Introduction

ClinMAVE is a curated and clinically-oriented database that transforms MAVE studies' results into structured, accessible, and evidence-grade annotations suitable for genetic testing applications. ClinMAVE bridges the gap between functional effect of variants and their clinical impact by systematically curating MAVE datasets and rigorously transforming the experimental assessment into strength of evidence supporting the pathogenicity classification. ClinMAVE also integrates available variant annotations from public repositories (e.g., ClinVar, gnomAD, TCGA) and provides a user-friendly interface to support visualization and analysis modules for further exploration.

Data Collection and Processing

1. Data collection

ClinMAVE compiles a comprehensive collection of Multiplexed Assays of Variant Effect (MAVE) studies, encompassing both deep mutational scanning and CRISPR-based genome editing techniques.

1.1 Deep mutation scanning MAVE

We collected deep mutational scanning data primarily from MAVEdb, a centralized repository of multiplexed functional assay results. Notably, raw protein-level variants are uniformly mapped to genomic DNA coordinates by Ensembl, enabling standardized genomic annotation for downstream analyses. To ensure data quality and relevance, only datasets with more than 100 variants and an associated peer-reviewed publication or publicly available preprint are retained. This filtering guarantees that included studies meet a minimum threshold of experimental scale, scientific rigor, and transparency.

1.2 CRISPR-based genome editing MAVE

For CRISPR-based genome editing datasets, we collect studies primarily through searches in PubMed. Only datasets that report more than 100 variants are retained to ensure adequate scale. For base-editing studies, the reported protein- or transcript-level changes are mapped to genomic DNA coordinates, rather than relying on guide RNA target sites. For Prime editing and Saturated Genome Editing (SGE) studies, the DNA-level coordinates are directly extracted from the supplementary materials provided with the publications.

2. Processing

ClinMAVE employs a unified variant processing suite designed to standardize and enrich variant-level data across all studies. This suite integrates two key components:

2.1 Genomic annotation:

Variants are mapped to genomic DNA coordinates and annotated using the MANE (Matched Annotation from NCBI and EMBL-EBI) transcript to ensure clinical consistency. Each variant is then formatted according to HGVS (Human Genome Variation Society)

nomenclature for structured and interoperable representation.

2.2 External Annotation Integration

To provide biological and clinical context, the pipeline incorporates external annotations from clinical relevant resources, including:

(1) ClinVar: clinical significance interpretations;
(2) gnomAD: population allele frequencies;
(3) TCGA: Somatic variant frequencies from cancer cohorts;
(4) In silico predictors: Scores from variant effect prediction tools such as CADD, EVE, and REVEL;
(5) dbSNP: standardized reference IDs that facilitate cross-referencing with external genomic databases.

In addition, in silico predictor scores were translated into ACMG-style pathogenicity classifications based on the recommended thresholds provided by VarSome.

Software	Benign Strong	Benign Moderate	Benign Supporting	VUS	Pathogenic Supporting	Pathogenic Moderate	Pathogenic Strong
AlphaMissense	<= 0.0853	<= 0.166	<= 0.316		>= 0.787	>= 0.956	>= 0.994
CADD	<= 16.1	<= 22	<= 23.2		>= 25.6	>= 28.8	>= 33
EVE		<= 0.162	<= 0.255		>= 0.603	>= 0.723	>= 0.905
MetaSVM		<= -0.677	<= -0.286		>= 0.794	>= 0.901	
REVEL	<= 0.133	<= 0.351	<= 0.471		>= 0.685	>= 0.798	>= 0.946
Polyphen2	<= -1.04	<= 1.08	<= 3.58		>= 7.52	>= 9.88	

3. Software and external resources:

Software/Database	Version
ANNOVAR	V2.3.0
MANE	GRCh38 v1.4
dbSNP	avSNP150
gnomAD	V2.1.1
AlphaMissense	dbNSFP v4.7a
CADD	dbNSFP v4.7a
EVE	dbNSFP v4.7a
MetaSVM	dbNSFP v4.7a
REVEL	dbNSFP v4.7a
Polyphen2	dbNSFP v4.7a
Cbioportal	Accessed on 2025/05/22
Varsome	Accessed on 2025/05/12

Data Curation

1. Curation model

We curated both quantitative information (e.g., score thresholds used to define functional classification) and qualitative information (e.g., experimental model, phenotype) from published MAVE studies. To standardize the interpretation of functional effects across diverse assays and reporting formats, we applied a structured curation model, as detailed below:

Data type	Description	Value
Gene	Gene identifier	*e.g. BRCA1*
Functional score	Quantification of the functional impact of a specific genetic variant	*e.g.* 0.34
Functional classification	Controlled vocabulary	Loss-of-function, Gain-of-function, Functional neutral
MAVE technique	Controlled vocabulary	Deep mutational scanning, CRISPR-based genome editing
Mutagenesis strategy	Controlled vocabulary	Prime editing, Base editing, Saturation genome editing, Oligonucleotide synthesis, Mutagenic PCR
Functional assay	Controlled vocabulary	Protein abundance and stability, Protein binding, Specialized molecular function, Cellular fitness
Experiment model	The cell types used to perform MAVE	e.g. HAP1
Phenotype	The specific biological outcome measured by a MAVE experiment	e.g. *MLH1* mediated mismatch repair function
Direction of Effect	how a variant alters the measured phenotype relative to the wild-type or reference allele	e.g. diminished, enhanced

