



OPEN

Functional mutation, splice, distribution, and divergence analysis of impactful genes associated with heart failure and other cardiovascular diseases

Ishani Mhatre^{1,6}, Habiba Abdelhalim^{1,6}, William Degroat^{1,6}, Shreya Ashok^{1,6}, Bruce T. Liang^{3,4} & Zeeshan Ahmed^{1,2,5}✉

Cardiovascular disease (CVD) is caused by a multitude of complex and largely heritable conditions. Identifying key genes and understanding their susceptibility to CVD in the human genome can assist in early diagnosis and personalized treatment of the relevant patients. Heart failure (HF) is among those CVD phenotypes that has a high rate of mortality. In this study, we investigated genes primarily associated with HF and other CVDs. Achieving the goals of this study, we built a cohort of thirty-five consented patients, and sequenced their serum-based samples. We have generated and processed whole genome sequence (WGS) data, and performed functional mutation, splice, variant distribution, and divergence analysis to understand the relationships between each mutation type and its impact. Our variant and prevalence analysis found *FLNA*, *CST3*, *LGALS3*, and *HBA1* linked to many enrichment pathways. Functional mutation analysis uncovered *ACE*, *MME*, *LGALS3*, *NR3C2*, *PIK3C2A*, *CALD1*, *TEK*, and *TRPV1* to be notable and potentially significant genes. We discovered intron, 5' Flank, 3' UTR, and 3' Flank mutations to be the most common among HF and other CVD genes. Missense mutations were less common among HF and other CVD genes but had more of a functional impact. We reported *HBA1*, *FADD*, *NPPC*, *ADRB2*, *ADBR1*, *MYH6*, and *PLN* to be consequential based on our divergence analysis.

Abbreviations

AI	Artificial intelligence
AF	Atrial fibrillation
AVD	Atheromatous vascular disease
BWA	Burrows–Wheeler aligner
CNV	Copy number variants
CAD	Coronary artery disease
CHD	Coronary heart disease
CVD	Cardiovascular disease
EHR	Electronic health records
ETL	Extraction, transfer, loading
Gal-3	Galectin-3
GATK	Genome analysis toolkit
GWAS	Genome-wide association studies
HF	Heart failure

¹Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick, NJ 08901, USA. ²Department of Genetics and Genome Sciences, UConn Health, 400 Farmington Ave, Farmington, CT, USA. ³Pat and Jim Calhoun Cardiology Center, UConn Health, 263 Farmington Ave, Farmington, CT, USA. ⁴UConn School of Medicine, University of Connecticut, 263 Farmington Ave, Farmington, CT, USA. ⁵Department of Medicine/Cardiovascular Disease and Hypertension, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson St, New Brunswick, NJ, USA. ⁶These authors contributed equally: Ishani Mhatre, Habiba Abdelhalim, William Degroat and Shreya Ashok. ✉email: zahmed@ifh.rutgers.edu

IRB	Institutional review board
JWES	Java based whole genome/exome sequence data processing pipeline
JSD	Jensen–Shannon divergence
JS-MA	Jensen–Shannon divergence-based method
MAV-clc	Management, analysis, and visualization of clinical data
ML	Machine learning
PolyPhen-2	Polymorphism phenotyping v2
Pfam domains	Functional protein domains
QC	Quality checking
SIFT	Scale-invariant feature transform
SV	Structural variants
STR	Short tandem repeats
SERCA2a	Ca ²⁺ ATPase
WGS	Whole-genome-sequencing

Cardiovascular disease (CVD) is the leading cause of death and mortality internationally, with as many as 655,000 deaths per-year^{1,2}. In 2015, there were approximately 422.7 million cases of CVD and 17.92 million deaths reported³. CVD include primary pathologies such as heart failure (HF), cardiac arrhythmias, venous thromboembolism, cerebrovascular and peripheral arterial disease, coronary heart disease (CHD), coronary artery disease (CAD), and atheromatous vascular disease (AVD)^{4,5}. The most common causes of CVD mortality include but are not limited to ischemic and nonischemic HF and stroke³. Hence, one of the focuses of life science involves investigating genetic epidemiology of CVD. Due to the complex nature, risk factors, inherent genetic makeup, and progression of CVD, personalized treatment is believed to be essential⁶. Precision medicine involves integrating clinical and multi-omics/genomics data for predictive and personalized medicine within a diverse CVD population⁷. It focuses on analyzing genetic composition of patients to identify the key biomarkers and increase understanding of the pathophysiology of CVD⁸.

CVD is a complex, partially heritable condition, encompassing a range of conditions from CHD to myocardial infarction⁹. By utilizing high-quality sequenced DNA of transcribed genes, we can be better informed of a CVD patient's inherent genetic makeup and factors that may contribute to increased susceptibility for CVD¹⁰. Whole-Genome-Sequencing (WGS) has been proven to be one of the most recommended techniques to sequence DNA and capture all genetic variations. Various WGS based studies have focused on investigating mutated genes with altered expression^{11–13}, and discovered underlying genetic etiology in CVD patients^{14,15}. State of the art studies have supported the claim that performing variant analysis will assist in understanding of the complex pathophysiology of CVD progression through the application of multiple biomarkers^{16–18}. However, we are still in the early stages of developing a comprehensive database of genetic biomarkers for CVD to assist in predictive analysis and deep phenotyping^{19–22}. Previously, we have explored and discussed diverse genomic strategies that investigate genes linked to AF, HF, and other CVDs²³. In this study, we aimed to investigate genes primarily associated with HF and other CVDs by analyzing genetic variants that correlate with CVD phenotype²⁴.

Material and methods

Achieving the goals of this study, we analyzed electronic health records (EHR) received from EPIC health system to build a cohort of thirty-five patients with CVD (Fig. 1). Our selection criteria mainly included adult and aging CVD patients with HF phenotype. In addition, we collected information centered on their age, gender, ethnicity, medical details, and demographics. We identified 21 male and 14 female individuals (60% male and 40% female population) aged between 24 and 94 years (details are attached in supplementary material S6). These patients were clinically diagnosed with CVD and CMS/HCC HF, as well as cardiomyopathy, hypertension, obesity, type 2 diabetes mellitus, asthma, high cholesterol, hernia, chronic kidney, joint pain, myalgia, dizziness and giddiness, osteopenia of multiple sites, chest pain, and osteoarthritis. We collected blood samples from these CVD patients and extracted DNA. We have utilized our in-house developed applications to support patient consenting, sample collection, data management, and EHR extraction, transfer, loading (ETL) and analysis^{25,26}. Written informed consent was obtained from all subjects. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institution and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All human samples were used in accordance with relevant guidelines and regulations, and all experimental protocols were approved by the Institutional Review Board (IRB) at UConn Health.

We performed high-throughput WGS of collected blood samples, and processed sequence data for quality checking (QC) and variant discovery (QC report is attached in supplementary material S7). We utilized our in-house built pipeline (JWES) for WGS data processing, management, visualization (Circos plots), and gene-variant discovery, annotation, prediction, and genotyping²⁷. JWES mainly utilizes the Burrows-Wheeler Aligner (BWA, version 0.7.17) for mapping sequence data against the reference human genome²⁸, and Genome Analysis Toolkit (GATK, version 3.8) for the variant discovery²⁹. We performed variant calling of the whole genome using JWES for all subjects but focused on targeted HF and other CVD genes for further analyses. Utilizing significant results of differentially regulated genes from our previous expression and enrichment analysis³⁰ that were validated through our gene-disease-variant database³¹, we generated a list of forty-one HF and twenty-three other CVD genes (Supplementary material S1). We calculated pLI scores for these genes using The Genome Aggregation Database (gnomAD) to better contextualize these mutation's effects on disease (Supplementary Tables S8, S9)³².

