Date: 11/04/2023

The Tenth Evidence-Based Pediatric Update Symposium

PSEG Children's Specialized Hospital, New Brunswick, NJ

Artificial Intelligence to Predict Cardiovascular Disease in Adults and Children

Zeeshan Ahmed, Ph.D.

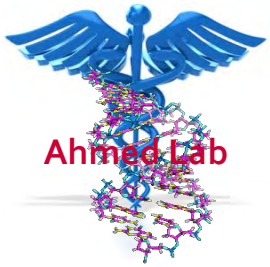
Assistant Professor – Tenure Track

Department of Medicine / Cardiovascular Disease and Hypertension

Institute for Health, Health Care Policy and Aging Research

Robert Wood Johnson Medical School

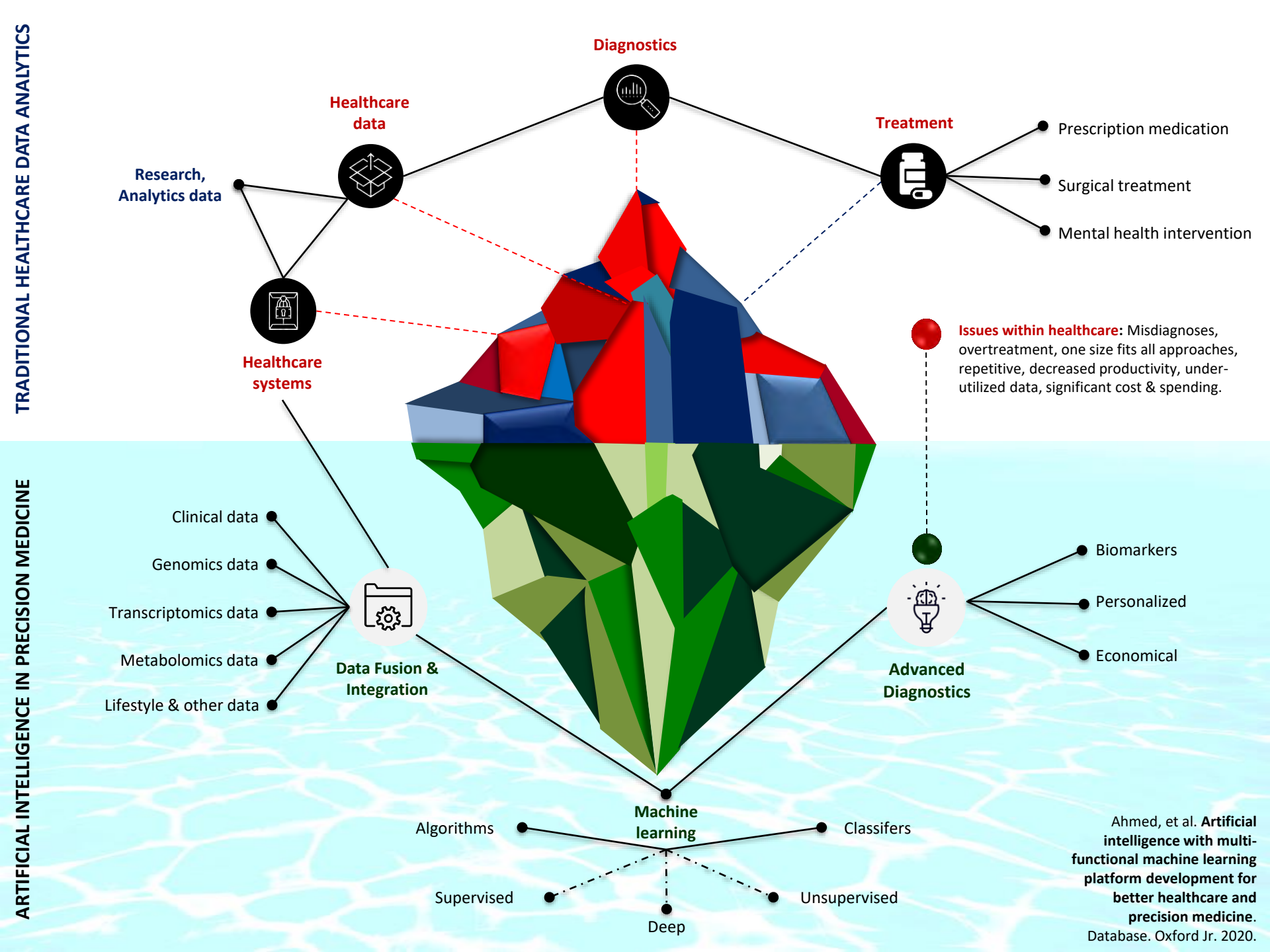
Rutgers University, NJ.



RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

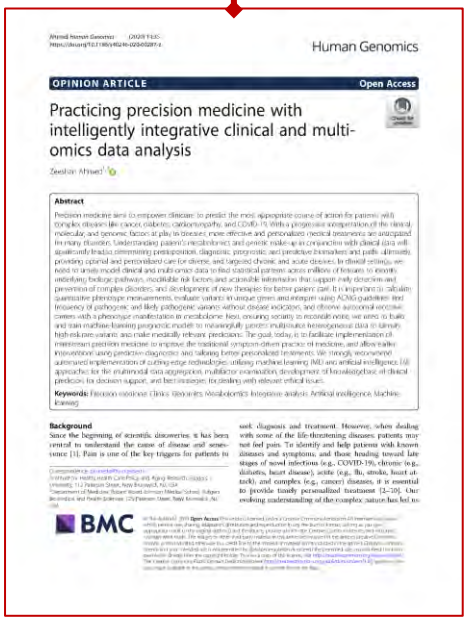
URL, Lab: < https://promis.rutgers.edu/ **>**

URL, Precision Medicine Project: < https://sites.rutgers.edu/precision-medicine/ **>**



Ahmed, et al. **Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine**. Database. Oxford Jr. 2020.

Related – Most Recent Peer Revised Publications by Ahmed Lab


Human Genomics

OPINION ARTICLE Open Access

Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis

Zeeshan Ahmed

Abstract

BMC


Progress in Molecular Biology and Translational Science

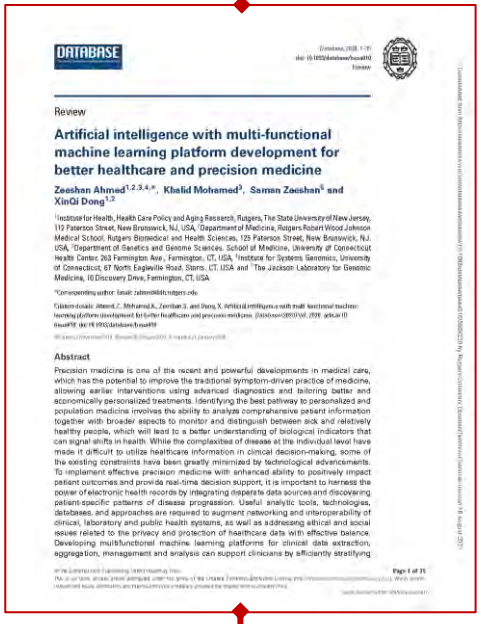
Volume 190, Issue 1, 2022, Pages 101-125

Precision medicine with multi-omics strategies, deep phenotyping, and predictive analysis

Zeeshan Ahmed

Contents

Abstract


DATABASE

Review

Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine

Zeeshan Ahmed, Khalid Mohamed, Saman Zeeshan and XinQi Dong

Abstract


frontiers | Frontiers in Genetics

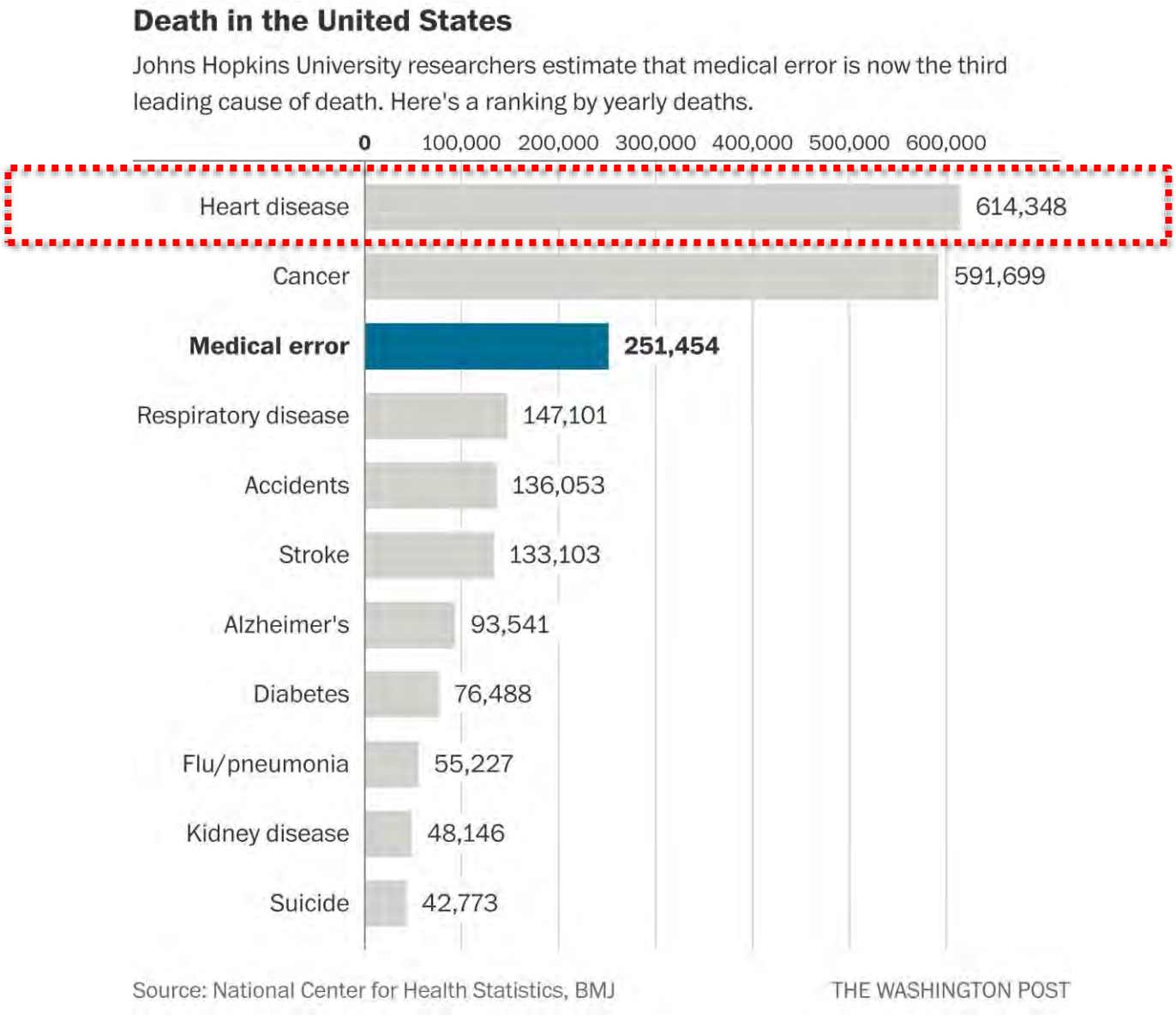
Artificial Intelligence, Healthcare, Clinical Genomics, and Pharmacogenomics Approaches in Precision Medicine

OPEN ACCESS

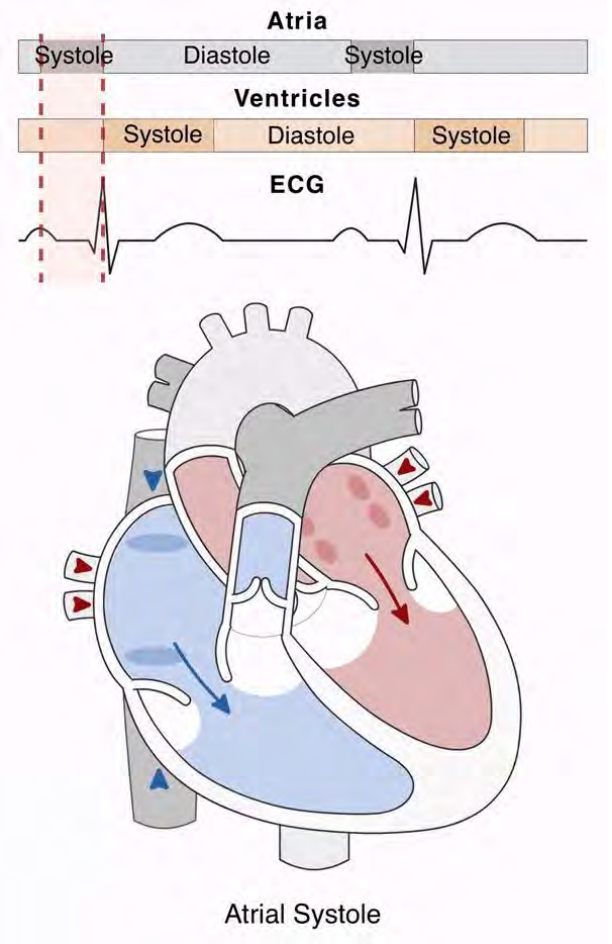
INTRODUCTION

The appropriate utilization of artificial intelligence (AI) and machine learning (ML) methodologies can yield novel understandings of complex traits, enabling improved personalized treatments through predictive analysis and deep phenotyping.

Multi/Disease Research @ Ahmed Lab



Cardiovascular disease (CVD)

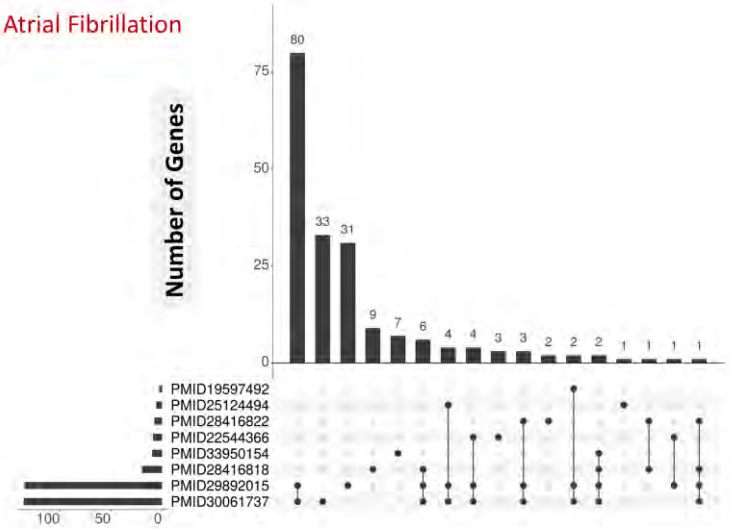


- **Heart Failure (HF)** and **Atrial Fibrillation (AF)** are among the most common manifestations of CVD and contribute to about 45% of all CVD deaths. *(Dickinson et al., 2014)*
  - **AF is an arrhythmic disorder** in the atrium of the heart, which can cause irregular heart rhythms. *(Staerk et al., 2017)*
  - **HF is a chronic disorder**, which weakens heart muscle and affects the regular function of the heart impairing its ability to pump enough oxygen-rich blood. *(Kalogirou et al., 2020)*

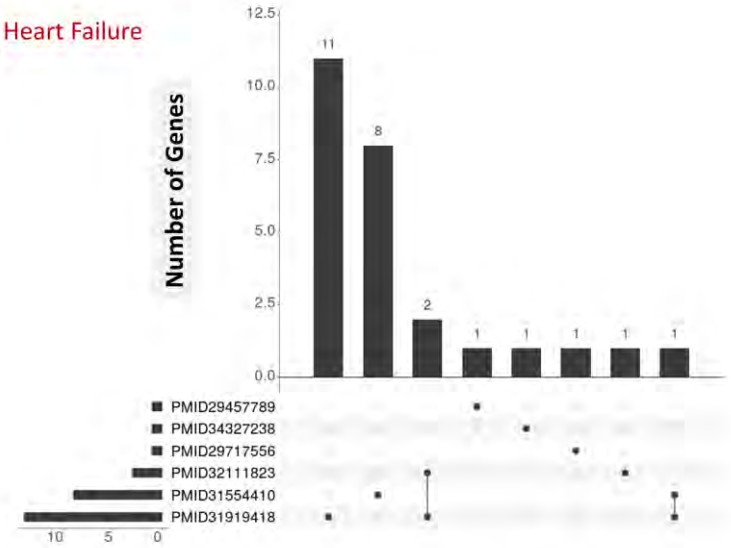
- Due to the **complex nature, progression, inherent genetic makeup, and heterogeneity in CVDs**, personalized treatments are critical for CVD patients.
- To improve the **deciphering of CVD mechanisms**, it will be necessary to systematically investigate known and identify novel genes that are responsible for the CVD development.
- Studying genetic insight with the application of **Artificial Intelligence (AI), Machine Learning (ML),** and state-of-the-art bioinformatics approaches can accelerate the processes of discovering disease causing variants and decode genetics of complex phenotypes to predict, prevent, and treat CVD.

