Folklore

READING THE GENOME

The genome is an inherited story.
Folklore reads it.

Folklore is knowledge passed from generation to generation. Never formally written, yet carrying the identity of a whole nation.

The genome is exactly that kind of inherited memory, a story recopied at every generation, with an occasional new addition. Every variant is a line added along the line of descent, a heritage carried through time.

The folklorist does not invent the tale. Only reads it.

Folklorists read, interpret and systematise them, gathering scattered oral knowledge and drawing out meaning, structure and significance.

That is exactly what Folklore does: reading the genome and interpreting the results. The platform takes called variant data, runs it through a deterministic interpreter backed by dozens of curated evidence sources, and returns every variant sorted and classified: by pathogenicity across the five ACMG classes, and by closeness to the patient's phenotype.

We do not invent the story. We read it.

One story.
Eight ways to read it.

Each module is one reading of the genome, and together they turn a raw VCF into a prioritised, review-ready result.

- **Variant classification** — *Millions of variants → five ACMG classes, ranked by priority.* — V3.39.1
- **Mitochondrial DNA** — *Rule-based mtDNA under the MMDWG 2020 framework.* — CLINGEN MT-VCEP
- **Structural variants · SV / CNV** — *Deletions and duplications, Riggs 2020, five tiers.* — RIGGS 2020
- **Phenotype matching** — *HPO terms → clinical priority, five tiers.* — HPO · 0–100
- **Clinical screening** — *Review order, not pathogenicity, Tier 1–4.* — 7 COMP · 9 BOOSTS
- **Cohort analysis** — *Groups of samples → burden, enrichment, candidates.* — FISHER · SKAT-O
- **Family & trio analysis** — *Child–mother–father → the origin of a variant.* — TRIO · DE NOVO
- **Literature evidence** — *Variant → relevant, anchored publications.* — PUBMED-ANCHORED

The modules sit within a navigable genome browser, an on-premise AI assistant, a review board, and a tiered clinical report. Pathogenicity is decided by the deterministic classifier, not by AI. Folklore is a decision-support tool for the qualified geneticist. (Research Use Only.)

Nuclear variants, under ACMG/AMP.

Every variant is annotated through Ensembl VEP, then classified against the 2015 ACMG/AMP guidelines as deterministic, rule-based logic.

Folklore implements the ACMG/AMP 2015 framework (Richards et al.) through the Tavtigian Bayesian point system, calibrated to the ClinGen Sequence Variant Interpretation recommendations.

Of the 28 evidence criteria, 19 are computed automatically against a curated reference layer of 45 databases: gnomAD, ClinVar, dbNSFP, SpliceAI, AlphaMissense, HPO, UniProt, and more, billions of records held locally on EU infrastructure. The remaining 9, which need segregation, functional, or de novo evidence, stay with the reviewing geneticist. Each applicable criterion carries a weight and the combined score maps onto a single verdict.

Every classification is recorded with its evidence and version, so the result is reproducible, auditable and resolves to one of five classes:

PATHOGENIC | LIKELY PATHOGENIC | VUS | LIKELY BENIGN | BENIGN

Every classification traces back to the rule and the source behind it.

Every result carries its evidence.

The variants arrange by gene and by ACMG priority, Pathogenic through Benign. Along the top, five class cards hold the live totals for each class and filter the list beneath them. A gene search narrows it to one.

3	7	214	1.1k	18k
PATHOGENIC	LIKELY PATH.	VUS	LIKELY BENIGN	BENIGN

#1	BRCA1	c.5266dupC · p.Gln1756fs	DE NOVO?	PATH
#2	MYH7	c.1988G>A · p.Arg663His	CURATED	LIK. PATH
#3	TTN	c.21A>G · p.Ser7=		VUS

Each variant carries its whole case. The ACMG criteria that fired, shown as colour-coded evidence. ClinVar significance and review stars, with any ClinGen expert-panel assertion. In-silico predictions, gnomAD frequency, conservation and constraint. The sequencing quality behind the call. Every value traces to the record that produced it.

The record is always one step beneath the result.

Mitochondrial variants, under MMDWG.

