Current Trends in Biotechnology and Pharmacy
Vol. 17(3) 1041 - 1051, JUly 2023, ISSN 0973-8916 (Print), 2230-7303 (Online)
DOI: 10.5530/ctbp.2023.3.42

Cardiomyopathy-Associated Genetic Variants in the South Asian Population: Insights from gnomAD Database Analysis

Saroja M K[1,2], Sudha Rao[2], Gothandam Kodiveri Muthukaliannan*[1]

[1]School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore –632014, Tamil Nadu, India
[2]Genotypic Technology Pvt Ltd., Bangalore – 560094, Karnataka, India
Corresponding author:gothandam@gmail.com

Abstract

Cardiomyopathies are usually inherited heart muscle disorders which mostly cause sudden death in young and adults. Early screening and diagnosis in families with risk will enable taking measures to prevent the sudden cardiac death. Analysis of the genome Aggregation Database (gnomAD), consisting of both whole genome and whole Exome data from unrelated individuals for variants in 58 genes implicated in primary cardiomyopathies revealed several insights. The gnomAD consisted of two hundred and Eighty-three pathogenic variants reported in ClinVar. Twenty-two of the variants were present in South Asian population and ten of them were exclusive to this population. Majority of pathogenic variants were observed in hypertrophic cardiomyopathy related genes which is the highly prevalent phenotype in India. Unannotated variants form ClinVar were predicted for their pathogenicity using multiple predictive tools specific for the variant types. Fewer predicted pathogenic variants in key genes like MYBPC3, PKP2 and LMNA suggest, mutations in these genes might be intolerant and hence less represented in the gnomAD database. Analysis of data resulted in identification of inflated numbers of pathogenic alleles for each of the phenotypes. Analysis establishes the value of population database in assessing the pathogenic alleles. However higher number of pathogenic variants than the phenotype prevalence in the population suggests that some of these variants might be less penetrant or might have disease modifying effect in presence of other pathogenic variants rather than being causative of the phenotype. Availability of clinical phenotypes for these variants might also result in better interpretation and conclusion on variant's pathogenicity.

Keywords: Primary Cardiomyopathy, Gene, Allele frequency, Variants, Exome, Next generation sequencing

Introduction

Cardiomyopathiesare a group of condition characterized by morphological abnormalities of the heart affecting the cardiac function. Based on the pathophysiological features, primary cardiomyopathies are of four major types hypertrophic (HCM), dilated (DCM), restrictive (RC) and arrhythmogenic right ventricular (ARVD) (1). Clinical manifestation of cardiomyopathies can be evidenced as a slow decline in the cardiac function, which is expressed as fatigue, breathlessness, exercise intolerance, chest pain finally leading to heart failure. Diagnosis is done by non-invasive methods like echocardiography for information on chamber dimension and function and cardiac Magnetic Resonance Imaging for analysis of myocardial tissue composition (2, 3). As clinical symptoms range from asymptomatic to sudden cardiac death, diagnostics based on symptoms is of very little value in the prognosis of the disease. 30-50% of cardiomyopathies are inherited and even in case of sporadic cases there is genetic factor involved. Genes with distinct functions have been identified to be responsible for specific cardiomyopathy phenotypes. HCM is majorly due to variants in sarcomere genes, ARVD is due to desmosome genes. They are highly heterogeneous group of disorders. Presence of multiple variants in one or more gene causing severe phenotype (4) to variants in the same gene responsible for diverse phenotypes, phenotype shows a wide spectrum.

Genome wide association (GWA) studies have successfully identified disease associated loci susceptible for complex diseases in Caucasian and European populations which were extensively sequenced. Genetic and demographic histories between different world populations vary greatly leading to differences in allele frequencies/presence of novel disease alleles and also variable penetrance of alleles (5). Due to these differences certain alleles might predispose a population specifically to disease susceptibility or resistance. Knowledge and screening for such variants has a high prognostic value. Large scale genome and exome sequencing efforts worldwide like 1000 genome project, Exome Aggregation Consortium (ExAC) have catalogued the variants in different populations and these databases have enabled researchers to examine allele frequencies of disease susceptibility gene/loci across different populations. Presence of large number of rare variants that are specific to ethnic populations (Allele frequencies <0.5%) serve as a source for discovery of new variants specific to phenotype (6). Interpretation of the pathogenicity of the rare variants is challenging due to unavailability of sufficient functional evidence on deleterious effect on protein structure and function. As per ACMG guidelines, analysis of allele frequency from population data is the first criterion to evaluate the pathogenicity of rare variant in pathological labs (7) Variants that are rare and deleterious in highly sequenced Caucasian population might be of common allele in populations that have been underrepresented in sequencing studies.

