The global prevalence and ethnic heterogenicity of iron-refractory iron deficiency anemia

Shanghua Fan[1*], Ting Zhao[2*], Liu Sun[3#]

[1]Department of Neurology, Renmin Hospital of Wuhan University, Wuhan, 430060, China
[2]Department of Neurology, Henan Provincial People's Hospital, People's Hospital of Zhengzhou University, Zhengzhou, 450003, China
[3]Department of Information Technology, School of Mathematics and Information Technology, Yuxi Normal University, Yuxi, 653100, China
*Shanghua Fan and Ting Zhao contributed equally to this work.
[#]Correspondence: sunliu@yxnu.edu.cn

Abstract

Background: Iron-Refractory Iron Deficiency Anemia (IRIDA) is an autosomal recessive iron deficiency anemia due to mutations in the *TMPRSS6* gene. Iron deficiency anemia is common, but IRIDA is rare. The prevalence of IRIDA is uncertain. The study aimed to estimate the carrier frequency and genetic prevalence of IRIDA using Genome Aggregation Database (gnomAD) data.

Methods: The pathogenicity of *TMPRSS6* variants were interpreted according to the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) standards and guidelines. the minor allele frequency (MAF) of disease-causing variants in TMPRSS6 gene in 141456 unique individuals to estimate the global prevalence of IRIDA in seven ethnicities: African/African American (afr), American Admixed/Latino (amr), Ashkenazi Jewish (asj), East Asian (eas), Finnish (fin), Non-Finnish European (nfe) and South Asian (sas). The global and population-specific carrier frequencies and genetic prevalence of IRIDA were calculated using the Hardy-Weinberg equation.

Results: According to the ACMG/AMP guidelines, 86 pathogenic/likely pathogenic variants (PV/LPV) were identified. The global carrier frequency and genetic prevalence of IRIDA are 2.02 per thousand and 1.02 per million, respectively.

Conclusions: The prevalence of IRIDA is greater than previous estimates.

Keywords: iron-refractory iron deficiency anemia (IRIDA), carrier frequency, genetic prevalence, genome aggregation database (gnomAD), transmembrane serine protease 6 (TMPRSS6), rare diseases

Background

Iron-refractory iron deficiency anemia (IRIDA) (OMIM: 206200, ORPHA: 209981) is a refractory iron deficiency anemia that is unresponsive to oral iron treatment[1-3], which leads to microcytic hypochromic anemia, low transferrin saturation and serum iron[4]. Symptoms of IRIDA are usually mild, including tiredness (fatigue), weakness, pale skin, dizziness, and exercise-induced dyspnea. These symptoms are most pronounced during childhood, Most IRIDA patients have normal growth and development. IRIDA is usually suspected in childhood due to the results of routine blood tests, diagnosis is based on laboratory tests that show hypochromic microcytic anemia with low serum iron and transferrin saturation levels, and then the diagnosis is confirmed by genetic testing. Treatment usually involves intravenous (IV) iron therapy[5].

IRIDA is an ultrarare autosomal recessive disorder caused by mutations in the transmembrane serine protease 6 (*TMPRSS6*) gene [6], which is a transmembrane serine protease that plays an essential role in down-regulating hepcidin antimicrobial peptide (HAMP), the key regulator of iron homeostasis[7-9], through the cleavage of the cell surface hemojuvelin BMP co-receptor (HJV), an activator of hepcidin expression [10-13]. The TMPRSS6 gene spans 18 exons and 51125 base pairs located at chromosome 22 and encodes a protein with 802 amino acids. To date, more than 100 TMPRSS6 variants have been identified in patients with IRIDA since 2008[14]. (full paper list and full variant list are available in the Science Data Bank (ScienceDB) repository)

Although iron deficiency anemia is relatively common[3], IRIDA is rare, and its prevalence is unknown. orphanet database (https://www.orpha.net/consor/cgi-bin/index.php) estimated less than 1 per 1 million. we then proceeded to attempt to obtain a more reliable estimate of global IRIDA prevalence and carrier frequency from Genome Aggregation Database (gnomAD) dataset using previous protocol[15]. Additionally, we aimed to generate a curated artificial intelligence training dataset of TMPRSS6 variants for pathogenicity classification.

