The global prevalence and ethnic heterogeneity of iron-refractory iron deficiency anaemia

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 anaemia (IRIDA) is an autosomal recessive iron deficiency anaemia caused by mutations in the *TMPRSS6* gene. Iron deficiency anaemia is common, whereas IRIDA is rare. The prevalence of IRIDA is unclear. This study aimed to estimate the carrier frequency and genetic prevalence of IRIDA using Genome Aggregation Database (gnomAD) data.

Methods: The pathogenicity of *TMPRSS6* variants was 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 *TMPRSS6* gene disease-causing variants in 141456 unique individuals was

examined 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: In total, 86 pathogenic/likely pathogenic variants (PV/LPV) were identified according to ACMG/AMP guideline. The global carrier frequency and genetic prevalence of IRIDA were 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 anaemia (IRIDA), carrier frequency, genetic prevalence, genome aggregation database (gnomAD), transmembrane serine protease 6 (*TMPRSS6*), rare diseases

Background

Iron-refractory iron deficiency anaemia (IRIDA) (OMIM: 206200, ORPHA: 209981) is a refractory iron deficiency anaemia that is unresponsive to oral iron treatment[1-4], leading to microcytic hypochromic anaemia, low transferrin saturation and serum iron[5]. Symptoms of IRIDA are usually mild, including tiredness (fatigue), weakness, pale skin, dizziness, and exercise-induced dyspnoea. These symptoms are most pronounced during childhood. Most IRIDA patients have normal growth and development. [5] IRIDA is usually suspected in childhood based on the results of routine blood tests. Diagnosis is according to laboratory tests showing hypochromic microcytic anaemia with low serum iron and transferrin saturation levels, and then the

diagnosis is confirmed by genetic testing[6]. Treatment usually involves intravenous (IV) iron therapy[7].

IRIDA is a rare autosomal recessive disorder caused by mutations in the *TMPRSS6* (transmembrane serine protease 6) gene [8]. The gene encodes a serine protease that plays an essential role in downregulating hepcidin, the key regulator of iron homeostasis[9-11], through the cleavage of the cell surface haemojuvelin BMP coreceptor (*HJV*), an activator of hepcidin expression [12-15]. The *TMPRSS6* gene, located on chromosome 22, spans 18 exons and comprises 51125 base pairs, encoding a protein of 802 amino acids. Since 2008, more than 100 *TMPRSS6* variants have been identified in patients with IRIDA[16-18]. (The full paper list and full variant list are available in the Science Data Bank (ScienceDB) repository)

As high-throughput sequencing technology has evolved, re-evaluation guidelines for interpreting and classifying the pathogenicity of identified variants have been implemented. In addition, large-scale population databases have become widely available and can be used for assessment of genetic variants in rare diseases. In fact, for several diseases, there is evidence that these databases have improved the interpretation and classification of variants in patients with monogenic disease and allowed better prediction of which variants are likely to cause disease. The most widely used large-scale reference dataset is from Genome Aggregation Database (gnomAD), which consists of exome sequencing and genome sequencing data from unrelated individuals of diverse ancestries. [19]

The World Health Organization (WHO) defines anaemia as a condition in which the

number of red cells is insufficient to meet the body's physiological needs. Iron deficiency anaemia (IDA) is the most common cause of anaemia worldwide. Although iron deficiency anaemia is relatively common[3], IRIDA is rare. Although its prevalence is unknown, the Orphanet database (https://www.orpha.net/consor/cgi-bin/index.php) estimates is to be less than 1 per 1 million. We attempted to obtain a more reliable estimate of the prevalence and carrier frequency of IRIDA from the Genome Aggregation Database (gnomAD) dataset using a previous protocol[20]. Additionally, we aimed to generate a curated artificial intelligence training dataset of *TMPRSS6* variants for pathogenicity classification and interpretation.

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 *TMPRSS6* gene variants. Searches were conducted in the PubMed or Scopus database using the following combinations of search terms: iron deficiency anaemia, iron deficiency, iron-refractory iron deficiency anaemia, 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 of 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 anaemia in genome-wide association studies were also excluded.

