Middle Eastern Genetic Variation Improves Clinical Annotation of the Human Genome

Sathishkumar Ramaswamy[1], Ruchi Jain[1], Maha El Naofal[1], Nour Halabi[1], Alan Taylor[1], Ahmad Abou Tayoun[1,2*]

[1]Al Jalila Genomics Center, Al Jalila Children's Hospital, Dubai, United Arab Emirates
[2]Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates

*Corresponding Author: Ahmad Abou Tayoun (Ahmad.Tayoun@ajch.ae)

Conflicts of Interest: The authors declare no conflicts of interest.

Abstract

Genetic variation in populations of Middle Eastern origin remains highly underrepresented in most comprehensive genomic databases. This underrepresentation hampers the functional annotation of the human genome, and also challenges accurate clinical variant interpretation. To highlight the importance of capturing genetic variation in the Middle East, we aggregate whole exome and genome sequencing data from 2,116 individuals in the Middle East and establish the Middle East Variation (MEV) database. Of the high impact coding variants in this database, 34% were absent from the most comprehensive Genome Aggregation Database (gnomAD), thus representing unique Middle Eastern variation which might directly impact clinical variant interpretation. We highlight 167 variants with MAF >1% in the MEV database which were previously reported as rare disease variants in ClinVar and the Human Gene Mutation Database (HGMD). Furthermore, the MEV database consisted of 365 homozygous loss of function (LoF) variants, the majority of which (239/365, 65.5%) were absent from gnomAD, representing complete knockouts of 229 unique genes in reportedly healthy individuals. Intriguingly, 58 of those genes have several clinically significant variants reported in ClinVar and HGMD. Our study shows that genetic variation in the Middle East improves functional annotation and clinical interpretation of the genome and emphasizes the need for expanding sequencing studies in the Middle East and other underrepresented populations.

Introduction

Cataloguing human genetic variation at an unprecedented scale has significantly improved clinical interpretation of genetic variants found in patients with Mendelian disorders [1]. However, current genomic databases still fall short on capturing the full representation of the human genomic diversity. For example, the Middle Eastern and African populations, among others, remain highly underrepresented in the Genome Aggregation Database (gnomAD) database, which is the most comprehensive compendium of human genetic variation to date [1], [2]. This lack of representation is a missed opportunity to fully understand the human genome, and to functionally and clinically annotate its variation.

The Middle Eastern population spanning North Africa, the Arabian Peninsula, and the Syrian desert has a long history of admixture and migration leading to a rich and highly diverse genetic architecture. In addition, this population is characterized by significant endogamy, relatively high consanguinity rates, extended family structure and advanced paternal and/or maternal age at conception [3]. As a result, a high prevalence of Mendelian recessive disorders is expected [4][5] given the higher burden of regions of homozygosity (ROH) in this population. Furthermore, these extended ROH regions can be enriched for biallelic gene knockouts in apparently healthy individuals, shedding light on the biological roles of several genes, and empowering the clinical interpretation of the genome.

Expanding sequencing studies in the Middle East would therefore be undoubtedly a unique opportunity for advancing the human genetics field. However, few attempts have been made [3], [6] to characterize the genetic variation in the Middle East population, while the impact of cataloguing this variation, albeit on a small scale, on the clinical interpretation of genetic variants remains to be elucidated.

In the present study we have assembled sequencing data from Qatar [6] and the Greater Middle East (GME) [3] to highlight the contribution of variants from this population to existing and commonly utilized genomic variation datasets, specifically gnomAD. We also capture disease pathogenicity assertions of rare (based on gnomAD) variants in the Human Gene Mutation

(HGMD) [7] and ClinVar databases [8], which we annotate with allele frequency in the Middle East cohort. These comprehensive variant sets comprise possibly common Middle East disease variants and add a unique set of gene knockouts. Furthermore, our analysis questions the pathogenicity of previously reported disease variants which might be putative polymorphisms. This study demonstrates the importance of capturing genetic variation in the Middle East and highlights the integration of different variant datasets to improve the clinical annotation of the human genome.

Materials and methods

Study sample

We compiled sequencing data from 1,005 individuals from Qatar (88 whole genomes, and 917 whole exomes) [6] and 1,111 individuals from The Greater Middle East (GME) exome sequencing study to characterize variation in the Middle East. Sequencing protocols, variant calling pipelines, and quality control metrics are detailed in the original studies [3], [6].

Individuals from Qatar were Bedouin (n = 490), Arabs (n = 193), Persian (n = 170), South Asian (n = 76), Sub-Saharan African (n = 70), European (n = 5), or African Pygmy (n=1). On the other hand, individuals from the GME dataset were from Northeast Africa (NEA, n=423), Northwest Africa (NWA, n=85), Arabian Peninsula (AP, n=214), Turkish Peninsula (TP, n=140), Syrian Desert (SD, n=81), and Persia and Pakistan (PP, n=168) (**Figure 1**).