Methods

Identification of previously reported TMPRSS6 disease-causing variants

To evaluate the genetic spectrum of IRIDA, a comprehensive search was performed to identify all previously reported disease-causing variants in the TMPRSS6 gene. Searches were conducted in PubMed or Scopus database and using the following combinations of search terms: iron deficiency anemia, iron deficiency, iron-refractory iron deficiency anemia, iron-refractory iron deficiency anaemia, Matriptase-2 (MT2), tmprss6, transmembrane serine protease 6, IRIDA, mutations, variants, variations, and mutants.

Two independent authors screened publications according to inclusion and exclusion criteria: original case reports reporting disease-causing variants within the TMPRSS6 gene were included, and variants in abstract, full-text, tables, figures or supplementary figures and tables were extracted. Non-English-language articles, reviews, comments, editorials, letters, etc. and *in vitro* and animal model studies were excluded. Common TMPRSS6 polymorphisms associated with iron deficiency/iron deficiency anemia in genome-wide association studies were excluded.

All publications were saved in the Medline format and stored in the MySQL and

MongoDB database using NCBI Entrez Programming Utilities [17] (E-utilities) with the Python package biopython [18] and mysql connector and pymongo implementation. reported TMPRSS6 variants were also identified from Leiden Open Variation Database (LOVD) (https://www.lovd.nl/), NCBI[16] ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), dbSNP (https://www.ncbi.nlm.nih.gov/snp/), VarSome[17] (https://varsome.com/), Ensembl [18] (https://www.ensembl.org/) UniProtKB[19] (https://www.uniprot.org/), Genomenon Mastermind[20] (https://mastermind.genomenon.com/), LitVar [21] (https://www.ncbi.nlm.nih.gov/research/litvar2/), Online Mendelian Inheritance in Man (OMIM) (https://www.omim.org/) and HGMD (http://www.hgmd.cf.ac.uk/ac/index.php) database.
prediction and functional annotation for all TMPRSS6 potential nonsynonymous single-nucleotide variants (nsSNVs) were compiled using dbNSFP 4. [22]

Identification and prediction of novel TMPRSS6 disease-causing variants

The gnomAD database (https://gnomad.broadinstitute.org/) was searched for novel disease-causing TMPRSS6 variants that had not yet been reported, and protein-truncating variants (PTVs) were examined (frameshifts, stop codons, initiator codons, splice donors and splice acceptors).

Interpretation of Pathogenicity of *TMPRSS6* variants

The pathogenicity of variants was interpreted followed the protocol by Zhang *et al.* [23] pathogenicity of all missense, synonymous and protein-truncating TMPRSS6 variants were classified according to Standards and Guidelines of the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) [24] with ClinGen Variant Curation Interface[25].
for missense and synonymous variants, we executed additional literature retrieval to curate *in vitro* or *in vivo* functional studies supportive of a damaging effect on TMPRSS6 missense/synonymous variants.
Variants classified as pathogenic and likely pathogenic were included. Variants classified as benign, likely benign, or uncertain significance were excluded. Pathogenic/likely pathogenic variants were included in the carrier frequency and genetic prevalence calculation.

Annotation of variants with minor allele frequencies (MAFs)

Each identified variant were normalized representation with HGVS nomenclature, SPDI [26], VCF, ClinVar, ClinGen Allele Registry[27] and dbSNP Reference SNP (rs or RefSNP) using Mutalyzer[28] (https://mutalyzer.nl/), NCBI Entrez Programming Utilities (E-utilities), VariantValidator[29] (https://variantvalidator.org/) and Ensembl REST API (https://rest.ensembl.org/).
canonical TMPRSS6 transcript (NM_001374504.1, 3,265 bp) and protein (NP_001361433.1, 802 aa) selected by Matched Annotation from NCBI and EMBL-EBI (MANE) [30] different from 811 aa isoform in the literature, The difference between the two isoforms is initiated translation at an alternate start codon MLLLFHSKRM. [31].
Parquet format of gnomAD v2.1.1 VCF files were downloaded from Microsoft Azure Open Datasets with Microsoft AzCopy tool, the minor allele frequency of gnomAD population for the following ethnic groups: African/African American (AFR),

American Admixed/Latino (AMR), East Asian (EAS), Non-Finish European (NFE) and South Asian (SAS) were retrieved with Apache Drill, Apache Spark and Apache Zeppelin.

