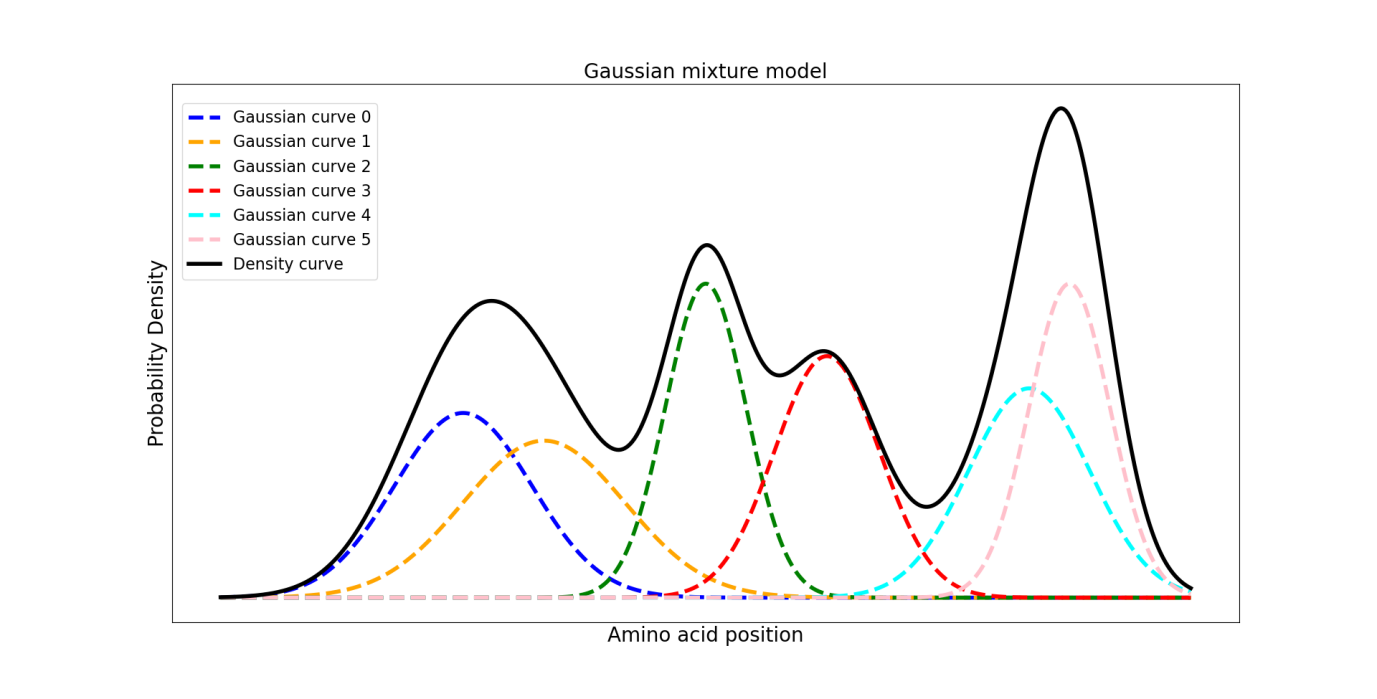
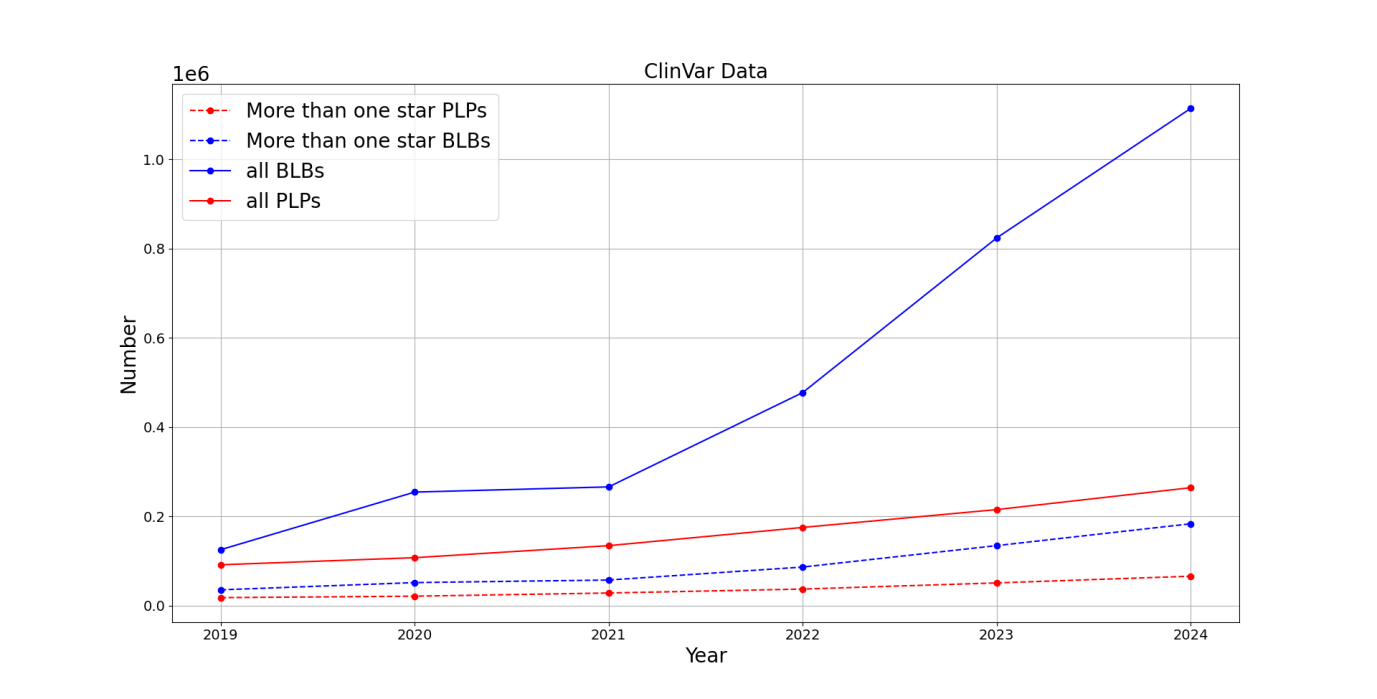