Bioinformatics & CVD

A. Atrial Fibrillation



B. Heart Failure



Patel et al. Human Genomics (2023) 17:47
https://doi.org/10.1186/s40246-023-00498-0

Human Genomics

REVIEW Open Access

Genomic approaches to identify and investigate genes associated with atrial fibrillation and heart failure susceptibility

Kush Ketan Patel[1†], Cynthia Venkatesan[1†], Habiba Abdelhalim[1], Saman Zeeshan[2], Yuichiro Arima[3], Suvi Linna-Kuosmanen[4,5,6] and Zeeshan Ahmed[7,8*]

Abstract

Atrial fibrillation (AF) and heart failure (HF) contribute to about 45% of all cardiovascular disease (CVD) deaths in the USA and around the globe. Due to the complex nature, progression, inherent genetic makeup, and heterogeneity of CVDs, personalized treatments are believed to be critical. To improve the deciphering of CVD mechanisms, we need to deeply investigate well-known and identify novel genes that are responsible for CVD development. With the advancements in sequencing technologies, genomic data have been generated at an unprecedented pace to foster translational research. Correct application of bioinformatics using genomic data holds the potential to reveal the genetic underpinnings of various health conditions. It can help in the identification of causal variants for AF, HF, and other CVDs by moving beyond the one-gene one-disease model through the integration of common and rare variant association, the expressed genome, and characterization of comorbidities and phenotypic traits derived from the clinical information. In this study, we examined and discussed variable genomic approaches investigating genes associated with AF, HF, and other CVDs. We collected, reviewed, and compared high-quality scientific literature published between 2009 and 2022 and accessible through PubMed/NCBI. While selecting relevant literature, we mainly focused on identifying genomic approaches involving the integration of genomic data; analysis of common and rare genetic variants; metadata and phenotypic details; and multi-ethnic studies including individuals from ethnic minorities, and European, Asian, and American ancestries. We found 190 genes associated with AF and 26 genes linked to HF. Seven genes had implications in both AF and HF, which are *SYNPO2L*, *TTN*, *MTSS1*, *SCN5A*, *PITX2*, *KLHL3*, and *AGAP5*. We listed our conclusion, which include detailed information about genes and SNPs associated with AF and HF.

Keywords Genes, Genetic loci, Heart failure, Atrial fibrillation, Cardiovascular diseases, Genomics, Multi-OMICS

†Kush Ketan Patel and Cynthia Venkatesan are equally contributing first authors.

*Correspondence:
Zeeshan Ahmed
zahmed@ifh.rutgers.edu
[1] Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson St, New Brunswick, NJ, USA
[2] Rutgers Cancer Institute of New Jersey, Rutgers University, 195 Little Albany St, New Brunswick, NJ, USA
[3] Developmental Cardiology Laboratory, International Research Center for Medical Sciences, Kumamoto University, 2-2-1 Honjo, Kumamoto City, Kumamoto, Japan
[4] A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, 70211 Kuopio, Finland
[5] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[7] Department of Genetics and Genome Sciences, UConn Health, 400 Farmington Ave, Farmington, CT, USA
[8] Department of Medicine, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson St, New Brunswick, NJ, USA

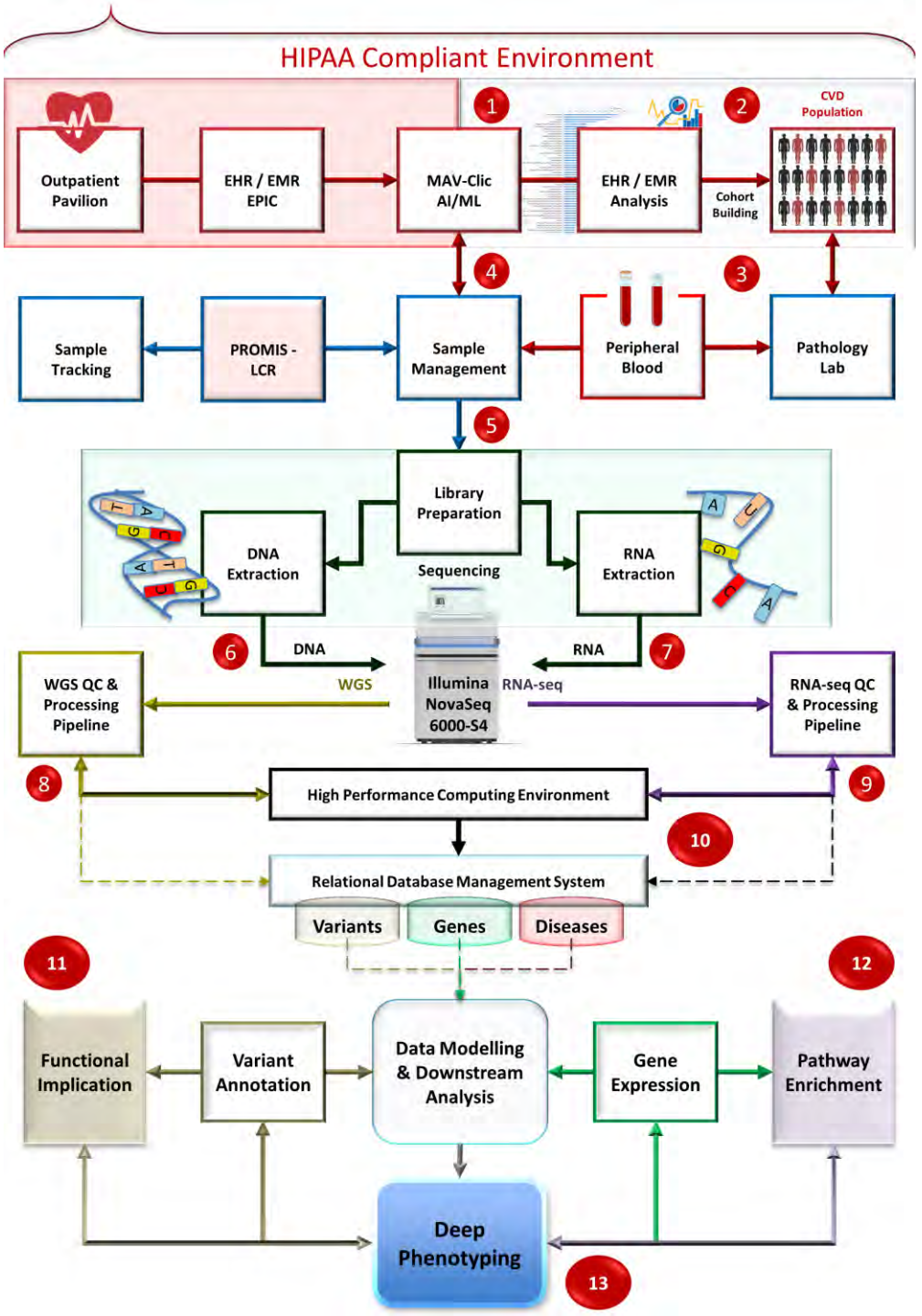
CVD Study Design

Bioinformatics Analyses:

1. EHR extraction from EPIC
2. CVD cohort building
3. Sample collection
4. Sample management & tracking
5. Library preparation
6. WGS data generation
7. RNA-seq data generation
8. WGS data QC and processing
9. RNA-seq data QC and processing
10. Gene-disease data annotation
11. Variant analysis & validation (WGS)
12. Gene expression analysis (RNA-seq)

AI/ML Analyses:

- Predict disease with high accuracy
- Novel biomarkers discovery
- Intelligent Gene Score



CVD Cohort Building & MAV-clis

JAMIA Open, 2(1), 2019, 23–28
doi: 10.1093/jamiaopen/ooy052
Advance Access Publication Date: 29 December 2018
Database Notes

AMIA OXFORD

Database Notes

MAV-clic: management, analysis, and visualization of clinical data

Zeeshan Ahmed,[1] Minjung Kim[2] and Bruce T. Liang[3]

[1]Department of Genetics and Genome Sciences, Institute for Systems Genomics, School of Medicine, University of Connecticut Health Center, Farmington, Connecticut, USA, [2]The Pat and Jim Calhoun Cardiology Center, School of Medicine, University of Connecticut Health Center, Farmington, Connecticut, USA and [3]Ray Neag Distinguished Professor of Cardiovascular Biology and Medicine, Director Pat and Jim Calhoun Cardiology Center, Dean UConn School of Medicine, University of Connecticut Health Center, Farmington, Connecticut, USA

Corresponding Author: Zeeshan Ahmed, PhD, Assistant Professor & Assistant Director Bioinformatics Precision Medicine Program, Department of Genetics and Genome Sciences, School of Medicine, University of Connecticut Health Center, 400 Farmington Ave, Farmington, CT 06032, USA (zahmed@uchc.edu)

Received 30 January 2018; Revised 18 July 2018; Editorial Decision 3 November 2018; Accepted 22 November 2018

ABSTRACT

Objectives: Develop a multifunctional analytics platform for efficient management and analysis of healthcare data.

Materials and Methods: Management, Analysis, and Visualization of Clinical Data (MAV-clic) is a Health Insurance Portability and Accountability Act of 1996 (HIPAA)-compliant framework based on the Butterfly Model. MAV-clic extracts, cleanses, and encrypts data then restructures and aggregates data in a deidentified format. A graphical user interface allows query, analysis, and visualization of clinical data.

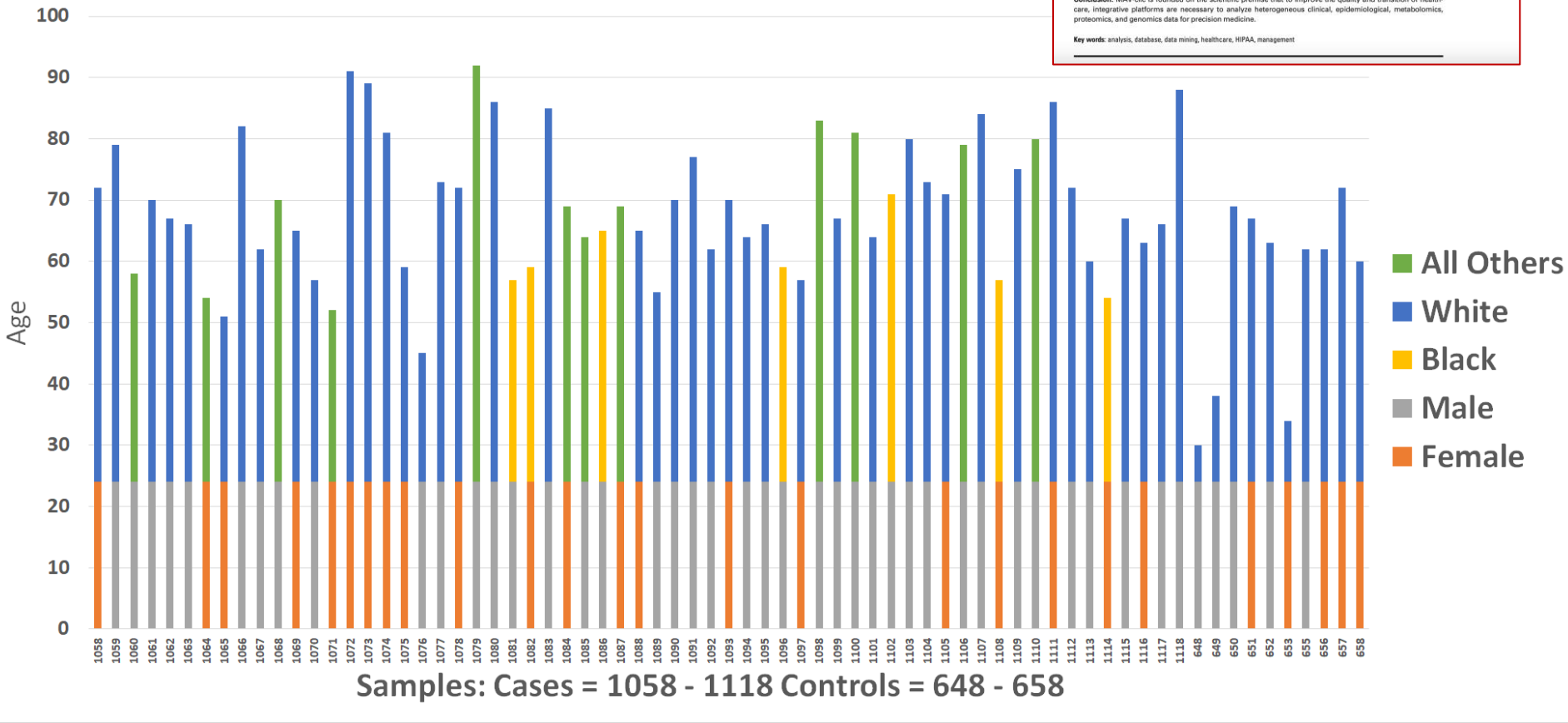
Results: MAV-clic manages healthcare data for over 800 000 subjects at UConn Health. Three analytic capabilities of MAV-clic include: creating cohorts based on specific criteria; performing measurement analysis of subjects with a specific diagnosis and medication; and calculating measure outcomes of subjects over time.

Discussion: MAV-clic supports clinicians and healthcare analysts by efficiently stratifying subjects to understand specific scenarios and optimize decision making.

Conclusion: MAV-clic is founded on the scientific premise that to improve the quality and transition of healthcare, integrative platforms are necessary to analyze heterogeneous clinical, epidemiological, metabolomics, proteomics, and genomics data for precision medicine.

Key words: analysis, database, data mining, healthcare, HIPAA, management

The CVD cohort include **40 male and 21 female individuals (n=61)**, aged between 45 to 92, with self-described race (42 Whites, 7 Blacks or African Americans, 1 Asian, and 11 of unknown race). In addition, the PI has built a control set, which included healthy individuals (n=10); 5 males and 5 females; out of which 9 were White race and 1 unknown race; aged between 28 to 78.



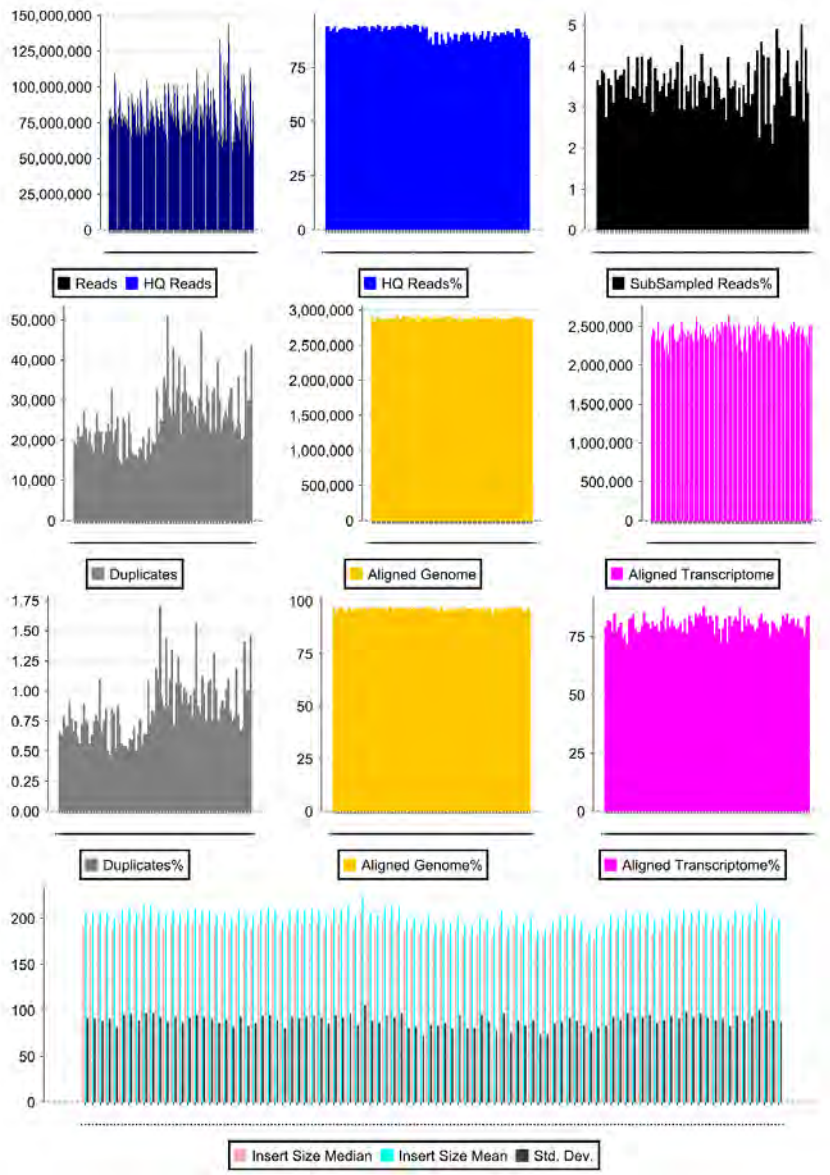
Sequence Data Generation

Illumina NovaSeq 6000-S4

RNA-seq → Expression Analysis

WGS → Variant Analysis

QC Report using PROMS-MED



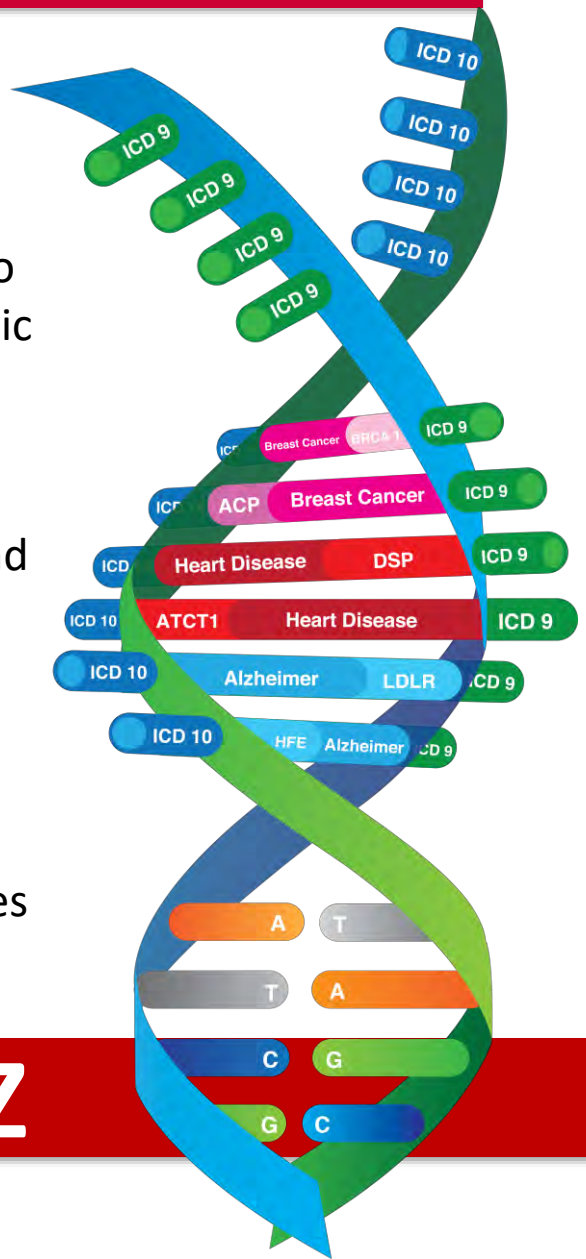
RNA-seq Driven Gene Expression and Gene-Disease Annotation

- **RNA-seq** has become the most used method for gene expression analysis

- **Gene expression** analysis is a widely adopted method to identify abnormalities in normal function and physiologic regulation.