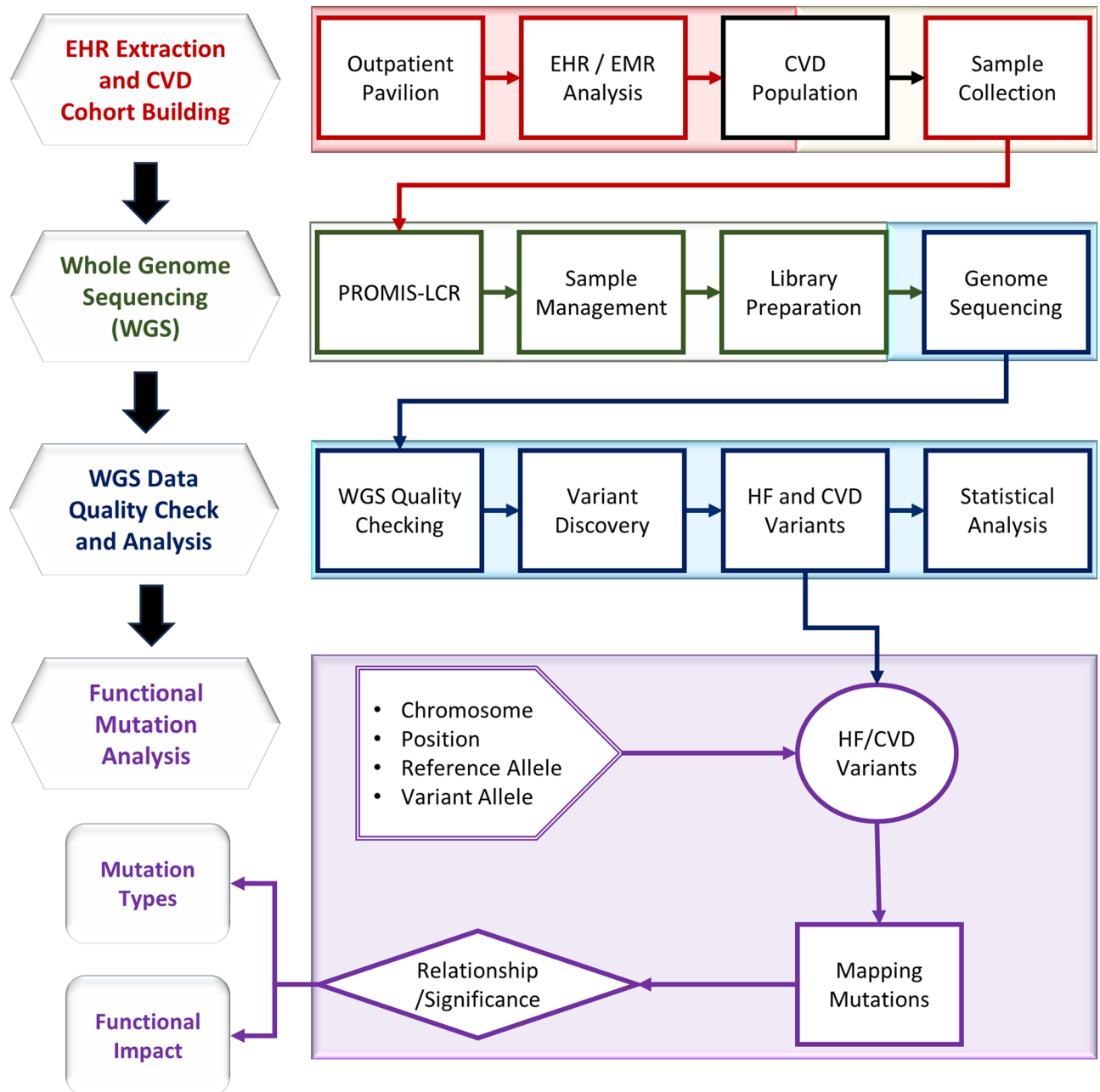


Figure 1. Study design. Overall research methodology includes four major steps: (1) EHR extraction and CVD cohort building; (2) Whole Genome Sequencing (WGS); (3) WGS data quality check and analysis; and (4) Functional mutation analysis to identify mutation types and their functional impact. A more detailed overview of our methodology is highlighted in the right panel. Different approaches grouped within the same major steps utilize identical colors.

We conducted functional mutation, splice, variant distribution, and divergence analysis to understand the relationships between each mutation type and its impact. We utilized Scale-Invariant Feature Transform (SIFT)^{33–35}, Polymorphism Phenotyping v2 (PolyPhen-2)³⁶, and MutationAssessor³⁷ to classify the biological and functional impacts of the variant data. SIFT supported in analyzing the impact of coding variants on the function of protein and identify variants that have a causal relationship to the manifestation of HF and other CVDs³⁴. PolyPhen-2 garnered a wide breadth of information about the substitution site of the coding variant and identified the specific gene sequences and structural features of the substitution site. It analyzed single-nucleotide polymorphism (SNP) substitutions and predicted the functional impact of the mutations. Then, MutationAssessor differentiated between specificity scores to account for functional shifts between subfamilies, proteins, and conserved patterns^{38,39}. Scores from SIFT, PolyPhen-2, and MutationAssessor are included in our supplementary material S1.

We performed splice mutation analysis and a Jensen-Shannon Divergence (JSD)-based Method (JS-MA) for the measurement and variant distribution analysis⁴⁰. We reported our findings on RNA, silent, 3' UTR, 3' Flank, 5' UTR, 5' Flank, intron, truncating, splice, and missense mutations for genes associated with HF and other CVDs. We analyzed RNA, truncating, missense, 3' UTR, and 5' UTR mutations to study the structural consequences of the cellular proteome. These mutations affect the functionality of the protein produced and can lead to a gain or loss of function^{41–45}. Mutations in RNA can lead to changes in the sequence of nucleotides, which can affect the structure and function of the RNA molecule and subsequently impact molecular processes⁴¹. RNA-based mutations include but are not limited to point, nonsense, silent and missense⁴¹. We observed the suppression or overexpression of a gene by investigating 3' Flank and 5' Flank mutations⁴⁵. By examining intro and splice mutations, we gained a better understanding of the effect that they can have on RNA splicing process resulting in a decrease efficiency of mRNA translation^{46–48}.

Utilizing JS-MA, we conducted a genome-wide search for complex gene-disease interactions, helping us better understand the effects that gene mutations can have on a phenotypic state⁴⁰. Divergence analysis involved comparing each gene's distribution of mutations to a weighted average of all genes in that disease type. Variance from this distribution indicates an overrepresented mutation type among HF and CVD patients. We calculated Jensen-Shannon Divergence (JSD) scores to evaluate the similarity between the two distributions. The JSD score measured the variance associated with two distributions and provided a statistical quantification on the influence of specific mutations on disease types⁴⁰. A JSD score closer to '1' indicates the highest variance denoting a unique mutation profile with greater impact. We identified notable and potentially significant genes based on whether the HF and other CVD genes met a certain threshold using their calculated JSD scores. We compared proportion distributions of unique genes and a weighted average distribution of all genes within the disease type. To ensure the validity of our results, we tried to account for confounding variables and found that the biological variables such as age of onset of HF, severity of disease, alcoholic cardiomyopathy and different aetiologies can be ruled out as they did not have any significant impact on the outcome of our study^{49–51}.

Ethical approval and consent to participate

Informed consent was obtained from all subjects. All human samples were used in accordance with relevant guidelines and regulations, and all experimental protocols were approved by the Institutional Review Board.

Results

Our variant analysis started with examining the variant distribution and prevalence of HF and CVD genes to better understand the frequency of these genetic variants. We generated Circos plots and observed a total of 229,963 variants for HF genes (Fig. 2A). For CVD genes, we visualized a total of 389,761 variants (Fig. 2B). The outer circle of the plot represents patient sample IDs, while the inner circle represents genes. Figure 2A has more HF genes along the inner circle compared to Fig. 2B which has fewer other CVD genes. Next, we conducted functional mutation analysis to evaluate the effects of disease-causing alleles for HF and other CVDs. We detected consistent distribution of mutation types for the mapped genes. These mutations included Missense, Splice, Truncating, Intron, 5' Flank, 5' UTR, 3' Flank, 3' UTR, Silent, and RNA-driven mutations for HF (Table 1) and other CVDs (Table 2). We generated lollipop plots for HF and other CVD genes to visualize the functional impact for each mutation type (Figs. 3 and 4). Currently, there are 373 datasets and a total of 162,055 mutations referenced in cBioPortal. These datasets referenced do not encompass all variants that we reported in our prevalence analysis. Due to this limitation, some genes were not annotated and visualized. These genes include *CDKN2B-AS1*, *HOTAIR*, *LSINCT5*, *RP11-451G4.2*, and *TUSC7*. Missense mutations had higher functional impacts and were more likely to be 'possibly or probably damaging.' We measured the effect of mutations using a score assigned to predict whether an amino acid substitution affects protein function. SIFT scores varied from 0.0 to 1.0. Mutations ranging from 0.0 to 0.5 were considered "deleterious" while those ranging from 0.5 to 1.0 were "tolerated/benign." Additionally, scores regarded as "deleterious low confidence" were less likely to have a phenotypic effect than "deleterious" while "tolerated low confidence" were more likely to have a phenotypic effect than "tolerated"³⁵.

Functional impact scores from PolyPhen-2 ranged from 0.0 to 1.0 with values closer to 1.0 being 'possibly or probably damaging' and those closer to 0 being 'benign'³⁶. The *AGTR1*, *AQP2*, *EDNRA*, *EPO*, *NPPC*, *PLN*, and *TNF* genes had no missense mutations and provided no further information regarding functional impact for the mutations. *ACE* had the highest number of missense mutations: twelve mutations in total. Five of those missense mutations were found to have some negative impact on the function of the protein. *NR3C2* had the highest number with a total of 2,057 intron mutations. *PIK3C2A* was the only gene with an RNA-based mutation. Aside from the RNA mutation, the rarest mutation type was truncating mutations. *AMPD1*, *KNG1*, *MYBPC3*, and *NPPA* were found to have a truncating mutation. Splice and 5' UTR were also found to be less common. Genes such as *CORIN*, *MMP2*, *MYBPC3*, *NOS3*, and *PIK3C2A* had more specific functional protein domains (Pfam domains), on average, compared to the other HF genes. From the genes investigated in our study, we found the *ACE*, *MME*, *LGALS3*, *NR3C2*, and *PIK3C2A* genes to be more significant based on various criteria such as the largest number of mutations mapped, rare mutation types, and highest number of mutations with functional impact. Previous literature has already linked or hypothesized *ACE*, *MME*, *LGALS3*, *NR3C2*, and *PIK3C2A* to be significant genes and potential biomarkers for CVDs^{52–56}. Further research must be conducted to solidify these claims and increase confidence regarding the significance of these genes. We reported different types of mutations and their impact on all HF genes in Supplementary material S2.

For other CVD genes, *CALD1*, *TEK*, *TRPVI*, *ATP2A2*, and *SMUG1* were discovered to be more significant based on the same criteria which includes genes with the highest number of mutations mapped, rare mutation types, and the largest number of mutations with functional impact. *CALD1*, *TEK*, and *TRPVI* all had the highest number of missense mutations, with eight missense mutations each. In the *CALD1* gene, the breakdown was one