Middle East Variation (MEV) database

Sequence variants from both Qatar and GME datasets (VCF/TSV files) were compared and merged using hg19 chromosomal coordinates, by in-house pipeline, to generate a non-redundant Middle East Variation (MEV) database (**Table 1**).

Functional and Clinical Annotation of variants in the Middle East Variation (MEV) database

Merged MEV variants were then annotated with gnomAD allele frequency using the exome dataset in gnomAD 2.1.1 which has 17,209,972 variants from unrelated individuals sequenced as

part of various disease-specific and population genetic studies [1]. Variants were then further annotated using multiple public and commercial databases including, HGMD v3 [7], ClinVar v29032021 [8], NCBI-RefSeq-105 [9], OMIM [10], Erichr [11] and uniprot database [14] using an in-house pipeline.

Our variant annotation pipeline overlays variant positions with extensive resources from the NCBI-RefSeq-105 database and assigns variant consequences [15] with respect to each transcript and protein within the NCBI-RefSeq-105 database. Additional gene annotation such as disease phenotype, pathogenicity and function were obtained from various data sources [10], [7], [8], [14]. High impact coding variants (missense, stop lost/gain, splice acceptor/donor (±1,2), frameshift) in MEV database were then classified as "unique" or "reported" if they were, respectively, absent from or reported at least one time in gnomAD 2.1.1 (**Table 1**).

Annotated variants were subsequently classified into two main classes I and II.

Class I: Common Middle East Disease Variants (CMEDVs)

To obtain this list, heterozygous variants with minor allele frequency (MAF) >1% or ≥1 homozygote in our MEV database were intersected with rare (<1% in gnomAD) HGMD – disease mutations (DM) and variants with pathogenic (P) and likely pathogenic (LP) classifications and ≥ 1 star in ClinVar.

Class II: Knockouts (KOs)

Homozygous loss of function (LoF) variants (nonsense, frameshift, and ±1,2 splice site) in the MEV database were filtered for further manual curation. High confidence LoF status was extracted from gnomAD 2.1.1 for LoF variants that were present in this database. To infer high confidence LoF impact for LoF variants absent from gnomAD 2.1.1, we excluded variants affecting initiator codons or those located in the last coding exon (CDS) or within 50 bp of the penultimate CDS. Variants located in alternatively spliced exons were excluded if the exon was not functionally or clinically relevant based on clinically curated transcripts in disease databases and/or expression data in the Genotype-Tissue Expression (GTEx) dataset [12]. In addition, we removed LoF variants with low quality associated with high homology regions or pseudogenes

as described [1313]. High confidence LoFs were manually verified using Alamut programme v2.11.

Enrichment analysis

We used Enrichr [11] to perform pathway enrichment analysis of genes with CMEDVs reported as P/LP with ≥ 1 star in ClinVar and no conflicting interpretations. Enrichr was run with default settings and significant pathways were identified using Fisher exact test p-value <0.05 that has been corrected for multiple testing across all genes (n = 25) using the Benjamini-Hochberg False Discovery Rate (FDR) procedure [16].

Results

Middle East Variation (MEV) database

A total of 26,756,405 non-redundant variants were merged from the GME and Qatar datasets to establish the MEV database. We focus on the set of high impact coding (missense, stop gain/loss, splice acceptor/donor (±1,2), frameshift) variants (n = 688,730) affecting RefSeq transcripts/exons (**Methods**) given such variants represent the majority of disease variants [7]. Of those, 234,438 (34%) variants were absent from gnomAD 2.1.1 exomes (**Table 1**) representing unique coding variation in this Middle Eastern cohort.

Table 1: Distribution of variants in MEV database

	Total variants	SNPs	Indels
Total MEVs	26,756,405	21,321,147	5,435,258
Total coding variants	688,730	686,205	2,525
Unique coding variants*	234,438 (34%)	232,783	1,655
Reported coding variants**	454,292 (66%)	453,422	870

*Unique coding variants = Variants not reported in gnomAD 2.1.1 Exomes
**Reported coding variants = Variants reported at least once gnomAD 2.1.1 Exomes

There were significantly more singleton or rare (MAF <0.1%) unique coding variants compared to the reported ones (43% vs 35%, p<0.00001, Fisher Exact Test), while the majority of the latter variants were relatively common (MAF >0.1%) (**Supplementary Figure 1**).