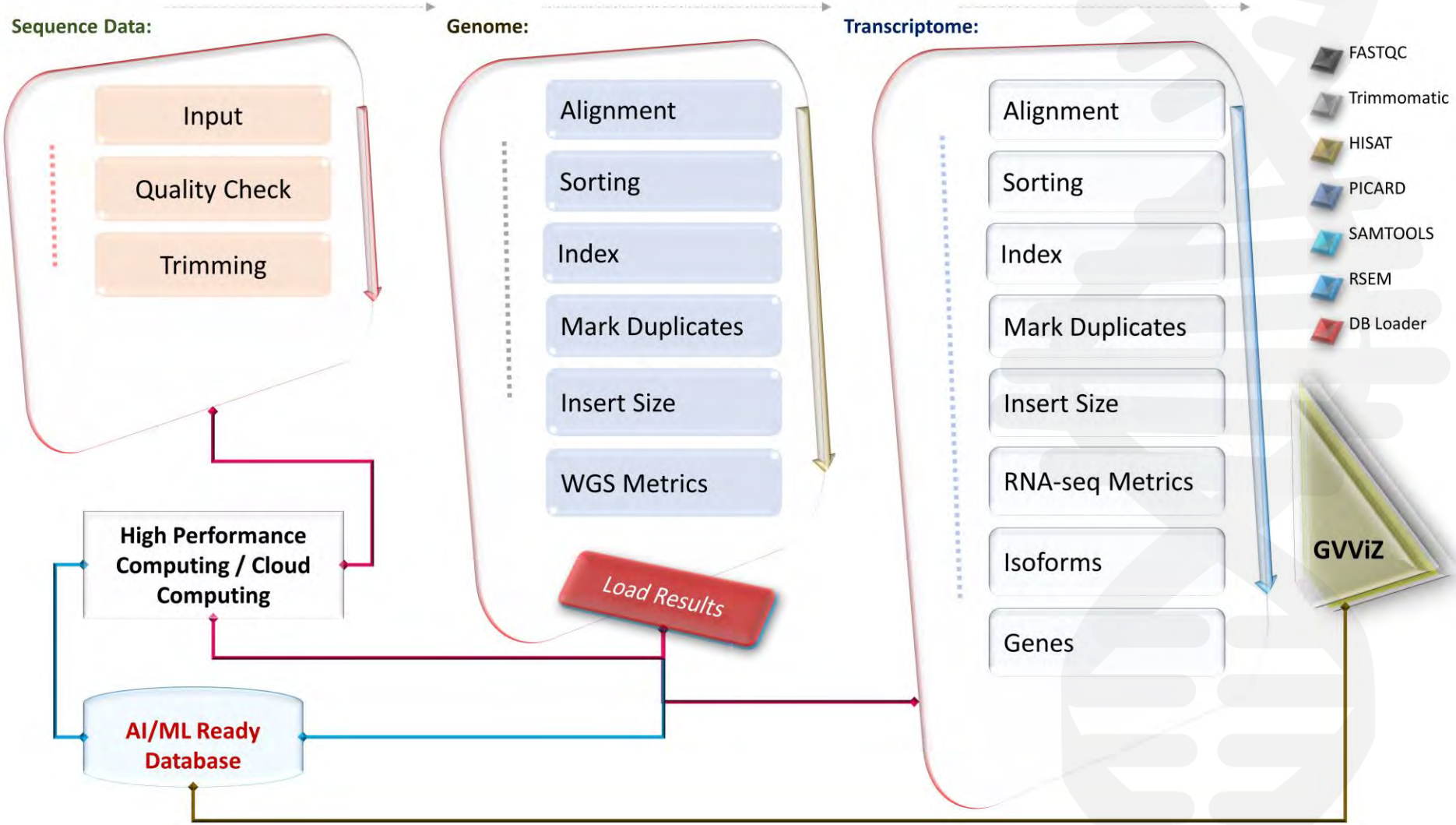
- Findable, accessible, interactive, and reusable (FAIR) bioinformatics platform for **RNA-seq**-driven variable and complex **gene-disease data annotation and expression analysis** with a dynamic heat map visualization.

- It supports **transcriptomic profiling and expression analyses** to identify measure and compare genes and transcripts in multiple conditions, and in different tissues and individuals.



GVViZ

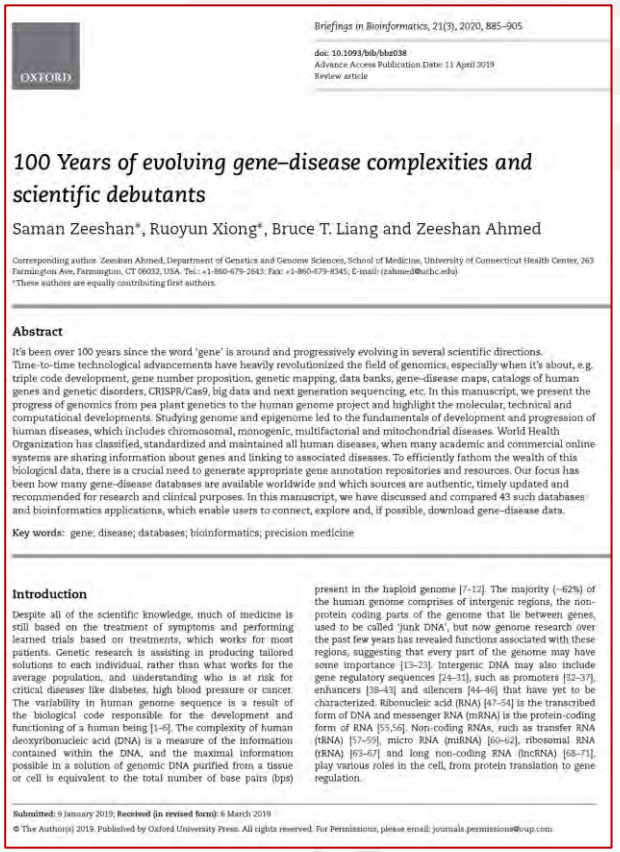
GVViZ – RNA-seq data quality checking and processing



Ahmed, Z., et al. (2021). **Advancing clinical genomics and precision medicine with GVViZ: FAIR bioinformatics platform for variable gene-disease annotation, visualization, and expression analysis.** Human Genomics. 15(1), 37. PMID: 34174938.

Type	Names	Gene → Disease	Multiple inputs	ICD	Free
Databases	ClinVar				
	CNVD				
	Cochrane Library				
	Cosmic				
	dbSNP				
	DGIdb				
	Disease Ontology				
	Diseasecard				
	DiseaseEnhancer				
	DISEASES				
	DrugBank				
	ExPASy				
	FDA Approved Drugs				
	FMA				
	GARD				
	GeneCards				
	GeneGo (Thomson Reuters)				
	GeneReviews				
	Genetics Home Reference				
	GenomeRNAi				
	GEO DataSets				
	GO				
	GTR				
	HMDB				
	HPO				
	IUPHAR				
	KEGG				
	LifeMap				
	LncRNADisease				
	LOVD				
	MedGen				
	MedlinePlus				
	MeSH				
	MGI				
	miR2Disease				
	NCBI Bookshelf				
	NCI				
	NCIt				
	NDF-RT				
	NIH Clinical Center				
	NINDS				
	Novus Biologicals				
	OMIM				
	Orphanet				
	R&D Systems				
	Reactome				
	Sino Biological				
	SNOMED-CT				
	UniProtKB/Swiss-Prot				
	Tocris				
Tools	MSigDB				
	DigSee				
	DAVID				
	DisGeNET				
	HGMD				
	Gene2Function				
	SwissVar				
	eDGAR				
	Gene Analytics				

Existing Gene-Disease Databases & Bioinformatics Tools.



Briefings in Bioinformatics, 21(3), 2020, 885–905

doi: 10.1093/bib/bbz038
Advance Access Publication Date: 11 April 2019
Review article

100 Years of evolving gene–disease complexities and scientific debutants

Saman Zeeshan*, Ruoyun Xiong*, Bruce T. Liang and Zeeshan Ahmed

Corresponding author. Zeeshan Ahmed, Department of Genetics and Genome Sciences, School of Medicine, University of Connecticut Health Center, 263 Farmington Ave, Farmington, CT 06032, USA. Tel.: +1-860-679-2643; Fax: +1-860-679-8345; E-mail: (zahmed@uchc.edu)
*These authors are equally contributing first authors.

Abstract

It's been over 100 years since the word 'gene' is around and progressively evolving in several scientific directions. Time-to-time technological advancements have heavily revolutionized the field of genomics, especially when it's about, e.g. triple code development, gene number proposition, genetic mapping, data banks, gene–disease maps, catalogs of human genes and genetic disorders, CRISPR/Cas9, big data and next generation sequencing, etc. In this manuscript, we present the progress of genomics from pea plant genetics to the human genome project and highlight the molecular, technical and computational developments. Studying genome and epigenome led to the fundamentals of development and progression of human diseases, which includes chromosomal, monogenic, multifactorial and mitochondrial diseases. World Health Organization has classified, standardized and maintained all human diseases, when many academic and commercial online systems are sharing information about genes and linking to associated diseases. To efficiently fathom the wealth of this biological data, there is a crucial need to generate appropriate gene annotation repositories and resources. Our focus has been how many gene–disease databases are available worldwide and which sources are authentic, timely updated and recommended for research and clinical purposes. In this manuscript, we have discussed and compared 43 such databases and bioinformatics applications, which enable users to connect, explore and, if possible, download gene–disease data.

Key words: gene; disease; databases; bioinformatics; precision medicine

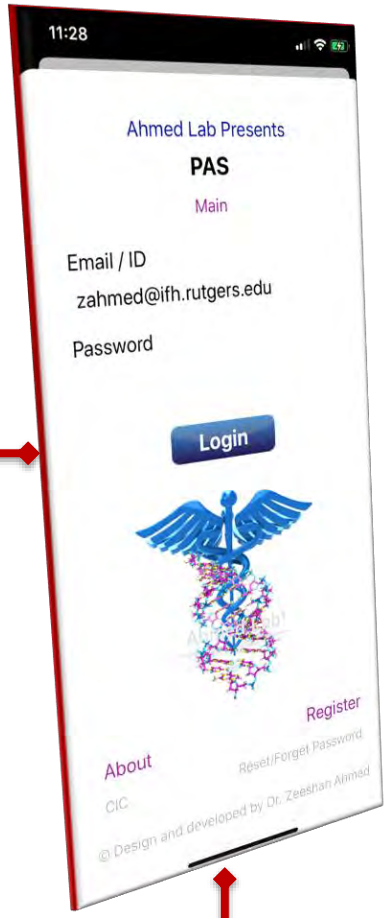
Introduction

Despite all of the scientific knowledge, much of medicine is still based on the treatment of symptoms and performing learned trials based on treatments, which works for most patients. Genetic research is assisting in producing tailored solutions to each individual, rather than what works for the average population, and understanding who is at risk for critical diseases like diabetes, high blood pressure or cancer. The variability in human genome sequence is a result of the biological code responsible for the development and functioning of a human being [1–6]. The complexity of human deoxyribonucleic acid (DNA) is a measure of the information contained within the DNA, and the maximal information possible in a solution of genomic DNA purified from a tissue or cell is equivalent to the total number of base pairs (bps) present in the haploid genome [7–12]. The majority (~62%) of the human genome comprises of intergenic regions, the non-protein coding parts of the genome that lie between genes, used to be called 'junk DNA', but now genome research over the past few years has revealed functions associated with these regions, suggesting that every part of the genome may have some importance [13–23]. Intergenic DNA may also include gene regulatory sequences [24–31], such as promoters [32–37], enhancers [38–43] and silencers [44–46] that have yet to be characterized. Ribonucleic acid (RNA) [47–54] is the transcribed form of DNA and messenger RNA (mRNA) is the protein-coding form of RNA [55,56]. Non-coding RNAs, such as transfer RNA (tRNA) [57–59], micro RNA (miRNA) [60–62], ribosomal RNA (rRNA) [63–67] and long non-coding RNA (lncRNA) [68–71], play various roles in the cell, from protein translation to gene regulation.

Submitted: 9 January 2019; **Received (in revised form):** 6 March 2019

Clinical – Genomics Database Development!

- **Genomics**
  - *Authentic Genes*
  - *Germline SNPs*
  - *Somatic SNPs*
- **Clinical Genomics**
  - *Gene to Disease*
  - *Germline SNP to Disease*
  - *Somatic SNP to Disease*



RESEARCH ARTICLE

WILEY

Human gene and disease associations for clinical-genomics and precision medicine research

Zeeshan Ahmed[1,2] | Saman Zeeshan[3] | Dinesh Mendhe[1] | XinQi Dong[1,2]

[1]Institute for Health, Health Care Policy and Aging Research, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

[2]Department of Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, New Brunswick, New Jersey, USA

[3]Rutgers Career Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

Correspondence
Zeeshan Ahmed, Institute for Health, Health Care Policy and Aging Research, Rutgers, The State University of New Jersey | 112 Paterson Street, New Brunswick, NJ 08901, USA.
Email: zahmed@ifh.rutgers.edu

Abstract

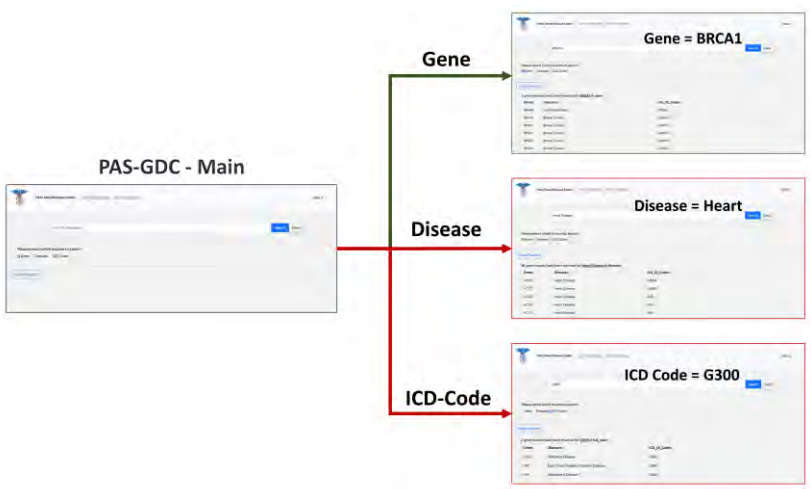
We are entering the era of personalized medicine in which an individual's genetic makeup will eventually determine how a doctor can tailor his or her therapy. Therefore, it is becoming critical to understand the genetic basis of common diseases, for example, which genes predispose and rare genetic variants contribute to diseases, and so on. Our study focuses on helping researchers, medical practitioners, and pharmacists in having a broad view of genetic variants that may be implicated in the likelihood of developing certain diseases. Our focus here is to create a comprehensive database with mobile access to all available, authentic and actionable genes, SNPs, and classified diseases and drugs collected from different clinical and genomics databases worldwide, including Ensembl, GenCode, ClinVar, GeneCards, DISEASES, HGMD, OMIM, GTR, CNVD, Novoseek, Swiss-Prot, LncRNADisease, Orphanet, GWAS Catalog, SwissVar, COSMIC, WHO, and FDA. We present a new cutting-edge gene-SNP-disease-drug mobile database with a smart phone application, integrating information about classified diseases and related genes, germline and somatic mutations, and drugs. Its database includes over 59 000 protein-coding and noncoding genes; over 67 000 germline SNPs and over a million somatic mutations reported for over 19 000 protein-coding genes located in over 1000 regions, published with over 3000 articles in over 415 journals available at the PUBMED; over 80 000 ICDs; over 123 000 NDCs; and over 100 000 classified gene-SNP-disease associations. We present an application that can provide new insights into the information about genetic basis of human complex diseases and contribute to assimilating genomic with phenotypic data for the availability of gene-based designer drugs, precise targeting of molecular fingerprints for tumor, appropriate drug therapy, predicting individual susceptibility to disease, diagnosis, and treatment of rare illnesses are all a few of the many transformations expected in the decade to come.

