproducible recalculation as technology changes rather than silently blending clinical evidence from different dates.

Conclusions

Our stringent integration of a frozen ClinVar resource with 6,968 high-confidence GenCC gene–disease–inheritance relationships identifies a substantial, but carefully bounded, molecular genome-editing opportunity landscape. Canonical ABE/CBE reaction chemistry was associated with 2,396 relationship records, a prespecified short-indel class with 1,763, and the nonredundant combined molecular evidence set with 2,732 (39.21%). The molecular fractions were highly dependent on variant ascertainment, disease linkage and source review confidence; genome sequence modeling further restricted and reclassified candidate designs within the chromosome 21 pilot.

These observations support systematic prioritization of variants for laboratory evaluation, but they do not demonstrate that 39.21% of genetic diseases are therapeutically editable. Biological mechanism, causal-versus-edited target, required target-cell fraction, deliverability, experimentally documented rescue, and clinical safety remain independent conditions. The clinical programs examined show that effective editing may modify a compensatory gene instead of a disease allele, and that human editing activity does not automatically establish acceptable risk.

A useful therapeutic genome-editing landscape is therefore a layered, evidence-versioned atlas of condition–genotype–intervention relationships. The present work provides a reproducible molecular and disease-validity foundation for that atlas while explicitly defining the additional evidence required to infer clinical feasibility.

Data availability

The frozen primary variant and gene–disease submissions originated from NCBI ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/docs/downloads/) and GenCC (https://thegencc.org/download). The exact uploaded ClinVar snapshot (SHA-256: 7c39f2d875f3cd1ccd2d568662f06a7df647dd91fc0f51a072a758b08004c72e), filtered variation inventory, 6,968-relationship matrix, 269,669 association edges, editor-screening outputs, analysis scripts, and quality-control results are represented in the accompanying review-stage data/code package. A permanent public third-party repository deposit for the study-specific outputs and code has not yet been completed; the accession or DOI must be supplied before publication. The source file timestamp could not be independently verified; 6 November 2026 is the analysis and evidence cutoff, not a claimed upstream ClinVar release date.

Acknowledgments

The author acknowledges contributors to ClinVar and the Gene Curation Coalition resources. OpenAI ChatGPT assisted with parts of code drafting, exploratory computational analyses, literature identification, figure generation, and manuscript language editing.

Funding

This research received no external funding.

Conflict of interest

The author declares no competing interests.

Author contributions

Achab-Awelana Ernest: Conceptualization, methodology, data curation, formal analysis, investigation, visualization, writing – original draft, and writing – review and editing.

Ethics and consent

This work reanalyzed published, non-identifiable clinical-variant annotations and public gene–disease curation records. No new human participants or animals were enrolled, and no identifiable individual-level data were collected. Consent for publication is not applicable.

References

Anzalone AV, Randolph PB, Davis JR, et al. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature. 576:149–157. doi:10.1038/s41586-019-1711-4.

Chen PJ, Hussmann JA, Yan J, et al. 2021. Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. Cell. 184:5635–5652.e29. doi:10.1016/j.cell.2021.09.018.

Cho SI, Lee S, Mok YG, et al. 2022. Targeted A-to-G base editing in human mitochondrial DNA with programmable deaminases. Cell. 185:1764–1776.e12. doi:10.1016/j.cell.2022.03.039.

Cho SI, Lim K, Hong S, et al. 2024. Engineering TALE-linked deaminases to facilitate precision adenine base editing in mitochondrial DNA. Cell. 187:95–109.e26. doi:10.1016/j.cell.2023.11.035.

Clinical Genome Resource (ClinGen). n.d. Gene–disease validity and dosage-sensitivity curation knowledge base. https://search.clinicalgenome.org/kb/ [accessed 2026 Nov 6].

Cohn DM, Gurugama P, Longhurst HJ, et al. 2026. Lonvoguran ziclumeran—In vivo CRISPR gene editing in hereditary angioedema. N Engl J Med. 395:1471–1482. doi:10.1056/NEJMoa2600931.

Dadush A, Merdler-Rabinowicz R, Gorelik D, et al. 2024. DNA and RNA base editors can correct the majority of pathogenic single nucleotide variants. npj Genom Med. 9:16. doi:10.1038/s41525-024-00397-w.

De Caneva A, Porro F, Bortolussi G, et al. 2019. Coupling AAV-mediated promoterless gene targeting to SaCas9 nuclease to efficiently correct liver metabolic diseases. JCI Insight. 4:e128863. doi:10.1172/jci.insight.128863.

Gaudelli NM, Komor AC, Rees HA, et al. 2017. Programmable base editing of A·T to G·C in genomic DNA without DNA cleavage. Nature. 551:464–471. doi:10.1038/nature24644.

Gene Curation Coalition. 2022. The Gene Curation Coalition: a global effort to harmonize gene–disease evidence resources. Genet Med. 24(8):1732–1742. doi:10.1016/j.gim.2022.04.017.