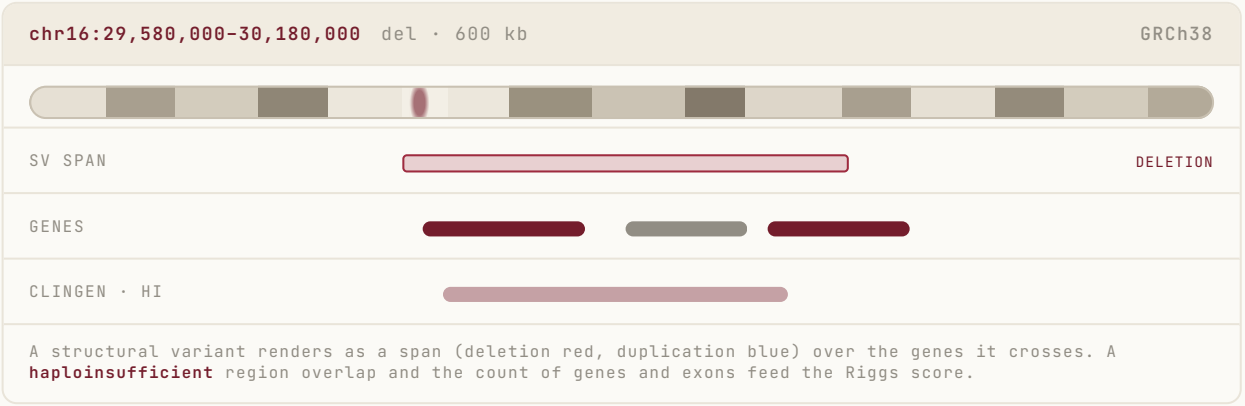
Mitochondrial variants are classified under MMDWG 2020 (McCormick), an independent module running alongside nuclear ACMG/AMP. Every variant in the output carries an explicit framework-provenance label, so the geneticist always knows which framework produced the class. Of the 28 nuclear criteria, seven are excluded for the mitochondrial genome with verbatim rationale. The rest are respecified with mtDNA-specific thresholds: 11 automated, 10 curated with strength tiers. Combining rules from Richards 2015 are preserved unchanged.

A curation panel exposes the ClinGen mtDNA VCEP criteria for the reviewer, alongside heteroplasmy assessment and haplogroup-aware interpretation (APOGEE2, MitoTIP, MITOMAP). The classification follows the genome's biology: PP1 cannot contribute when a variant is homoplasmic across all maternal members (VCEP §5.2.7). The panel then shows the pipeline class and the curated class side by side:

Two frameworks, each named on every variant.

Structural variants, under ClinGen/ACMG.

Copy-number variants are classified under the ClinGen/ACMG 2020 semiquantitative point system (Riggs 2020): losses on Table 1, gains on Table 2, as two distinct point systems. Dosage sensitivity comes from pHaplo / pTriplo (Collins 2022) and overlap with established haploinsufficient (HI) and triplosensitive (TS) regions from the ClinGen dosage map. Population frequency comes from gnomAD-SV. Points sum through a single five-tier ladder into the same P / LP / VUS / LB / B labels.



A structural variant renders as a span (deletion red, duplication blue) over the genes it crosses. A **haploinsufficient** region overlap and the count of genes and exons feed the Riggs score.

16p13.11 deletion · 600 kb **Pathogenic**

The **CNV classification (Riggs)** panel shows the total score and tier with each criterion met or not met: established-HI overlap, gene count, dosage, population frequency, alongside copy state, breakpoint precision and paired/split-read support.

Its own tab, its own metric, landing on the same five-class scale.

Variants, ranked by phenotype.

ACMG classification tells you how pathogenic a variant is. It does not tell you which variant explains your patient. Phenotype matching answers that second question.

The patient's HPO terms are correlated with gene–disease associations by semantic similarity, scored 0–100 for each term, and folded into a clinical priority score. Genes are ranked and placed into five tiers (Tier 1, Tier 2, Incidental Findings, Tier 3, Tier 4). Each variant shows its strongest HPO-term matches.

> THE CORRECT BEHAVIOUR
>
> *A VUS with strong phenotypic relevance ranks above a Pathogenic variant for an unrelated condition. Relevance to this patient, not pathogenicity in the abstract.*

The HPO terms entered at intake become filters over the gene list. A selected phenotype collapses the list to the genes that match it.

#1	SCN1A	HPO match 0.92 · seizures, DEE	TIER 1
#2	KCNQ2	HPO match 0.71	TIER 2
#3	TTN	HPO match 0.08 · unrelated	TIER 4

The story is read for the patient in front of you, not for a textbook.

Clinical screening, tier by tier.

Screening sets the review order from the patient's full context: age, sex, ethnicity, family history, sample structure, and the chosen panels.