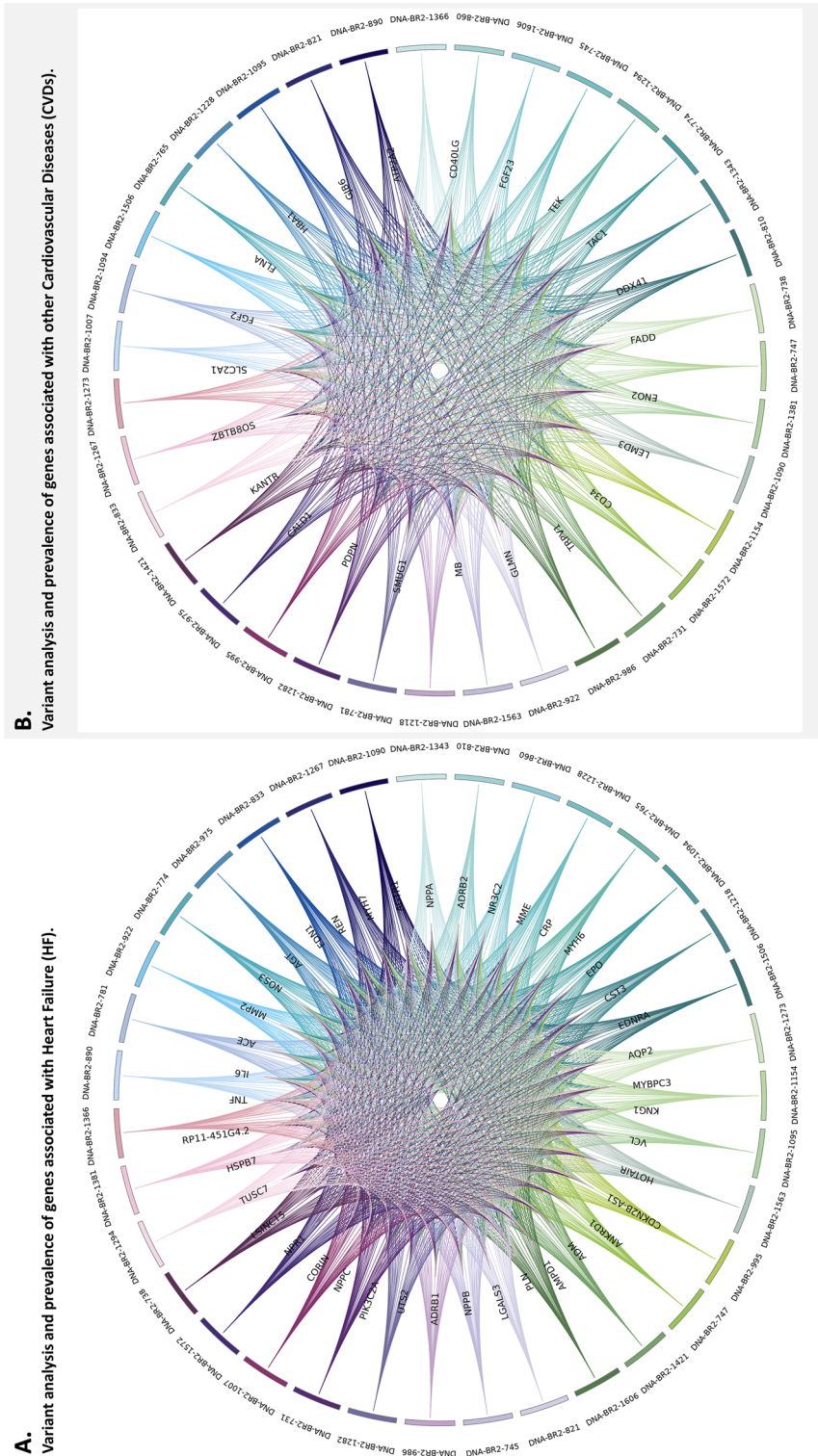


Figure 2. Gene-variant analysis of significant protein-coding genes associated with heart failure (HF) and other cardiovascular diseases (CVD) discovered through mutation analysis. (A) Variant analysis and prevalence of significant HF genes. (B) Variant analysis and prevalence of significant other CVD genes. These Circos plots are produced using the visualization module of the JWES pipeline²⁶.

Gene Name	Missense	Splice	Truncating	Intron	5' Flank	5' UTR	3' Flank	3' UTR	Silent	RNA	Total	Failed Mutations
ACE	12	4	0	120	24	N/A	66	6	15	N/A	247	60
ADM	1	0	0	3	28	N/A	16	2	1	N/A	51	29
ADRB1	5	0	0	N/A	30	2	31	8	5	N/A	81	1779
ADRB2	5	0	0	N/A	38	3	29	2	4	N/A	81	1866
AGTR1	0	0	0	257	40	1	31	12	2	N/A	343	479
AGT	3	0	0	67	37	N/A	24	2	2	N/A	135	718
AMPD1	4	1	1	103	9	N/A	N/A	N/A	1	N/A	119	78
ANKRD1	3	0	0	31	15	1	29	2		N/A	81	61
AQP2	0	1	0	28	25	1	8	17	2	N/A	82	85
CDKN2B-AS1	X	X	X	X	X	X	X	X	X	X	X	X
CORIN	5	2	0	1234	17	N/A	2	9	6	N/A	1275	404
CRP	1	0	0	2	32	N/A	8	5	2	N/A	50	291
CST3	2	0	0	16	33	2	67	6	2	N/A	128	254
EDN1	1	0	0	33	32	N/A	38	3	1	N/A	108	1320
EDNRA	0	0	0	296	34	3	21	8	2	N/A	364	438
EPO	0	0	0	9	25	3	15	4	N/A	N/A	56	124
HOTAIR	X	X	X	X	X	X	X	X	X	X	X	X
HSPB7	1	0	0	19	66	3	46	19	4	N/A	158	14
IL6	2	0	0	22	15	N/A	31		1	N/A	71	95
KNG1	3	1	1	15	N/A	2	2	1	4	N/A	29	307
LGALS3	4	1	0	83	61	1	N/A	N/A	N/A	N/A	150	474
LSINCT5	X	X	X	X	X	X	X	X	X	X	X	X
MME	3	3	0	953	63	1	1	18	2	N/A	1044	1096
MMP2	1	1	0	456	68	2	N/A	9	9	N/A	546	741
MYBPC3	9	1	1	72	11	N/A	N/A	N/A	10	N/A	104	84
MYH6	3	1	0	74	N/A	N/A	N/A	N/A	11	N/A	89	51
MYH7	4	5	0	105	35	N/A	29	1	16	N/A	195	231
NOS3	7	0	0	143	39	1	1	2	6	N/A	199	116
NPPA	3	0	1	16	25	N/A	N/A	4	N/A	N/A	49	52
NPPB	2	0	0	5	28	N/A	40	N/A	1	N/A	76	244
NPPC	0	0	0	1	23	N/A	36	N/A	N/A	N/A	60	457
NPR1	5	0	0	53	13	3	37	N/A	1	N/A	112	151
NR3C2	3	2	0	2057	27	1	2	6	5	N/A	2103	1079
PIK3C2A	1	1	0	423	37	2	29	4	5	1	503	326
PLN	0	0	0	49	N/A	N/A	N/A	9	N/A	N/A	58	46
REN	1	1	0	108	34	3	N/A	N/A	2	N/A	149	150
RP11-451G4.2	X	X	X	X	X	X	X	X	X	X	X	X
TNF	0	0	0	7	15	N/A	N/A	2	N/A	N/A	24	780
TUSC7	X	X	X	X	X	X	X	X	X	X	X	X
UTS2	3	0	0	30	37	N/A	19	N/A	1	N/A	90	806
VCL	2	0	0	522	42	1	N/A	5	7	N/A	579	875

Table 1. Functional mutation analysis of genes associated with heart failure disease. Tabulated information includes, gene name, missense, splice, truncating, intron, 5' flank, 5' UTR, 3' flank, 3' UTR, silent, RNA, total and failed mutations.

tolerated low confidence and benign, two deleterious and benign, two deleterious low confidence and benign, one deleterious and possibly damaging, one tolerated and benign, and one deleterious and probably damaging; hence, six of the eight mutations had some negative functional impact on the protein. In the *TEK* gene, the breakdown was five tolerated and benign, one tolerated and probably damaging, one deleterious and possibly damaging, and one deleterious and benign; hence, three of the eight mutations had some negative functional impact on the protein. In the *TRPV1* gene, the breakdown was seven tolerated and benign and one deleterious and benign; hence, only one of the eight mutations had some negative functional impact on the protein. *CALD1*, *TEK*, and *TRPV1* were found to be the most significant of the investigated genes as they have the largest number of functional mutations. *CALD1* and *TEK* also had the highest number of mutations mapped in total. Other CVD genes mutations including *ATP2A2* and *SMUG1* were discovered to have rare mutation types. We reported no missense mutations for multiple genes, therefore no further information regarding functional impact scores could be found. These genes included *ATP2A2*, *CD34*, *CD40LG*, *DDX41*, *FADD*, *FGF2*, *FLNA*, *HBA1*, *KANTR*,

Gene name	Missense	Splice	Truncating	Intron	5' UTR	5' Flank	3' UTR	3' Flank	Silent	RNA	Total	Failed Mutation
ATP2A2	0	2	0	270	6	11	14	24	3	0	330	478
CALD1	8	0	0	1100	1	66	7	26	2	0	1210	492
CD34	1	1	0	109	N/A	30	39	1	1	0	182	701
CD40LG	0	0	0	32	N/A	18	9	N/A	2	0	61	56
DDX41	0	2	0	17	N/A	4	2	N/A	1	0	26	13
ENO2	1	1	0	51	1	N/A	8	N/A	1	0	63	79
FADD	0	0	0	9	2	36	3	20	N/A	0	70	706
FGF23	2	0	0	68	N/A	33	5	17	N/A	0	125	451
FGF2	0	0	0	394	N/A	37	N/A	N/A	3	0	434	694
FLNA	2	6	0	82	N/A	N/A	2	10	11	0	113	73
GJB6	1	0	0	88	3	42	4	36	1	0	175	2133
GLMN	3	1	0	172	N/A	1	1	4	N/A	0	182	96
HBA1	0	0	0	2	N/A	12	1	15	N/A	0	30	28
KANTR	0	0	0	145	4	17	7	20	1	3	197	165
LEMD3	2	3	0	271	N/A	19	4	19	4	0	322	525
MB	0	0	0	95	1	39	4	49	2	0	190	738
PDPN	1	0	0	203	1	35	20	1	2	0	263	351
SLC2A1	0	1	0	216	2	N/A	5	30	7	0	261	230
TAC1	0	0	0	36	N/A	24	5	38	N/A	0	103	2518
TEK	8	1	0	1120	N/A	46	8	69	9	0	1261	518
TRPV1	8	0	1	207	3	68	7	38	5	0	337	256
SMUG1	1	1	1	61	1	17	4	26	N/A	0	112	433
ZBTB80S	0	0	0	197	N/A	N/A	12	23	1	0	233	205

Table 2. Functional mutation analysis of genes associated with other cardiovascular diseases. Tabulated information includes, gene name, missense, splice, truncating, intron, 5' flank, 5' UTR, 3' flank, 3' UTR, silent, RNA, total and failed mutations.