India which has served as a major corridor for the human migration is inhabited by 20% of the global population. Health burden due to CVDs are very high in this region leading to significant loss of productive years. Death and disability rates in India due to cardiomyopathies are higher than the global burden of the disease estimates (8). High disease burden of CVDs is partly attributed to high prevalence of endogamy. Isolated studies conducted on few Indian samples have identified novel variants in the Indian population (9) on MYH7 and MYBPC3; (10) on MYBPC3. Screening by NGS method will assist in early diagnosis before the clinical symptoms. The current study aims at identifying the mutationsby mining alleles in genes implicated in cardiomyopathy that might predispose the Indian population to high disease burden of cardiomyopathy. To achieve this, we have made use of publicly available data from Genome Aggregation Database (gnomAD). Genome data of 123,136 individuals sequenced for whole-exome sequencing and 15,496 individuals sequenced for whole-genome sequencing was made available by gnomAD resource. The database includes 15391 exomes sequenced from South Asian population.

Materials and Methods

Preparation of Gene list

Genes reported to be associated with primary cardiomyopathies were obtained from databases Online Mendelian Inheritance in Man (OMIM) (11), ClinVar (12) and atlas of cardiac genetic variation (13). List from disease databases OMIM and ClinVar were searched and filtered to obtain the subset of genes implicated in primary cardiomyopathies- Hypertrophic, Dilated, Arrhythmogenic right ventricular, Restrictive and Left ventricular non-compaction. Post manual curation of the list from literature survey a list of fifty-eight genes was finalized for study. Owing to the pleotropic nature of cardiomyopathies, there were few genes implicated in more than one cardiomyopathy phenotype.

Table 1: List of genes implicated in primary cardiomyopathy phenotypes studied from gnom AD data base

Genes implictaed in Arrythmogenic right ventricular cardiomyopathy								
CTNNA3	DSC2	DSG2	DSP	JUP	PKP2	RYR2	TGFB3	TMEM43
Genes implictaed in Hypertrophic cardiomyopathy								
ACTC1	ACTN2	CALR3	CAV3	CSRP3	DTNA	GLA	ILK	JPH2
LAMP2	MIB1	MYBPC3	MYH6	MYH7	MYL2	MYL3	MYLK2	MYOZ2
PLN|CE-P85L	PRDM16	PRKAG2	TNNC1	TNNI3	TNNT2	TPM1	TRIM63	TTR
Genes implictaed in Dilated cardiomyopathy								
ABCC9	ANKRD1	BAG3	CRYAB	CTF1	DES	EYA4	FHL2	GATAD1
LAMA4	LDB3	LMNA	MYPN	NEXN	PDLIM3	RAF1	RBM20	TCAP
TMPO	TTN	VCL	TXNRD2					

Dataset

To estimate the mutational load of cardiomyopathies, dataset used in this study is from gnomAD database (gnomAD v2.1.1) (14) consisting of 15708 whole genome and 125748 Exome data from unrelated individuals sequenced as part of disease cohorts or population studies from different parts of the world. This is one of the most comprehensive dataset representing different world populations.

Analysis

A computational analysis pipeline was followed to predict the deleterious effects of variants based on functional annotations as well as annotations from different databases. For genes that were included in the analysis, chromosomal coordinates were obtained from UCSC genome browser. Complete genomic regions including the intronic regions were analysed. Along with the allele frequencies for seven major populations (American, African, European, Finland, East Asian, South Asian and Others). Functional effect of variants was assessed with several tools like REVEL (15), CADD (16, 17), Eigen (18), PolyPhen-2 (19), and Sorting Intolerant from Tolerant (SIFT) (20). These tools predict the effect of variation on protein structure and function by taking into consideration several parameters that include conservation of the base in homologous sequences, conservative and non-conservative change in the amino acid position, effect of modified amino acid on structure and function of the protein. In order to complement the limitation of these predictor models, we used multiple prediction tools to complement and identify variants with higher confidence. Few specific tools were used to predict effect of variant in splice region, Splice AI (21) as well as non-coding regions (Eigen).