A seven-component score (constraint, deleteriousness, phenotype, dosage, consequence, compound-het, age relevance) is adjusted by up to nine clinical boosts (ACMG strength, ethnicity, family history, de novo, and more), then placed into four tiers. Each gene carries a clinical-actionability signal: immediate, monitoring, or future. Screening modes tune the run for neonatal, paediatric, adult diagnostic, or carrier testing.

T1	RYR1	score 0.86 · malignant hyperthermia	IMMEDIATE
T1	BRCA2	score 0.79	MONITORING
T2	MYBPC3	score 0.55	FUTURE

Screening sets the order of reading. Pathogenicity stays with ACMG.

Cohort analysis, across samples.

A study gathers classified samples into one matrix and runs population-level statistics, from two samples upward.

Gene-burden testing reports Fisher exact p, FDR q and odds ratios (with CMC and SKAT-O), read off a volcano plot. A loss-of-function pass adds pLI and LOEUF. Pathway enrichment and cross-sample compound-heterozygote detection round out the signal. A candidate-gene nomination fuses seven independent lines of evidence (burden, pLoF, disease, constraint, pathway, GWAS, compound-het) into a ranked list and an evidence-matrix heatmap. Where samples disagree on a class, the discordance is surfaced, not hidden.

#1	LDLR	Fisher p 3.1e-5 · OR 6.4 · 5/7 evidence axes	SIGNIFICANT
#2	APOB	Fisher p 8.0e-4 · OR 3.2	FDR Q<0.05
#3	PCSK9	pathway + GWAS + constraint	CANDIDATE

The same reading, now across a population, with every disagreement between samples shown.

Inheritance, read from the trio.

A trio (proband, mother, father) adds what a single sample can never carry: the parental genotypes. Three inheritance workflows run on them: **de novo**, **compound heterozygous**, and **segregation**.

De novo. The proband carries the alternate allele; both parents are reference. A per-member table shows the genotype, depth, quality and class behind each call. A clinical-grade filter scopes the drill.

MEMBER	GENOTYPE	DP · QUAL
Proband	het (0/1)	42x · 99
Father	hom_ref (0/0)	38x · 99
Mother	hom_ref (0/0)	45x · 99

Compound het. Two variants in one gene, one inherited from each parent, are phased from the trio genotypes: paternal when father is het and mother reference, maternal for the reverse. Only a trio-phased pair is called *in trans*.

PATERNAL
GENE · variant A
father het · mother 0/0

IN TRANS ↔

MATERNAL
GENE · variant B
father 0/0 · mother het

Cis-ambiguous pairs are shown as such. The trio phases only what the genotypes support.

Segregation, and its ceiling.

Segregation testing computes a LOD score for each variant. But a trio (one proband and two parents) has too few informative meioses. The maximum LOD attainable is about **0.30**, which sits exactly at the ClinGen SVI 2016 *supporting* evidence threshold.

> THE CEILING
>
> *A trio reaches PP1_Supporting at most, never PP1 Moderate or Strong. The platform states the ceiling and recommends extending the pedigree for stronger evidence, rather than borrowing a strength the data cannot support.*

Before any of this, cross-sample QC checks that the family is the family: PLINK identity-by-descent relatedness catches sample swaps, unexpected consanguinity, and duplicates. A critical alert holds the line: inheritance findings should not drive clinical interpretation until the pedigree checks out.

MAX LOD 0.30	Trio ceiling = ClinGen SVI 2016 supporting threshold. Extend the pedigree for more.
CROSS-SAMPLE QC	PLINK IBD (PI_HAT): sample-swap, consanguinity and duplicate detection.

The limit is named, not worked around. A stated ceiling holds where a borrowed strength would not.

Literature, anchored to the variant.

A local, genetics-filtered PubMed mirror with pre-extracted gene, variant and phenotype mentions turns literature search into a sub-second, variant-anchored query.

Publications rank by a six-component model aligned to ACMG evidence categories, each badged by strength: strong, moderate, supporting, weak. An exact-match badge flags a paper that names *this* variant. A functional badge flags experimental data. The gene ranking fuses clinical priority and literature relevance (60/40), and every hit keeps full PMID / PMC / DOI traceability.

EXACT MATCH	The paper reports this exact variant, not just the gene.
SIX-COMPONENT SCORE	Phenotype, publication type, gene focus, functional data, variant, recency.
TRACEABLE	PMID / PMC / DOI on every citation, evidence you can open.

Every citation opens to its source. The folklorist names where the claim came from.

AI assistant, under the deterministic gate.