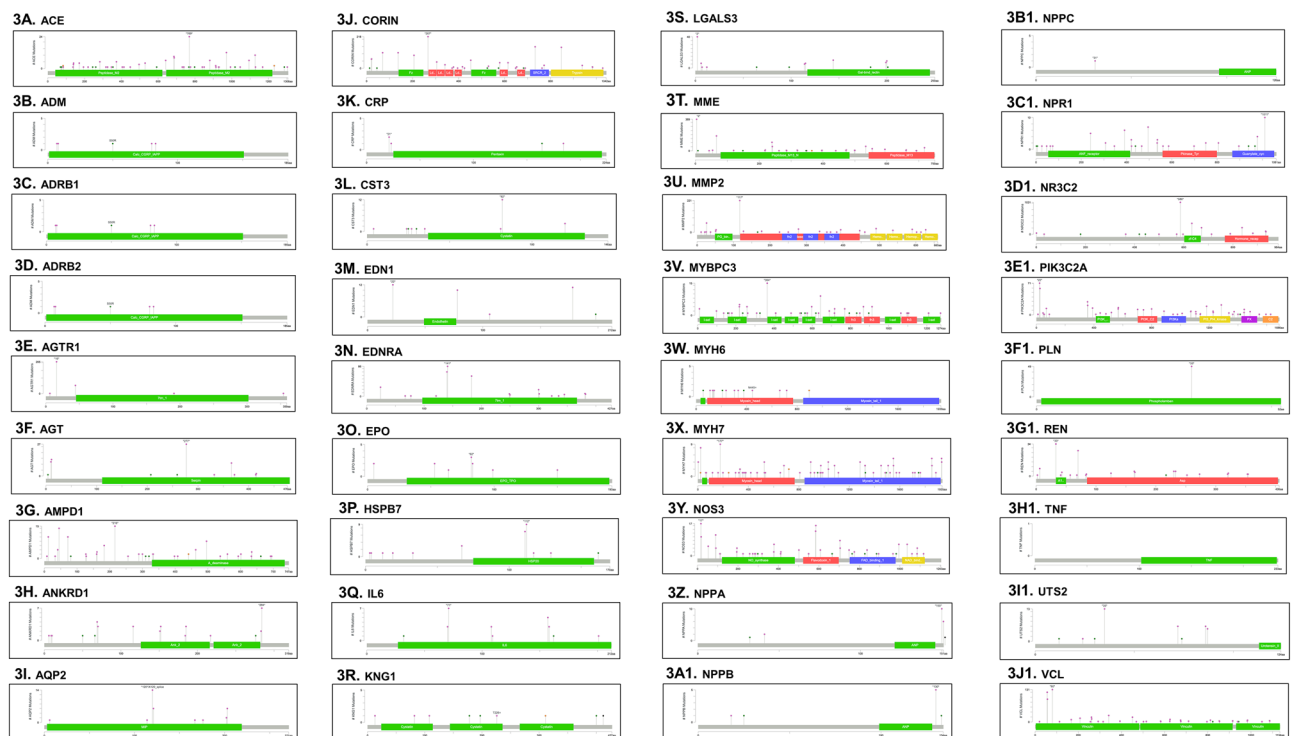


Figure 3. Functional mutation analysis of heart failure (HF) genes. Lollipop graphs of *ACE*, *ADM*, *ADRB1*, *ADRB2*, *AGTR1*, *AGT*, *AMPD1*, *ANKRD1*, *AQP2*, *CORIN*, *CRP*, *CST3*, *EDN1*, *EDNRA*, *EPO*, *HSPB7*, *IL6*, *KNG1*, *LGALS3*, *MME*, *MMP2*, *MYBPC3*, *MYH6*, *MYH7*, *NOS3*, *NPPA*, *NPPB*, *NPPC*, *NPR1*, *NR3C2*, *PIK3C2A*, *PLN*, *REN*, *TNF*, *UTS2*, and *VCL*. Green represents Missense mutations; black represents Truncating mutations; brown represents Inframe mutations; and purple represents Fusion mutations.



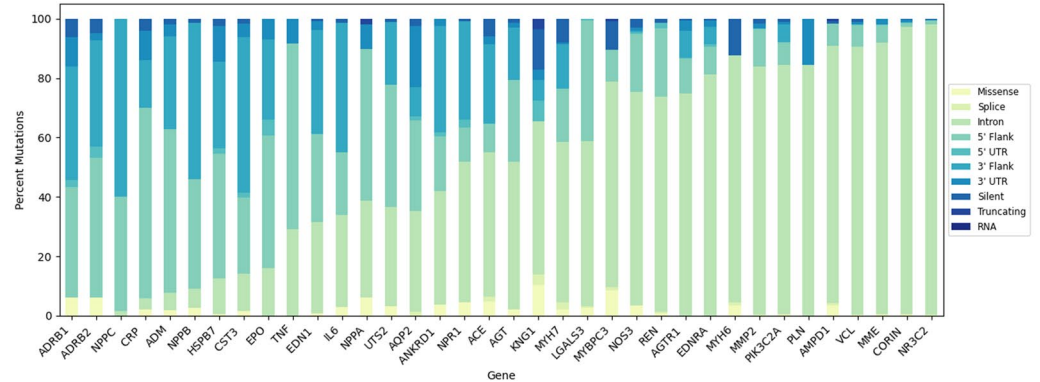
Figure 4. Functional mutation analysis of other cardiovascular disease (CVDs) genes. Lollipop graphs of: *ATP2A2*, *CALD1*, *CD34*, *CD40LG*, *DDX41*, *ENO2*, *FADD*, *FGF23*, *FGF2*, *FLNA*, *GJB6*, *GLMN*, *HBA1*, *KANTR*, *LEMD3*, *MB*, *PDPN*, *SLC2A1*, *SMUG1*, *TAC1*, *TEK*, *TRPV1*, and *ZBTB8OS*. Green represents Missense mutations; black represents truncating mutations; brown represents inframe mutations; and purple represents fusion mutations.

MB, *SLC2A1*, *TAC1*, and *ZBTB8OS*. Previous literature has linked *CALD1*, *TEK*, *TRPV1*, *ATP2A2*, and *SMUG1* to CVDs, supporting the findings from our functional mutation analysis^{57–61}. Further research must be conducted to solidify these claims regarding the significance of these genes. We reported different types of mutations and their impact on all CVD genes in Supplementary material S3.

Next, our splice mutation analysis uncovered mutation frequencies for the list of significant mutated genes generated after performing high-throughput WGS and utilizing JWES for WGS data processing and gene-variant discovery²⁷. We were able to analyze the percentages of each mutation (missense, splice, truncating, intron, 5' flank, 5' UTR, 3' flank, 3' UTR, silent and RNA) in comparison to each other (Fig. 5). We reported that intron, 5' Flank and 3' Flank mutations were present in high frequencies in genes associated with HF (Fig. 5A) and other CVDs (Fig. 5B). *NR3C2* had the highest number of intron mutations with a total of 2,057. *PIK3C2A* was the only gene with an RNA-based mutation. Aside from the RNA mutation, the rarest mutation type was truncating mutations. *AMPD1*, *KNG1*, *MYBPC3*, and *NPPA* were found to have a truncating mutation. Splice and 5' UTR were also less common or rarer mutation types (Fig. 5A). Among the genes associated with other CVDs, *TEK* had the highest number of intron mutations, with a total of 1,120. RNA mutations were the rarest in CVD genes as well, with *KANTR* being the only gene possessing RNA mutations. Truncating mutations were also very rare. *TRPV1* and *SMUG1* possessed truncating mutations (Fig. 5B).

We implemented JS-MA and the computed JSD scores highlighted the variance for all genes in relation to the disease (HF or other CVDs). The JSD scores for both HF and other CVD genes ranged from 0.09 to 0.49 with the diameter of each circle representing the score (Fig. 6). For the genes associated with HF, we observed five genes to be highly variant compared to others. These included *NPPC*, *ADRB2*, *ADRB1*, *MYH6* and *PLN* with JSD scores of 0.489, 0.474, 0.473, 0.453, and 0.449 respectively. *NR3C2*, *CRP*, *CORIN*, *NPPB*, *KNG1*, and *ADM* had moderate JSD for HF (Fig. 6A). For genes associated with other CVDs, we identified one gene, *HBA1*, to be extremely significant with a JSD of 0.493. We found *FADD* to have the second highest variance with a score of 0.425. Other genes with moderate JSD included *ENO2*, *GLMN*, *FLNA*, *CD40LG*, *FGF2*, *TAC1*, *CD34*, *DDX41*,

A. Mutation analysis of genes associated with heart failure (HF) disease.



B. Mutation analysis of genes associated with other cardiovascular disease (CVD).

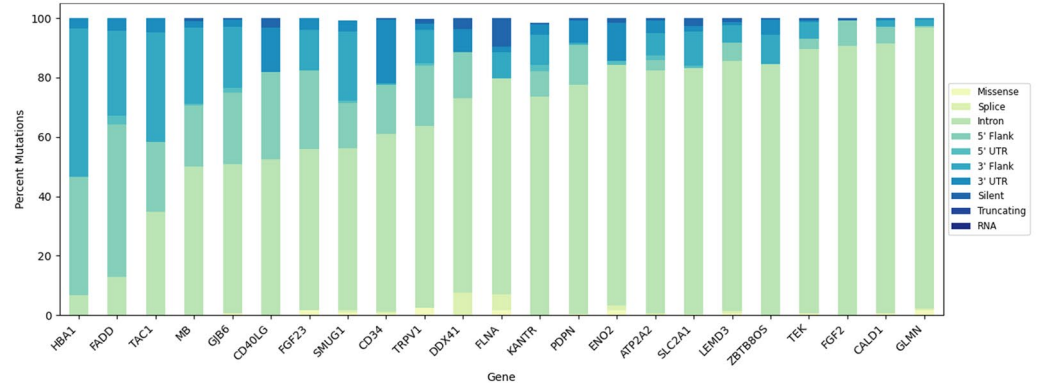


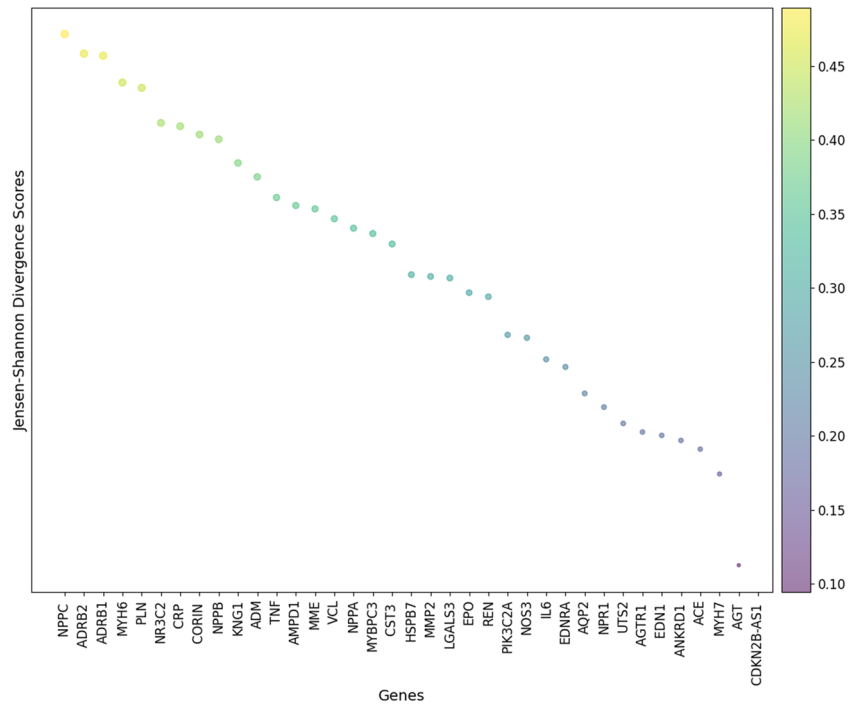
Figure 5. Splice mutation analysis of genes associated with heart failure (HF) and other cardiovascular disease (CVDs). Figure reports findings on missense, splice, truncating, intron, 5' flank, 5' UTR, 3' flank, 3' UTR, silent and RNA mutations for genes associated with HF (A) and other CVDs (B).