Variant classification

Minor allele frequency cut off of 0.001 was considered based on recommendation for dominant allele detection in case of Mendelian disorders. This allele frequency cut off allows inclusion of estimated prevalence of Hypertrophic cardiomyopathy which is the most prevalent among the primary cardiomyopathies under this study. However, for other cardiomyopathies where the disease prevalence is lower that is

1 in 1000 to 1 in 2500, variants with low penetrance might be possibly selected. Based on the allele frequencies (MAF < 0.001) and annotation from different tools, variants were classified as Pathogenic and uncertain significance.Tools used use different statistical models to predict the effect of variants. Only Variants that had high conservational score, predicted as damaging or disease causing by all the tools for specific variant consequence (Missense or Splice variant) are classified as Pathogenic to ensure increased confidence.

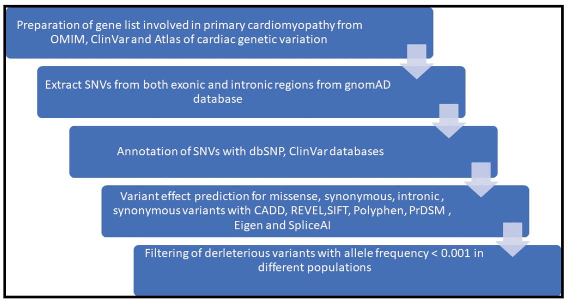
Preparation of gene list involved in primary cardiomyopathy from OMIM, ClinVar and Atlas of cardiac genetic variation
Extract SNVs from both exonic and intronic regions from gnomAD database
Annotation of SNVs with dbSNP, ClinVar databases
Variant effect prediction for missense, synonymous, intronic , synonymous variants with CADD, REVEL,SIFT, Polyphen, PrDSM , Eigen and SpliceAI
Filtering of derleterious variants with allele frequency < 0.001 in different populations

Fig. 1: Workflow followed in the analysis of variants extracted from gnomAD database for genes implicated in primary cardiomyopathy

Results and Discussion

Cardiomyopathy Variants distribution in gnomAD database:

Querying of gnomAD database with fifty-eight genes implicated in primary cardiomyopathies resulted in 76662 variants that included exonic, intronic, exon flanking regions and untranslated regions in these genes, 593 variants were stop gained variants, 748 were frame shift and 25614 were missense variants. Distribution of Single nucleotide variants has been summarized in table 2.

Table 2: Distribution of single nucleotide variants in 58 primary cardiomyopathy genes

Data set	Variants in gnomAD
Cardiomyopathy genes assessed in the study	58
Exonic+Intronic variants	76662
Exonic	42446
Upstream gene variant	574
Downstream Gene variant	2273
5' UTR	562
3' UTR	982
Mis-sense variants	26470
Loss of function variants (Frameshift, splice site, Stop gained or lost)	2415
ClinVar annotated	27551
ClinVar Pathogenic/Likely Pathogenic	283

Variant analysis based on ClinVar and custom annotations:

gnomAD database consisted of 27551 variants annotated in ClinVar. Query for Likely/pathogenic variants found in gnomAD database (Ascertained for pathogenicity) identified 283variants. MYBPC3 (48 variants), MYH7 (37) and PKP2 (38) genes carried highest number of pathogenic variants. PKP2 gene had maximum number of stop gain (16/51) variants.

Analysis of variants annotated as pathogenic and present at an allele frequency of >0.001 resulted in only 1 variant each in East Asian and African populations(EAS-rs77856833 and AFR rs76992529). There were no alleles at higher allele frequencies found in SAS population. It was noted while pathogenic variants were shared between the populations few variants were exclusively present in each of these population. 22 pathogenic variants were found only in SAS population.Ten of these variants have been associated with HCM phenotype which is a most frequently seen phenotype in Indian population. Nine pathogenic variants were associated with ARVD phenotype.