All publications were saved in Medline format and stored in the MySQL and MongoDB databases using NCBI Entrez Programming Utilities [17] (E-utilities) with the Python package biopython [18] and MySQL/MongoDB database implementation.

Reported *TMPRSS6* variants were also identified from Leiden Open Variation Database (LOVD) (https://www.lovd.nl/), NCBI[21] ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), dbSNP (https://www.ncbi.nlm.nih.gov/snp/), VarSome[22] (https://varsome.com/), Ensembl [23] (https://www.ensembl.org/) UniProtKB[24] (https://www.uniprot.org/), Genomenon Mastermind[25] (https://mastermind.genomenon.com/), LitVar [26] (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).

Prediction and functional annotation for all *TMPRSS6* potential nonsynonymous single-nucleotide variants (nsSNVs) were compiled using dbNSFP 4. [27]

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 the pathogenicity of *TMPRSS6* variants

The pathogenicity of variants was interpreted following the protocol by Zhang *et al*.[28] The pathogenicity of all missense, synonymous, and protein-truncating *TMPRSS6* variants was classified according to Standards and Guidelines of the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) [29] with the ClinGen Variant Curation Interface[30]. For missense and synonymous variants, we performed 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 was normalized with HGVS nomenclature, SPDI [31], VCF, ClinVar, ClinGen Allele Registry[32], and dbSNP Reference SNP (rs or RefSNP) using Mutalyzer[33] (https://mutalyzer.nl/), NCBI Entrez Programming Utilities (E-utilities), VariantValidator[34] (https://variantvalidator.org/) and Ensembl REST API (https://rest.ensembl.org/).

The 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) [35] differ from the 811 aa isoform in the literature. The difference between the two isoforms is that translation is initiated at an alternate start codon MLLLFHSKRM. [36].

gnomAD v2.1.1 VCF files in Parquet format were downloaded from Microsoft Azure Open Datasets with the Microsoft AzCopy tool. The minor allele frequency of the gnomAD population for the following ethnic groups were retrieved with Apache Drill, Apache Spark, and Apache Zeppelin: African/African American (AFR), American Admixed/Latino (AMR), East Asian (EAS), Non-Finish European (NFE) and South Asian (SAS).

Carrier frequency and genetic prevalence calculation

The carrier frequency and genetic prevalence of IRIDA were calculated based on the Hardy–Weinberg equation[37]. For a monogenic autosomal recessive disorder, the genetic prevalence is $[1-\prod_i(1-q_i)]^2$ based on the theory of probability, where q_i stands for each pathogenic/likely pathogenic variant minor allele frequency. The genetic prevalence was approximately equal to $(\sum q_i)^2$, and the carrier frequency was $2(1-\sum q_i)\sum q_i \approx 2\sum q_i$. The disease prevalence was estimated by using the observed allele frequency of a pathogenic/likely pathogenic variant in the 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, 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 employed for calculating the 95% confidence interval (95% CI) for

the binomial proportion of carrier frequency and the genetic prevalence with the Wilson Score interval. Graphics were plotted using the R graphic package ggplot2.

Results

Identification of *TMPRSS6* missense, synonymous and protein-truncating variants

Comprehensive retrieval of IRIDA disease-causing variants resulted in the identification of 813 articles, of which 39 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), are classified as pathogenic/likely pathogenic according to the Standards and guidelines of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (missense variants are shown in Table 1; all variants are listed in the Science Data Bank (ScienceDB) online data repository). The most common mutation consequence is missense, accounting for 46.5% (40/86) of all pathogenic/likely pathogenic variants and more than half of the total allele frequency.

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

Genomic position	Nucleotide change	Amino acid change	gnomAD MAF †
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

Genomic position	Nucleotide change	Amino acid change	gnomAD MAF †
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

Genomic position	Nucleotide change	Amino acid change	gnomAD MAF †
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
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

One synonymous variant (c.1086G>A/c.1086G>T, p.Thr362=) is classified as likely pathogenic [38] (PP3, PP4, PM2, and PS3 criteria). However, this variant was excluded from the gnomAD dataset because of low confidence pLoF.