ZBTB80S, *SLC2A1*, *CALD1*, *TEK*, and *PDPN* (Fig. 6B). We found the following genes to have the highest variance: *HBA1*, *FADD*, *NPPC*, *ADRB2*, *ADRB1*, *MYH6*, and *PLN*. The exact JSD scores for all genes can be found in Supplementary material S1. Processed variant data of genes associated with HF and other CVDs are attached in the supplementary material (S4, S5, and S10).

We utilized a variety of analyses to identify notable genes including variant and prevalent analysis, functional mutation analysis, splice, and divergence analysis. Next, we performed comparative analysis to identify which genes were found to be notable and potentially significant in more than one method of analyses. The *HBA1* gene had a high JSD score and was observed in multiple enrichment pathways using our variant analysis and prevalence analysis. Hemoglobin subunit alpha 1 is involved in controlling pathways such as oxygen-carbon dioxide exchange in erythrocytes as well as cellular response to stimuli⁶². Mutations in *HBA1* have been found to be associated with multiple CVDs including but not limited to CAD⁶². Loss of function in *HBA1* can lead to Hemoglobin H disease, more commonly known as Alpha-thalassemia⁶². We found *LGALS3* reported in our variant as well as functional mutation analysis. *LGALS3* codes for Galectin-3 (Gal-3), a protein that plays an important role in cell proliferation, adhesion, differentiation, and apoptosis. Recent studies have linked Gal-3 levels to organ health and increase in Gal-3 levels have been associated with fibrotic and inflammatory diseases⁶³. *CALD1* and *TEK* were found to be highly significant based on our functional mutation analysis and had moderate JSD scores. *CALD1* is a protein coding gene that affects myosin in the smooth muscle. Mutations in *CALD1* have been associated with CVDs including but not limited to cardiomyopathy⁶⁴. *TEK* is involved in many biological pathways such as influencing the growth of blood vessels. Mutations in this gene can lead to abnormal formation of blood vessels and the heart⁶⁵. From the HF and other CVD genes, *HBA1*, *LGALS3*, and *TEK* had the strongest evidence of being significant and linking to CVDs based on the multiple analyses conducted as well as previous literature.

Comparing the results between HF and other CVD genes, we discovered many trends and distribution of mutation types and variations to be similar for both HF and other CVD genes (Fig. 2A,B). Most lollipop plots for HF and other CVDs had only one type of Pfam domain mapped for the corresponding gene (Figs. 3, 4). For HF genes, eleven genes in total (*CORIN*, *MME*, *MMP2*, *MYBPC3*, *MYH6*, *MYH7*, *NOS3*, *NPR1*, *NR3C2*, *PIK3C2A*, and *REN*) had two or more Pfam domains mapped (Fig. 3). For other CVD genes, the following seven genes were discovered to have two or more Pfam domains: *ATP2A2*, *LEMD3*, *ENO2*, *FADD*, *TEK*, *TRPV1*, *FLNA* (Fig. 4). HF genes, on average, had more Pfam domains that were able to be mapped. The most common mutation type for both HF and other CVDs was intron mutations with the least common being RNA, silent, and truncating mutation types. One major difference was that HF genes had an overall greater number of mutation types including RNA and truncating, both of which were not found in the other CVD genes (Fig. 5A,B).

A. Jensen-Shannon Divergence (JSD) of genes associated with heart failure (HF) disease.



B. JSD of genes associated with other cardiovascular disease (CVD).

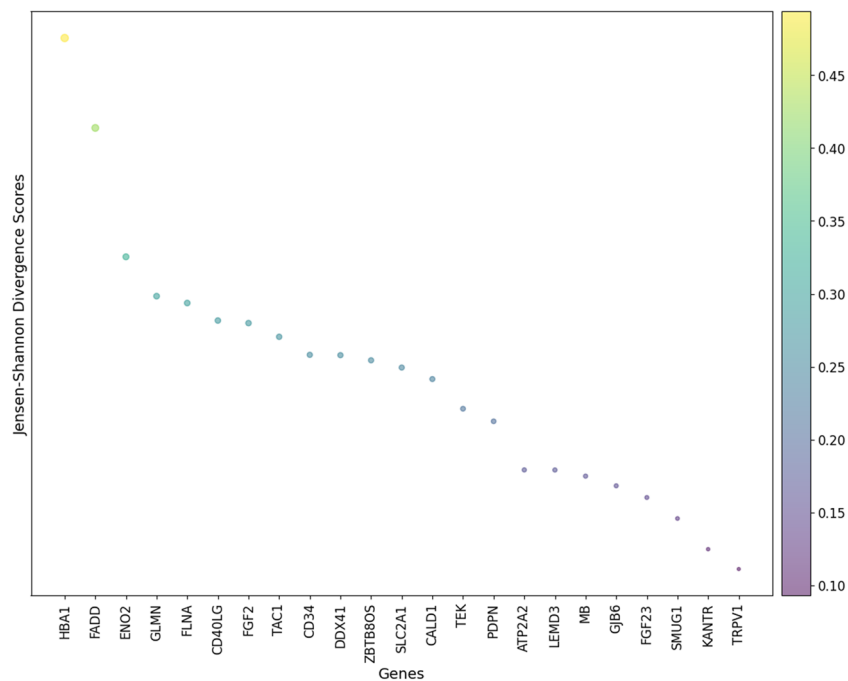


Figure 6. Jensen-Shannon divergence-based (JSD) statistical analysis and variant distribution analysis of genes associated with heart failure (HF) and other cardiovascular diseases (CVDs). Figure reports JSD scores of genes associated with HF (A) and other CVDs (B).

Understanding the common trends and variations in mutation distributions for HF and other CVDs can reveal similarities between the pathophysiology of multiple diseases and highlight the importance of further research to understand the relationship between HF and other CVD genes.

Discussion

LGALS3 codes for Gal-3 and recent studies have linked Gal-3 levels to organ health as well as fibrotic and inflammatory diseases⁶³. *LGALS3* had four missense mutations; the mutation mapped to P64H had a high functional impact (deleterious and probably damaging were SIFT and PolyPhen-2 scores), and the other two missense mutations were mapped to T98P and R183K; both mutations had low functional impact. Our analysis suggests *LGALS3* could also be linked to CVDs in addition to fibrotic diseases. Further studies are needed to confirm this relationship. A previous trial linked *MME* with CVDs and found HF patients had less chances of being hospitalized if treated with an angiotensin receptor neprilysin inhibitor⁶⁶. Although the remaining genes (*CST3*, *NR3C2*, *PIK32CA*, *TNF*, and *VCL*) had low functional impact for mutations, *PIK32CA* was also significant since it was the only gene out of thirty-six HF genes that produced a lollipop graph with an RNA mutation type. Additionally, we found *NPPC*, *ADRB2*, *ADRB1*, *MYH6* and *PLN* genes to have high variance based on JS-MA.

When conducting mutation analysis, our study was able to generate functional mutation scores for *LEMD3* and *SMUG1*; for the other genes, no functional mutation information could be found, as there were no missense mutations present. *LEMD3* had one mutation mapped with a high functional impact (deleterious and possibly damaging for the SIFT and PolyPhen-2 scores) and one mutation with low functional impact. Mutations in *LEMD3* have been linked to various conditions such as Buschke–Ollendorff⁶⁷ and our study suggests the gene can have further links to CVDs. Less gene expression of *SMUG1* has been linked to breast cancer⁶⁸. *SMUG1* had one mutation with low functional impact, which suggests further research should be conducted to assess its association with CVDs as well. We found *HBA1* and *FADD* were found to be extremely significant using JS-MA. Mutations in *HBA1* have been found to be associated with multiple CVDs including but not limited to CAD⁶⁴. While mutations in *FADD* have been associated with post-ischemic HF, further studies are needed to study if *FADD* can be used in gene therapy for HF treatment⁶⁵. Further research is needed for *LEMD3*, *SMUG1*, *HBA1*, *FLNA*, *ZBTB80S*, and *SLC2A1* since they were found significant in multiple analyses conducted.

Additional genes from our variant and functional mutation analysis were reported to be significant. From the HF genes, *ACE* was found to have the largest number of missense mutations with a high functional impact; in the CVD genes, *CALDI*, *TEK*, and *TRPV1* genes had the largest number of mutations with high functional impact. Future studies are needed to be better informed and targeted towards certain genes for mutation analysis and disease-specific variants. Findings from our functional mutation analysis warrant further study of the gene-disease causal relationships involving HF and CVD genes, especially *ACE*, *CALDI*, *TEK*, and *TRPV1*. Significant genes noted in our current study were also supported by findings from our previous RNA-seq driven gene differential expression and pathway enrichment analysis. Genes such as *FADD*, *HBA1* and *LGALS3* were found to be differentially expressed in HF patients³⁰. While *CALDI*, *TEK*, and *TRPV1* showed low expression in HF patients compared to healthy controls³⁰. Most of our biological findings for significant genes are thus validated by previous gene-disease annotation, phenotyping as well as mRNA abundance analysis³⁰. We found *ADRB1*, *ADRB2*, and *NPPC* to have great variance and significance based on JS-MA from our previous variant analysis from a separate ensemble of CVD patients⁶⁹. Thus, supporting our claim that these genes have significant or altered expression in CVD patients. Additionally, we observed that *ACE* and *CALDI* were highly associated with CVDs and played a major role in disease prediction based on our Artificial Intelligence (AI) and Machine Learning (ML) driven analysis⁷⁰.