Probing of individual population allele frequency information resulted in identification of comparable number of each class of variants across populations. Data is presented in Fig 2:

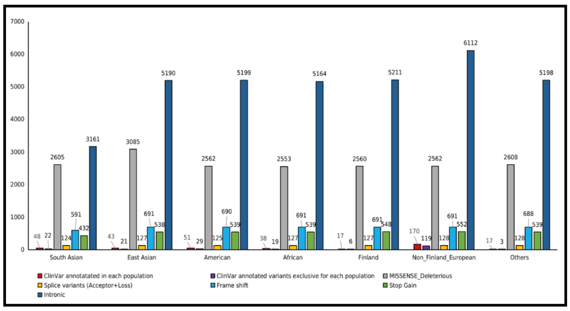

Fig 2: Variant allele distribution across major world populations predicted as damaging effect. Allele frequency cutoff of 0.001 was considered. Custom annotation with multiple tools like SIFT, POlyPhen, REVEL, CADD, Eigen, Phylop were applied to classify variants as damaging.

Few genes like MYBPC3 on chromosome 11, MYH7 on chromosome 14, PKP2 on chromosome 12 and LAMA4, DES, LMNA on chromosome 1 were found to be enriched with clinically significant pathogenic variants across all seven populations. A Variants with allele frequencies > 0.001 were unique in each population. Only a single variant in MYH6 gene with dbSNP ID rs267606904 found to be shared between East Asian and Finland population at higher frequencies. (0.00130477 and 0.00161678 respectively).

Variant allele distribution in SAS population:

Extraction of variants present in south Asian population in fifty-eight primary cardiomyopathy genes yielded 22514 variants. We found 48pathogenic/likely pathogenic alleles (as annotated in ClinVar) present in South Asian populations. South Asian population shares 17% of the variants with other populations and 8% of the variants are exclusive. Investigating the variants with allele frequencies >0.001 in south Asian population identified 101 exclusive variants (1671 shared across populations).

Variants with allele frequency less than 0.001 resulted in 9840 records. As these are too large number of low frequency variants, analysis was further performed on subsets of genes specific to phenotypes for better interpretation.

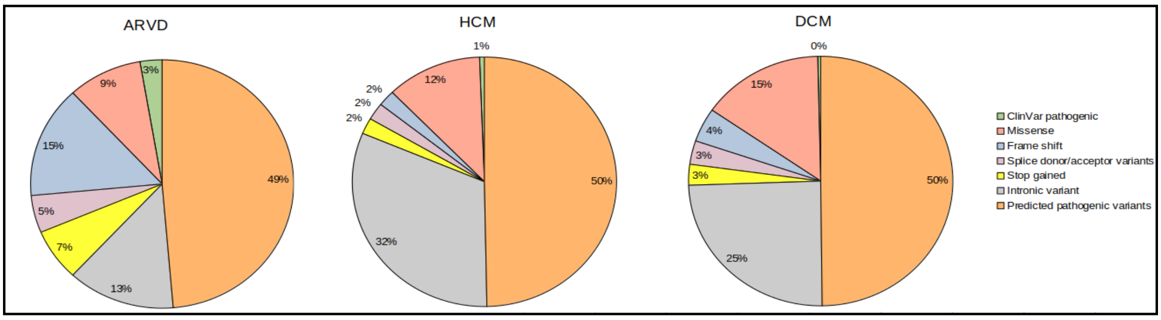

Fig. 3: Variants across HCM, DCM and ARVD phenotypes exclusively present in SAS population

ARVD associated allele distribution in SAS population:

gnomAD database was queried for variants in key ARVD genes (PKP2, DSP, DSG2, JUP, DSC2). along with the genes RYR2, TMEM43, CTNNA3 and TGFB3 which are reported to carry causative pathogenic variant in <1% of cases. Of the 15 pathogenic/likely pathogenic variants as annotated by ClinVar, 10 variants are exclusively found in south Asian population. Variants are present in key ARVD genes PKP2, DSP, DSG2, DSC2.