Common Middle East Disease Variants (CMEDVs)

Of the total HGMD-DM and ClinVar P/LP variants that were relatively rare in gnomAD (MAF <1%), 4,185 were observed in the MEV database and 167 of those variants were common (MAF >1%) or had at least 1 homozygote in the MEV database (**Supplementary Table 1** and **Figure 2a**). Those common Middle East disease variants (CMEDVs), which were mostly missense (n = 138, 83%) affected 144 genes (**Figure 2b**) with P/LP assertions in ClinVar (n = 113) or DM status in HGMD (n = 54) (**Figure 2c**).

While it is highly likely that a significant proportion of those 167 variants can be reclassified to benign or likely benign based on the MEV allele frequency, it is also possible that a subset can still be clinically significant. In fact, while 51 of the 113 ClinVar variants had conflicting interpretations and can possibly be reclassified, the remaining 62 ClinVar variants had no conflicting interpretations and might be common founder mutations in the Middle East. We therefore examined pathway/gene enrichment analysis on this latter set (non-conflicting P/LP with ≥1 star) and identified significant enrichment (adjusted P value <0.05) in 'Blood related disease', 'Hereditary deafness', and 'Familial hypercholesterolemia' pathways, among others (**Supplementary Table 2** and **Figure 2d**), which are consistent with the disease landscape in this part of the world. This further supports the notion that some variants in the MEV database represent common pathogenic variants for prevalent disorders in the Middle East.

Knockouts in the MEV database

There were 365 high confidence homozygous LoF variants in the MEV database and 239 of those (65.5%) were not present in the homozygous state in gnomAD 2.1.1 exomes, thus representing unique Middle Eastern knockouts (**Supplementary Table 3** and **Figure 3a**). This unique homozygous variant set were mostly nonsense (n = 228, 95%) (**Figure 3b**) impacting 229 genes where at least 145 of those genes had some disease association in disease databases (**Figure 3c**).

Particularly, 58 of the 239 unique knockouts, which were identified in reportedly healthy individuals in the GME dataset, had several DM and P/LP reported in HGMD and ClinVar, respectively (**Supplementary Table 3**). Examples include *SPG11* (MIM# 610844) which is associated with autosomal recessive spastic paraplegia, *RAB3GAP2* (MIM# 609275) which is associated with autosomal recessive Marsolf syndrome, *OTOGL* (MIM# 614925) which associated with autosomal recessive deafness, *NPHP4* (MIM# 607215) which associated with autosomal recessive nephronophthisis, and *SZT2* (MIM# 615463) which associated with autosomal recessive epileptic encephalopathy).

Interestingly, our dataset includes a homozygous LoF in *SHOC2*, a gene where gain of function variants have been strongly associated with autosomal dominant Noonan syndrome-like disease (MIM# 602775). While this suggests that complete absence of this *SHOC2* has no clinical outcomes, the extreme intolerance of this gene to LoF variants in gnomAD remains to be investigated.

Discussion

We have aggregated the largest variant database from 2,116 individuals of Middle Eastern origin and characterized the impact of this dataset on the functional and clinical annotation of the human genome. We therefore focus on coding variants which represent the majority of reported disease variants to date [7] and show that 34% of those variants in the MEV database are absent from gnomAD 2.1.1 exomes and are thus specific to the Middle East population.

Using the MEV database, we highlight 167 variants, which were previously reported as rare clinically significant variants in disease databases (ClinVar and HGMD) yet were common (MAF 1%) or present in the homozygous state at least once in Middle Eastern individuals. While this information might question the pathogenicity of a proportion of those variants, specifically some of the HGMD-DM variants (n = 54) and those with conflicting interpretations in ClinVar (n =51), others might actually be clinically significant founder mutations. In fact, our pathway enrichment analysis of the P/LP variant set, with ≥ 1 star and no conflicting interpretations in ClinVar (n = 62), supports this notion given that most significant pathway enrichments were for highly prevalent genetic diseases in the Middle East. Similarly, we expect a subset of the 54

HGMD-DM variants to be common mutations underlying somewhat common disorders in the Middle East.

Our MEV database consists of 365 high confidence homozygous LoF variants, the majority of which (239/365, 65.5%) were absent from gnomAD exomes 2.1.1. This unique set affects 229 genes, of which 58 had several clinically significant variants reported in ClinVar and HGMD yet were identified in reportedly healthy individuals in our dataset. While it might question the clinical validity of some of the impacted genes, this information might in fact refine our understanding of penetrance, expressivity, and severity for the diseases caused by those genes. Furthermore, similar to *SHOC2*, it is possible the disease mechanism for some of the genes is gain of function and heterozygous or homozygous loss of function is clinically and biologically irrelevant. Finally, it is also possible that the current exon and transcript structure and expression for some genes should be revisited in light of this information. Nonetheless, there is no question that this unique MEV variant dataset improves the clinical and functional annotation of the human genome.