Forty-six protein-truncating variants were divided into three categories: frameshift

(n=20), nonsense (n=17), and splice acceptor/donor (n=9). PTVs accounted for more than half of the number of pathogenic/likely pathogenic variants after filtering (low confidence pLoF, LCR, Not LoF) but less than half of the total allele frequency.

Estimation of population prevalence and the carrier frequency of IRIDA

Pooling of the allele frequencies for all pathogenic/likely pathogenic variants provided 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]) and a carrier frequency of 2.02 per 1 thousand (95% CI: [1.8, 2.28]). The African population had the highest prevalence, at 3.55 per million (95% CI: [1.76, 7.18]), with carrier frequency at 3.77 per thousand. Both non-Finish European and East Asian populations had a prevalence of greater than 1 per million, whereas the prevalence of American/Admixed Latino and South Asian populations was smaller than 1 per million. Ashkenazi Jewish and Finnish populations were not analysed because the sample size was too small to be calculated (Table 2 and Fig. 2).

Table 2 Estimated prevalence and carrier frequency among ethnicities and ancestries

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 a rare autosomal recessive disorder caused by homozygous or compound heterozygous mutations in the transmembrane serine protease 6 (*TMPRSS6*) gene. IRIDA patients have microcytic hypochromic anaemia, 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 to deepen our understanding of *TMPRSS6* genetic variation. Using gnomAD data, we found that IRIDA affects 1.0228 per 1 million in the global population, which is higher than previous Orphanet database estimates of less than 1 per 1 million.

The main strength of this study was the use of 86 *TMPRSS6* variants from a population of more than 100 k for prevalence estimates. In addition, we produced a curated variant dataset for *TMPRSS6* that will be useful for further artificial intelligence-based classification of germline variant pathogenicity.

IRIDA can be difficult to distinguish from acquired iron deficiency unresponsive to oral iron therapy. Overall, IRIDA is a genotypically and phenotypically heterogeneous disease. Allelic *TMPRSS6* mutations are found in most patients with IRIDA. An algorithm has been developed based on blood and plasma values to distinguish patients with IRIDA resulting from biallelic *TMPRSS6* mutations [40-42].

Combined heterozygous mutations in *TMPRSS6* and activin A receptor type 1 (ACVR1) have been reported to cause IRIDA[43]. High hepcidin expression in the proband was reported with ACVR1 R258S and *TMPRSS6* I212T. This digenic model suggests that combined heterozygous mutations in *TMPRSS6* and other BMP/SMAD signalling pathway genes 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.

TMPRSS6 polymorphisms are more frequent than germline *TMPRSS6* mutations, and common *TMPRSS6* variants are associated with erythrocyte traits, including haemoglobin concentration (Hb), haematocrit (Hct), mean corpuscular volume (MCV), mean corpuscular haemoglobin (MCH), mean corpuscular haemoglobin concentration (MCHC) and red blood cell (RBC) count[44, 45]. Additionally, *TMPRSS6* polymorphisms in iron deficiency anaemia are partially responsive to oral iron treatment[46, 47]. A 3-tier classification system for variants of the *TMPRSS6* gene has been proposed to better understand and define the IRIDA spectrum. [48]

Synonymous mutations are also called neutral or silent mutations. Synonymous mutations in protein-coding genes do not alter protein sequences and are thus generally presumed to be neutral. Shen *et al.* constructed 8,341 yeast mutants and measured their fitness relative to wildtype. They found that three-quarters of synonymous mutations resulted in a significant reduction in fitness; the distribution of fitness effects was overall similar, albeit nonidentical[49].