**Figure 1.**Density curves calculated with different kernel functions



**Figure 2.**Gaussian Mixture Model

**Figure 3.**Distribution of various types of mutations reported in the Clinvar database from 2019 to 2024

**Bandwidth Calculation Steps.**

Obtaining the optimal fixed bandwidth for a specific amino acid sequence by minimizing the *Mean Integrated Squared Error (MISE)*:

And *,*where *AMISE* stands for the Asymptotic Mean Integrated Squared Error. *AMISE* has:

Where:

Minimizing *MISE(h)* is equivalent to minimizing *AMISE(h)*. Taking the derivative and setting it to zero gives:

Yield:

Once the kernel function is determined, the values of *R, m* and in the formula for *h* can be established, and *h* will have an analytical solution. We have chosen the Gaussian kernel function, so the optimal bandwidth we select is:

Where σ is the standard deviation of the sample.

**Expectation Maximization Algorithm**

The EM algorithm consists of three steps: the E-step (calculating the expected values of *μ* and *σ*), the M-step (computing the maximum likelihood estimates of these two parameters from the E-step for the Gaussian model), and the iteration step (repeating the E-step and M-step until these parameters converge). This process determines the boundaries of mutation clusters, through which a set of mutation clusters can be identified.

Specific steps:

1. Initialize *μ* and *σ*. We set the maximum values of the kernel density curve to be *(k=1,2,...,K)*, where *K* is the number of maximum values. Each maximum value corresponds to two minimum values on either side,, and in the edge case where there are no local minimum values before or after *k*, the values *(0,f(0))* and *(L,f(L))* are used as local minimum values depending on the situation. where *L* is the length of the amino acids in the gene. Then the initial parameters and for each Gaussian distribution are given by the following formula:
2. Calculate the probabilities of all samples belonging to each category:
3. Update

4.Repeat steps 2 and 3 until the maximum number of iterations is reached or until *μ, σ,* and *ω* converge during the iterations.

After completing the above iterations, you will obtain *J (J ≤ K)* values for*μ*and*σ*, which correspond to *J* candidate hotspot regions [].

**Table 1.**Number of Genes Under Various Conditions.

|  |  |
| --- | --- |
| type | Number of gene |
| Genes containing all hotspots and coldspots | 1523 |
| Genes containing only hotspots | 138 |
| Genes containing only coldspots | 1338 |
| Genes containing both hotspots and coldspots | 47 |
| Genes with PLP counts exceeding 80% of the total number of PLPs and BLBs among genes with hotspots | 25 |
| Genes with hotspots that contain only PLPs | 0 |
| Genes with BLB counts exceeding 80% of the total number of PLPs and BLBs among genes with coldspots | 1238 |
| Genes with coldspots that contain only BLBs | 623 |
| Genes where the number of PLPs exceeds 80% of the total count of PLPs and BLBs among all genes | 98 |
| Genes where the number of BLBs exceeds 80% of the total count of PLPs and BLBs among all genes | 4318 |
| Genes that contain only PLPs | 0 |
| Genes that contain only BLBs | 3149 |

**Table 2.** Comparison of Hotspot Regions in This Study with Those in AutoPVS1, Including critical Functional Areas and disease-specific expert-curated functional areas