Our study is limited by its size (n = 2,116 individuals) which might not capture the full genetic diversity in the Middle East. Despite its small size, however, the value of cataloguing genetic variation in this population, as demonstrated in this study, should encourage the expansion of sequencing studies in this Middle East and other underrepresented populations, to maximize our understanding of the human genome.

References:

1. Karczewski, K. J.; Francioli, L. C.; Tiao, G.; Cummings, B. B.; Alföldi, J.; Wang, Q.; et al. The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans. *Nature.* **2020**, 581 (7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7.
2. Tayoun, A. N. A.; Rehm, H. L. Genetic Variation in the Middle East — an Opportunity to Advance the Human Genetics Field. *Genome Med.* **2020**, 7, 12–15. https://doi.org/10.1186/s13073-020-00821-7.

3. Scott, E. M.; Halees, A.; Itan, Y.; Spencer, E. G.; He, Y.; Azab, M. A.; et al. Characterization of Greater Middle Eastern Genetic Variation for Enhanced Disease Gene Discovery. *Nat. Genet.* **2016**, 48 (9), 1071–1079. https://doi.org/10.1038/ng.3592.
4. Mahfouz, N. A.; Kizhakkedath, P.; Ibrahim, A.; El Naofal, M.; Ramaswamy, S.; Harilal, D.; Qutub, Y.; et al. Utility of Clinical Exome Sequencing in a Complex Emirati Pediatric Cohort. *Comput. Struct. Biotechnol. J.* **2020**, 18, 1020–1027. https://doi.org/10.1016/j.csbj.2020.04.013.
5. Alsalem, A. B.; Halees, A. S.; Anazi, S.; Alshamekh, S.; Alkuraya, F. S. Autozygome Sequencing Expands the Horizon of Human Knockout Research and Provides Novel Insights into Human Phenotypic Variation. *PLoS Genet.* **2013**, 9 (12). https://doi.org/10.1371/journal.pgen.1004030.
6. Fakhro, K. A.; Staudt, M. R.; Ramstetter, M. D.; Robay, A.; Malek, J. A.; Badii, R.; Al-Marri, A. A. N.; et al. The Qatar Genome: A Population-Specific Tool for Precision Medicine in the Middle East. *Hum. Genome Var.* **2016**, 3, 1–7. https://doi.org/10.1038/hgv.2016.16.
7. Stenson, P. D.; Mort, M.; Ball, E. V; Chapman, M.; Evans, K.; Azevedo, L.; et al. The Human Gene Mutation Database (HGMD ®): Optimizing Its Use in a Clinical Diagnostic or Research Setting. *Hum. Genet.* **2020**, 139 (10), 1197–1207. https://doi.org/10.1007/s00439-020-02199-3.
8. Landrum, M. J.; Lee, J. M.; Riley, G. R.; Jang, W.; Rubinstein, S.; Church, D. M.; Maglott, D. R. ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype. *Nucleic Acids Res.* **2014**, 42, 980–985. https://doi.org/10.1093/nar/gkt1113.
9. Pruitt, K. D.; Brown, G. R.; Hiatt, S. M.; Thibaud-Nissen, F.; Astashyn, A.; Ermolaeva, O.; et al. RefSeq: An Update on Mammalian Reference Sequences. *Nucleic Acids Res.* **2014**, 42 (D1), 756–763. https://doi.org/10.1093/nar/gkt1114.
10. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). https://omim.org/
11. Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D., Evangelista, J. E., Jenkins, S. L., Lachmann, A., et al. Gene Set Knowledge Discovery with Enrichr. *Current protocols.* **2021**, 1(3), e90. https://doi.org/10.1002/cpz1.90

12. Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; et al. The Genotype-Tissue Expression (GTEx) Project. *Nat. Genet.* **2013**, 45 (6), 580–585. https://doi.org/10.1038/ng.2653.
13. Blueprint Genetics' approach to pseudogenes and other duplicated genomic regions. https://blueprintgenetics.com/pseudogene/.
14. UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Res*. **2021**, 49, D1.
15. Eilbeck, K.; Lewis, S. E.; Mungall, C. J.; Yandell, M.; Stein, L.; Durbin, R.; Ashburner, M. The Sequence Ontology: A Tool for the Unification of Genome Annotations. *Genome Biol.* **2005**, *6* (5). https://doi.org/10.1186/gb-2005-6-5-r44.
16. Hochberg, Y., & Benjamini, Y. More powerful procedures for multiple significance testing. *Statistics in medicine*. **1990**, 9(7), 811–818. https://doi.org/10.1002/sim.4780090710.