IRIDA is not on the first national list of rare diseases issued by China, and the

prevalence of IRIDA in China has not been estimated. [50] The National Rare Diseases Registry System (NRDRS) has not registered any IRIDA cases. [51] We demonstrated the power and limitations of the 100k population genome database [52, 53] to calculate the prevalence of rare diseases, but gnomAD and the Trans-Omics for Precision Medicine (TOPMed) population are of predominantly non-Asian ancestry. The Chinese population genome project will fill this gap. Currently, NyuWa[54], Westlake BioBank for Chinese (WBBC)[55], and PGG. Han[56], China Metabolic Analytics Project (ChinaMAP)[57] have been released. We will estimate the prevalence of IRIDA and other rare diseases through the aggregation of Chinese population genome data.

The number of *TMPRSS6* curated variants in the ClinVar database is far less than the number of variants we manually curated, and the state-of-the-art model EVE (evolutionary model of variant effect) uses the ClinVar dataset as the training set[58]. ClinGen Variant Curation Expert Panels (VCEPs) [59] define biocuration application of ACMG/AMP guidelines for sequence variant interpretation for specific genes or diseases. Additionally, the ClinGen Variant Curation Interface [30] is 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 expanded our recognition of disease-causing mutations to 86 variants. IRIDA is a rare disease, and these data can help clinicians to diagnose IRIDA that is unlikely to represent a significant proportion of patients presenting with refractory iron

deficiency or iron-deficiency anaemia.

Abbreviations

TMPRSS6: transmembrane serine protease 6

IRIDA: iron refractory iron deficiency anaemia

MAF: minor allele frequency

ClinGen: Clinical Genome

NCBI: National Center for Biotechnology Information

ACMG: American College of Medical Genetics and Genomics

AMP: Association for Molecular Pathology

VCEP: Variant Curation Expert Panel

CI: confidence interval

gnomAD: Genome Aggregation Database

ID: iron deficiency

IDA: iron deficiency anaemia

PTVs: protein-truncating variants

AC: allele count

AF: allele frequency

AN: allele number

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets are available online in the Science Data Bank (ScienceDB) repository. https://doi.org/10.11922/sciencedb.01515

Competing interests

The authors declare no conflicts of interest.

Funding

This study was supported by Yunnan Fundamental Research Projects (grant No. 202101AU070007). The funding bodies had no role in the design of the study; the collection, analysis, and interpretation of data; or in writing the manuscript.

Author Contributions

FSH and ZT retrieved the literature and wrote the manuscript text, and SL designed the project and revised the manuscript and data analysis. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge Genome Aggregation Database (gnomAD).

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. Lopez A, Cacoub P, Macdougall IC, Peyrin-Biroulet L: **Iron deficiency anaemia**. *Lancet* 2016, **387**(10021):907-916.
5. De Falco L, Sanchez M, Silvestri L, Kannengiesser C, Muckenthaler MU, Iolascon A, Gouya L, Camaschella C, Beaumont C: **Iron refractory iron deficiency anemia**. *Haematologica* 2013, **98**(6):845-853.
6. Fletcher A, Forbes A, Svenson N, Wayne Thomas D: **Guideline for the laboratory diagnosis of iron deficiency in adults (excluding pregnancy) and children**. *Br J Haematol* 2022, **196**(3):523-529.
7. Snook J, Bhala N, Beales ILP, Cannings D, Kightley C, Logan RP, Pritchard DM, Sidhu R, Surgenor S, Thomas W *et al*. **British Society of Gastroenterology guidelines for the management of iron deficiency anaemia in adults**. *Gut* 2021, **70**(11):2030-2051.
8. 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.
9. Du X, She E, Gelbart T, Truksa J, Lee P, Xia Y, Khovananth K, Mudd S, Mann N, Moresco EM *et al*. **The serine protease *TMPRSS6* is required to sense iron deficiency**.

Science 2008, **320**(5879):1088-1092.

10. 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.

11. Folgueras AR, de Lara FM, Pendas AM, Garabaya C, Rodriguez F, Astudillo A, Bernal T, Cabanillas R, Lopez-Otin C, Velasco G: **Membrane-bound serine protease matriptase-2 (*Tmprss6*) is an essential regulator of iron homeostasis**. *Blood* 2008, **112**(6):2539-2545.