The carrier frequency and genetic prevalence calculation

the carrier frequency and genetic prevalence of IRIDA were calculated based on the Hardy–Weinberg equation[32]. For monogenic autosomal recessive disorder, the genetic prevalence would be $[1-\prod_i(1-q_i)]^2$ based on the theory of probability, q_i stands for each pathogenic/likely pathogenic variant minor allele frequency. the genetic prevalence was approximately equal $(\sum q_i)^2$ and carrier frequency would be $2(1-\sum q_i)\sum q_i \approx 2\sum q_i$. estimated disease prevalence by using the observed allele frequency of a pathogenic/likely pathogenic variant in gnomAD database as the direct estimator for q_i. the disease prevalence can be estimated by $[1-\prod_i(1-\frac{AC_i}{AN_i})]^2 \approx \sum AF_i^2$. where AC is the allele count and AN is the corresponding allele number and AF is the corresponding allele frequency.

The Python statistics package statsmodels and scientific computing package NumPy and pandas were calculated for the 95% confidence interval (95% CI) for the binomial proportion of carrier frequency and genetic prevalence with the Wilson Score interval. Graphics were plotted using the ProteinPaint [33].

Results

Identification of TMPRSS6 missense, synonymous and protein-truncating variants

Comprehensive retrieving for IRIDA disease-causing variants resulted in the identification of 813 articles, of which 39 articles were considered eligible according to the exclusion and inclusion criteria. All disease-causing variant ETL pipelines and tools are shown in Fig. 1. From these articles, 86 disease-causing variants in the TMPRSS6 gene, including 40 missense variants and 46 protein-truncating variants (PTVs) were classified as pathogenic/likely pathogenic (missense variants are showed in Table 1 and all variants are shown in Fig. 2). The most common mutation consequence was missense, accounting for 46.5% (40/86) of all pathogenic/likely pathogenic variants and contributing more than half of the total allele frequency.

One synonymous variant (c.1086G>A / c.1086G>T, p.Thr362=) was classified as likely pathogenic [34] (PP3, PP4, PM2 and PS3 criteria). But the variants were excluded from gnomAD dataset because low confidence pLoF.

46 protein-truncating variants were divided into three consequences: frameshift (n=20), nonsense (n=17) and splice (n=9). PTVs accounted for more than half of the number of pathogenic/likely pathogenic variants, but less than half of the total allele frequency.

Estimation of population prevalence of IRIDA

Pooling of the allele frequencies for all pathogenic/likely pathogenic variants gave a global minor allele frequency of 0.001, which is equivalent to a genetic prevalence of 1.0228 per 1 million (95% CI: [0.82, 1.28]). The African population had the highest prevalence, at 3.55 per million population (95% CI: [1.76, 7.18]), and non-Finish

European and East Asian population had a prevalence of greater than 1 per million, the prevalence of American/Admixed Latino and South Asian smaller than 1 per million. Ashkenazi Jewish and Finish population are ignored because the sample size is too small to be calculated (Table 2).

Table 1 TMPRSS6 pathogenic/likely pathogenic missense variants present in gnomAD with global minor allele frequencies