Figure Legends

Graphical abstract: Graphical representation of the analysis workflow

Figure 1: Samples used for this study. a) Data from a total of 88 whole genomes and 2028 whole exomes from Qatar and the Greater Middle East (GME) study were aggregated in this study. b) Ancestry distribution of samples from Qatar dataset. c) Ancestry distribution of samples from GME dataset. NWA: Northwest Africa, NEA: Northeast Africa, TP: Turkish Peninsula, SD: Syrian Desert, AP: Arabian Peninsula, PP: Persia and Pakistan.

Figure 2: Characterization of common Middle East disease variants (CMEDVs). a) Percentage of CMDEVs in the MEV data set with MAF >1%. MEV MAF <1% represents variants DM, or P/LP which are also rare (<1%) in gnomAD. b) Effect of CMEDVs and total number of genes impacted by those variants. c) Distribution of CMEDVs which are reported in at different star levels in ClinVar and in HGMD. d) Gene enrichment analysis using the P/LP ClinVar variant set with ≥1 star and no conflicting interpretations.

Figure 3: Characterization of high confidence homozygous LoFs or knockouts (KOs) in the mEV database. a) Distribution of unique (present in MEV database only) and reported (present in both MEV database and gnomAD) knockouts. b) Effects of unique KOs variants. c) Distribution of unique KOs genes in different disease database (ClinVar, HGMD and OMIM).

Supplementary Figure 1: Distribution of Unique and Reported MEVs based on MAF. Compared to the reported coding variants, a larger proportion of the unique variants were singletons or had allele frequency of <0.1%.

Unique coding variants = Variants not reported in gnomAD 2.1.1 Exomes

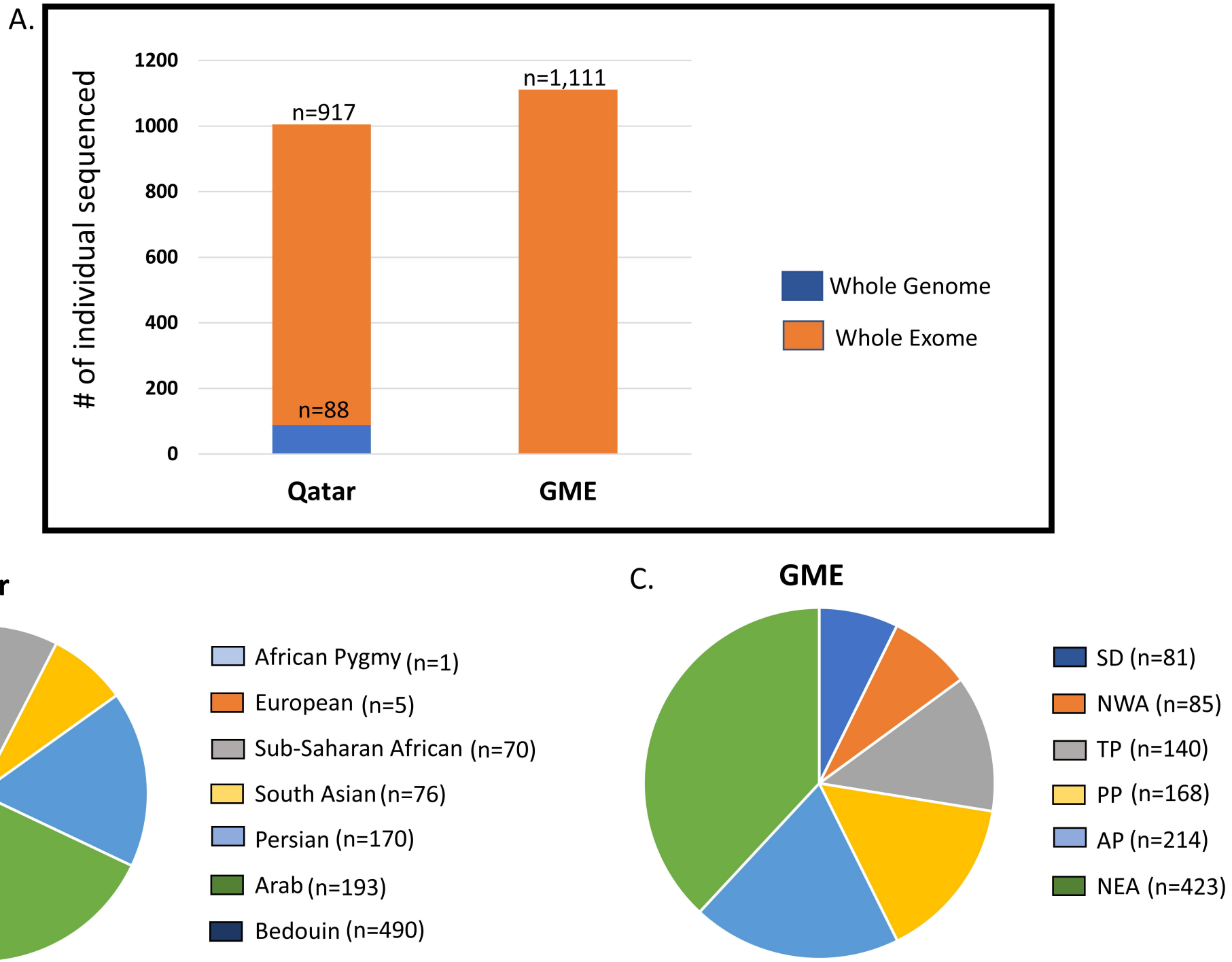
Reported coding variants = Variants reported at least once gnomAD 2.1.1 Exomes