An on-premise AI clinical assistant, a self-hosted open-weight model on EU infrastructure where no external AI API ever sees patient data, integrates classification, phenotype and literature into a structured clinical narrative. You can ask it questions in plain language: it runs queries against the session database and charts them, searches the literature, and explains which ACMG criteria support a class.

The division of labour is the whole point: the AI drafts interpretation text, the deterministic classifier decides pathogenicity. The assistant carries its own caution. The finished output is a single tiered PDF: Tier 1 (actionable) and Tier 2 (potentially actionable) variants, each with a complete evidence chain, ready for the geneticist's review and sign-off.

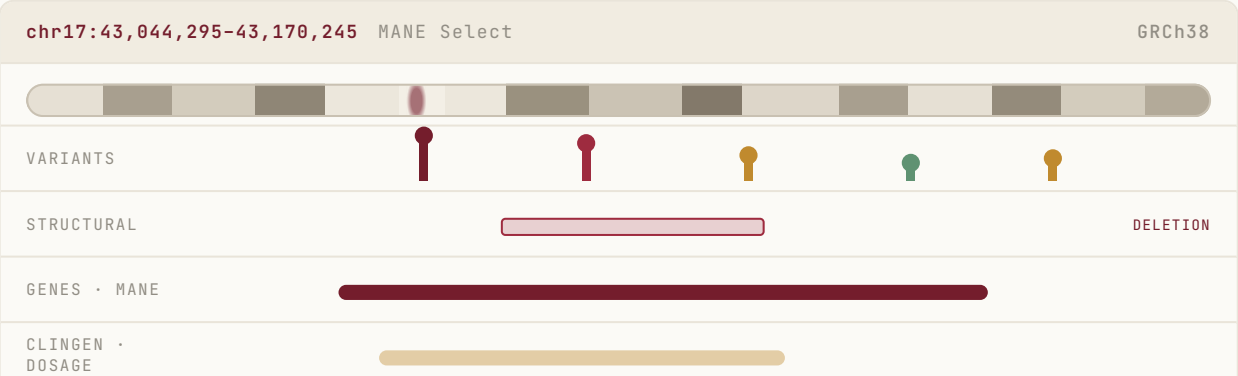
> THE DIVISION OF LABOUR
>
> *AI writes the sentence; rules assign the class. The geneticist makes the call. Research Use Only. Not for diagnostic use without professional review.*

CLINICAL REPORT · TIERED · EVIDENCE CHAIN			RUO
TIER 1	BRCA1	c.5266dupC · Pathogenic · PVS1 PS4 PM2 · ClinVar ★★★	
TIER 2	MYH7	c.1988G>A · Likely path. · PM1 PM2 PP3	
Reviewed & signed: _____ · qualified clinical geneticist			

A draft is a draft until a human signs it. Folklore never forgets which is which.

The genome browser.

A navigable per-session window over the patient's GRCh38: variants as ACMG-coloured lollipops sized by clinical salience, structural variants as spans, genes as MANE models with strand direction, ClinGen dosage regions, alongside evidence lanes for BayesDel, SpliceAI, pext, conservation, ClinVar and constraint. Zoom, pan, jump to a locus, an HGVS change, or a gene. The browser adds no reading of its own. It shows what was already computed.



Variants as **lollipops** by ACMG class, height tracks salience. Structural variants as spans, genes with direction along MANE. A graph mode reads insertions and deletions as divergence. Where phase is not measured, it says **"inferred"**.

BRCA1 · c.5266dupC (p.Gln1756fs) **Pathogenic**

PVS1 PS4 PM2

Click a feature for the full readout: fired ACMG criteria as badges, **BayesDel** and **SpliceAI**, the **pext** level, **MANE**, and the **ClinVar** stars. Only what has a value is shown.

One coordinate frame holds every finding, each shown where the analysis placed it.

The review board.

A case carries the exact classifier version it was run with, so every result stays tied to the rules that produced it. The review board is where a qualified geneticist takes over from the pipeline.

A reviewer can reclassify any variant, with a written justification that is attributed and revertable, or leave threaded notes for a colleague. The reviewer’s call always overrides the pipeline’s. A case-level outcome (solved, likely solved, unsolved) is tracked with its full change history.

Every reclassification is signed, sourced, and reversible. The audit trail is the product.

From VCF to clinical report.