Gene Curation Coalition. n.d. GenCC data downloads and evidence-classification documentation. https://thegencc.org/download [accessed 2026 Nov 6].

GeneReviews. n.d. Hypohidrotic ectodermal dysplasia. NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK1112/ [accessed 2026 Nov 6].

Gillmore JD, Gane E, Taubel J, et al. 2021. CRISPR-Cas9 in vivo gene editing for transthyretin amyloidosis. N Engl J Med. 385:493–502. doi:10.1056/NEJMoa2107454.

Human Phenotype Ontology Consortium. 2026. Human Phenotype Ontology: versioned release v2026-09-01. https://github.com/obophenotype/human-phenotype-ontology/releases/tag/v2026-09-01

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. doi:10.1038/nature17946.

Landrum MJ, Chitipiralla S, Kaur K, et al. 2025. ClinVar: updates to support classifications of both germline and somatic variants. Nucleic Acids Res. 53(D1):D1313–D1321. doi:10.1093/nar/gkae1090.

Li G, Han F, Wang C, et al. 2026. In vivo genome-editing therapy for Duchenne muscular dystrophy: preclinical studies and first-in-human trial. Cell. Published online 2026 Oct 26. doi:10.1016/j.cell.2026.09.059.

Mok BY, de Moraes MH, Zeng J, et al. 2020. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. Nature. 583:631–637. doi:10.1038/s41586-020-2477-4.

Musunuru K, Grandinette SA, Wang X, et al. 2025. Patient-specific in vivo gene editing to treat a rare genetic disease. N Engl J Med. 392:2235–2243. doi:10.1056/NEJMoa2504747.

National Center for Biotechnology Information. n.d.a. ClinVar documentation: review status and clinical significance. https://www.ncbi.nlm.nih.gov/clinvar/docs/ [accessed 2026 Nov 6].

National Center for Biotechnology Information. n.d.b. Matched Annotation from NCBI and EMBL-EBI (MANE), GRCh38 release 1.5. https://www.ncbi.nlm.nih.gov/refseq/MANE/ [accessed 2026 Nov 6].

Ngo W, Wu JLY, Wasko KM, Doudna JA. 2026. Targeted delivery of genome editors in vivo. Nat Biotechnol. 44:49–59. doi:10.1038/s41587-025-02945-w.

Pierce EA, Aleman TS, Jayasundera KT, et al. 2024. Gene editing for CEP290-associated retinal degeneration. N Engl J Med. 390:1972–1984. doi:10.1056/NEJMoa2309915.

Sousa AA, Hemez C, Lei L, et al. 2025. Systematic optimization of prime editing for the efficient functional correction of CFTR F508del in human airway epithelial cells. Nat Biomed Eng. 9:7–21. doi:10.1038/s41551-024-01233-3.

US Food and Drug Administration. 2026. FDA approves first gene therapy for young children with sickle cell disease; supplemental CASGEVY approval. https://www.fda.gov/news-events/press-announcements/fda-approves-first-gene-therapy-young-children-sickle-cell-disease.

Vafai SB, Täubel J, Ashdown T, et al. 2026. In vivo base editing of PCSK9 with VERVE-102 for hypercholesterolemia. N Engl J Med. 395:648–659. doi:10.1056/NEJMoa2601283.

Walton RT, Christie KA, Whittaker MN, Kleinstiver BP. 2020. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science. 368:290–296. doi:10.1126/science.aba8853.

Supplementary-data inventory and reproducibility

Supplementary Data S1: Frozen 6,968-row HGNC–MONDO–MOI relationship matrix with direct-variant counts, chemistry and short-indel evidence flags.

Supplementary Data S2: The 269,669 variant–relationship association edges, including exact identifiers, sequence class, review tier, match source and unique-relationship assignment.

Supplementary Data S3: Variant-level ClinVar inclusion and chemistry tables, review-tier sensitivity and source provenance checksums.

Supplementary Data S4: Chromosome 21 GRCh38 reference-checked sequence/PAM/window and nominal bystander data.

Supplementary Data S5: Disease-mechanism and clinical/preclinical intervention–indication source ledger; the 14-case catalog is purposive and not a trial census.

Supplementary Data S6: The complete 336 stringent short-indel-only gene–disease relationship list and 1,493 relationships without direct variant links.

Supplementary Data S7 (revision): Recomputed sequential crosswalk attrition and inheritance-stratified strict-route results, accompanied by an independent quantitative QA record.

Supplementary Methods S1: Frozen scripts, selection predicates, class definitions, checksum manifests and validation test suites.

For journal submission, deposit the permissible versioned outputs and original analysis code in an accessible repository, archive the exact version used in this manuscript and append stable DOI or accession links. Large third-party source files may require reference to official downloads instead of redistribution.

Supplementary Files

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

- SupplementaryFigure1.png
- SupplementaryFigure2.png