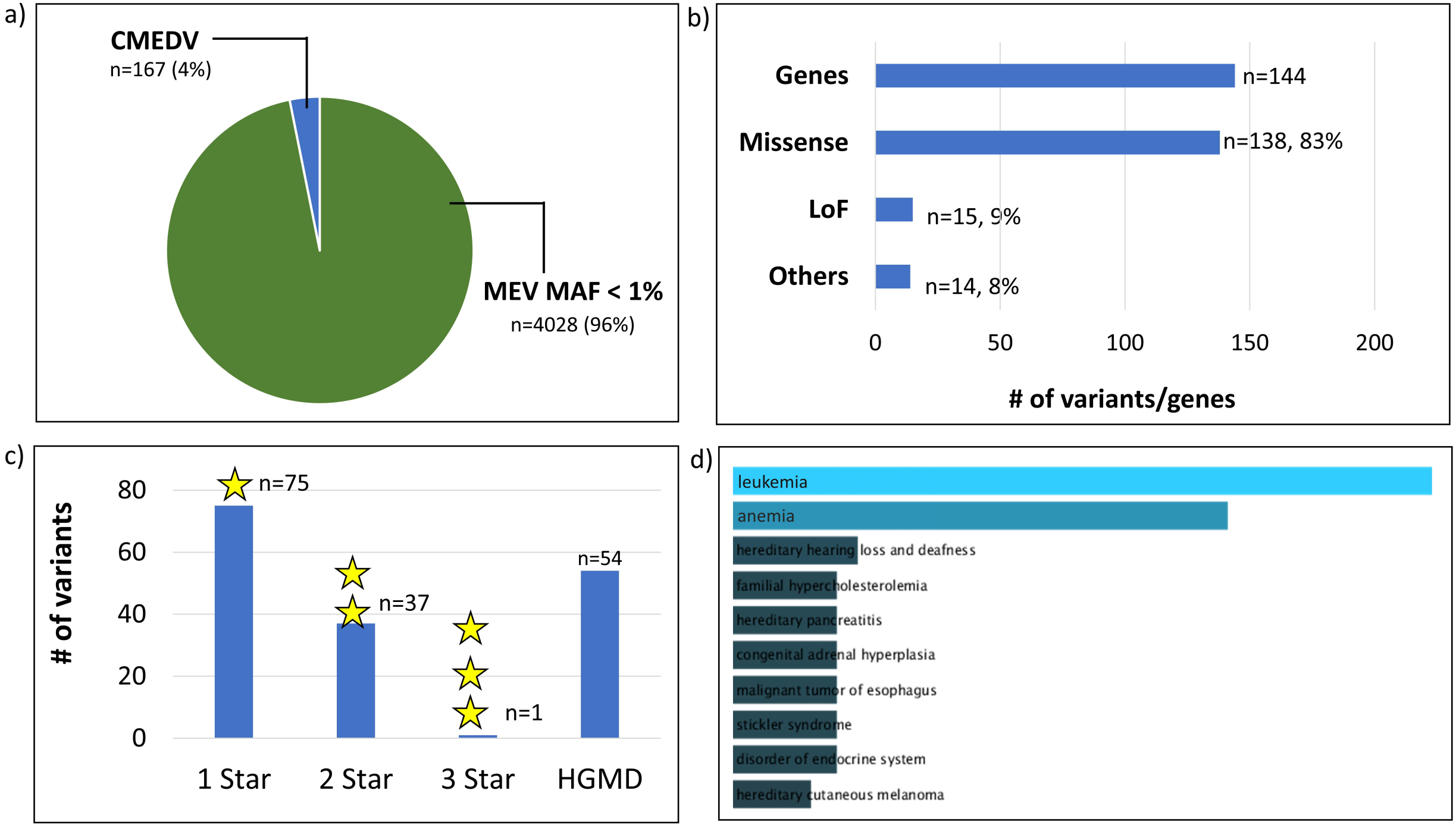
Singletons = single alleles observed in only one individual in the MEV dataset.

Supplementary Table 1: Functional annotation of the common Middle East Disease variants (CMEDVs).