KEYWORDS
clinical-genomics, database, diseases, drugs, genes, germline mutations, precision medicine, somatic mutations

URL: https://apps.apple.com/us/app/pas/id1447589546

Clinical – Genomics Database Development!

Genes approved by the American College of Medical Genetics (ACMG)



URL: https://promis.rutgers.edu/pas

Database, 2023, 1–10
DOI: https://doi.org/10.1093/database/baad033
Database tool

DATABASE

Integrated ACMG-approved genes and ICD codes for the translational research and precision medicine

Raghunandan Wable[1,‡], Achuth Suresh Nair[1,‡], Anirudh Pappu[1,‡], Widnie Pierre-Louis[1,‡], Habiba Abdelhalim[1,‡], Khushbu Patel[1,‡], Dinesh Mendhe[1], Shreyas Bolla[1], Sahil Mittal[1] and Zeeshan Ahmed[1,2,*]

[1]Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson St, New Brunswick, NJ 08901, USA
[2]Department of Medicine, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson St, New Brunswick, NJ 08901, USA

*Corresponding author: Tel: +848-932-5866; Fax: +732-932-0069; Email: zahmed@ifh.rutgers.edu
‡Those authors contributed equally to this work.

Citation details: Wable, R., Nair, A.S., Pappu, A. *et al.* Integrated ACMG-approved genes and ICD codes for the translational research and precision medicine. *Database* (2023) Vol. 2023: article ID baad033; DOI: https://doi.org/10.1093/database/baad033

Abstract

A timely understanding of the biological secrets of complex diseases will ultimately benefit millions of individuals by reducing the high risks for mortality and improving the quality of life with personalized diagnoses and treatments. Due to the advancements in sequencing technologies and reduced cost, genomics data are developing at an unmatched pace and levels to foster translational research and precision medicine. Over 10 million genomics datasets have been produced and publicly shared in 2022. Diverse and high-volume genomics and clinical data have the potential to broaden the scope of biological discoveries and insights by extracting, analyzing and interpreting the hidden information. However, the current and still unresolved challenges include the integration of genomic profiles of the patients with their medical records. The definition of disease in genomics medicine is simplified, whereas in the clinical world, diseases are classified, identified and adopted with their International Classification of Diseases (ICD) codes, which are maintained by the World Health Organization. Several biological databases have been produced, which include information about human genes and related diseases. However, still, there is no database that exists, which can precisely link clinical codes with relevant genes and variants to support genomic and clinical data integration for clinical and translational medicine. In this project, we focused on the development of an annotated gene–disease–code database, which is accessible through an online, cross-platform and user-friendly application, i.e. PROMIS-APP-SUITE-Gene-Disease-Code. However, our scope is limited to the integration of ICD-9 and ICD-10 codes with the list of genes approved by the American College of Medical Genetics and Genomics. The results include over 17 000 diseases and 4000 ICD codes, and over 11 000 gene–disease–code combinations.

Database URL: https://promis.rutgers.edu/pas/

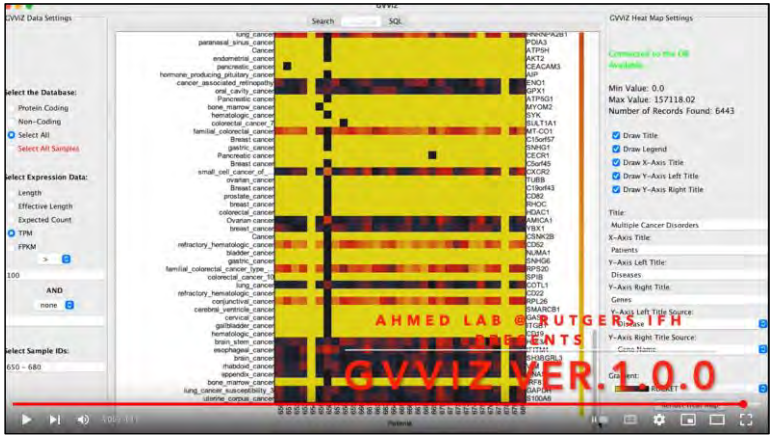
Introduction

Symptom-driven medicine has become the domain of medical research in the past decade (1, 2). However, some challenges arise when focusing on the symptoms rather than the disease. Patients with life-threatening diseases might not feel pain and seek professional help. Thus, personalized treatment to help manage and identify those patients using precision medicine is needed to effectively diagnose and provide the most optimal actions needed for such patients (3–5). Precision medicine is a multi-disciplinary field that utilizes the clinical and multi-omics data of an individual to create patient-specific treatment plans and diagnoses (4, 7, 8). Clinical data are most familiar to clinicians and patients as a medium that communicates personal and health information between the provider and the patient. Genomic information is stored within various databases that include but are not limited to ClinVar, CNVD, Cochrane Library, Disease Ontology and Disease Enhancer, which allow for gene annotation (4). However, there is a lack of standardized, comprehensive databases that consolidate the known gene–disease relationships. Furthermore, there is no known database that connects International Classification of Diseases (ICD), mediated by the World Health Organization (WHO), with the list of 73 genes compiled by the American College of Medical Genetics and Genomics (ACMG), whose mutations are known to be causative of disorders and diseases (9).

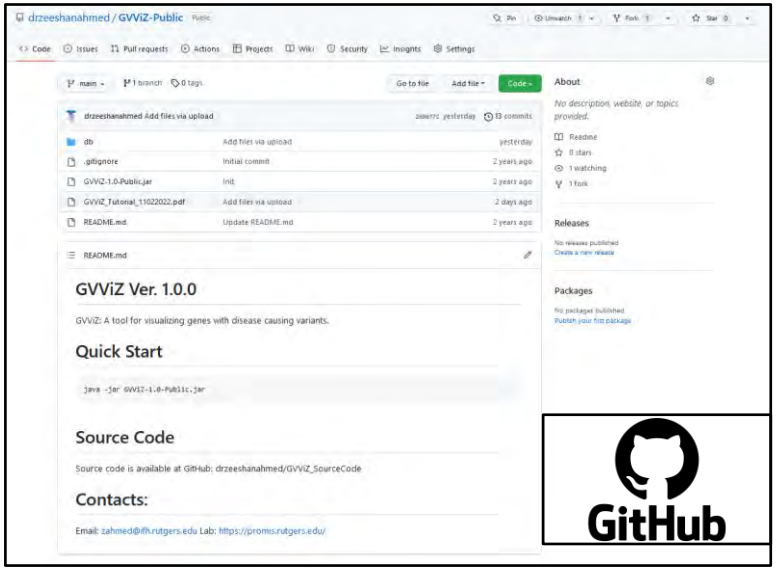
The evolution from the first use of the word 'gene' to our current understanding has launched a new scientific age. On an introductory level, the chemical structure of the genome is in the form of deoxyribose nucleic acid (DNA), which is composed of a double helix with pairs of nucleotides connected by hydrogen bonds (1, 10, 11). These alternating patterns of nucleotides (adenine, cytosine, guanine and thymine) encode the instructions for all the proteins in our body,

Received 4 December 2022; Revised 19 February 2023; Accepted 14 April 2023

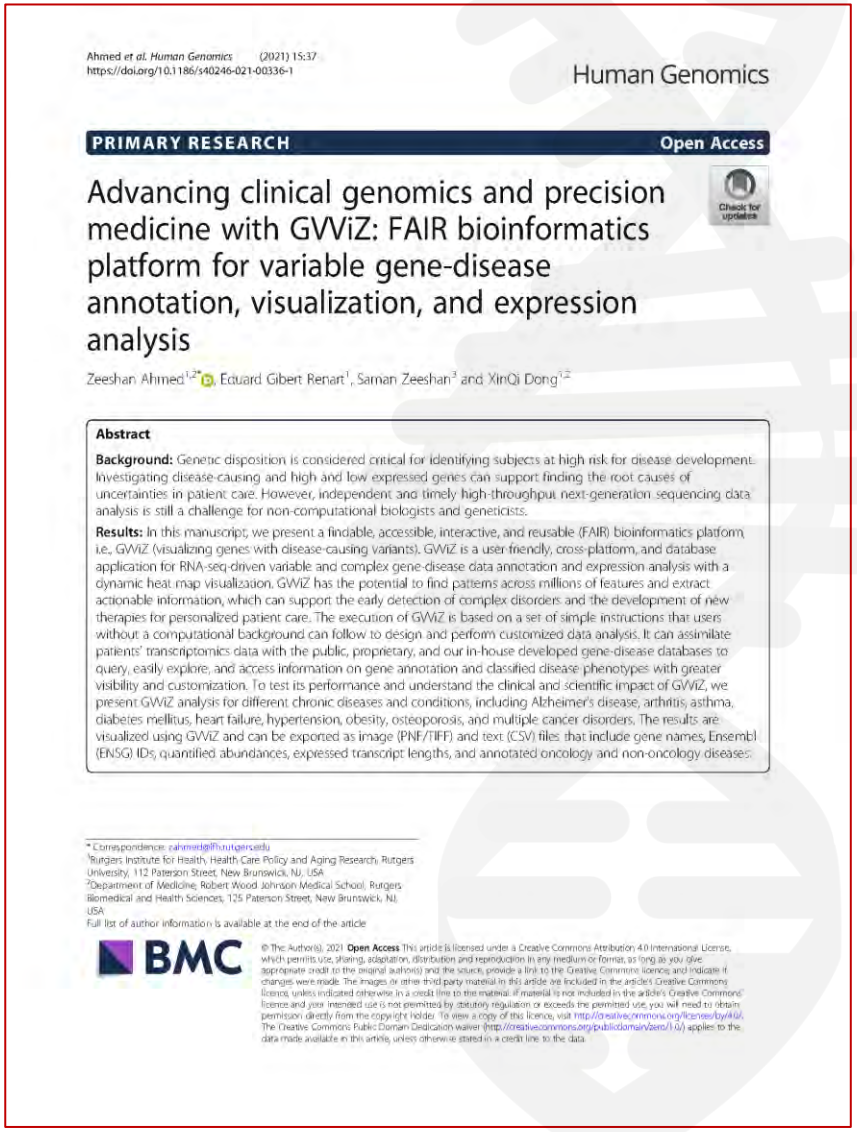
GVViZ – Demo and Download Information & Publication



URL: https://www.youtube.com/watch?v=x0RroYpk8Nw



URL: https://github.com/drzeeshanahmed/GVViZ-Public



Ahmed *et al. Human Genomics* (2021) 15:37
https://doi.org/10.1186/s40246-021-00336-1

Human Genomics

PRIMARY RESEARCH — Open Access

Advancing clinical genomics and precision medicine with GVViZ: FAIR bioinformatics platform for variable gene-disease annotation, visualization, and expression analysis

Zeeshan Ahmed[1,2*], Eduard Gibert Renart[1], Saman Zeeshan[3] and XinQi Dong[1,2]

Abstract

Background: Genetic disposition is considered critical for identifying subjects at high risk for disease development. Investigating disease-causing and high and low expressed genes can support finding the root causes of uncertainties in patient care. However, independent and timely high-throughput next-generation sequencing data analysis is still a challenge for non-computational biologists and geneticists.

Results: In this manuscript, we present a findable, accessible, interactive, and reusable (FAIR) bioinformatics platform i.e., GVViZ (visualizing genes with disease-causing variants). GVViZ is a user-friendly, cross-platform, and database application for RNA-seq-driven variable and complex gene-disease data annotation and expression analysis with a dynamic heat map visualization. GVViZ has the potential to find patterns across millions of features and extract actionable information, which can support the early detection of complex disorders and the development of new therapies for personalized patient care. The execution of GVViZ is based on a set of simple instructions that users without a computational background can follow to design and perform customized data analysis. It can assimilate patients' transcriptomics data with the public, proprietary, and our in-house developed gene-disease databases to query, easily explore, and access information on gene annotation and classified disease phenotypes with greater visibility and customization. To test its performance and understand the clinical and scientific impact of GVViZ, we present GVViZ analysis for different chronic diseases and conditions, including Alzheimer's disease, arthritis, asthma, diabetes mellitus, heart failure, hypertension, obesity, osteoporosis, and multiple cancer disorders. The results are visualized using GVViZ and can be exported as image (PNF/TIFF) and text (CSV) files that include gene names, Ensembl (ENSG) IDs, quantified abundances, expressed transcript lengths, and annotated oncology and non-oncology diseases.

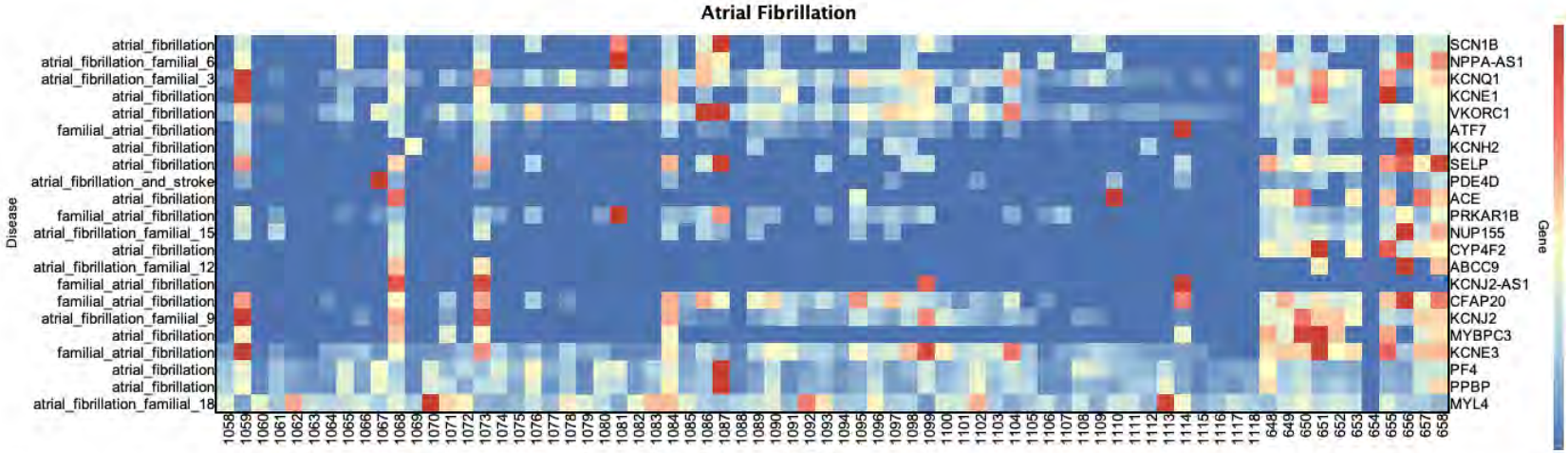
* Correspondence: zahmed@ifh.rutgers.edu
[1]Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick, NJ, USA
[2]Department of Medicine, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson Street, New Brunswick, NJ, USA
Full list of author information is available at the end of the article

BMC

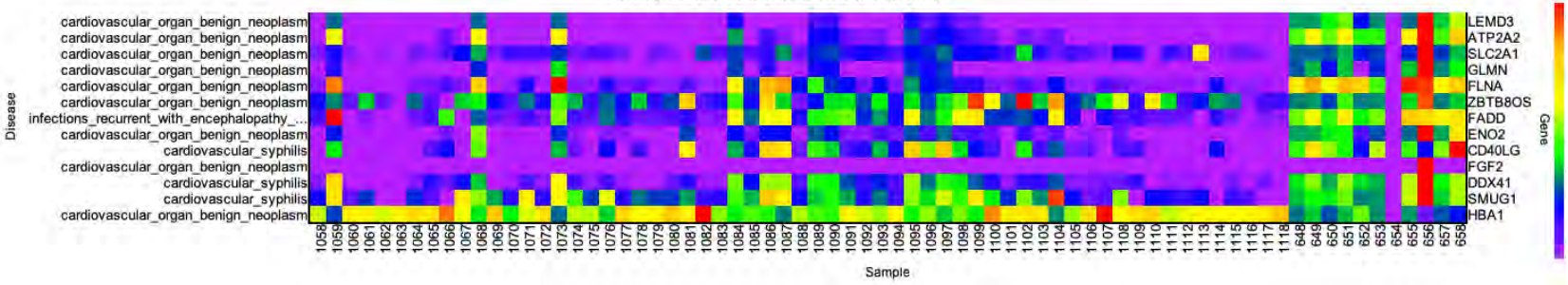
GVViz & Gene-disease annotation & expression analysis for CVDs

A.

B.



C.