Genomic position	Nucleotide change	Amino acid change	gnomAD AF
22-37494466-G-T	326C>A	Ala109Asp	0.000007953
22-37494467-C-T	325G>A	Ala109Thr	0.000003976
22-37492697-T-C	395A>G	Tyr132Cys	0.000005758
22-37491614-A-G	608T>C	Ile203Thr	0.00001608
22-37491613-T-C	609A>G	Ile203Met	0.000004025
22-37491614-A-T	608T>A	Ile203Lys	0.000004021
22-37485778-G-T	676C>A	Leu226Met	0.000004077
22-37485747-C-G	707G>C	Cys236Ser	0.00004799
22-37485747-C-T	707G>A	Cys236Tyr	0.000004102
22-37485646-C-T	808G>A	Gly270Arg	0.00002176
22-37482412-G-A	884C>T	Ser295Leu	0.000007375
22-37480876-C-T	977G>A	Cys326Tyr	0.000006391
22-37480819-G-A	1034C>T	Pro345Leu	0.00001131
22-37471316-A-G	1201T>C	Cys401Arg	0.000008036
22-37471291-T-C	1226A>G	Tyr409Cys	0.000004
22-37471220-C-T	1297G>A	Gly433Arg	0.000003981
22-37469593-C-T	1534G>A	Asp512Asn	0.00003183
22-37469591-G-C	1536C>A	Asp512Glu	0.000003979
22-37469590-C-T	1537G>A	Glu513Lys	0.00001591
22-37466665-C-A	1700G>T	Arg567Leu	0.000004018
22-37466665-C-T	1700G>A	Arg567His	0.00003134
22-37466666-G-A	1699C>T	Arg567Cys	0.00003215
22-37466624-A-G	1741T>C	Trp581Arg	0.000003997
22-37466603-G-A	1762C>T	Arg588Trp	0.00001599
22-37466602-C-T	1763G>A	Arg588Gln	0.00001598
22-37466585-C-G	1780G>C	Gly594Arg	0.00005675
22-37466584-C-G	1781G>C	Gly594Ala	0.000003991
22-37466585-C-A	1780G>T	Gly594Trp	0.00003193
22-37466578-G-A	1787C>T	Ala596Val	0.00003188
22-37466579-C-T	1786G>A	Ala596Thr	0.000004002
22-37466576-G-A	1789C>T	Leu597Phe	0.000007982
22-37465232-A-C	1994T>G	Leu665Arg	0.000004545
22-37465149-A-G	2077T>C	Cys693Arg	0.000006261
22-37462272-A-G	2257T>C	Ser753Pro	0.000007297
22-37462263-G-C	2266C>G	Pro756Ala	0.00002173
22-37462262-G-A	2267C>T	Pro756Leu	0.00001448

Genomic position	Nucleotide change	Amino acid change	gnomAD AF
22-37462236-G-A	2293C>T	Arg765Cys	0.00002154
22-37462235-C-T	2294G>A	Arg765His	0.00002154
37499456A>T	2T>A	Met1Lys	0.000003977
37499455C>T	3G>A	Met1Ile	0.000003977

MAF, minor allele frequency

Table 2 estimated prevalence and carrier frequency among ethnicity and ancestry

Population	**Genetic prevalence (per million)**		**Carrier frequency (per thousand)**	
		95% CI		**95% CI**
Non-Finnish European	1.54	0.87, 2.5	2.48	2.11, 2.92
American Admixed/Latino	0.14	0.05, 0.41	0.75	0.44, 1.29
African/African American	3.55	1.76, 7.18	3.77	2.65, 5.36
South Asian	0.75	0.35, 1.6	1.73	1.19, 2.53
East Asian	1.51	0.67, 3.42	2.46	0.69, 2.18
All	1.02	0.82, 1.28	2.02	1.8, 2.28

CI: confidence interval

Discussion

IRIDA is ultrarare autosomal recessive disorder caused by homozygous or compound heterozygous mutations on the transmembrane serine protease 6 (TMPRSS6) gene. IRIDA patients have microcytic hypochromic anemia, low serum iron and transferrin saturation. The prevalence is unknown. In this study, we sought to estimate the prevalence of IRIDA using 100k scale population genome data and deepen our understanding of the genetic variation of TMPRSS6. Using gnomAD data, we found IRIDA to affect 1.0228 per 1 million across the global population, which is higher than previous orphanet database estimates that have been lower than 1 per 1 million.

The main strength of this study is the use of 86 TMPRSS6 variants from a population of more than 100k for prevalence estimate. in addition to, we have also produced a curated variants dataset in TMPRSS6 that may prove useful for further artificial intelligence-based classification of germline variants pathogenicity.

combined heterozygous mutations of TMPRSS6 and activin A receptor type 1 (ACVR1) caused IRIDA has been reported[35]. ACVR1 R258S and TMPRSS6 I212T maintain high hepcidin expression in the proband. This digenic model suggests that combined TMPRSS6 and other BMP/SMAD signaling pathway gene heterozygous mutations are likely to cause IRIDA. This may explain index cases in which there is only one pathogenic mutation and one or more single nucleotide polymorphism combination.