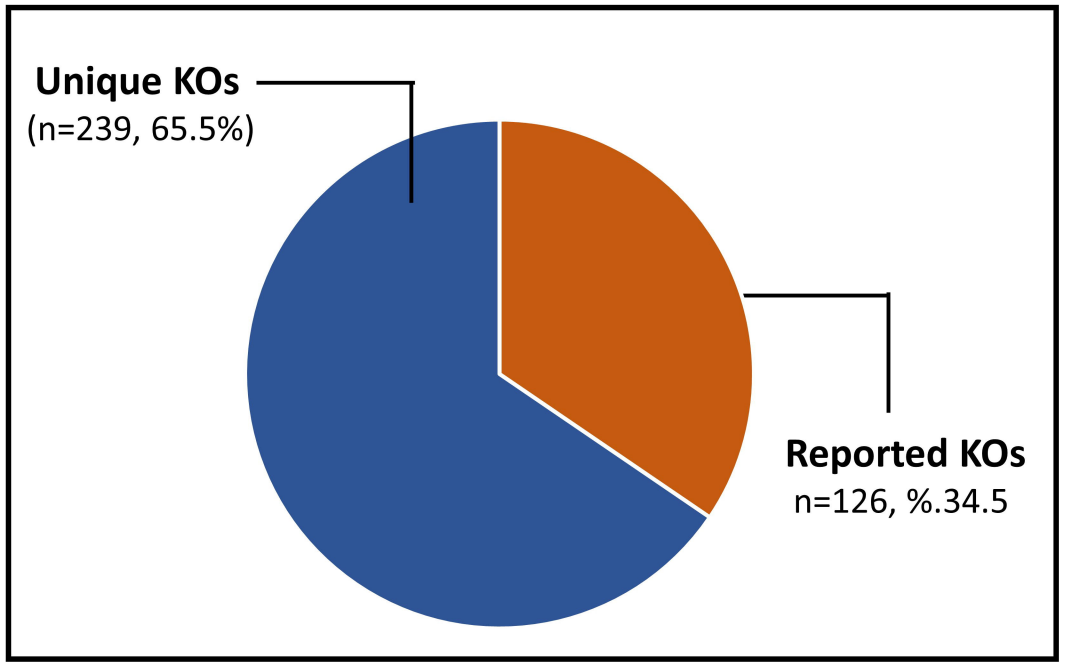
Supplementary Table 2: Gene enrichment analysis for CMEDV variants

Supplementary Table 3: Unique high confidence homozygous loss of function variants.