12. Ganz T: **Systemic iron homeostasis**. *Physiol Rev* 2013, **93**(4):1721-1741.

13. Silvestri L, Guillem F, Pagani A, Nai A, Oudin C, Silva M, Toutain F, Kannengiesser C, Beaumont C, Camaschella C *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.

14. Wang CY, Babitt JL: **Liver iron sensing and body iron homeostasis**. *Blood* 2019, **133**(1):18-29.

15. Camaschella C, Nai A, Silvestri L: **Iron metabolism and iron disorders revisited in the hepcidin era**. *Haematologica* 2020, **105**(2):260-272.

16. Finberg KE, Heeney MM, Campagna DR, Aydinok Y, Pearson HA, Hartman KR, Mayo MM, Samuel SM, Strouse JJ, Markianos K *et al*: **Mutations in *TMPRSS6* cause iron-refractory iron deficiency anemia (IRIDA)**. *Nat Genet* 2008, **40**(5):569-571.

17. Melis MA, Cau M, Congiu R, Sole G, Barella S, Cao A, Westerman M, Cazzola M, Galanello R: **A mutation in the *TMPRSS6* gene, encoding a transmembrane serine**

protease that suppresses hepcidin production, in familial iron deficiency anemia refractory to oral iron. *Haematologica* 2008, **93**(10):1473-1479.

18. Guillem F, Lawson S, Kannengiesser C, Westerman M, Beaumont C, Grandchamp B: **Two nonsense mutations in the *TMPRSS6* gene in a patient with microcytic anemia and iron deficiency**. *Blood* 2008, **112**(5):2089-2091.

19. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP *et al*. **The mutational constraint spectrum quantified from variation in 141,456 humans**. *Nature* 2020, **581**(7809):434-443.

20. Zhao T, Fan S, Sun L: **The global carrier frequency and genetic prevalence of Upshaw-Schulman syndrome**. *BMC Genom Data* 2021, **22**(1):50.

21. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S *et al*. **Database resources of the national center for biotechnology information**. *Nucleic Acids Res* 2022, **50**(D1):D20-D26.

22. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, Massouras A: **VarSome: the human genomic variant search engine**. *Bioinformatics* 2019, **35**(11):1978-1980.

23. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R *et al*. **Ensembl 2022**. *Nucleic Acids Res* 2022, **50**(D1):D988-D995.

24. UniProt C: **UniProt: the universal protein knowledgebase in 2021**. *Nucleic Acids Res* 2021, **49**(D1):D480-D489.

25. Chunn LM, Nefcy DC, Scouten RW, Tarpey RP, Chauhan G, Lim MS,

Elenitoba-Johnson KSJ, Schwartz SA, Kiel MJ: **Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation**. *Front Genet* 2020, **11**:577152.

26. 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).

27. 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.

28. Zhang J, Yao Y, He H, Shen J: **Clinical Interpretation of Sequence Variants**. *Curr Protoc Hum Genet* 2020, **106**(1):e98.

29. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E *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.

30. Preston CG, Wright MW, Madhavrao R, Harrison SM, Goldstein JL, Luo X, Wand H, Wulf B, Cheung G, Mandell ME *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.

31. 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.

32. Pawliczek P, Patel RY, Ashmore LR, Jackson AR, Bizon C, Nelson T, Powell B, Freimuth RR, Strande N, Shah N *et al*: **ClinGen Allele Registry links information about genetic variants**. *Hum Mutat* 2018, **39**(11):1690-1701.

33. Lefter M, Vis JK, Vermaat M, den Dunnen JT, Taschner PEM, Laros JFJ: **Next Generation HGVS Nomenclature Checker**. *Bioinformatics* 2021, **37**(18):2811-2817.

34. 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.

35. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, Cox E, Davidson C, Ermolaeva O, Farrell CM *et al*: **A joint NCBI and EMBL-EBI transcript set for clinical genomics and research**. *Nature* 2022, **604**(7905):310-315.