Our query for Variants exclusive to SAS population, with allele frequency >0.001 did not return any variants. Query for < 0.001 and non-pathogenic by clinvar resulted in 745 missense variants (411 in 5 key genes). As per multiple prediction tools 34 of these variants as damaging. Two of the variants in CTNNA3 were eliminated as they were present as homozygous alleles. Variant in CTNNA3 (c.1007 C>A) (p.R336L) was found be present in 27 individuals and in one individual as homozygous allele. It would be most probably benign allele in the population.A variant in PKP2 (c. 713 G>A) (p.P238L) and in TMEM43 (C.1085 T>G) (P. L362R) were also found in large number of individuals. Post filtering of high frequency variants 30 variants were retained that were carried by 36 individuals. Both PKP2 and JUP identified seven missense variants that are predicted as damaging. As JUP can contribute to disease phenotype only in presence of digenic heterozygous status and there is no way we can assess that status from gnomAD data, variants in JUP cannot be assessed further. Analysis also identified 18 splice site variants that are predicted as pathogenic. Based on the disease frequency of ARVD (1 in 2500) in a population we would expect to see less than 10 pathogenic variants in 15391 individuals. This high number of predicted pathogenic variants suggests that some of them are common alleles. Data is presented in supplementary document ARVD variants.

Table 3: Annotated as Pathogenic Variants in ClinVar for genes implicated in ARVD phenotype exclusive to South Asian population

Gene name	CHR	POS	REF	ALT	OMIM_ INTERITANCE	Consequence	ENSP	HGVSc	HGVSp
PKP2	12	32955433	G	A	AD	stop_gained	NP_004563.2	NM_004572.3 :c.2203C>T	NP_004563.2:p. Arg735Ter
PKP2	12	33021869	G	A	AD	missense_ variant	NP_004563.2	NM_004572.3 :c.1162C>T	NP_004563.2:p. Arg388Trp
PKP2	12	33030779	C	T	AD	splice_donor_variant	NP_004563.2	NM_004572.3 :c.1034+1G>A	--
PKP2	12	33030892	C	CA	AD	frameshift_ variant	NP_004563.2	NM_004572.3 :c.921dupT	NP_004563.2:p. Val308CysfsTer28
DSG2	18	29099774	CA	C	AD, Het	frameshift_ variant	NP_001934.2	NM_001943.3 :c.91delA	NP_001934.2:p. Thr31GlnfsTer14
DSG2	18	29101207	G	T	AD, Het	splice_donor_variant	NP_001934.2	NM_001943.3 :c.523+1G>T	--
DSG2	18	29118820	T	TA	AD, Het	frameshift_ variant	NP_001934.2	NM_001943.3 :c.1759dupA	NP_001934.2:p. Thr587AsnfsTer6
DSP	6	7584981	CTG	C	AD, AR	frameshift_ variant	NP_004406.2	NM_004415.2: c.7491_7492delTG	NP_004406.2:p. Cys2497Ter
DSC2	18	28651612	C	T	AD, AR	stop_gained	NP_077740.1	NM_024422.3: c.2084G>A	NP_077740.1:p. Trp695Ter
DSC2	18	28667658	A	G	AD, AR	missense_ variant	NP_077740.1	NM_024422.3: c.749T>C	NP_077740.1:p. Phe250Ser

HCM associated allele distribution in SAS population:

gnoMAD database returned Twenty-six ClinVar annotated pathogenic variants in ten HCM genes (Twenty-seven genes probed). All twenty-six variants are present in South Asian population and is present in total of thirty-five individuals. MYBPC3 and MYH7 that are implicated as major HCM genes carried highest number of pathogenic variants (eight in each gene). Eight of the variants are exclusive for South Asian population. Variant in *MYBPC3*, a 25bp deletion in intron 32 (MYBPC3 ΔInt32) was at highest allele frequency (0.032067) and nineteen individuals carried this allele in homozygous status.

Filtering of the entire HCM gene variants with allele frequencies >0.001 and exclusively present in South Asian population resulted in 33 variants. Two variants, MYH6 (C.1243 C>T) p(G415R) and MYH7 (C.4399 G > C) p(L1467V) are predicted as damaging by multiple prediction tools. Analysis of the variants <0.001 returned 2779 variants. Search for missense alleles with less than 0.001 frequency, exclusively present in south Asian population and predicted as deleterious in multiple predictive tools resulted in 158 variants. These variants were carried by 272 individuals. MYH6 and MYH7 genes were enriched for deleterious variants. Twenty-eight stop gain variants are carried by 35 individuals. Analysis also identified 30 splice variants predicted as highly pathogenic. Filtering of alleles with frequency <0.0001 did not reduce the number of deleterious variants drastically leading to a conclusion that all these variants might not be pathogenic and might have lower penetrance. Data is provided in supplementary information (Supplementary sheet: HCM variants)