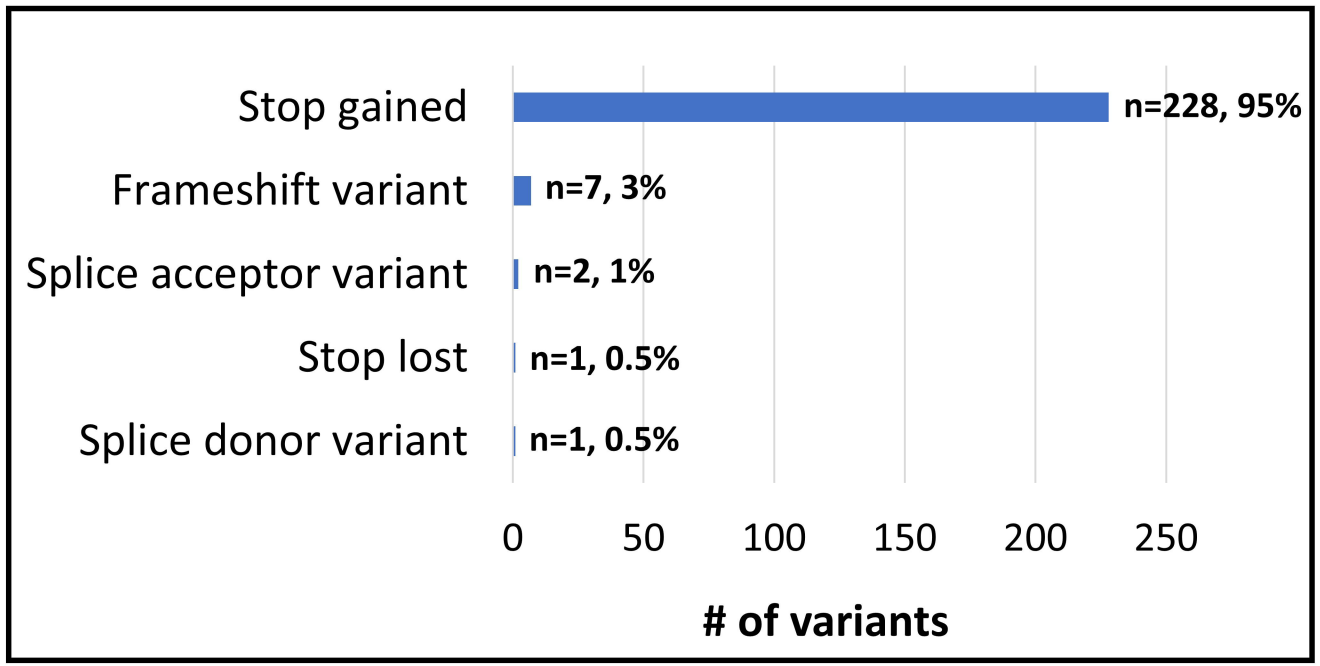

A.
# of individual sequenced
1200
1000
800
600
400
200
0
n=917
n=88
n=1,111
Qatar
GME
Whole Genome
Whole Exome
B.
Qatar
African Pygmy (n=1)
European (n=5)
Sub-Saharan African (n=70)
South Asian (n=76)
Persian (n=170)
Arab (n=193)
Bedouin (n=490)
C.
GME
SD (n=81)
NWA (n=85)
TP (n=140)
PP (n=168)
AP (n=214)
NEA (n=423)


a)
CMEDV
n=167 (4%)
MEV MAF < 1%
n=4028 (96%)
b)
Genes
n=144
Missense
n=138, 83%
LoF
n=15, 9%
Others
n=14, 8%
0
50
100
150
200
# of variants/genes
c)
# of variants
80
60
40
20
0
n=75
n=37
n=1
n=54
1 Star
2 Star
3 Star
HGMD
d)
leukemia
anemia
hereditary hearing loss and deafness
familial hypercholesterolemia
hereditary pancreatitis
congenital adrenal hyperplasia
malignant tumor of esophagus
stickler syndrome
disorder of endocrine system
hereditary cutaneous melanoma

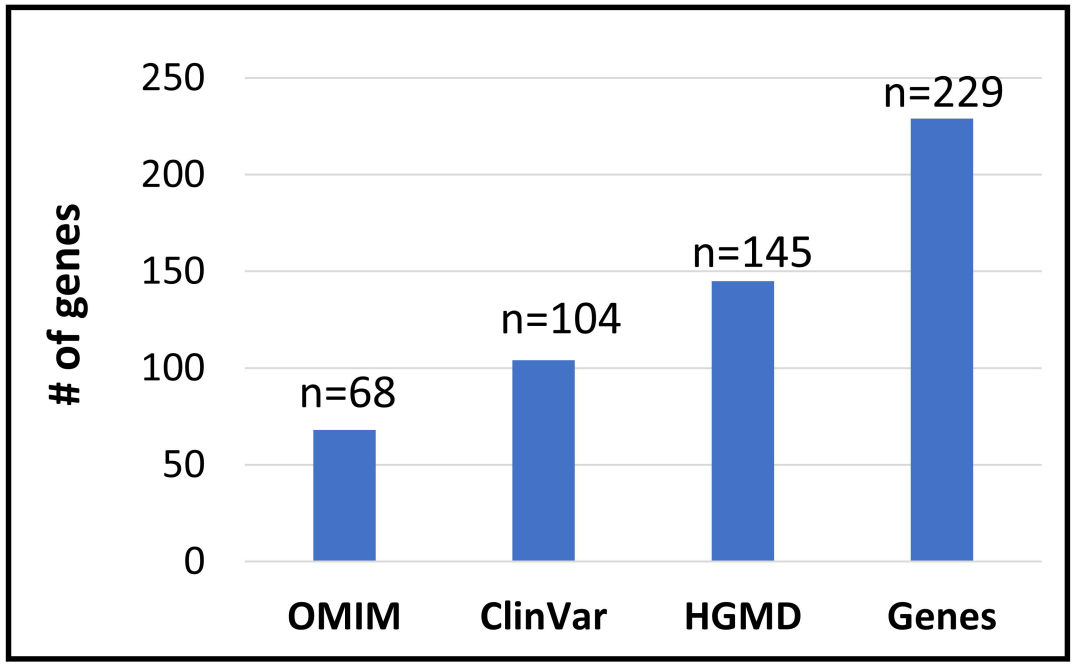
a)


Unique KOs
(n=239, 65.5%)
Reported KOs
n=126, %.34.5

b)


Stop gained
n=228, 95%
Frameshift variant
n=7, 3%
Splice acceptor variant
n=2, 1%
Stop lost
n=1, 0.5%
Splice donor variant
n=1, 0.5%
0
50
100
150
200
250
# of variants

c)


# of genes
250
200
150
100
50
0
n=68
n=104
n=145
n=229
OMIM
ClinVar
HGMD
Genes