1	**Ingest & build detection** VCF in; automatic GRCh37/38 detection and liftover to GRCh38.	GRCH37 / 38
2	**Quality control** Quality filters; ClinVar-pathogenic variants protected from being dropped.	CLINVAR-SAFE
3	**Annotation** Against the reference layer: 45 databases, billions of rows.	45 DATABASES
4	**Classification** ACMG/AMP for nuclear variants; MMDWG for mtDNA; Riggs for CNVs.	ACMG · MMDWG · RIGGS
5	**Prioritisation** Phenotype matching and screening; down to 3–20 candidates.	3–20
6	**Report** A clinical-interpretation draft for review by a geneticist.	RUO

Every stage traceable, every stage reproducible.

The specification.

Genome builds	**GRCh38** native; **GRCh37** auto-lifted
Variant classes	SNV, indel, mtDNA, SV / CNV (Riggs 2020), all active
Reference layer	**45** databases; SpliceAI ~3.4B rows, gnomAD ~909M, dbNSFP ~80.7M, ClinVar ~4.1M
Classification	ACMG/AMP 2015, deterministic; MMDWG (ClinGen mt-VCEP); Riggs 2020; 5 classes
Genome browser	Navigable GRCh38 view; variants, SVs, genes, dosage; evidence drawer; graph mode
Speed	Analysis pipeline ~**7–12** min (WGS)
EU data residency	Helsinki, Finland; data never leaves the EU; GDPR Art. 9
Audit & security	SHA-256 append-only chain; encryption in transit and at rest; RBAC
Artificial intelligence	Self-hosted open-weight model (Qwen) on EU infra; no external AI API
Status	**Research Use Only**. A decision-support tool, not a diagnostic device

EU DATA RESIDENCY GDPR ART. 9 SHA-256 AUDIT RBAC RUO

One reading, traceable down to the last rule.

The tale.

Every capability carries both a clinical proof and a signal of scale.

THE READING

Clinical-grade, traceable, reproducible. Five classes and an audit trail on every action. A decision-support tool that leaves the last word to the qualified geneticist.

Rare disease Newborn screening

Carrier screening

THE MOAT

Category. A European clinical-genomics software pure-player, no wet lab, no capex. **Moat.** Curation, not code: a reference layer (45 databases, billions of rows) and a discordance-gated classifier, built over years. **Defensibility.** Live rule-based mtDNA and live Riggs CNV, capabilities others treat as afterthoughts. **Sovereignty.** Self-hosted EU AI where no external API sees patient data; GDPR by design.

European software Self-hosted AI

Reference layer EU data

The tool reads the story. The last word is the human's.

Validation, against an external lab.

0

BENIGN ⇔ PATHOGENIC INVERSIONS (OPPOSITE POLES), ACROSS ALL THREE COHORTS

68.4%

EXACT FIVE-CLASS AGREEMENT · COHORT 2, CLASSIFIER V3.28.0, 19 EVALUABLE

3

REAL-WORLD COHORTS · 20 / 20 / 57 CASES, ACROSS CLASSIFIER GENERATIONS

Five-class agreement sits in the **60–75%** range that the literature reports *between* laboratories (Amendola 2016 ~66%, Harrison 2017 72–76%). Where Folklore differs, it differs by one step and in one direction: it returns VUS rather than asserting a pathogenicity it cannot fully evidence. No benign becomes pathogenic, nor the reverse.

WHAT WE HAVE VALIDATED

Concordance against an external clinical laboratory across three real-world cohorts. Zero opposite-pole inversions. A conservative skew toward VUS.

WHAT WE HAVE NOT

Large-scale prospective outcomes, multi-laboratory benchmarking, SV/CNV phenotype correlation. The claim does not exceed the evidence.

External clinical laboratory as a peer comparator, not a reference standard. An ongoing, iterative real-world audit.

Regulatory status.

Folklore is a Research Use Only tool. It is a clinical decision-support system, not a diagnostic device, and it is not a CE-marked IVD.

Every classification Folklore produces requires review and confirmation by a qualified clinical geneticist. The platform drafts and orders evidence. The clinical decision belongs to the reviewing professional.

Folklore does not replace clinical judgement, established laboratory procedure, or the reporting standards of the laboratory that uses it. Results are intended to support qualified interpretation, not to stand alone.

Research Use Only Not a diagnostic device Not a CE-marked IVD

Geneticist sign-off

The tool reads the evidence. The clinical decision is the geneticist's.

Folklore

The genome is an inherited story. Folklore reads it.

A HELENA BIOINFORMATICS PLATFORM · SOFIA, BULGARIA

CONTACT@HELENA.BIO

HELENA.BIO