Table 4: Annotated as Pathogenic Variants in ClinVar for genes implicated in HCM phenotype exclusive to South Asian population

Gene name	CHR	POS	REF	ALT	Consequence	ENSP	HGVSc	HGVSp
MYH6	14	23866396	T	C	missense_ variant	NP_002462.2	NM_002471.3 :c.2033A>G	NP_002462.2: p.Asn678Ser
MYH7	14	23894990	G	C	missense_ variant	NP_000248.2	NM_000257.2: c.2200C>G	NP_000248.2: p.Gln734Glu
MYH7	14	23896032	A	T	missense_ variant	NP_000248.2	NM_000257.2: c.1998T>A	NP_000248.2: p.His666Gln
MYBPC3	11	47364296	C	G	splice_acceptor_ variant	NP_000247.2	NM_000256.3: c.1458-1G>C	--
MYBPC3	11	47364479	AGG	A	frameshift_ variant	NP_000247.2	NM_000256.3: c.1357_1358delCC	NP_000247.2: p.Pro453CysfsTer21
MYL2	12	111353525	C	G	missense_ variant	NP_000423.2	NM_000432.3: c.163G>C	NP_000423.2: p.Ala55Pro
TNNI3	19	55665514	G	A	upstream_ gene_variant	NP_003274.3	0	--
CEP85L,PLN	6	118880187	TTC	T	intron_variant	NP_001171506.1	NM_001178035.1: c.1029+6503_ 1029+6504delGA	--

DCM associated allele distribution in SAS population:

Analysis of 22 genes implicated in DCM returned 50 variants annotated as pathogenic in ClinVar database. Of these seven variants were present in south Asian population and four were exclusive. Of the 38098 variants from all populations, 4977 variants were exclusively present in the SAS population. Probing for variants with allele frequencies >0.001 identified 48 variants which are all benign variants. Filtering for allele frequencies <0.001 and missense variants predicted as damaging by multiple prediction tools returned 106 variants carried by 192 individuals. 1 variant in gene DES(c.944 G>T) (p.R315L) was found be homozygous allele hence might not be pathogenic. LAMA4 and DES genes had highest number of variants (13) predicted as damaging. Two variants in TTN gene(Chr2:179505358 C>T and Chr2:179629539 C> A) and LDB3 (Chr10:88477720G>A) were assessed as highly pathogenic splice acceptor gain variants. One donor gain variant in TTN gene (Chr2: 179637835 C>T) was also assessed as pathogenic. Analysis also identified acceptor loss and donor loss splice variants that have been listed in supplementary data DCM variants. Thirty-one stop codon variants identified were predicted as pathogenic by both Eigen and CADD tools.

LAMA4 and DES genes carried highest number of pathogenic variants associated with DCM phenotype. Of the unannotated variants one hundred and sixty-seven missense variants were predicted as highly damaging. Detailed tables are provided in supplementary information (sheet: DCM variants)

Table 5: Annotated as Pathogenic Variants in ClinVar for genes implicated in DCM phenotype exclusive to South Asian population

Gene name	CHR	POS	REF	ALT	Consequence	ENSP	HGVSc	HGVSp
EYA4	6	133703563	C	T	stop_ gained	NP_004091.3	NM_004100.4: c.67C>T	NP_004091.3:p.Gln23Ter
TTN	2	179596570	G	A	stop_ gained	NP_001254479.1	NM_001267550.1: c.17032C>T	NP_001254479.1:p. Arg5678Ter
TTN	2	179622344	C	T	missense_ variant	NP_001254479.1	NM_001267550.1: c.10603G>A	NP_001254479.1:p. Gly3535Arg
TCAP	17	37821623	CAG	C	downstream_ gene_ variant	NP_006795.3	0	--

Advancements in next generation sequencing technologies have enabled large scale exome and genome sequencing that in turn has identified large number of novel variants. Clinical interpretation of these variants is still a challenge. Information on variant allele distribution in different ethnic groups and cohorts are of great help to assess the causal effect of a variant.gnomAD contains largest number of dataset from different world populations. We used this data set to study cardiomyopathy allele distribution in different populations and also specifically in South Asian population. Few studies earlier have analysed the gnomAD database to understand the disease allele distribution across populations (23,24).We analysed this dataset to get an insight on reported and novel disease allele distribution in gnomAD database specific to South Asian population that has not been addressed earlier.