IRIDA was not on the first national list of rare diseases issued by China, the prevalence of IRIDA has not been estimated in China. [36] National Rare Diseases Registry System (NRDRS) had not registered any IRIDA case. [37]we have demonstrated the power and limitations of the 100k population genome database [38,39] to calculate the prevalence of rare diseases, but gnomAD and the Trans-Omics for Precision Medicine (TOPMed) population are of predominantly non-Asian ancestry. Chinese population genome project will fill the gap. Currently, NyuWa[40], Westlake BioBank for Chinese (WBBC) [41], PGG.Han[42], China Metabolic Analytics Project (ChinaMAP) [43] had been released. We will estimate the prevalence of IRIDA and other rare diseases through the aggregation Chinese population genome data.

The number of TMPRSS6 curated variants in the ClinVar database was far less than the number of the variants we manually curated, but the state-of-the-art model EVE (evolutionary model of variant effect) used the ClinVar dataset as the training set[44]. ClinVar variants curated insufficient and incomplete. ClinGen Variant Curation Expert Panels (VCEPs) [45] define biocuration application of the ACMG/AMP guidelines for sequence variant interpretation in specific genes or diseases. Additionally, ClinGen Variant Curation Interface [25]as a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines for curating variant classifications.

In conclusion, through a comprehensive analysis of genetic variation in TMPRSS6, we have expanded our recognition of disease-causing mutants to 86 variants. IRIDA is an ultrarare disease, and these data can help clinicians diagnose that IRIDA is unlikely to represent a significant proportion of patients presenting with refractory iron deficiency or iron-deficiency anemia.

References

1. Camaschella C. Iron-deficiency anemia. *N Engl J Med*. 2015;372(19):1832-1843.
2. Camaschella C. Iron deficiency. *Blood*. 2019;133(1):30-39.
3. Pasricha SR, Tye-Din J, Muckenthaler MU, Swinkels DW. Iron deficiency. *Lancet*. 2021;397(10270):233-248.
4. De Falco L, Sanchez M, Silvestri L, et al. Iron refractory iron deficiency anemia. *Haematologica*. 2013;98(6):845-853.
5. Hershko C, Camaschella C. How I treat unexplained refractory iron deficiency anemia. *Blood*. 2014;123(3):326-333.
6. Ramsay AJ, Hooper JD, Folgueras AR, Velasco G, Lopez-Otin C. Matriptase-2 (TMPRSS6): a proteolytic regulator of iron homeostasis. *Haematologica*. 2009;94(6):840-849.
7. Du X, She E, Gelbart T, et al. The serine protease TMPRSS6 is required to sense iron deficiency. *Science*. 2008;320(5879):1088-1092.
8. Silvestri L, Pagani A, Nai A, De Domenico I, Kaplan J, Camaschella C. The serine protease matriptase-2 (TMPRSS6) inhibits hepcidin activation by cleaving membrane hemojuvelin. *Cell Metab*. 2008;8(6):502-511.
9. Folgueras AR, de Lara FM, Pendas AM, et al. Membrane-bound serine protease matriptase-2 (Tmprss6) is an essential regulator of iron homeostasis. *Blood*. 2008;112(6):2539-2545.
10. Ganz T. Systemic iron homeostasis. *Physiol Rev*. 2013;93(4):1721-1741.
11. Silvestri L, Guillem F, Pagani A, et al. Molecular mechanisms of the defective hepcidin inhibition in TMPRSS6 mutations associated with iron-refractory iron deficiency anemia. *Blood*. 2009;113(22):5605-5608.
12. Wang CY, Babitt JL. Liver iron sensing and body iron homeostasis. *Blood*. 2019;133(1):18-29.
13. Camaschella C, Nai A, Silvestri L. Iron metabolism and iron disorders revisited in the hepcidin era. *Haematologica*. 2020;105(2):260-272.
14. Finberg KE, Heeney MM, Campagna DR, et al. Mutations in TMPRSS6 cause iron-refractory iron deficiency anemia (IRIDA). *Nat Genet*. 2008;40(5):569-571.
15. Zhao T, Fan S, Sun L. The global carrier frequency and genetic prevalence of Upshaw-Schulman syndrome. *BMC Genom Data*. 2021;22(1):50.
16. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20-D26.
17. Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. *Bioinformatics*. 2019;35(11):1978-1980.
18. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. *Nucleic Acids Res*. 2022;50(D1):D988-D995.
19. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49(D1):D480-D489.
20. Chunn LM, Nefcy DC, Scouten RW, et al. Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation. *Front Genet*. 2020;11:577152.
21. Lee K, Wei CH, Lu Z. Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. *Brief Bioinform*. 2021;22(3).
22. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome*