There were some limitations to using the cBioPortal Mutation Mapper. The total amount of mutations discovered by our previous study for each significant HF and other CVD gene were not all able to be mapped onto the lollipop graphs^{26,27}. There were a significant number of mutations that failed to be annotated due to insufficient information in the reference database. Results showed that seven HF genes studied possessed mutations whose functional impacts could not be tracked due limitations of the software; the same was true for thirteen CVD genes. The cBioPortal software was unable to support this information since the mutations discovered were novel and the database has not been updated yet. These limitations prevented a complete lollipop plot of mutation distributions from being generated for each HF and CVD gene. However, based on the numerous mutations that were mapped, significant patterns were discovered. Another limitation of our study was the sample size utilized that can limit the generalization of our findings. To partially address this limitation, we have conducted an additional whole genome and variant analysis on an alternative group of consented CVD patients to support and validate our findings⁵³. Additionally, we plan on expanding our cohort in the future to include diverse individuals based on race, ethnicity, and socioeconomic factors to better highlight the importance and frequency of mutations linked to frequently studied HF and CVD genes.

Our methodology involved using JWES for WGS data processing and utilizing GATK for the identification of point mutations. Moving forward, the inclusion of other variation types including copy number variations (CNV), structural variants (SV), and short tandem repeats (STR) may increase or decrease the significance of genes depending on a variety of factors. Unlike SNPs which are variations of single nucleotide in a specific genome location, STRs are variations of the number of repeating DNA sequences. A previous study found that SNPs are considered a viable replacement for STRs to detect the structure of a population⁷¹. SVs are defined as a DNA region of about one kilobase (kb) and can include inversions or insertions and deletions, also known as CNVs⁷². While SNPs affect splicing or transcription and are present in coding or non-coding regions, CNVs are defined as sequence variants that can be as large as several megabases (Mbs) in size. CNVs have been linked to the pathogenesis of complex diseases; studies reveal that when associations exist between CNVs and SNPs, the coexistence frequency, and the type of CNV can lead to an overestimation or underestimation of the gene significance. The application of a joint analysis of CNVs and SNPs may address these current limitations and provide more accuracy in identifying significant genes moving forward⁷⁰.

To study chronic diseases such as CVDs with complex pathophysiology, conducting multiple analyses with over-compassing methodologies is essential. The overall goal of the study was to conduct a combination of

variant distribution and prevalence, functional mutation, splice mutation and divergence analysis to identify the significant impact of these mutations on the pathology of CVDs. Our results reinforce the established relationship between significant genes highlighted in previous literature and their impact on CVDs. Further research can be conducted to validate our claims regarding potentially significant genes by widening the sample size of consented patients to estimate trends within a population. This is a goal we hope to accomplish in the future. It is of paramount importance to fully understand the genetic basis of diseases, especially common ones, distinguish the genes which predispose an individual to medical conditions, and how rare genetic variations play a role in disease manifestation⁷⁴. Further inquiry into these genes may foster the development of novel clinical tools that will improve personalized medical treatment for HF and other CVD patients. Once the individual's genetic makeup is considered, medical providers will be able to formulate a more personalized treatment plan⁷⁵. Several studies have successfully employed integrative multi-omics approaches to investigate novel mechanisms and plasma biomarkers associated with cardiovascular diseases, ultimately speeding up the identification of new therapeutic targets and pathways⁷⁶. These studies serve as evidence that sophisticated integration techniques can yield dependable biological signals across various molecular levels and phenotypes⁷⁶.

Our research underscores the critical need for an integrative approach that combines gene variant data with clinical information. We employed a multifaceted analysis, including functional mutation, splice variant, variant distribution, and divergence analysis, to discern the significance and prevalence of variants linked to well-studied genes associated with HF and CVD. Our variant analysis revealed the significance of additional genes, such as *ACE*, *CALDI*, *TEK*, and *TRPVI*. Among HF and other CVD genes, we observed that mutations in introns, the 5' flank, 3' UTR, and 3' flank regions were the most prevalent. Although missense mutations were infrequent, they were more likely to exert a functional impact. By employing JS-MA, we pinpointed *NPPC*, *ADRB2*, *ADBR1*, *MYH6*, *PLN*, *HBA1*, and *FADD* as the genes exhibiting the highest degree of variability. Previously, we have examined state-of-the-art genomic approaches to identify and investigate genes associated with atrial fibrillation (AF) and HF susceptibility²³. We found multiple genes such as *PLN*⁷⁷, *MYH6*⁷⁷, *NPPA*⁷⁷, and *MYH7*⁷⁸ to be significant, all of which were discovered to be notable in this study as well. The wide range of patients from various ages, ethnicities, demographics, and geographic locations as well as the variety of methods from these previous studies contributes to a randomized sample size²³.

We expanded our research regarding these significant genes by exploring the clinical relevance of gene expression by leveraging RNA-seq data^{30,79}. Our analysis focused on discerning the disparities between healthy and afflicted conditions, aiming to gain insights into the underlying disease pathology. We performed age and gender-based analyses to further understand shared and unique expressions across different ethnic and racial profiles^{30,79}. Our previous and current studies have uncovered *ACE* to be a critical gene in CVD etiology and progression across all age groups. These findings hold significant importance for future research endeavors, as they indicate the opportunity to delve deeper into these genes opening a novel avenue that emphasizes a more personalized approach to therapy and treatment. The findings from previous studies corroborate our current results in this study. In conjunction, the variety of analyses performed including variant and prevalent analysis, functional mutation analysis, splice, and divergence analysis identified similar patterns and notable genes which suggests other confounding risk factors are not significant enough to overturn the conclusions reached in our study.

A multitude of genomic and statistical studies have similarly utilized phenotypic attributes such as gender, age, ethnicity, and diagnoses to determine gene causality in disease advancement^{49–51}. While the age at which patients developed HF, severity of disease, alcoholic cardiomyopathy, different aetiologies of their HF, treatments received are important risk factors, recent approaches now focus on the heritability component that supports the clinical manifestation of the disease^{50,51}. In this study, we utilized a cohort of only adult and aging CVD patients with HF phenotype. The data centered on age, gender, ethnicity, medical details, and demographics and added controls as the sample size was targeted and specific utilizing the restriction method designed to mitigate the effects of other confounding factors⁸⁰. Our claims are supported by cutting-edge research, leading us to conclude that these confounding risk factors can be ruled out from the context of our study and have little relevance to our overall findings. In the future, we hope to expand our cohort of our healthy controls and patient cohorts to investigate and solidify the association between significant genes and the development of HF and CVDs.

For cardiovascular genomic medicine to become both predictive and preventive, it is crucial to accurately assess the risk of associated disease, properly report the variants, and implement clinical management to prevent or reduce the disease. Currently, multi-omics data are not available in formats that are useful for the AI/ML analysis. In the future, AI/ML-ready genomic data sets should be more widely available to integrate AI/ML algorithms in predictive analysis. ML can help identify a predictive response and model clinical data for association of genetic variants to treatment outcomes in HF and other CVDs⁷⁵. We can process large volumes of clinical and variant data to identify biomarkers or gene sets associated with chronic diseases and improve diagnosis. With greater availability of AI/ML-ready datasets, the genomic data can be analyzed on a deeper level, with implications both in predictive analysis as well as deep phenotyping⁸¹. Additionally, growing evidence now suggests that there might be a direct link between infectious oral diseases and CVDs. The proposed mechanisms that explain the correlation between these two diseases consist of predisposing and precipitating aspects such as genetic and environmental factors, medications, and the individual's microbiome⁸². Further studies have suggested that maladaptive inflammatory reactivity, which may be influenced by SNPs in pathway genes, could act as pleiotropic genes and effect the link between oral infections and CVDs^{83,84}.

Conclusion

Our study emphasizes the importance of an integrative approach with gene variant and clinical data and utilizes functional mutation, splice, variant distribution, and divergence analysis to identify the significance and prevalence of variants associated with commonly investigated HF and CVD genes. Our variant analysis uncovered additional genes to be significant including *ACE*, *CALD1*, *TEK*, and *TRPV1*. We discovered intron, 5' Flank, 3' UTR, and 3' Flank mutations to be the most common among HF and other CVD genes. Missense mutations were rare but more likely to have functional impact. We implemented JS-MA and identified *NPPC*, *ADRB2*, *ADBR1*, *MYH6*, *PLN*, *HBA1*, and *FADD* genes to have the highest variance. The identification of the functional impact of these mutations will help us understand CVD progression and pathophysiology. Further studies are needed to determine if the genes with notable mutations can be used as potential biomarkers to improve early diagnosis and disease prediction.

Data availability

Processed variant data of genes associated with HF and other CVDs are attached in the supplementary material. All the source code reproducing the experiments of this study are available at GitHub, following web links: JWES <<https://github.com/drzeeshanahmed/JWES-Variant>>, and JSD-Variant-Distribution-Analysis <<https://github.com/drzeeshanahmed/JSD-Variant-Distribution-Analysis>>.