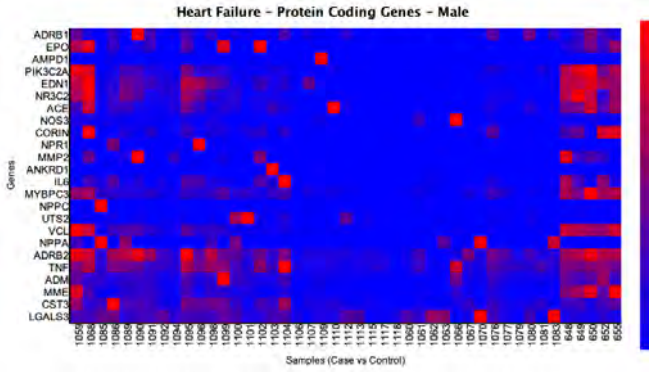


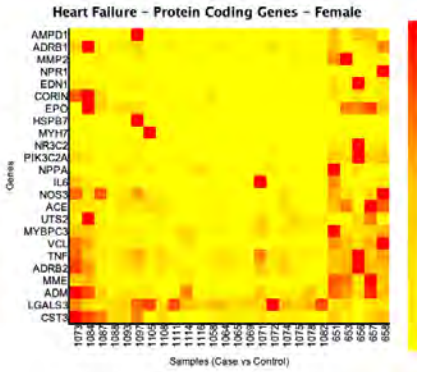
Gender-based gene expression analysis

MALE

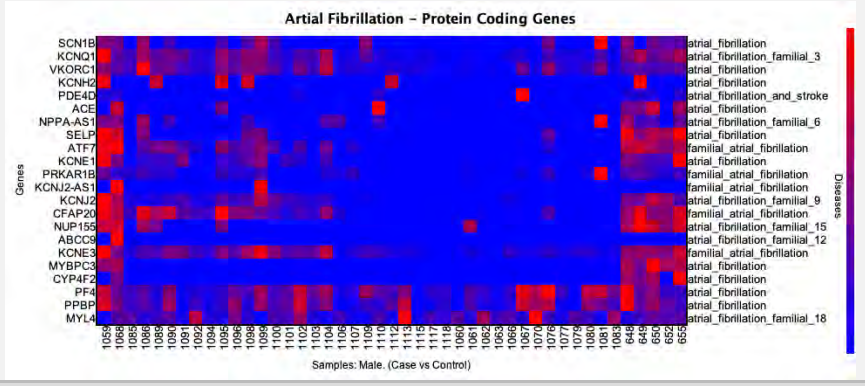
FEMALE

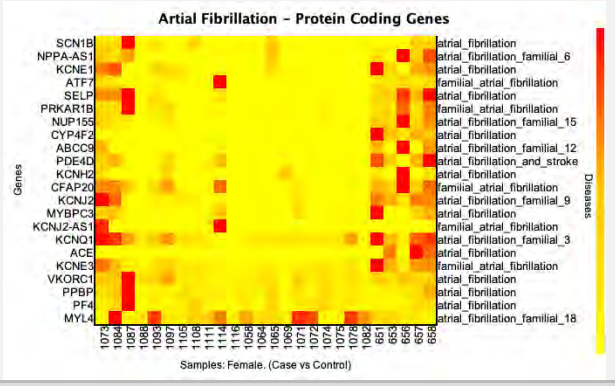
A. HF



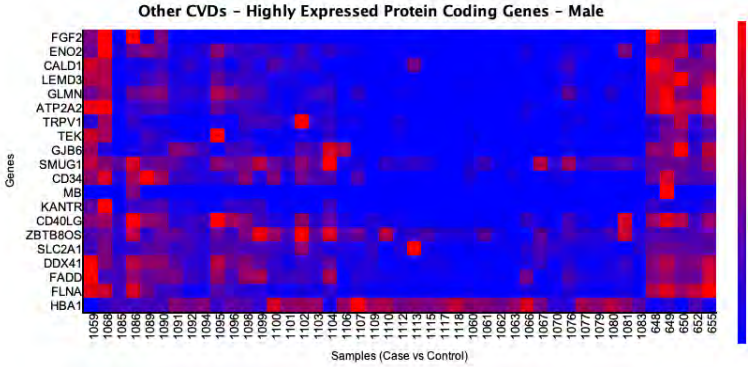


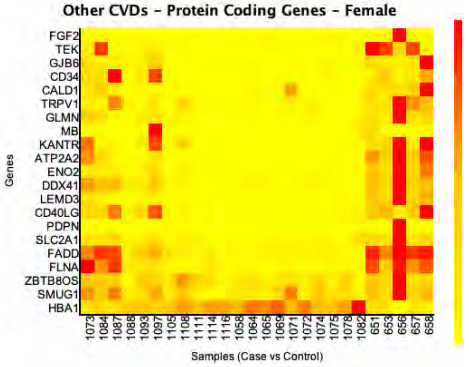
B. AF



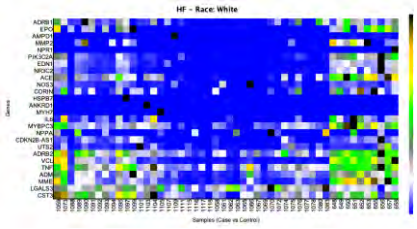
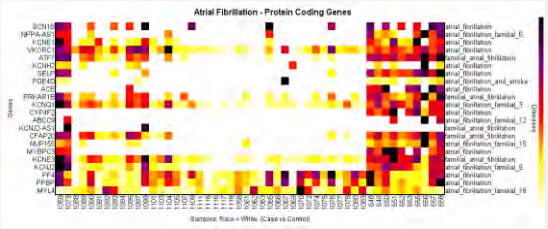
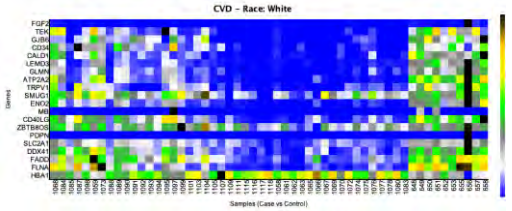
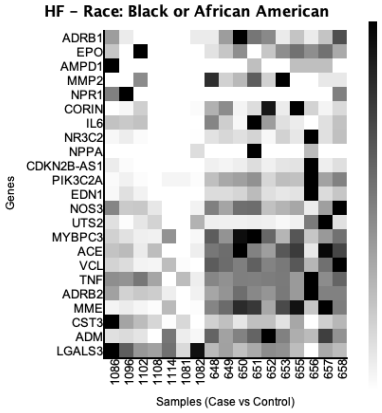
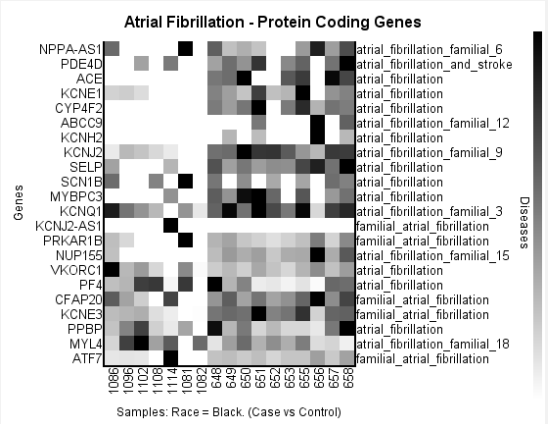
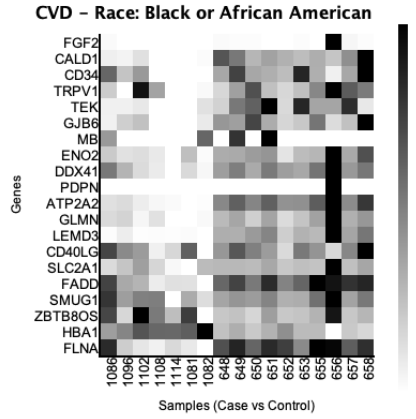
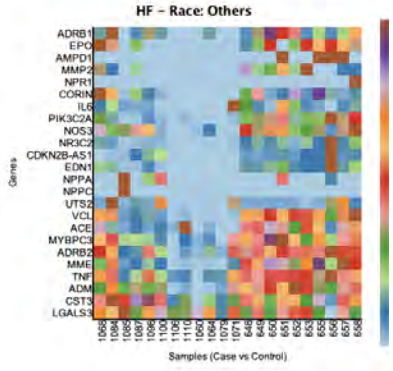
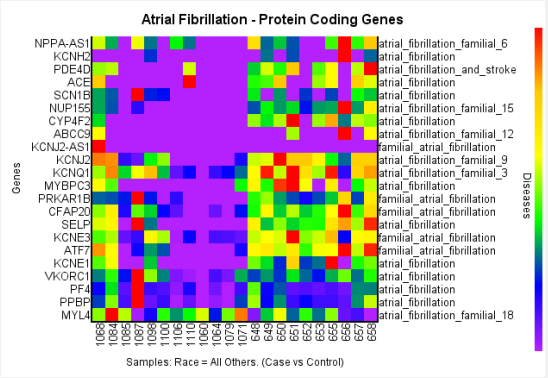
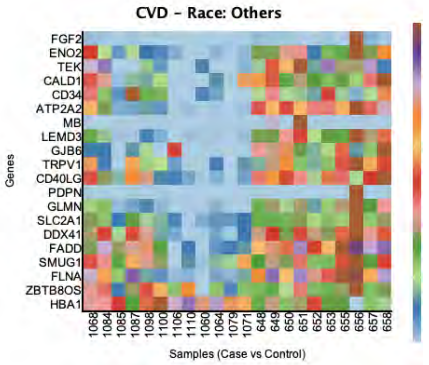


C. other CVDs

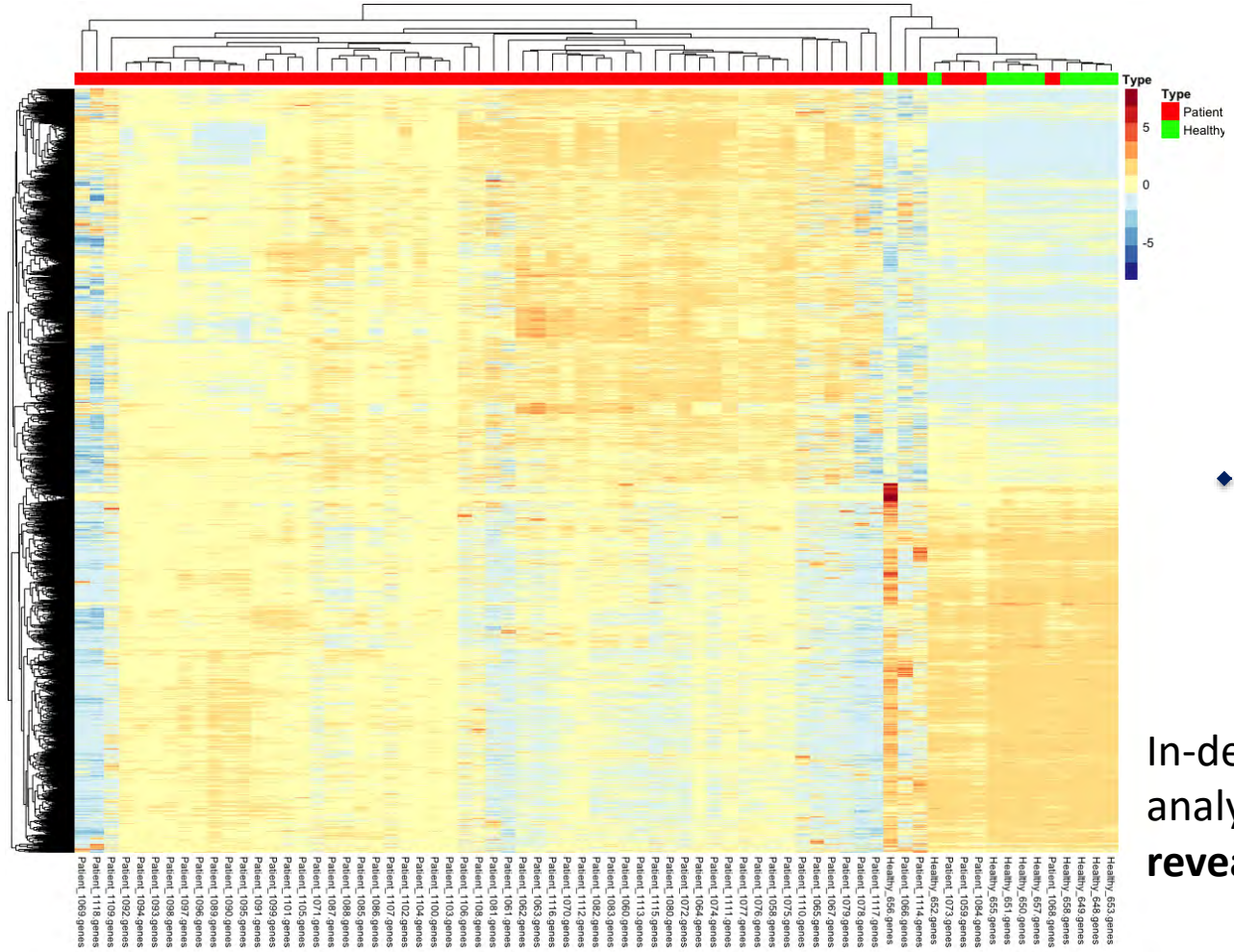


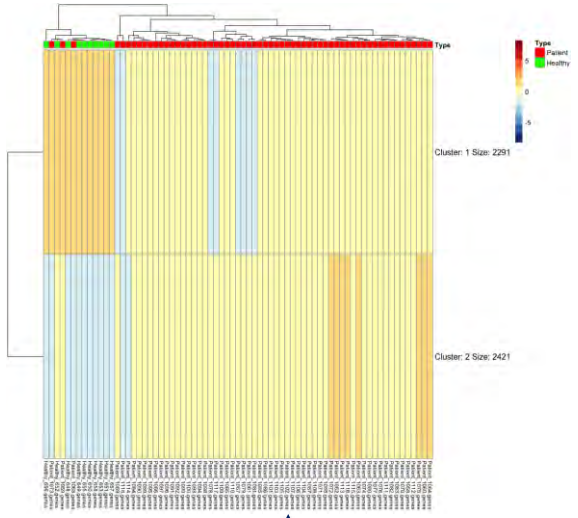


Race-based gene expression analysis

	HF	AF	Other CVDs
A. White			
B. Black			
C. All others			

Differential Gene Expression Analysis

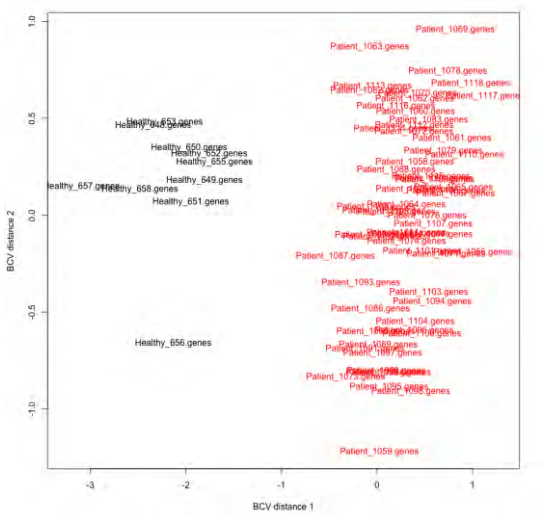




Clusters: 1 and 2

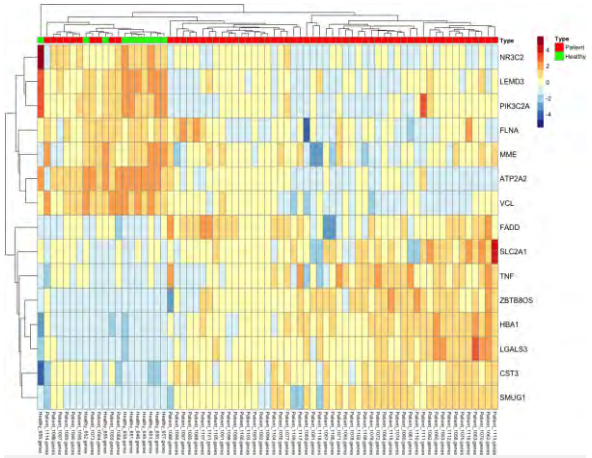
In-depth gene expression analysis and annotation **revealed 4,885 DEGs.**

Gene Enrichment and Pathways Analysis

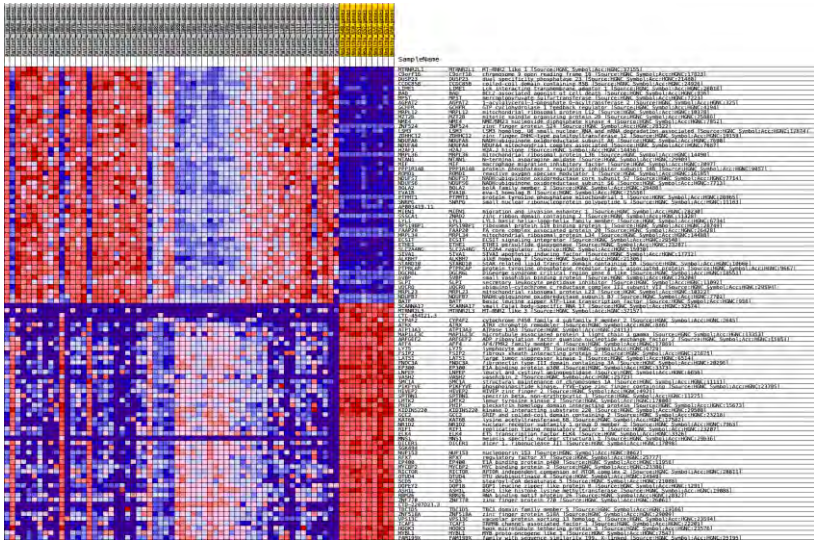


MDS plot showing biological distance between case-control samples based on BCV.