Med. 2020;12(1):103.
23. Zhang J, Yao Y, He H, Shen J. Clinical Interpretation of Sequence Variants. *Curr Protoc Hum Genet*. 2020;106(1):e98.
24. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424.
25. Preston CG, Wright MW, Madhavrao R, et al. ClinGen Variant Curation Interface: a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines. *Genome Med*. 2022;14(1):6.
26. Holmes JB, Moyer E, Phan L, Maglott D, Kattman B. SPDI: data model for variants and applications at NCBI. *Bioinformatics*. 2020;36(6):1902-1907.
27. Pawliczek P, Patel RY, Ashmore LR, et al. ClinGen Allele Registry links information about genetic variants. *Hum Mutat*. 2018;39(11):1690-1701.
28. Lefter M, Vis JK, Vermaat M, den Dunnen JT, Taschner PEM, Laros JFJ. Next Generation HGVS Nomenclature Checker. *Bioinformatics*. 2021;37(18):2811-2817.
29. Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R. VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. *Hum Mutat*. 2018;39(1):61-68.
30. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*. 2022;604(7905):310-315.
31. Dion SP, Béliveau F, Désilets A, Ghinet MG, Leduc R. Transcriptome analysis reveals TMPRSS6 isoforms with distinct functionalities. *J Cell Mol Med*. 2018;22(4):2498-2509.
32. Liu W, Pajusalu S, Lake NJ, et al. Estimating prevalence for limb-girdle muscular dystrophy based on public sequencing databases. *Genet Med*. 2019;21(11):2512-2520.
33. Zhou X, Edmonson MN, Wilkinson MR, et al. Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet*. 2016;48(1):4-6.
34. Xiong Y, Wu Z, Yang W, et al. A novel splicing mutation of TMPRSS6 in a Chinese child with iron-refractory iron deficiency anaemia. *Br J Haematol*. 2015;171(4):647-649.
35. Pagani A, Colucci S, Bocciardi R, et al. A new form of IRIDA due to combined heterozygous mutations of TMPRSS6 and ACVR1A encoding the BMP receptor ALK2. *Blood*. 2017;129(25):3392-3395.
36. Shi X, Liu H, Zhan S, et al. Rare diseases in China: analysis of 2014-2015 hospitalization summary reports for 281 rare diseases from 96 tertiary hospitals. *Orphanet J Rare Dis*. 2019;14(1):160.
37. Guo J, Liu P, Chen L, et al. National Rare Diseases Registry System (NRDRS): China's first nation-wide rare diseases demographic analyses. *Orphanet J Rare Dis*. 2021;16(1):515.
38. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021;590(7845):290-299.
39. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443.
40. Zhang P, Luo H, Li Y, et al. NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep*. 2021;37(7):110017.
41. Cong P-K, Bai W-Y, Li J-C, et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nature Communications*. 2022;13(1):2939.

42. Gao Y, Zhang C, Yuan L, et al. PGG.Han: the Han Chinese genome database and analysis platform. *Nucleic Acids Res*. 2020;48(D1):D971-d976.
43. Cao Y, Li L, Xu M, et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res*. 2020;30(9):717-731.
44. Frazer J, Notin P, Dias M, et al. Disease variant prediction with deep generative models of evolutionary data. *Nature*. 2021;599(7883):91-95.
45. Rivera-Muñoz EA, Milko LV, Harrison SM, et al. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum Mutat*. 2018;39(11):1614-1622.