Received: 22 April 2023; Accepted: 4 October 2023

Published online: 05 October 2023

References

1. Mc Namara, K., Alzubaidi, H. & Jackson, J. K. Cardiovascular disease as a leading cause of death: How are pharmacists getting involved?. *Integr. Pharm. Res. Pract.* **2019**(8), 1–11. <https://doi.org/10.2147/IPRPS133088> (2019).
2. Virani, S. S., Alonso, A., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Khan, S. S., Kissela, B. M., Knutson, K. L., Kwan, T. W., Lackland, D. T., Lewis, T. T., American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics–2020 Update: A report from the American heart association. *Circulation* **141**(9), e139–e596. <https://doi.org/10.1161/CIR.0000000000000757> (2020).
3. Roth, G. A. *et al.* Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70**(1), 1–25. <https://doi.org/10.1016/j.jacc.2017.04.052> (2017).
4. Stewart, J., Manmathan, G. & Wilkinson, P. Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. *JRSM Cardiovasc. Dis.* **6**, 2048004016687211. <https://doi.org/10.1177/2048004016687211> (2017).
5. Walden, R., & Tomlinson, B. Cardiovascular Disease. In I. Benzie (Eds.) *et al.*, 935 *Herbal Medicine: Biomolecular and Clinical Aspects*. (2nd ed.) (CRC Press/Taylor & Francis, 2011)
6. Doran, S., Arif, M., Lam, S., Bayraktar, A., Turkez, H., Uhlen, M., Boren, J., & Mardinoglu, A. Multi-omics approaches for revealing the complexity of cardiovascular disease. *Brief. Bioinf.* **22**(5), bbab061. <https://doi.org/10.1093/bib/bbab061> (2021).
7. Ahmed, Z. Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. *Human Genom.* **14**(1), 35. <https://doi.org/10.1186/s40246-020-00287-z> (2020).
8. Currie, G. & Delles, C. Precision medicine and personalized medicine in cardiovascular disease. *Adv. Exp. Med. Biol.* **1065**, 589–605. https://doi.org/10.1007/978-3-319-77932-4_36 (2018).
9. Kathiresan, S. & Srivastava, D. Genetics of human cardiovascular disease. *Cell* **148**(6), 1242–1257. <https://doi.org/10.1016/j.cell.2012.03.001> (2012).
10. Seo, D., Ginsburg, G. S. & Goldschmidt-Clermont, P. J. Gene expression analysis of cardiovascular diseases: Novel insights into biology and clinical applications. *J. Am. College Cardiol.* **48**(2), 227–235. <https://doi.org/10.1016/j.jacc.2006.02.070> (2006).
11. Dumény, L. *et al.* NR3C2 genotype is associated with response to spironolactone in diastolic heart failure patients from the Aldo-DHF trial. *Pharmacotherapy* **41**(12), 978–987. <https://doi.org/10.1002/phar.2626> (2021).
12. Heliste, J. *et al.* Genetic and functional implications of an exonic TRIM55 variant in heart failure. *J. Mol. Cell. Cardiol.* **138**, 222–233. <https://doi.org/10.1016/j.yjmcc.2019.12.008> (2020).
13. Min, K. D. *et al.* Identification of genes related to heart failure using global gene expression profiling of human failing myocardium. *Biochem. Biophys. Res. Commun.* **393**(1), 55–60. <https://doi.org/10.1016/j.bbrc.2010.01.076> (2010).
14. Vrablik, M., Dlouha, D., Todorovova, V., Stefler, D. & Hubacek, J. A. Genetics of cardiovascular disease: How far are we from personalized CVD risk prediction and page 33 of 148 clinical and translational medicine management?. *Int. J. Mol. Sci.* **22**(8), 4182. <https://doi.org/10.3390/ijms22084182> (2021).
15. Wain, L. V. Rare variants and cardiovascular disease. *Brief. Funct. Genom.* **13**(5), 384–391. <https://doi.org/10.1093/bfgp/elu010> (2014).
16. Kazmi, N. & Gaunt, T. R. Diagnosis of coronary heart diseases using gene expression profiling; stable coronary artery disease, cardiac ischemia with and without myocardial necrosis. *PLoS One* **11**(3), e0149475. <https://doi.org/10.1371/journal.pone.0149475> (2016).
17. Ataklte, F. & Vasan, R. S. Heart failure risk estimation based on novel biomarkers. *Expert Rev. Mol. Diagn.* **21**(7), 655–672. <https://doi.org/10.1080/14737159.2021.1933446> (2021).
18. Pei, S., Liu, T., Ren, X., Li, W., Chen, C., & Xie, Z. Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Brief. Bioinf.* **22**(3), bbaa148. <https://doi.org/10.1093/bib/bbaa148> (2021).
19. Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database J. Biol. Databases Curation* baaa010. <https://doi.org/10.1093/database/baaa010> (2020).
20. Leopold, J. A. & Loscalzo, J. Emerging role of precision medicine in cardiovascular disease. *Circ. Res.* **122**(9), 1302–1315. <https://doi.org/10.1161/CIRCRESAHA.117.310782> (2018).
21. Leopold, J. A., Maron, B. A. & Loscalzo, J. The application of big data to cardiovascular disease: Paths to precision medicine. *J. Clin. Investig.* **130**(1), 29–38. <https://doi.org/10.1172/JCI129203> (2020).
22. Antman, E. M. & Loscalzo, J. Precision medicine in cardiology. *Nat. Rev. Cardiol.* **13**(10), 591–602. <https://doi.org/10.1038/nrcardio.2016.101> (2016).
23. Patel, K. K. *et al.* Genomic approaches to identify and investigate genes associated with atrial fibrillation and heart failure susceptibility. *Hum. Genomics* **17**(1), 47. <https://doi.org/10.1186/s40246-023-00498-0> (2023).
24. Wung, S. F., Hickey, K. T., Taylor, J. Y. & Gallek, M. J. Cardiovascular genomics. *J. Nurs. Scholar.* **45**(1), 60–68. <https://doi.org/10.1111/jnu.12002> (2013).

25. Ahmed, Z., Kim, M. & Liang, B. T. MAV-clic: Management, analysis, and visualization of clinical data. *JAMIA open* **2**(1), 23–28. <https://doi.org/10.1093/jamiaopen/ooy052> (2018).
26. Ahmed, Z. Intelligent health system for the investigation of consenting COVID-19 patients and precision medicine. *Person. Med.* **18**(6), 573–582 (2021).
27. Ahmed, Z., Renart, E. G., Mishra, D. & Zeeshan, S. JWES: A new pipeline for whole genome/exome sequence data processing, management, and gene-variant discovery, annotation, prediction, and genotyping. *FEBS Open Bio* <https://doi.org/10.1002/2211-5463.13261> (2021).
28. Keel, B. N. & Snelling, W. M. Comparison of burrows-wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: Application to Illumina data for livestock genomes. *Front. Genet.* **9**, 35. <https://doi.org/10.3389/fgene.2018.00035> (2018).
29. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110> (2010).
30. Ahmed, Z., Zeeshan, S. & Liang, B. T. RNA-seq driven expression and enrichment analysis to investigate CVD genes with associated phenotypes among high-risk heart failure patients. *Human Genom.* **15**(1), 67. <https://doi.org/10.1186/s40246-021-00367-8> (2021).
31. Ahmed, Z., Renart, E. G., Zeeshan, S. & Dong, X. Advancing clinical genomics and precision medicine with GVViz: FAIR bioinformatics platform for variable gene-disease annotation, visualization, and expression analysis. *Hum. Genom.* **15**(1), 37. <https://doi.org/10.1186/s40246-021-00336-1> (2021).
32. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**(7809), 434–443. <https://doi.org/10.1038/s41586-020-2308-7> (2020).
33. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**(13), 3812–3814 (2003).
34. Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**(Web Server issue), W452–W457 (2012).
35. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protocols* **4**(7), 1073–1081 (2009).
36. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protocols Hum. Genet. Chapter 7*(Unit7), 20 (2013).
37. Montenegro, L. R., Lerário, A. M., Nishi, M. Y., Jorge, A. & Mendonca, B. B. Performance of mutation pathogenicity prediction tools on missense variants associated with 46, XY differences of sex development. *Clinics (Sao Paulo, Brazil)* **76**, e2052 (2021).
38. Vohra, S. & Biggin, P. C. Mutationmapper: A tool to aid the mapping of protein mutation data. *PLoS One* **8**(8), e71711. <https://doi.org/10.1371/journal.pone.0071711> (2013).
39. Zhang, W., Wang, C. & Zhang, X. Mutplot: An easy-to-use online tool for plotting complex mutation data with flexibility. *PLoS one* **14**(5), e0215838. <https://doi.org/10.1371/journal.pone.0215838> (2019).
40. Guo, X. JS-MA: A Jensen–Shannon divergence based method for mapping genome-wide associations on multiple diseases. *Front. Genet.* **11**, 507038. <https://doi.org/10.3389/fgene.2020.507038> (2020).
41. Stojković, V. & Fujimori, D. G. Mutations in RNA methylating enzymes in disease. *Curr. Opin. Chem. Biol.* **41**, 20–27. <https://doi.org/10.1016/j.cbpa.2017.10.002> (2017).
42. Hong, D., & Jeong, S. 3'UTR Diversity: Expanding repertoire of RNA alterations in human mRNAs. *Mol. Cells* **46**(1), 48–56. <https://doi.org/10.14348/molcells.2023.0003> (2023).
43. Schuster, S. L. & Hsieh, A. C. The Untranslated regions of mRNAs in cancer. *Trends Cancer* **5**(4), 245–262. <https://doi.org/10.1016/j.trecan.2019.02.011> (2019).
44. Herman, D. S. *et al.* Truncations of titin causing dilated cardiomyopathy. *N. Engl. J. Med.* **366**(7), 619–628. <https://doi.org/10.1056/NEJMoa1110186> (2012).
45. Guo, L. *et al.* A missense mutation in ISPD contributes to maintain muscle fiber stability. *Poult. Sci.* **101**(11), 102143. <https://doi.org/10.1016/j.psj.2022.1021> (2022).
46. Rose, A. B. Introns as gene regulators: A brick on the accelerator. *Front. Genet.* **9**, 672. <https://doi.org/10.3389/fgene.2018.00672> (2019).
47. Anna, A. & Monika, G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* **59**(3), 253–268. <https://doi.org/10.1007/s13353-018-0444-7> (2018).
48. Harrigan, P. R. *et al.* Silent mutations are selected in HIV-1 reverse transcriptase and affect enzymatic efficiency. *AIDS (London, England)* **22**(18), 2501–2508. <https://doi.org/10.1097/QAD.0b013e328318f16c> (2008).
49. Staerk, L., Sherer, J. A., Ko, D., Benjamin, E. J. & Helm, R. H. Atrial fibrillation: Epidemiology, pathophysiology, and clinical outcomes. *Circ. Res.* **120**(9), 1501–1517. <https://doi.org/10.1161/CIRCRESAHA.117.309732> (2017).
50. Backer, J. D. & Braverman, A. C. Heart failure and sudden cardiac death in heritable thoracic aortic disease caused by pathogenic variants in the SMAD3 gene. *Mol. Genet. Genomic Med.* **6**(4), 648–652. <https://doi.org/10.1002/mgg3.396> (2018).
51. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat. Commun.* **11**(1), 163. <https://doi.org/10.1038/s41467-019-13690-5> (2020).
52. Montecucco, F. & Mach, F. Statins, ACE inhibitors and ARBs in cardiovascular disease. *Best Pract. Res. Clin. Endocrinol. Metab.* **23**(3), 389–400. <https://doi.org/10.1016/j.beem.2008.12.003> (2009).
53. Pereira, N. L. *et al.* Natriuretic peptide pharmacogenetics: membrane metallo-endopeptidase (MME): Common gene sequence variation, functional characterization and degradation. *J. Mol. Cell. Cardiol.* **49**(5), 864–874. <https://doi.org/10.1016/j.yjmcc.2010.07.020> (2010).
54. Blanda, V., Bracale, U. M., Di Taranto, M. D. & Fortunato, G. Galectin-3 in cardiovascular diseases. *Int. J. Mol. Sci.* **21**(23), 9232. <https://doi.org/10.3390/ijms21239232> (2020).
55. Bauersachs, J. & López-Andrés, N. Mineralocorticoid receptor in cardiovascular diseases—Clinical trials and mechanistic insights. *Br. J. Pharmacol.* **179**(13), 3119–3134. <https://doi.org/10.1111/bph.15708> (2022).
56. Tan, B., Liu, M., Yang, Y., Liu, L. & Meng, F. Low expression of PIK3C2A gene: A potential biomarker to predict the risk of acute myocardial infarction. *Medicine* **98**(14), e15061. <https://doi.org/10.1097/MD.00000000000015061> (2019).
57. Kim, N. Y. *et al.* Quantitative proteomic analysis of human serum using tandem mass tags to predict cardiovascular risks in patients with psoriasis. *Sci. Rep.* **13**(1), 2869. <https://doi.org/10.1038/s41598-023-30103-2> (2023).
58. Heliste, J. *et al.* Receptor tyrosine kinase profiling of ischemic heart identifies ROR1 as a potential therapeutic target. *BMC Cardiovasc. Disord.* **18**, 196. <https://doi.org/10.1186/s12872-018-0933-y> (2018).
59. Pilic, L. & Mavrommatis, Y. Genetic predisposition to salt-sensitive normotension and its effects on salt taste perception and intake. *Br. J. Nutr.* **120**(7), 721–731. <https://doi.org/10.1017/S0007114518002027> (2018).
60. Angrisano, T. *et al.* Epigenetic switch at atp2a2 and myh7 gene promoters in pressure overload-induced heart failure. *PLoS One* **9**(9), e106024. <https://doi.org/10.1371/journal.pone.0106024> (2014).
61. Kroustallaki, P. *et al.* SMUG1 promotes telomere maintenance through telomerase RNA processing. *Cell Rep.* **28**(7), 1690–1702. e10. <https://doi.org/10.1016/j.celrep.2019.07.040> (2019).
62. Chonchol, M. & Nielson, C. Hemoglobin levels and coronary artery disease. *Am. Heart J.* **155**(3), 494–498. <https://doi.org/10.1016/j.ahj.2007.10.031> (2008).