Top 20 enriched pathways showing up-regulation and down-regulation in CVD based on their normalized enrichment scores (NES).



Differential gene expression of annotated CVD genes.



Gene enrichment heatmap of differentially expressed genes.

Summary: Gene Expression Analysis

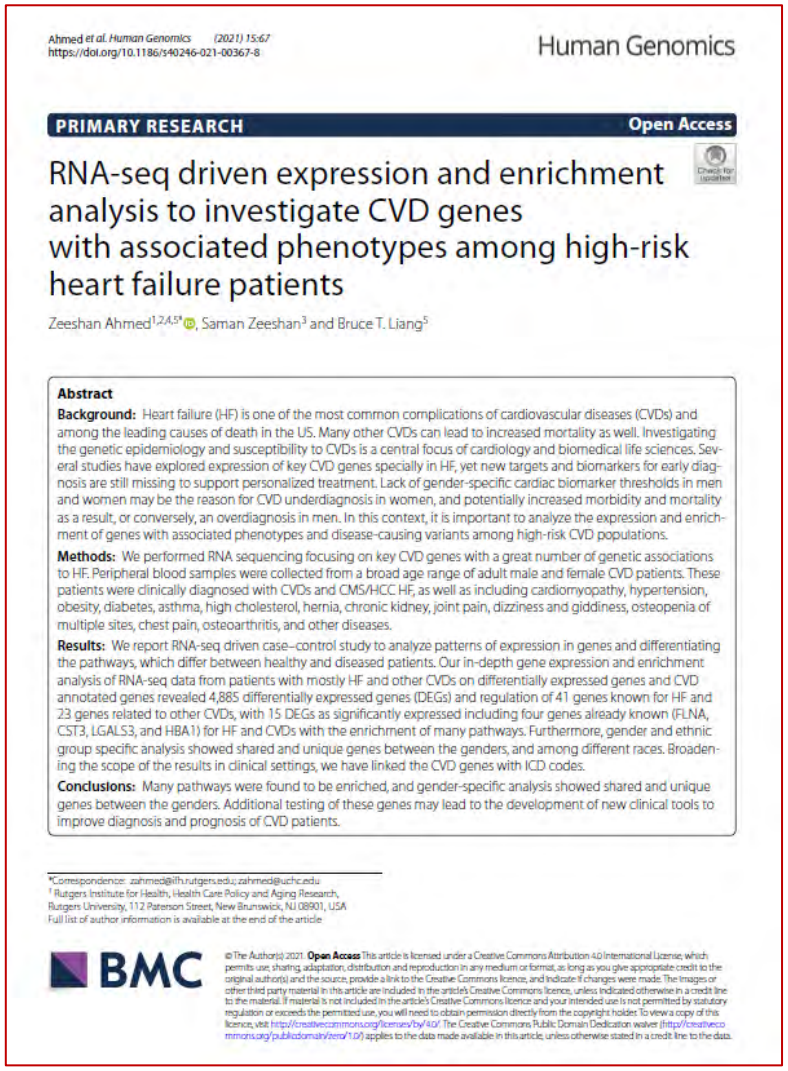
- We report RNA-seq driven case-control study to analyze patterns of expression in genes and differentiating the pathways, which differ between healthy and diseased patients.

- Our in-depth gene expression and enrichment analysis of RNA-seq data from patients with mostly HF and other CVDs on differentially expressed genes and CVD annotated genes revealed 4,885 differentially expressed genes and regulation of **41 genes known for HF, 22 genes associated with AF, and 23 genes related to other CVDs**.

- 15 DEGs as significantly expressed including four altered genes known (***FLNA, CST3, LGALS3***, and ***HBA1***) for HF and CVDs with the enrichment of many pathways. We found that ***PF4, PPBP, MYL4, KCNE3, VKORC1, KCNQ1*** and ***CYP4F2*** genes are highly expressed in AF.

- Gender and ethnic group specific analysis showed shared and unique genes between the genders, and among different races. Subsequent analyses were performed based on gender. Our analysis identified altered expression pathways of genes with gender differences in middle-aged to frail CVD patients.

Related – Most Recent Peer Revised Publications by Ahmed Lab



Ahmed et al. Human Genomics (2021) 15:67
https://doi.org/10.1186/s40246-021-00367-8

Human Genomics

PRIMARY RESEARCH Open Access

RNA-seq driven expression and enrichment analysis to investigate CVD genes with associated phenotypes among high-risk heart failure patients

Zeeshan Ahmed[1,2,4,5*], Saman Zeeshan[3] and Bruce T. Liang[5]

Abstract

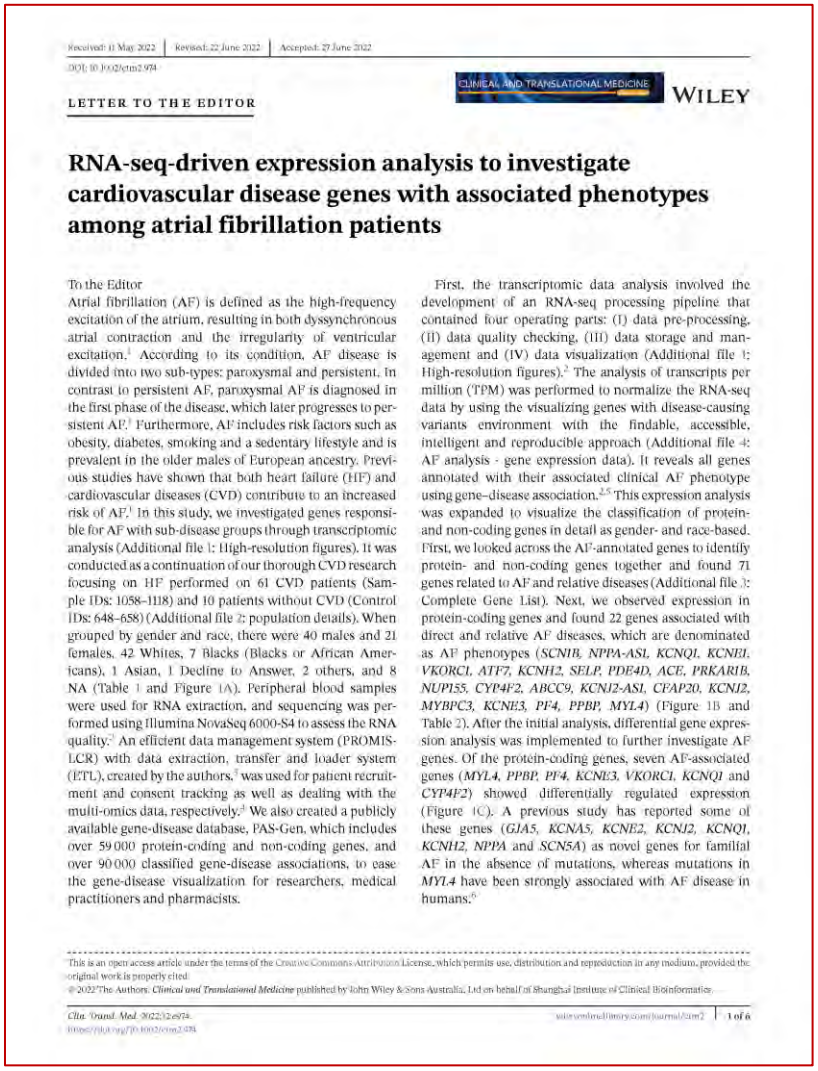
Background: Heart failure (HF) is one of the most common complications of cardiovascular diseases (CVDs) and among the leading causes of death in the US. Many other CVDs can lead to increased mortality as well. Investigating the genetic epidemiology and susceptibility to CVDs is a central focus of cardiology and biomedical life sciences. Several studies have explored expression of key CVD genes specially in HF, yet new targets and biomarkers for early diagnosis are still missing to support personalized treatment. Lack of gender-specific cardiac biomarker thresholds in men and women may be the reason for CVD underdiagnosis in women, and potentially increased morbidity and mortality as a result, or conversely, an overdiagnosis in men. In this context, it is important to analyze the expression and enrichment of genes with associated phenotypes and disease-causing variants among high-risk CVD populations.

Methods: We performed RNA sequencing focusing on key CVD genes with a great number of genetic associations to HF. Peripheral blood samples were collected from a broad age range of adult male and female CVD patients. These patients were clinically diagnosed with CVDs and CMS/HCC HF, as well as including cardiomyopathy, hypertension, obesity, diabetes, asthma, high cholesterol, hernia, chronic kidney, joint pain, dizziness and giddiness, osteopenia of multiple sites, chest pain, osteoarthritis, and other diseases.

Results: We report RNA-seq driven case–control study to analyze patterns of expression in genes and differentiating the pathways, which differ between healthy and diseased patients. Our in-depth gene expression and enrichment analysis of RNA-seq data from patients with mostly HF and other CVDs on differentially expressed genes and CVD annotated genes revealed 4,885 differentially expressed genes (DEGs) and regulation of 41 genes known for HF and 23 genes related to other CVDs, with 15 DEGs as significantly expressed including four genes already known (FLNA, CST3, LGALS3, and HBA1) for HF and CVDs with the enrichment of many pathways. Furthermore, gender and ethnic group specific analysis showed shared and unique genes between the genders, and among different races. Broadening the scope of the results in clinical settings, we have linked the CVD genes with ICD codes.

Conclusions: Many pathways were found to be enriched, and gender-specific analysis showed shared and unique genes between the genders. Additional testing of these genes may lead to the development of new clinical tools to improve diagnosis and prognosis of CVD patients.

*Correspondence: zahmed@ilh.rutgers.edu; zahmed@uchc.edu
[1] Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick, NJ 08901, USA
Full list of author information is available at the end of the article



Received: 11 May 2022 | Revised: 22 June 2022 | Accepted: 27 June 2022
DOI: 10.1002/ctm2.974

CLINICAL AND TRANSLATIONAL MEDICINE WILEY

LETTER TO THE EDITOR

RNA-seq-driven expression analysis to investigate cardiovascular disease genes with associated phenotypes among atrial fibrillation patients

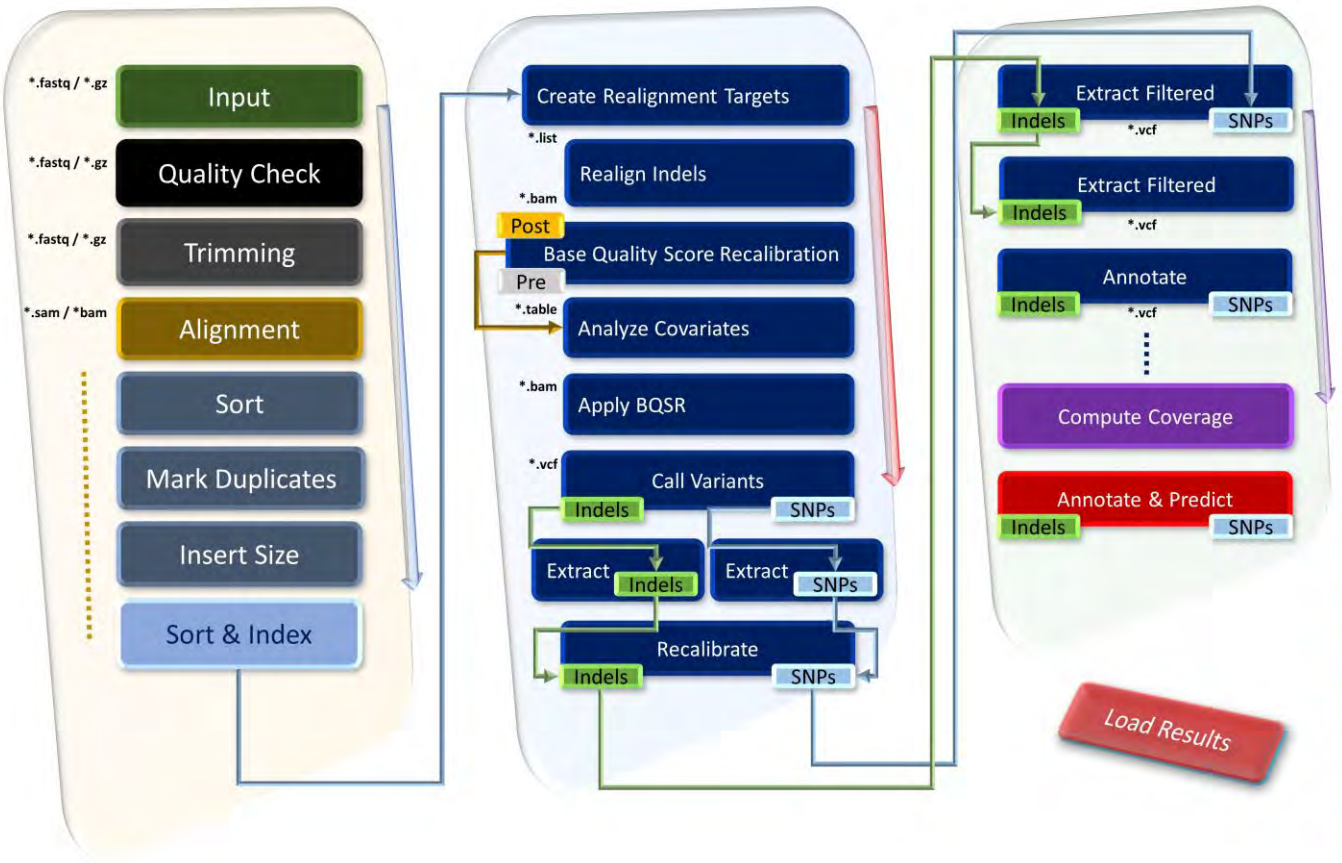
To the Editor

Atrial fibrillation (AF) is defined as the high-frequency excitation of the atrium, resulting in both dyssynchronous atrial contraction and the irregularity of ventricular excitation.[1] According to its condition, AF disease is divided into two sub-types: paroxysmal and persistent. In contrast to persistent AF, paroxysmal AF is diagnosed in the first phase of the disease, which later progresses to persistent AF.[1] Furthermore, AF includes risk factors such as obesity, diabetes, smoking and a sedentary lifestyle and is prevalent in the older males of European ancestry. Previous studies have shown that both heart failure (HF) and cardiovascular diseases (CVD) contribute to an increased risk of AF.[1] In this study, we investigated genes responsible for AF with sub-disease groups through transcriptomic analysis (Additional file 1: High-resolution figures). It was conducted as a continuation of our thorough CVD research focusing on HF performed on 61 CVD patients (Sample IDs: 1058–1118) and 10 patients without CVD (Control IDs: 648–658) (Additional file 2: population details). When grouped by gender and race, there were 40 males and 21 females, 42 Whites, 7 Blacks (Blacks or African Americans), 1 Asian, 1 Decline to Answer, 2 others, and 8 NA (Table 1 and Figure 1A). Peripheral blood samples were used for RNA extraction, and sequencing was performed using Illumina NovaSeq 6000-S4 to assess the RNA quality.[2] An efficient data management system (PROMIS-LCR) with data extraction, transfer and loader system (ETL), created by the authors,[3] was used for patient recruitment and consent tracking as well as dealing with the multi-omics data, respectively.[4] We also created a publicly available gene-disease database, PAS-Gen, which includes over 59 000 protein-coding and non-coding genes, and over 90 000 classified gene-disease associations, to ease the gene-disease visualization for researchers, medical practitioners and pharmacists.