2. MAVE quantitative calibration

ClinMAVE organizes functional data to support variant interpretation at both the dataset level and the individual variant level, aligning with ACMG/AMP recommendations for evaluating experimental evidence (e.g., PS3/BS3 criteria). To compute this, we define the control sets as follows:

- Clinvar-classified Pathogenic/Likely Pathogenic (P/LP) variants as true positives (TPs);
- Clinvar-classified Benign/Likely Benign (B/LB) variants as true negatives (TNs);

To ensure statistical reliability, we only conduct evidence strength evaluations for datasets that meet the following minimum requirements:

- A total of at least 11 benchmark variants (TP + TN ≥ 11);
- At least 4 TPs and 4 TNs, respectively;

2.1 Dataset-level strength of evidence:

The ClinGen Sequence Variant Interpretation (SVI) Working Group (2022) introduced the use of the Odds of Pathogenicity (OddsPath) to quantify the strength of functional evidence at the dataset level. This approach enables a standardized, quantitative mapping of functional assay results to ACMG/AMP evidence strength levels (e.g., PS3/BS3).

OddsPath is calculated using two conditional probabilities:

- P1: The probability that a variant is pathogenic *given* an abnormal (damaging) functional result
- P2: The probability that a variant is pathogenic *given* a normal (benign) functional result

The OddsPath formula is defined as:

$$OddsPath = \frac{P1}{P2} \times \frac{1 - \mathrm{P2}}{1 - \mathrm{P1}}$$

2.2 Per-variant level strength of evidence:

ClinMAVE assigns functional evidence strength at the per-variant level by quantifying how similar a variant's functional score is to known pathogenic (TP) and benign (TN) variants within the same dataset. This is achieved using a distance–based scoring method, which accounts for score distribution and variance. Each variant's score is transformed into a normalized "distance_score" on a [–1, 1] scale, where:

- +1 indicates high similarity to pathogenic variants (strong functional abnormality)
- –1 indicates high similarity to benign variants (normal function)

For datasets eligible for ACMG-style strength scaling, variants classified as functionally abnormal are assigned one of the following evidence levels:

- Strong: distance_score ≥ 0.9
- Moderate: distance_score ≥ 0.75
- Weak: distance_score < 0.75

Database Usage

The ClinMAVE database provides an interactive platform for browsing, searching, visualizing, and analyzing curated functional evidence from MAVE. The portal is organized into four core modules.

1. Browser module

The Browser Module in ClinMAVE provides a flexible and user-friendly interface for navigating curated MAVE datasets. Users can explore data through multiple entry points—from gene- and technique-level overviews to detailed variant-level records—using either the homepage search bar or the dedicated browser interface.

2. Search Module

ClinMAVE offers robust search capabilities to help users efficiently locate relevant functional variant data across genes, variants, and MAVE techniques. Accessible from the homepage, the quick search bar supports intuitive, keyword-based searches. Users can search using:

- Gene symbol (e.g., *BRCA1*, *TP53*)
- Gene ID (e.g., Ensembl Gene ID)
- HGVS-formatted variant notation (e.g., NM_000059.3:c.4035del)
- MAVE technique (e.g., "deep mutational scanning", "CRISPR-based genome editing")

3. Visualization module

ClinMAVE provides intuitive and informative visualizations to help users explore functional variant data in both dataset-specific and gene-centric contexts.

3.1 Visualization by Dataset:

Each MAVE dataset includes interactive plots that integrate functional scores with key biological annotations:

- MAVE score distribution plots, color-coded by: (1). molecular consequence (e.g., missense, nonsense, synonymous); (2). ClinVar classification (e.g., pathogenic, benign, VUS).
- Scatterplots or violin plots illustrate how variant functional scores correlate with population allele frequency from gnomAD.

3.2 Visualization by Gene:

ClinMAVE offers comprehensive gene-centric visualizations that map variant effects across the full length of a protein, with integrated domain annotations and support for cross-dataset comparison.

4. Analysis module

The Analysis Module in ClinMAVE provides tools for evaluating how well functional scores from MAVE experiments align with computational predictions and clinical classifications. This supports benchmarking of variant effect predictors and assessment of functional assay reliability.

4.1 MAVE score vs. In-silico Predictors

Users can compare MAVE-derived functional scores with widely used in silico predictors.

4.2 Clinical classification analysis

ClinMAVE enables quantitative benchmarking of functional assays by measuring their ability to distinguish clinically classified variants. Receiver Operating Characteristic (ROC) curves are generated to evaluate the ability of MAVE scores to separate (Likely-) pathogenic from (Likely-) benign variants.