Analysis of the dataset and comparison of pathogenic allele distribution as annotated by ClinVar database across major populations revealed shared between as well as unique variants to each population. There is no pathogenic variant shared by all the populations under the study. Few variants are shared between 2 or more population. Every population in the study showed unique pathogenic variants(Soth Asian -22, American-29, East Asian 21, African-19, Finland-11, Non Finland European-119 and in others-3 variants). This further established the importance of analysis of disease allele distribution specific to ethnic populations from large databases.

Addition of custom annotation on pathogenicity for unannotated variants by type specific computational tools enabled better interpretation of the variants. Stringency in selection reduced the number of predicted pathogenic variants for each class of variant. Few specific variants with high allele frequencies (>0.001) might also have modifying effect rather than contributing primarily to the disease phenotype. Considering the allele frequencies of <0.001 allowed inclusion of such variants. For each of the cardiomyopathy phenotype enrichment of pathogenic alleles were observed in the implicated genes. In case of ARVD, JUP had equivalent pathogenic missense alleles as compared to PKP2, the key gene. PKP2 was not enriched for pathogenic alleles from gnomAD database. Similar observations were recorded in HCM gene MYBPC3, DCM gene LMNA. This finding could suggest intolerance for the deleterious variants in primary or key genes compared to other genes.

Disease allele frequency distribution calculation estimated it as 1 individual in every 402 as carrying pathogenic allele as annotated by ClinVar. (with a MAF of <0.001). This frequency is very closer to prevalence of HCM phenotype (1 in 500) most frequent phenotype in the population. However, assessing the frequencies for ARVD and DCM independently resulted in higher estimations(1 in 1177 for ARVC and 1 in 1391 for DCM). This suggests slightly higher representation of clinVar annotated pathogenic alleles in gnomAD database.

Analysis through computational predictive tools identified large number of variants as disease causing in-spite of stringent MAF cutoff and filtering criterion. The estimated disease prevalence was far higher than the reported numbers. This might indicate either higher representation of pathogenic alleles in the database or need to reassess the disease prevalence in the population. Findings might also suggest that may be not all variants predicted as damaging, they might be bystanders or contribute primarily to phenotype but might have a modifying effect or they may be less penetrant.

Pleiotropy of certain genes like TTN (HCM and DCM phenotypes) and PKP2 (ARVD and channelopathy) makes it challenging to analyse them under a specific phenotype and estimate disease allele frequency for that phenotype. Few time less penetrant variants might also be contributing to disease modifying effect and might contribute to unique phenotype. This finding also emphasizes on the importance of clinical phenotype correlation with the genetic variants.

gnomAD database though is a collection of human data from diverse population, an absence of allele in the database might not necessarily make the variant pathogenic. Unavailability of the phenotype makes it challenging to interpret the data. European population has huge entries whereas South Asian and East Asian populations are highly under represented.

Even with the above limitations, this study provides the evidence for using the gnomAD database by combining allele frequency, ClinVar annotation as well as computational annotations to obtain an insight on probable pathogenic variants in the South Asian population. Addition of clinical phenotype information to the predicted disease variants will help in better association of genotype-phenotype correlation and would increase the reliability of such analysis.

Statements and Declarations

Funding

The funding support for the study was from DSIR certified R&D unit of Genotypic Technology Pvt Ltd

Competing Interests

We declare that there is no competing interest

Author contributions

Saroja M K contributed to primary analysis and interpretation of the data. Dr. Sudha Rao and Dr. K M Gothandam have contributed in reviewing, correction and discussion.

Acknowledgements

We sincerely acknowledge the contribution of Guatam Singh, Dhitiomics technologies for annotation support.

References

1. Maron BJ, Towbin JA, Thiene G, Antzelevitch C, Corrado D, Arnett D, Moss AJ, Seidman C, Young JB. (2006) Contemporary definitions and classification of the cardiomyopathies. An American Heart Association scientific statement from the Council on Clinical Cardiology, Heart Failure and Transplantation Committee; Quality of Care and Outcomes Research and Functional Genomics and Translational Biology Interdisciplinary Working Groups; and Council on Epidemiology and Prevention. Circulation 113:1807–1816.