First, the transcriptomic data analysis involved the development of an RNA-seq processing pipeline that contained four operating parts: (I) data pre-processing, (II) data quality checking, (III) data storage and management and (IV) data visualization (Additional file 1: High-resolution figures).[2] The analysis of transcripts per million (TPM) was performed to normalize the RNA-seq data by using the visualizing genes with disease-causing variants environment with the findable, accessible, intelligent and reproducible approach (Additional file 4: AF analysis - gene expression data). It reveals all genes annotated with their associated clinical AF phenotype using gene–disease association.[2,5] This expression analysis was expanded to visualize the classification of protein- and non-coding genes in detail as gender- and race-based. First, we looked across the AF-annotated genes to identify protein- and non-coding genes together and found 71 genes related to AF and relative diseases (Additional file 3: Complete Gene List). Next, we observed expression in protein-coding genes and found 22 genes associated with direct and relative AF diseases, which are denominated as AF phenotypes (*SCN1B, NPPA-AS1, KCNQ1, KCNE1, VKORC1, ATF7, KCNH2, SELP, PDE4D, ACE, PRKAR1B, NUP155, CYP4F2, ABCC9, KCNJ2-AS1, CFAP20, KCNJ2, MYBPC3, KCNE3, PF4, PPBP, MYL4*) (Figure 1B and Table 2). After the initial analysis, differential gene expression analysis was implemented to further investigate AF genes. Of the protein-coding genes, seven AF-associated genes (*MYL4, PPBP, PF4, KCNE3, VKORC1, KCNQ1* and *CYP4F2*) showed differentially regulated expression (Figure 1C). A previous study has reported some of these genes (*GJA5, KCNA5, KCNE2, KCNJ2, KCNQ1, KCNH2, NPPA* and *SCN5A*) as novel genes for familial AF in the absence of mutations, whereas mutations in *MYL4* have been strongly associated with AF disease in humans.[6]

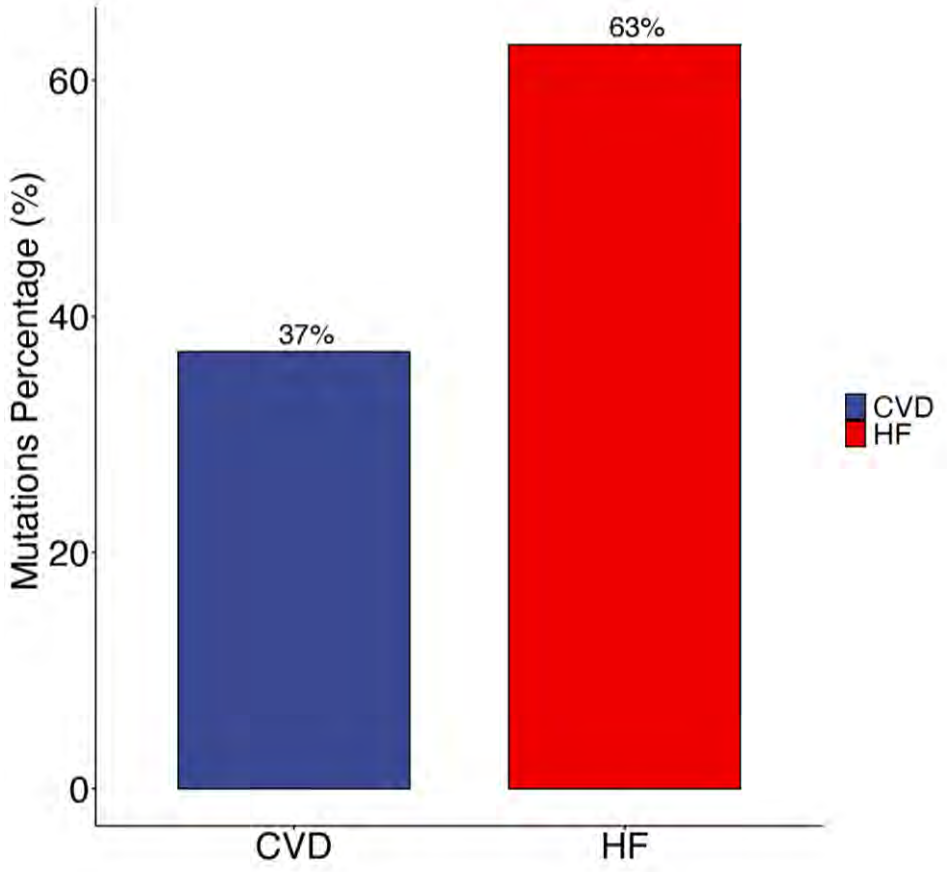
Variant Analysis & Validation using WGS Data of Same Patients

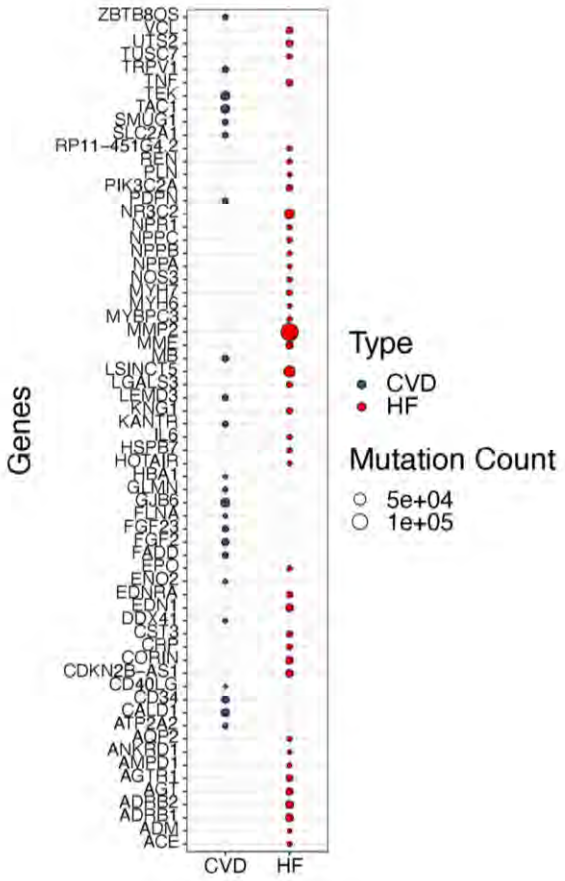
JWES: a new pipeline for gene-variant discovery, annotation, prediction, visualization, and genotyping.



URL: https://github.com/drzeeshanahmed/JWES-DB

Variant Analysis: Mutation % and Count Per Gene



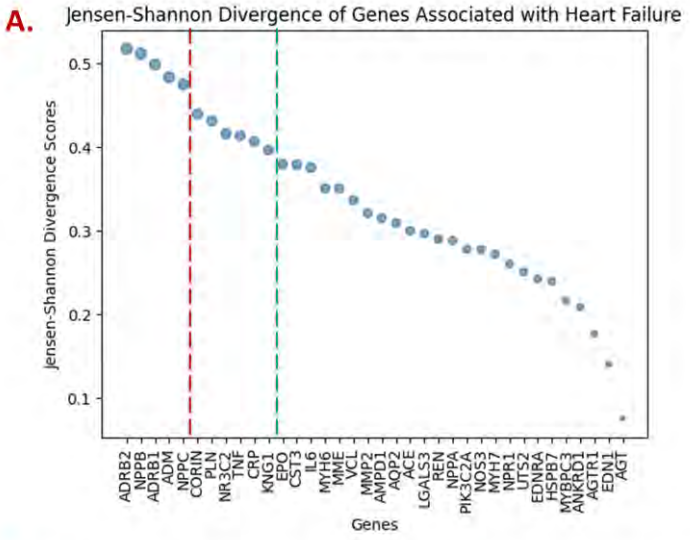


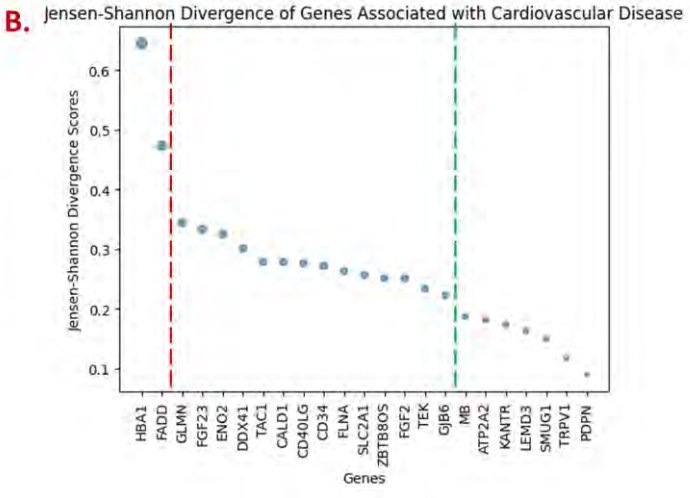
Mutation-percentage of HF (63%) and other CVD (37%) genes.

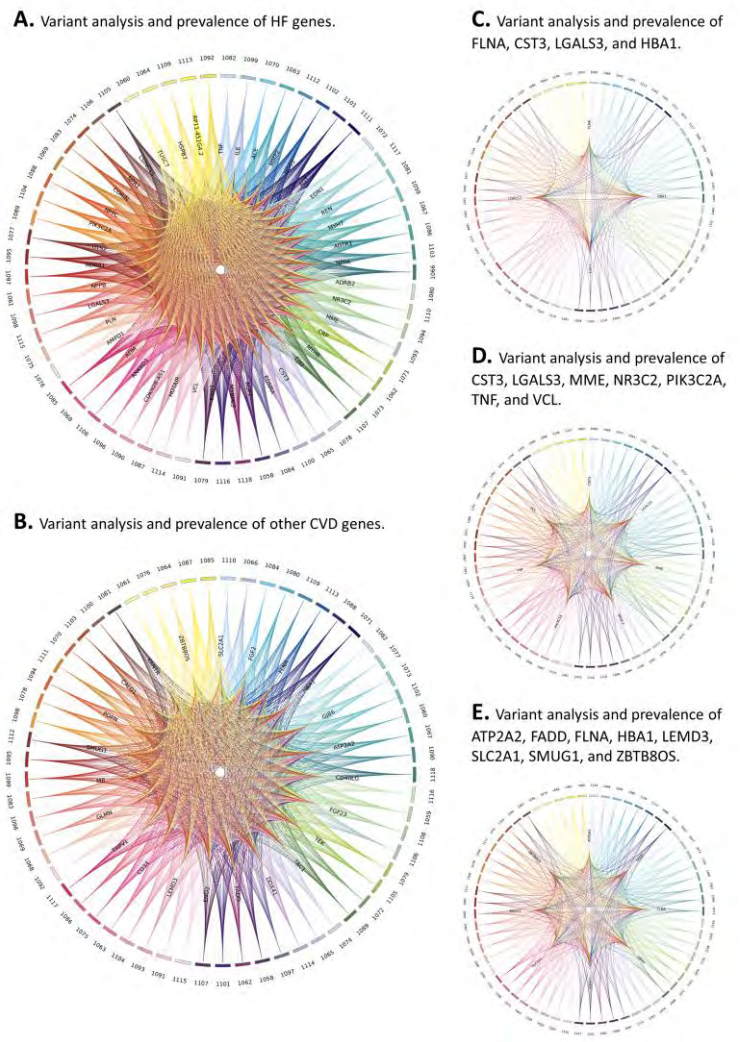
Mutation-count per gene associated with HF and other CVDs.

Variant and Prevalence Analysis, and JSD

Jensen-Shannon Divergence (JSD) measurement and variant distribution analysis of genes associated with heart failure (HF) and other cardiovascular diseases (CVDs). Figure presents JSD scores of genes associated with HF, and other CVDs.



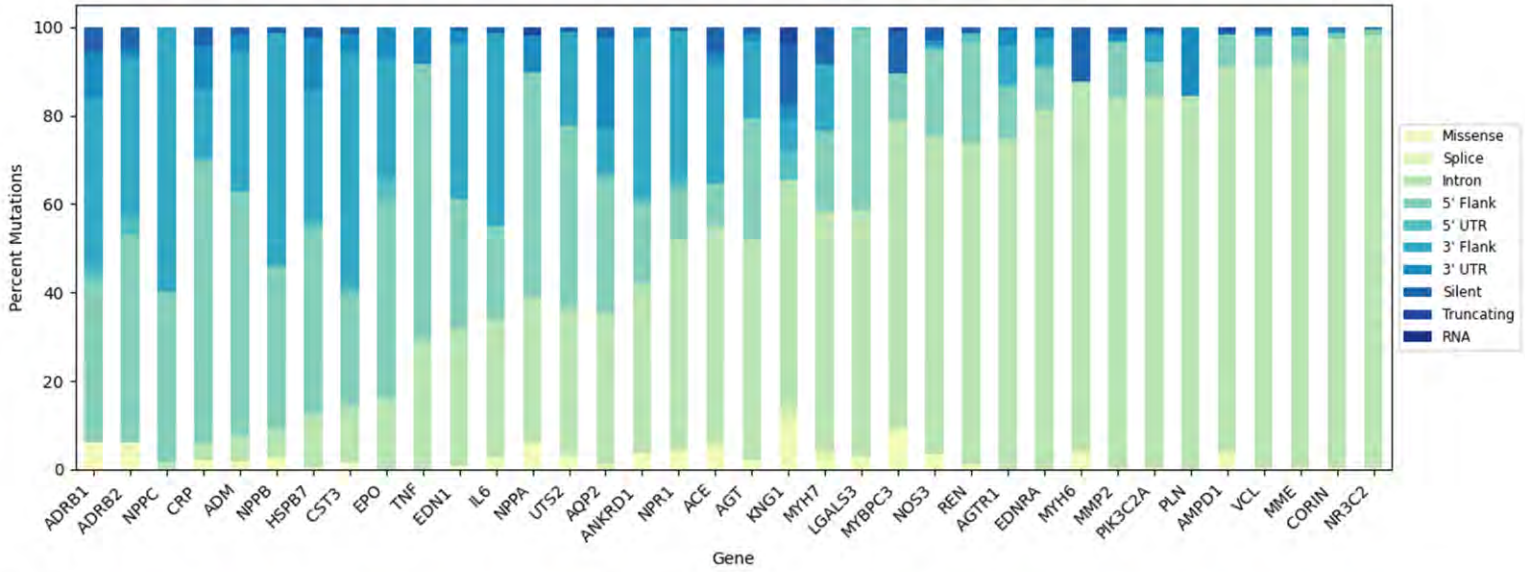




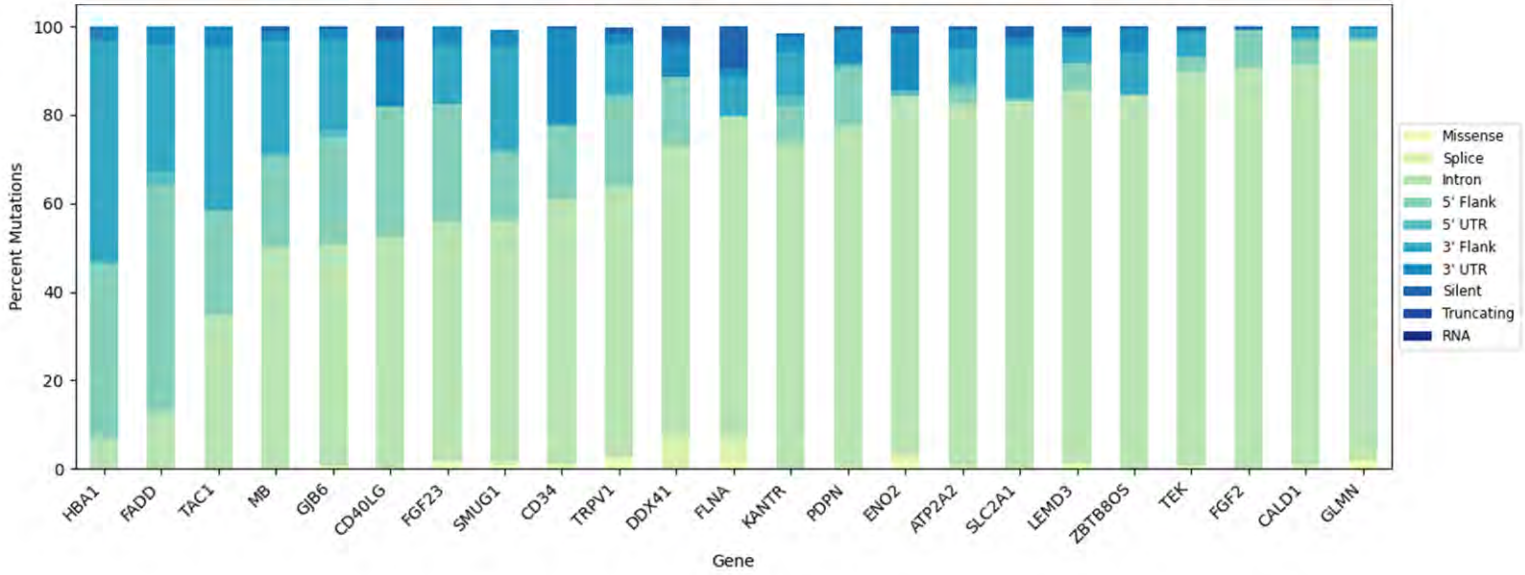
A) Variant analysis and prevalence of HF genes. **B)** Variant analysis and prevalence of other CVD genes. **C)** Variant analysis and prevalence of *FLNA, CST3, LGALS3*, and *HBA1*. **D)** Variant analysis and prevalence of *CST3, LGALS3, MME, NR3C2, PIK3C2A, TNF,* and *VCL*. **E)** Variant analysis and prevalence of *ATP2A2, FADD, FLNA, HBA1, LEMD3, SLC2A1, SMUG1,* and *ZBTB8OS*.