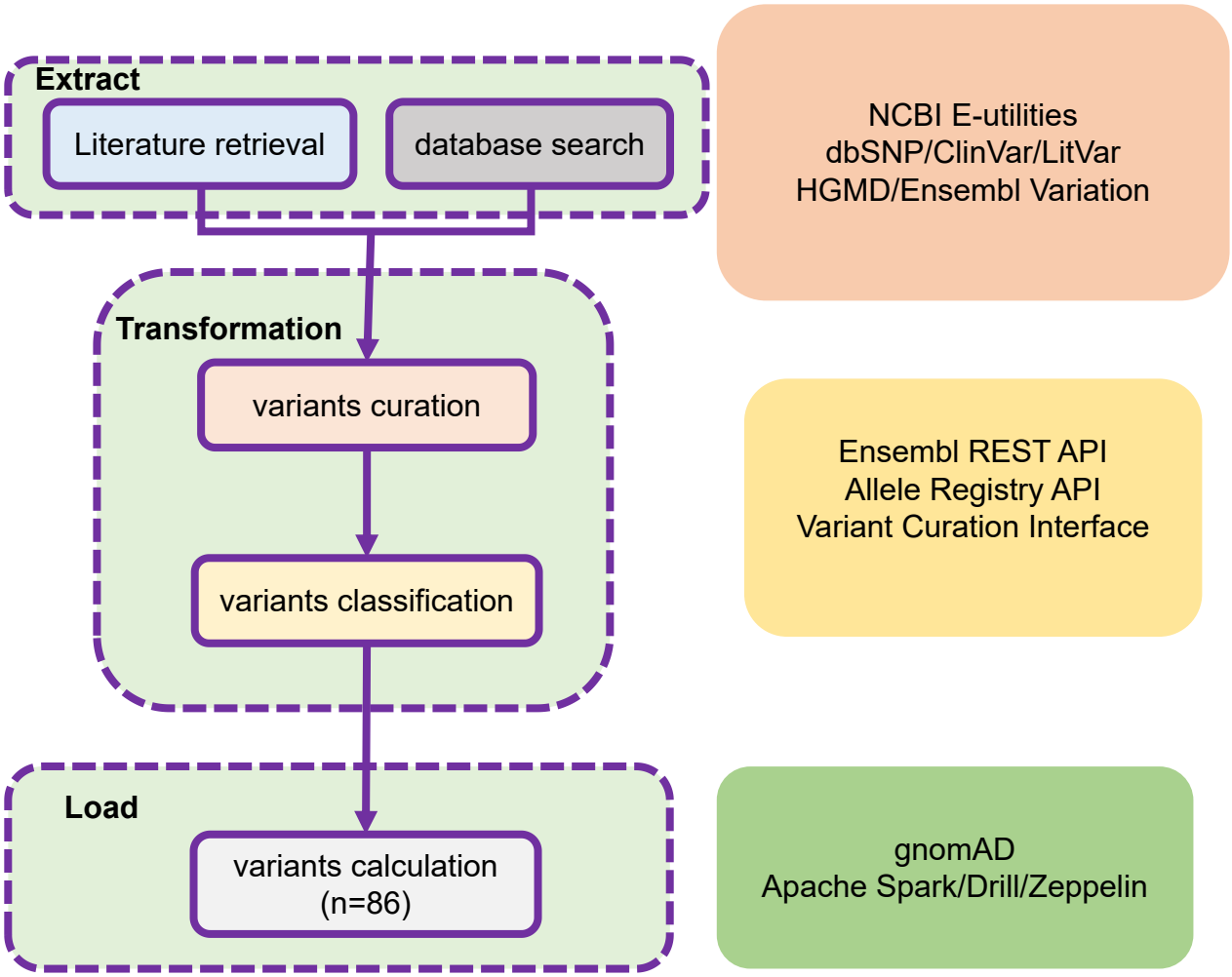

Figure 1. ETL pipeline of curation and classification of TMPRSS6 pathogenic/likely pathogenic variants. Extract: variants from literature retrieval and database search. **Transformation**: raw variant normalized HGVS nomenclature and curated. **Load**: based on the Hardy-Weinberg equation and binomial distribution model calculation