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Comparis on Category | describe | Number of all variants | Number of PLPs | Number of BLBs | PLP/(PLP+BLB)(%) |
| Compared with the AutoPVS1 hotspot interval | The number of all mutations in all genes within this regions (18,605 genes) | 5676 | 1299 | 237 | 84.57% |
| The number of variants on genes that have PLP and are located within this region（2476 genes） | 5643 | 1299 | 237 | 84.57% |
| The number of variants on genes that have hotspots in this study and are located within this region（186 genes） | 3907 | 979 | 139 | 87.61% |
| The number of variants located both in the hotspot areas of this study and within this region | 1888 | 640 | 41 | 93.94% |
| The number of variants in the hotspot intervals calculated in this study | 16226 | 3455 | 281 | 92.38% |
| Compared with the AutoPVS1 critical functional areas | The number of all mutations in all genes within this regions (18,605 genes) | 53772 | 3992 | 4032 | 49.62% |
| The number of variants on genes that have PLP and are located within this region（2476 genes） | 51037 | 3992 | 4032 | 50.61% |
| The number of variants on genes that have hotspots in this study and are located within this region（186 genes） | 24556 | 2823 | 1633 | 63.21% |
| The number of variants located both in the hotspot areas of this study and within this region | 3164 | 871 | 62 | 94.44% |
| The number of variants in the hotspot intervals calculated in this study | 16226 | 3455 | 281 | 92.38% |
| Compared with the AutoPVS1 functional areas curated by disease-specific experts | The number of all mutations in all genes within this regions (18,605 genes) | 2072 | 294 | 200 | 59.85% |
| The number of variants on genes that have PLP and are located within this region（2476 genes） | 2072 | 294 | 200 | 59.85% |
| The number of variants on genes that have hotspots in this study and are located within this region（186 genes） | 1689 | 230 | 157 | 59.59% |
| The number of variants located both in the hotspot areas of this study and within this region | 203 | 88 | 5 | 94.58% |
| The number of variants in the hotspot intervals calculated in this study | 16226 | 3455 | 281 | 92.38% |

**Table 3.** Summary of integrated data sources in HCSeer.

|  |  |
| --- | --- |
| Category | Data sources |
| **Part one: variation-level implication** | |
| Allele frequency | gnomAD, ExAC, 1000 Genomes, ESP, Kaviar, HRC, Mitomap |
| In silico function and pathogenicity prediction | ReVe, CADD, DANN, Eigen, Fathmm-MKL, FATHMM, FitCons, GenoCanyon, REVEL, SIFT, PolyPhen2-HDIV, PolyPhen2-HVAR, LRT, MutationTaster, MutationAssessor, PROVEAN, VEST4, MetaSVM, MetaLR, M-CAP, GERP++, phyloP100way-vertebrate, phastCons100way-vertebrate, SiPhy, Eigen-PC, Fathmm-XF, SIFT4G, LINSIGHT, MutPred2,MVP,MPC,PrimateAI,DEOGEN2,BayesDel-addAF,BayesDel-noAF,ClinPred,LIST-S2,ALoFT,bStatistic,phyloP470way-mammal, phyloP17way-primate, phastCons470way-mammal, phastCons17way-primate, gMVP, VARITY-R, VARITY-ER, VARITY-R-LOO, VARITY-ER-LOO, AlphaMissense, FitCons2, Funseq2, ReMM, CScape, Orion, FIRE, PAFA, CDTS, DVAR, ncER, regBase-REG, regBase-CAN, regBase-PAT, Divan-TSS, Divan-Region, CADD-splice, SCAP, spliceAI, dpsi-max-tissue, dpsi-zscore, dbscSNV-ADA-SCORE, dbscSNV-RF-SCORE, MaxEntScan, GeneSplicer, ESRseq, Spliceogen, Squirl, RegSNPs-intron, MMSplice, KipoiSplice, Synvep, SPiCE-MES, SPiCE-SSF, SPiCE, FatHmmW, EFIN-SP, EFIN-HD, PANTHER, PhD-SNP, SNAP, Mitoclass1, SNPDryad, Meta-SNP, CAROL, Condel, COVEC-WMV, MtoolBox, APOGEE, MitoTIP, PON-Classification |
| Disease-related | ClinVar, InterVar, ICGC, COSMIC, GWAS Catalog |
| Variant information | Gene4Denovo, SPCards |
| Regulatory information | GTEx, VARAdb, GREEN-DB, EPimap EPigenomics |
| **Part two: gene-level implication** | |
| Basic information | NCBI Gene, Entrez, OMIM, HGNC, Ensembl, GeneCards, UniProtKB |
| Genic intolerance | RVIS, LoFtool, GDI, Episcore, heptanucleotide context intolerance score, pLI score |
| Gene function | Gene Ontology, UniProtKB, InterPro, NCBI BioSystems, InBio Map™ |
| Disease-related | OMIM, ClinVar, GeneReviews, ClinGen, Human Phenotype Ontology, GenCC, DECIPHER, Orpha data, DisGeNET, GTR, Noncode, MGI, Gene4Denovo |
| Gene expression | BrainSpan, GTEx, Allen Brain Atlases, The Human Protein Atlas |
| Target drug | DGIdb, PharmGKB, CTD, Drug Central, Drug Target Commons |