36. 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.

37. Liu W, Pajusalu S, Lake NJ, Zhou G, Ioannidis N, Mittal P, Johnson NE, Weihl CC, Williams BA, Albrecht DE *et al*: **Estimating prevalence for limb-girdle muscular dystrophy based on public sequencing databases**. *Genet Med* 2019, **21**(11):2512-2520.

38. Xiong Y, Wu Z, Yang W, Zhao X, Peng G, Tang K, Tian Z, Xing H, Rao Q, Wang M *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.

39. Hershko C, Camaschella C: **How I treat unexplained refractory iron deficiency anemia**.

Blood 2014, **123**(3):326-333.

40. Heeney MM, Guo D, De Falco L, Campagna DR, Olbina G, Kao PP, Schmitz-Abe K, Rahimov F, Gutschow P, Westerman K *et al*: **Normalizing hepcidin predicts *TMPRSS6* mutation status in patients with chronic iron deficiency**. *Blood* 2018, **132**(4):448-452.

41. Donker AE, Schaap CC, Novotny VM, Smeets R, Peters TM, van den Heuvel BL, Raphael MF, Rijneveld AW, Appel IM, Vlot AJ *et al*: **Iron refractory iron deficiency anemia: a heterogeneous disease that is not always iron refractory**. *Am J Hematol* 2016, **91**(12):E482-E490.

42. Girelli D, Nemeth E, Swinkels DW: **Hepcidin in the diagnosis of iron disorders**. *Blood* 2016, **127**(23):2809-2813.

43. Pagani A, Colucci S, Bocciardi R, Bertamino M, Dufour C, Ravazzolo R, Silvestri L, Camaschella C: **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.

44. Chambers JC, Zhang W, Li Y, Sehmi J, Wass MN, Zabaneh D, Hoggart C, Bayele H, McCarthy MI, Peltonen L *et al*: **Genome-wide association study identifies variants in *TMPRSS6* associated with hemoglobin levels**. *Nat Genet* 2009, **41**(11):1170-1172.

45. Benyamin B, Ferreira MA, Willemsen G, Gordon S, Middelberg RP, McEvoy BP, Hottenga JJ, Henders AK, Campbell MJ, Wallace L *et al*: **Common variants in *TMPRSS6* are associated with iron status and erythrocyte volume**. *Nat Genet* 2009, **41**(11):1173-1175.

46. Poggiali E, Andreozzi F, Nava I, Consonni D, Graziadei G, Cappellini MD: **The role of**

TMPRSS6 **polymorphisms in iron deficiency anemia partially responsive to oral iron treatment**. *Am J Hematol* 2015, **90**(4):306-309.

47. Buerkli S, Pei SN, Hsiao SC, Lee CT, Zeder C, Zimmermann MB, Moretti D: **The *TMPRSS6* variant (SNP rs855791) affects iron metabolism and oral iron absorption - a stable iron isotope study in Taiwanese women**. *Haematologica* 2021, **106**(11):2897-2905.

48. Sharma P, Bhatia P, Singh M, Das R, Jain R, Bansal D, Attri SV, Trehan A: **A case series highlighting structured hematological, biochemical and molecular approach to clinical oral iron refractoriness in children: A pressing need for a 3-tier system for classification of variants in *TMPRSS6* gene**. *Blood Cells Mol Dis* 2021, **89**:102569.

49. Shen X, Song S, Li C, Zhang J: **Synonymous mutations in representative yeast genes are mostly strongly non-neutral**. *Nature* 2022.

50. Shi X, Liu H, Zhan S, Wang Z, Wang L, Dong C, Wang Y, Yao C, Ding J, Li Y: **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.

51. Guo J, Liu P, Chen L, Lv H, Li J, Yu W, Xu K, Zhu Y, Wu Z, Tian Z *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.

52. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM *et al*: **Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program**. *Nature* 2021, **590**(7845):290-299.

53. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL,

Laricchia KM, Ganna A, Birnbaum DP *et al*. **The mutational constraint spectrum quantified from variation in 141,456 humans**. *Nature* 2020, **581**(7809):434-443.

54. Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, Niu Y, Shi Y, Zhou H, Song T *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.

55. Cong P-K, Bai W-Y, Li J-C, Yang M-Y, Khederzadeh S, Gai S-R, Li N, Liu Y-H, Yu S-H, Zhao W-W *et al*. **Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project**. *Nature Communications* 2022, **13**(1):2939.

56. Gao Y, Zhang C, Yuan L, Ling Y, Wang X, Liu C, Pan Y, Zhang X, Ma X, Wang Y *et al*. **PGG.Han: the Han Chinese genome database and analysis platform**. *Nucleic Acids Res* 2020, **48**(D1):D971-d976.

57. Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, Xu Y, Du P, Wang T, Hu R *et al*. **The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals**. *Cell Res* 2020, **30**(9):717-731.

58. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y, Marks DS: **Disease variant prediction with deep generative models of evolutionary data**. *Nature* 2021, **599**(7883):91-95.

59. Rivera-Muñoz EA, Milko LV, Harrison SM, Azzariti DR, Kurtz CL, Lee K, Mester JL, Weaver MA, Currey E, Craigen W *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.