2. Mahrholdt H, Wagner A, Judd RM, Sechtem U, and Kim RJ. (2005) Delayed enhancement cardiovascular magnetic resonance assessment of non-ischaemic cardiomyopathies. Eur Heart J 26, 1461–1474.

3. Rickers C, Wilke NM, Jerosch-Herold M, Casey SA, Panse P, Panse N, Weil J, Zenovich AG, and Maron BJ (2005) Utility of cardiac magnetic resonance imaging in the diagnosis of hypertrophic cardiomyopathy. Circulation 112:855–861.

4. Bick AG, Flannick J, Ito K, et al. (2012) Burden of rare sarcomere gene variants in the Framingham and Jackson Heart Study cohorts. Am J Hum Genet91(3):513-9. doi: 10.1016/j.ajhg.2012.07.017.

5. Adeyemo A, Rotimi C. (2010) Genetic Variants Associated with Complex Human Diseases Show Wide Variation across Multiple Populations. Public Health Genomics 13:72–79, DOI: 10.1159/000218711

6. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–73.

7. Richards S, Aziz N, Bale S, et al. (2015) 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: Off J Am Coll Med Genet. 17:405–24.

8. Prabhakaran D, Jeemon P, Roy A. (2016) Cardiovascular diseases in India: Current Epidemiology and Future Directions. Circulation. 133(16):1605-20. doi: 10.1161/CIRCULATIONAHA.114.008729

9. Waldmüller S, Sakthivel S, Saadi AV, et al. (2003) Novel deletions in MYH7 and MYBPC3 identified in Indian families with familial hypertrophic cardiomyopathy. J Mol Cell Cardiol 35(6):623-36. doi: 10.1016/s0022-2828(03)00050-6.

10. Dhandapany PS, Sadayappan S, Xue Y, et al. (2009) A common MYBPC3(cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. Nat Genet 41:187–191)

11. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 33 (suppl 1):D514±D7.

12. Landrum MJ, Lee JM, Riley GR, et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research 42(D1): D980±D5.

13. Walsh R, Thomson KL, Ware JS, et al.

(2017) Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genet Med. 19(2):192-203. doi: 10.1038/gim.2016.90.

14. Karczewski KJ, Francioli LC, Tiao G. et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443. https://doi.org/10.1038/s41586-020-2308-7

15. Ioannidis NM, Rothstein J H, Pejaver V, et al. (2016) REVEL: An ensemble method for predicting the pathogenicity of rare Missense variants. Am J Hum Genet 99(4): 877-885.

16. Kircher M, D.M. Witten, P. Jain, B.J. O'Roak, G.M. Cooper, J. ShendureA general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet., 46 (2014), pp. 310-315

17. Rentzsch P, Schubach M, Shendure J, Kircher M. (2021) CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med doi: 10.1186/s13073-021-00835-9.

18. Ionita-Laza I, McCallum K, Xu B. et al. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 48, 214–220. https://doi.org/10.1038/ng.3477

19. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. (2010) A method and server for predicting damaging missense mutations. Nature Methods 7(4):248-9. https://doi.org/10. 1038/nmeth0410-248

20. Kumar P, Henikoff S, Ng PC. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4(7):1073±81. https://doi.org/10.1038/nprot.1.2009.86

21. Jaganathan K, Kyriazopoulou Panagiotopoulou S, et al. (2019) Predicting Splicing from Primary Sequence with Deep Learning. Cell 176(3):535-548.e24. doi: 10.1016/j.cell.2018.12.015

22. Hall CL, Sutanto H, Dalageorgou C, McKenna WJ, Syrris P, Futema M. (2018) Frequency of genetic variants associated with arrhythmogenic right ventricular cardiomyopathy in the genome aggregation database. Eur J Hum Genet. 2018 26(9):1312-1318. doi: 10.1038/s41431-018-0169-4.

23. Shakeel M, Irfan M, Khan IA (2018) Estimating the mutational load for cardiovascular diseases in Pakistani population. PLoSOne 13(2): e0192446. https://doi.org/10.1371/journal. pone.0192446