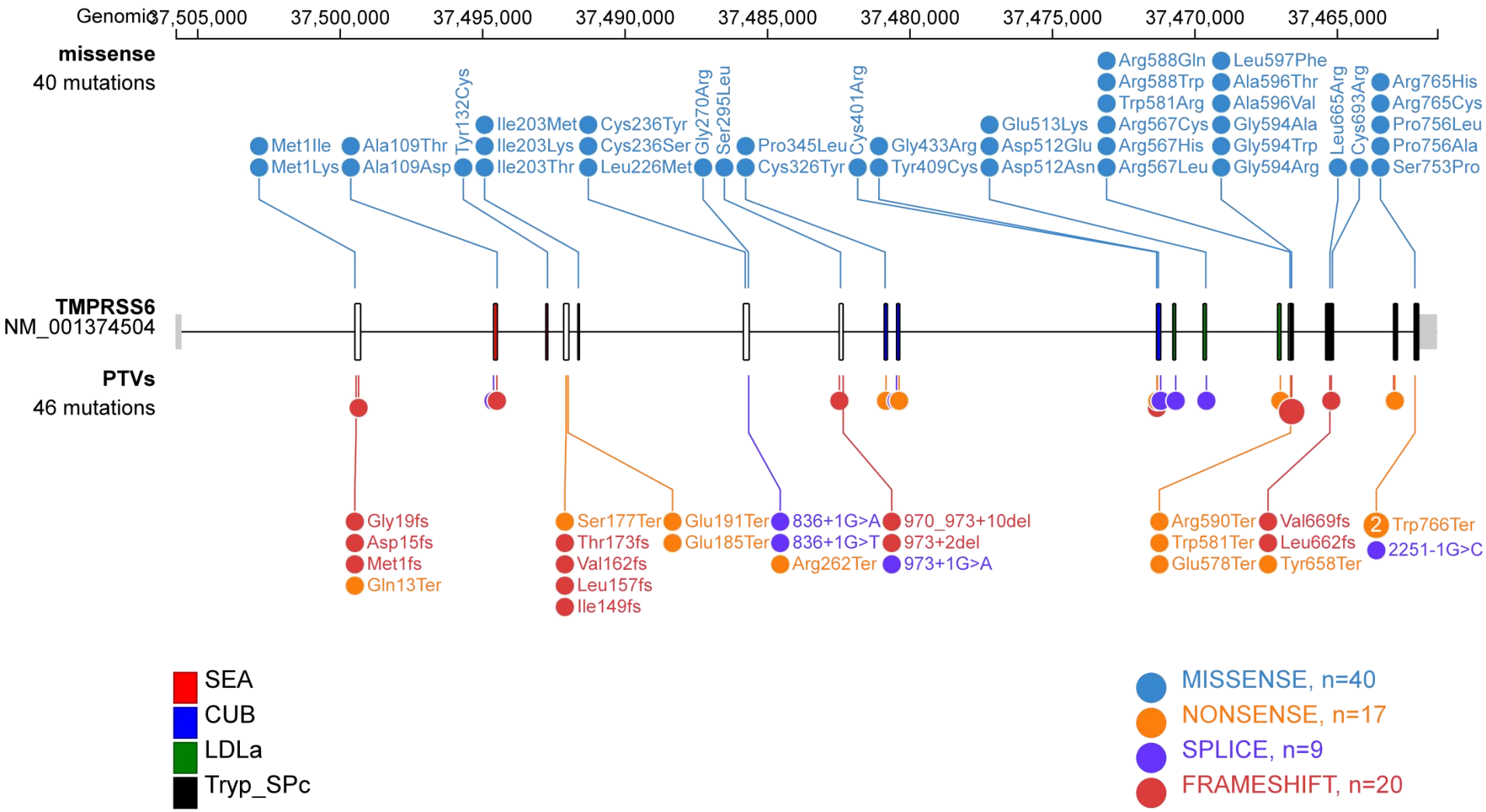

Figure 2. Disease-causing mutations were spread along the length of the TMPRSS6 gene. Missense variants are above and protein-truncating variants (PTVs) are below. TMPRSS6 gene based on Matched Annotation from NCBI and EMBL-EBI (MANE) select transcript and protein. and compared with the variants in the literature, canonical transcript translate protein with offset of 9.