1

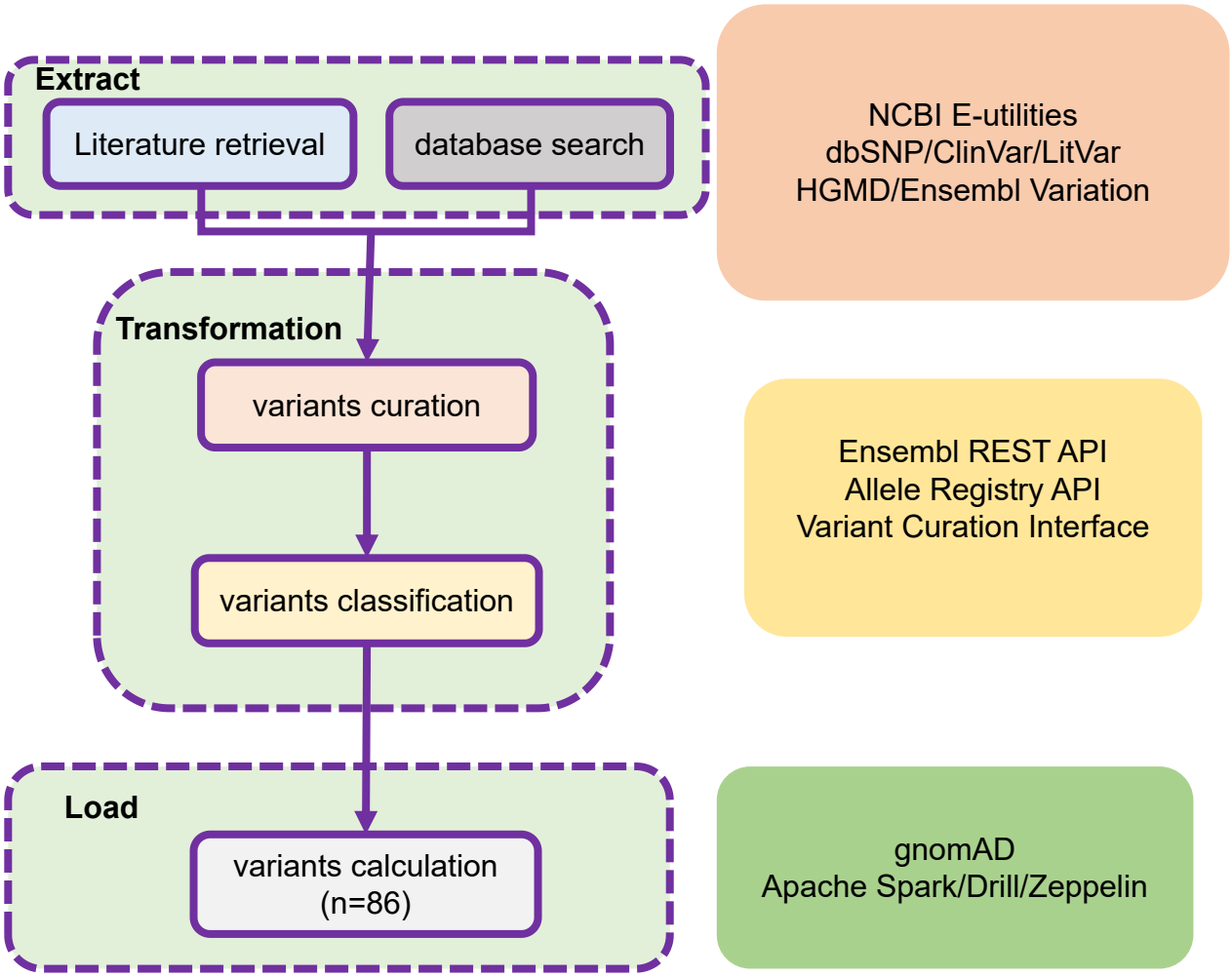

Figure 1. ETL pipeline of curation and classification of TMPRSS6 pathogenic/likely pathogenic variants. Extract: variants from literature retrieval and database search. **Transformation**: raw variant normalized HGVS nomenclature and curated. **Load**: based on the Hardy-Weinberg equation and binomial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

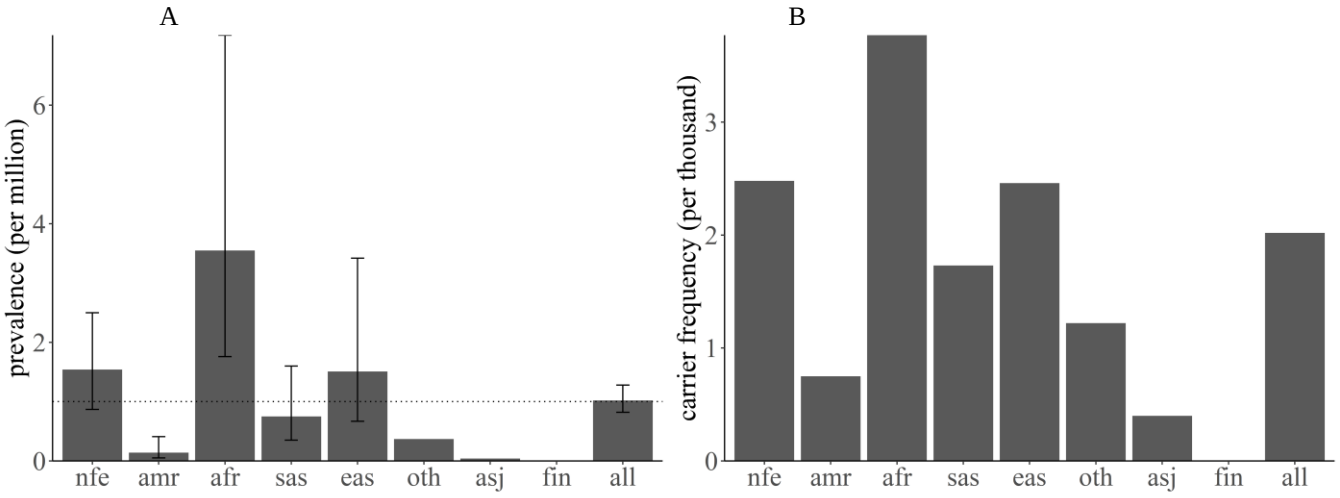

Fig. 2 genetic prevalence and carrier frequency of IRIDA. A. IRIDA genetic prevalence estimated from gnomAD allele frequencies. B. IRIDA carrier frequency among diverse population. African/African American (afr), American Admixed/Latino (amr), Ashkenazi Jewish (asj), East Asian (eas), Finnish (fin), Non-Finnish European (nfe) and South Asian (sas), Other(oth)