63. Hara, A. *et al.* Galectin-3 as a next-generation biomarker for detecting early stage of various diseases. *Biomolecules* **10**(3), 389. <https://doi.org/10.3390/biom10030389> (2020).
64. Zheng, P. P., Severijnen, L. A., van der Weiden, M., Willemsen, R. & Kros, J. M. A crucial role of caldesmon in vascular development in vivo. *Cardiovasc. Res.* **81**(2), 362–369. <https://doi.org/10.1093/cvr/cvn294> (2009).
65. Eklund, L., Kangas, J. & Saharinen, P. Angiotensin-Tie signalling in the cardiovascular and lymphatic systems. *Clin. Sci.* **131**(1), 87–103. <https://doi.org/10.1042/CS20160129> (2017).
66. Krittanawong, C. & Kitai, T. Pharmacogenomics of angiotensin receptor/neprilysin inhibitor and its long-term side effects. *Cardiovasc. Ther.* **35**(4), 1. <https://doi.org/10.1111/1755-5922.12272> (2017).
67. Lin, F., Morrison, J. M., Wu, W. & Worman, H. J. MAN1, an integral protein of the inner nuclear membrane, binds Smad2 and Smad3 and antagonizes transforming growth factor-beta signaling. *Hum. Mol. Genet.* **14**(3), 437–445. <https://doi.org/10.1093/hmg/ddi040> (2005).
68. Abdel-Fatah, T. M. *et al.* Single-strand selective monofunctional uracil-DNA glycosylase (SMUG1) deficiency is linked to aggressive breast cancer and predicts response to adjuvant therapy. *Breast Cancer Res. Treatm.* **142**(3), 515–527. <https://doi.org/10.1007/s10549-013-2769-6> (2013).
69. Ahmed, Z. *et al.* Investigating genes associated with cardiovascular disease among heart failure patients for translational research and precision medicine. *Clin. Transl. Discov.* **3**(3), e206. <https://doi.org/10.1002/ctd2.206> (2023).
70. Venkat, V., Abdelhalim, H., DeGroat, W., Zeeshan, S. & Ahmed, Z. Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine. *Genomics* **115**(2), 110584. <https://doi.org/10.1016/j.ygeno.2023.110584> (2023).
71. Kauwe, J. S., Bertelsen, S., Bierut, L. J., Dunn, G., Hinrichs, A. L., Jin, C. H., & Suarez, B. K. The efficacy of short tandem repeat polymorphisms versus single-nucleotide polymorphisms for resolving population structure. *BMC Genet.* **6**(Suppl 1), S84. <https://doi.org/10.1186/1471-2156-6-S1-S84> (2005).
72. U.S. National Library of Medicine. (n.d.). Overview of structural variation. National Center for Biotechnology Information. <https://www.ncbi.nlm.nih.gov/dbvar/content/overview/>.
73. Liu, J. *et al.* The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Hum. Genet.* **137**(6–7), 553–567. <https://doi.org/10.1007/s00439-018-1910-3> (2018).
74. Ahmed, Z. Multi-omics strategies for personalized and predictive medicine: Past, current, and future translational opportunities. *Emerg. Top. Life Sci.* **6**(2), 215–225. <https://doi.org/10.1042/ETLS20210244> (2022).
75. Vadapalli, S., Abdelhalim, H., Zeeshan, S., & Ahmed, Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. Briefings in bioinformatics, bbac191. <https://doi.org/10.1093/bib/bbac191> (2022).
76. Leon-Mimila, P., Wang, J. & Huertas-Vazquez, A. Relevance of multi-omics studies in cardiovascular diseases. *Front. Cardiovasc. Med.* **6**, 91. <https://doi.org/10.3389/fcvm.2019.00091> (2019).
77. Christophersen, I. E., Rienstra, M., Roselli, C., Yin, X., Geelhoed, B., Barnard, J., Lin, H., Arking, D. E., Smith, A. V., Albert, C. M., Chaffin, M., Tucker, N. R., Li, M., Klarin, D., Bihlmeyer, N. A., Low, S. K., Weeke, P. E., Müller-Nurasyid, M., Smith, J. G., Brody, J. A., AFGen Consortium. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **49**(6), 946–952. <https://doi.org/10.1038/ng.3843> (2017).
78. Chalazan, B. *et al.* Association of rare genetic variants and early-onset atrial fibrillation in ethnic minority individuals. *JAMA Cardiol.* **6**(7), 811–819. <https://doi.org/10.1001/jamacardio.2021.0994> (2021).
79. Berber, A. *et al.* RNA-seq-driven expression analysis to investigate cardiovascular disease genes with associated phenotypes among atrial fibrillation patients. *Clin. Transl. Med.* **12**(7), e974. <https://doi.org/10.1002/ctm2.974> (2022).
80. Jager, K. J., Zoccali, C., Macleod, A. & Dekker, F. W. Confounding: What it is and how to deal with it. *Kidney Int.* **73**(3), 256–260. <https://doi.org/10.1038/sj.ki.5002650> (2008).
81. Jiang, F. *et al.* Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2**(4), 230–243. <https://doi.org/10.1136/svn-2017-000101> (2017).
82. Kapila, Y. L. Oral health's inextricable connection to systemic health: Special populations bring to bear multimodal relationships and factors connecting periodontal disease to systemic diseases and conditions. *Periodontology* **87**(1), 11–16. <https://doi.org/10.1111/prd.12398> (2021).
83. Bezamat, M. An updated review on the link between oral infections and atherosclerotic cardiovascular disease with focus on phenomics. *Front. Physiol.* **13**, 1101398. <https://doi.org/10.3389/fphys.2022.1101398> (2022).
84. Yu, H. *et al.* Association of carotid intima-media thickness and atherosclerotic plaque with periodontal status. *J. Dent. Res.* **93**(8), 744–751. <https://doi.org/10.1177/0022034514538973> (2014).

Acknowledgements

We appreciate great support by the Pat and Jim Calhoun Cardiology Center, and Department of Genetics and Genome Sciences, at the UConn School of Medicine, UConn Health; Rutgers Institute for Health, Health Care Policy and Aging Research (IFH), and Rutgers Robert Wood Johnson Medical School (RWJMS), Rutgers Biomedical and Health Sciences (RBHS) at the Rutgers, The State University of New Jersey. We thank members and collaborators of Ahmed Lab at Rutgers (IFH, RWJMS, RBHS) for their support, participation, and contribution to this study. We appreciate all colleagues and institutions who provided direct and indirect insight and expertise that greatly assisted the research and development of this project.

Author contributions

Z.A. led and supervised this study. Z.A. participated in sample collection, sequencing, W.G.S. data processing, quality checking, and downstream analysis. I.M. and S.A. performed functional mutation analysis; H.A. conducted splice mutation analysis; and W.D. implemented JS-MA to calculate J.S.D. scores for genes associated with HF and other CVDs. B.L. supported the study. I.M., H.A., and Z.A. drafted the paper. All authors have participated in writing, review, and have approved it for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-44127-1>.

Correspondence and requests for materials should be addressed to Z.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023