Splice Mutation Analysis of Genes

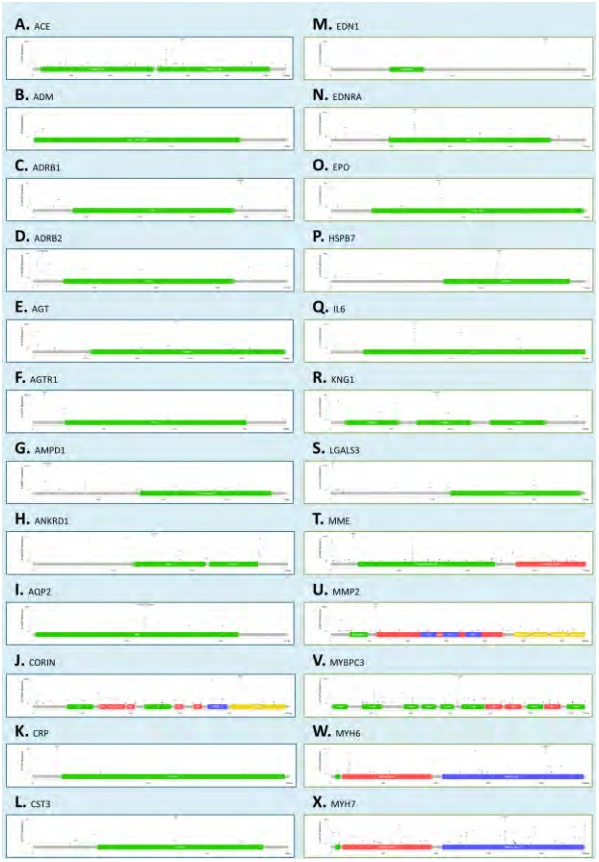
A. Mutation analysis of genes associated with heart failure (HF) disease.



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

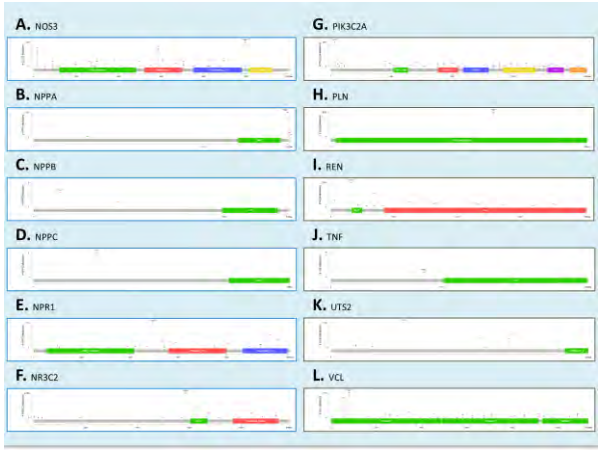


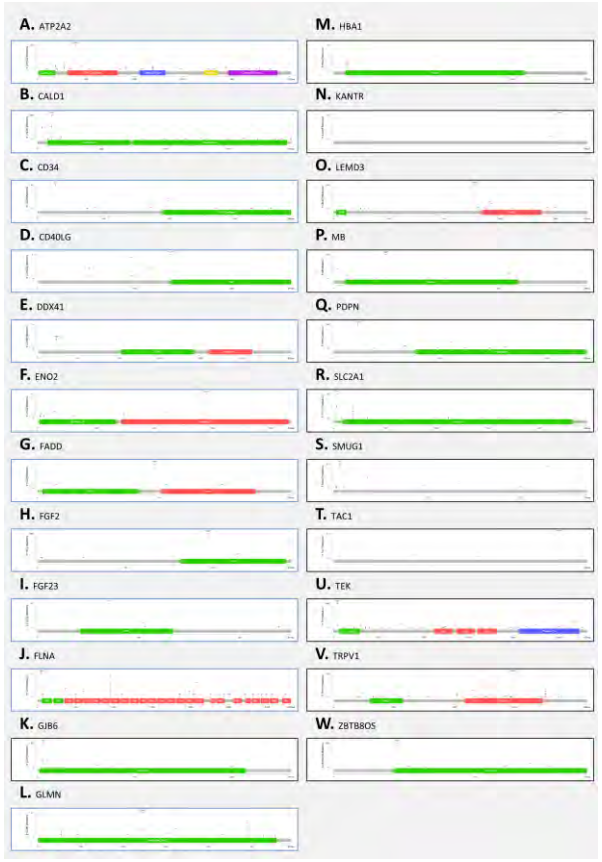
Functional and non-functional mutation analysis



Lollipop plots of *ACE, ADM, ADRB1, ADRB2, AGT, AGTR1, AMPD1, ANKRD1, AQP2, CORIN, CRP, CST3, EDN1, EDNRA, EPO, HSPB7, IL6, KNG1, LGALS3, MME, MMP2, MYBPC3, MYH6,* and *MYH7*.

Lollipop plots of *NOS3, NPPA, NPPB, NPPC, NPR1, NR3C2, PIK3C2A, PLN, REN, TNF, UTS2*, and *VCL*.





Lollipop plots of *SLC2A1, FGF2, FLNA, HBA1, GJB6, ATP2A2, CD40LG, FGF23, TEK, TAC1, DDX41, FADD, ENO2, LEMD3, CD34, TRPV1, GLMN, MB, SMUG1, PDPN, CALD1, KANTR,* and *ZBTB8OS*.

Green color represents Missense Mutations, black represents Truncating Mutations, brown represents Inframe Mutations, and purple represents Fusion Mutations.

Summary

- We performed variant analysis and **verified mutations among the annotated genes** for HF, AF, and other CVDs, and identified missense mutations among genes with altered expression.

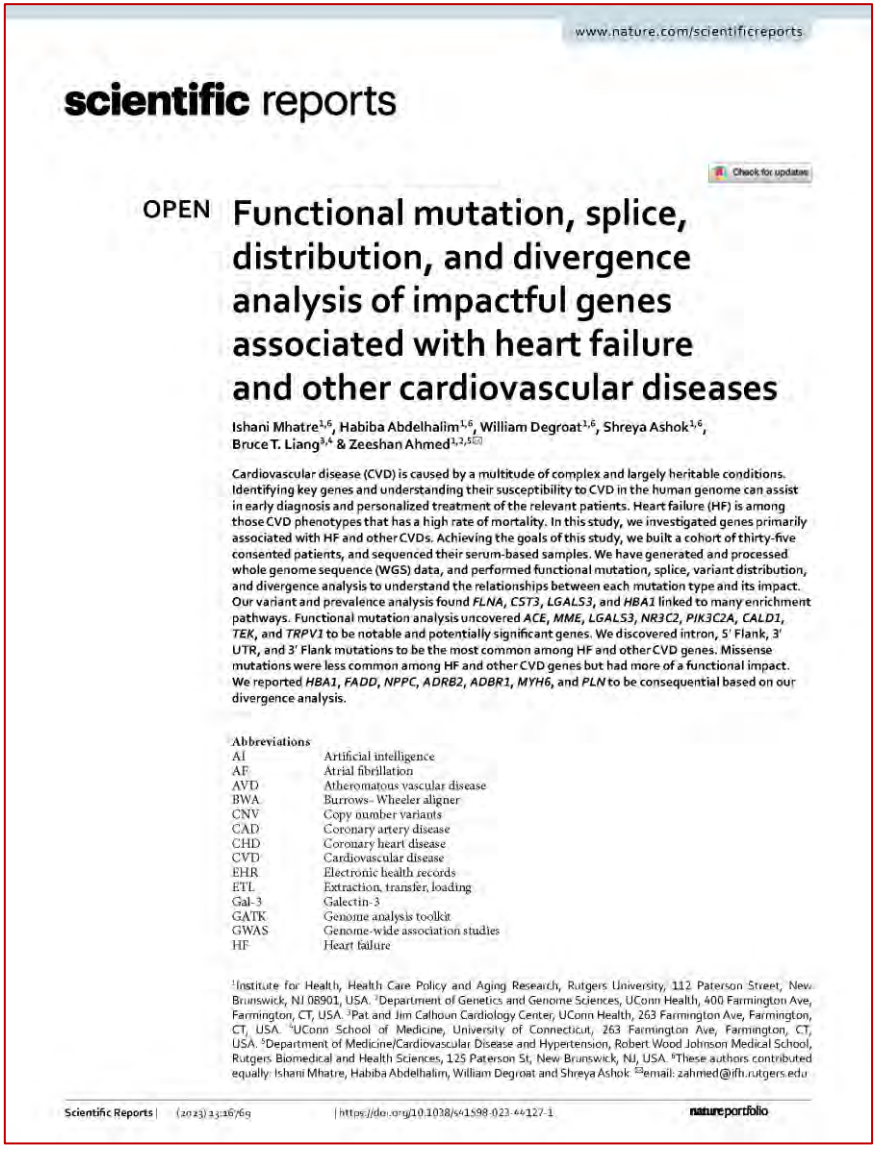
- We annotated these mutations to identify functional and nonfunctional mutations in genes associated with HF, AF, and other CVDs. **We detected over a million SNV and insertion and deletion events.**

- We implemented Jensen-Shannon Divergence (JSD) Based Method and **identified *HBA1, FADD, ADRB2, NPPB, ADRB1, ADB*, and *NPPC* genes with the greatest variance** based on their JSD scores

- The most common mutation types in HF, AF, and other CVD genes were intronic, 5' flank, and 3' flank. **Mutations in these genes have been linked to aberrant expression in CVD** and observed having low functional impact among common missense mutations.

- WGS allowed us to do in-depth analysis of CVD genes as RNA-seq cannot detect any of the variants located in noncoding DNA regions.

Related – Most Recent Peer Revised Publications by Ahmed Lab



www.nature.com/scientificreports

scientific reports

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

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

Received: 5 April 2023 | Revised: 11 May 2023 | Accepted: 11 May 2023

DOI: 10.1002/ctd2.206

RESEARCH ARTICLE

CLINICAL AND TRANSLATIONAL DISCOVERY WILEY

Investigating genes associated with cardiovascular disease among heart failure patients for translational research and precision medicine

Zeeshan Ahmed[1,2,4] | Saman Zeeshan[3] | Nicholas Persaud[2] | William Degroat[1] | Habiba Abdelhalim[1] | Bruce T. Liang[5,6]

[1]Institute for Health, Health Care Policy and Aging Research, Rutgers University, New Brunswick, New Jersey, USA

[2]Department of Medicine, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, New Brunswick, New Jersey, USA

[3]Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, New Jersey, USA

[4]Department of Genetics and Genome Sciences, UConn Health, Farmington, Connecticut, USA

[5]Pat and Jim Calhoun Cardiology Center, UConn Health, Farmington, Connecticut, USA

[6]UConn School of Medicine, University of Connecticut, Farmington, Connecticut, USA

Correspondence
Zeeshan Ahmed, Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick, 08901, NJ, USA.
Email: zahmed@ifh.rutgers.edu

Funding information
School of Medicine, UConn Health, CT

Abstract

Background: Cardiovascular disease (CVD) is a leading cause of premature mortality in the United States and the world. CVD comprises several complex and mostly heritable conditions, which range from myocardial infarction to congenital heart disease. The risk factors contributing to the development of CVD and response to therapy in an individual patient are highly variable. Here, we report our findings from an integrative analysis of gene expression, disease-causing gene variants and associated phenotypes among CVD populations, with a focus on high-risk heart failure (HF) patients.

Methods: We built a cohort using electronic health records of consented patients with available samples and then performed high-throughput whole genome and RNA sequencing of key genes responsible for HF and other CVD pathologies. Our in-depth gene expression analysis revealed differentially expressed genes associated with HF and other CVDs. We performed a variant analysis of whole genome sequence data of CVD patients and identified genes with altered gene expression with functional and non-functional mutations in these genes.

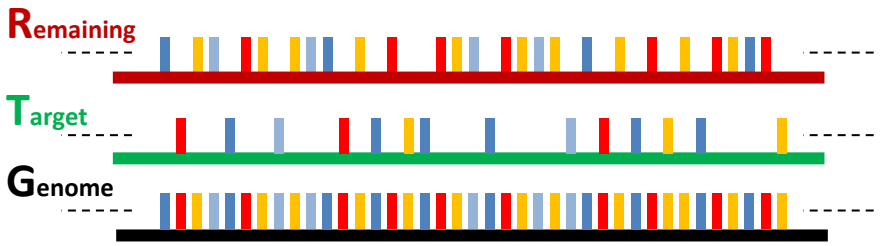
Results: Our results highlight the importance of investigating the mechanisms of CVD progression through multi-omics datasets. Next, we performed splice mutation and variant distribution analysis of genes associated with HF and other CVD. We implemented Jensen–Shannon divergence (JSD)-based method and identified *HBA1*, *FADD*, *ADRB2*, *NPPB*, *ADRB1*, *ADB* and *NPPC* genes with the greatest variance based on their JSD scores. Our study provided evidence that applying integrative data analysis approach involving genomics and transcriptomics data will not only help understand the pathophysiology of CVD diseases but also reduce heterogeneity in disease subtypes.

KEYWORDS
cardiovascular disease, expression, gene, genome, heart failure, RNA-seq, variant

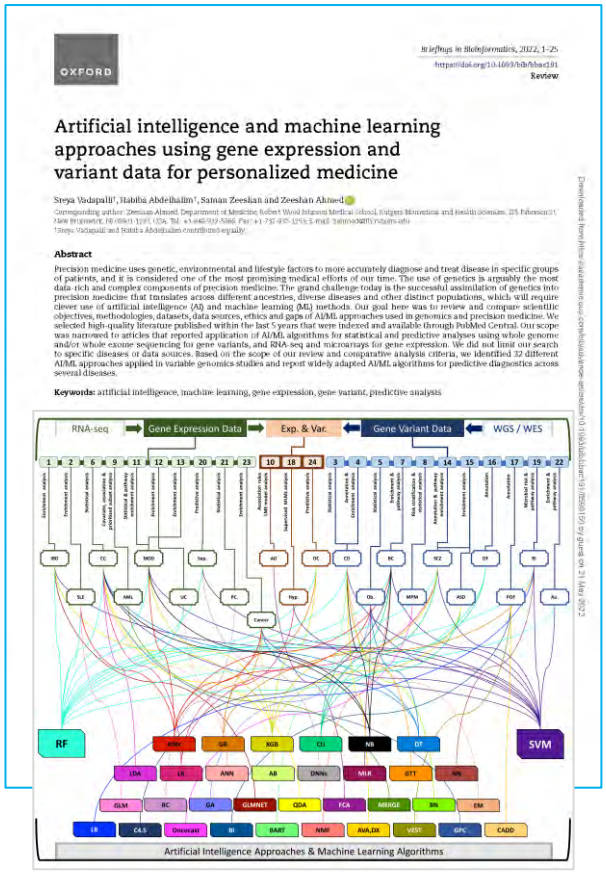
Implementing Artificial Intelligence (AI) & Machine Learning (ML)

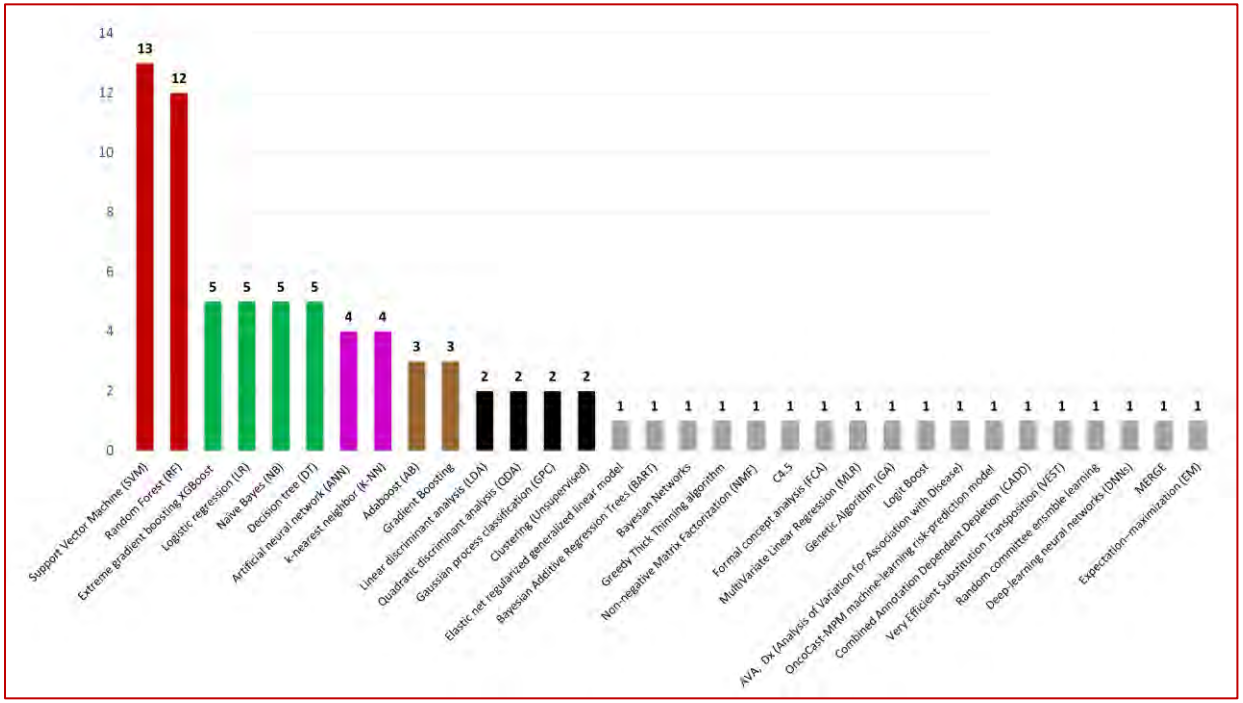
1. **Predict CVD with high accuracy** with knowledge-driven approach based on known genetic evidence establishing association by implementing best fitting AI/ML algorithms for deep phenotyping and predictive analytics.

2. **Identify new predictive biomarkers** using data-driven approach with a novel nexus of AI/ML algorithms to reveal biological patterns explaining the causal relationship based on genomic, transcriptomic, and phenotypic data of patients with CVD.



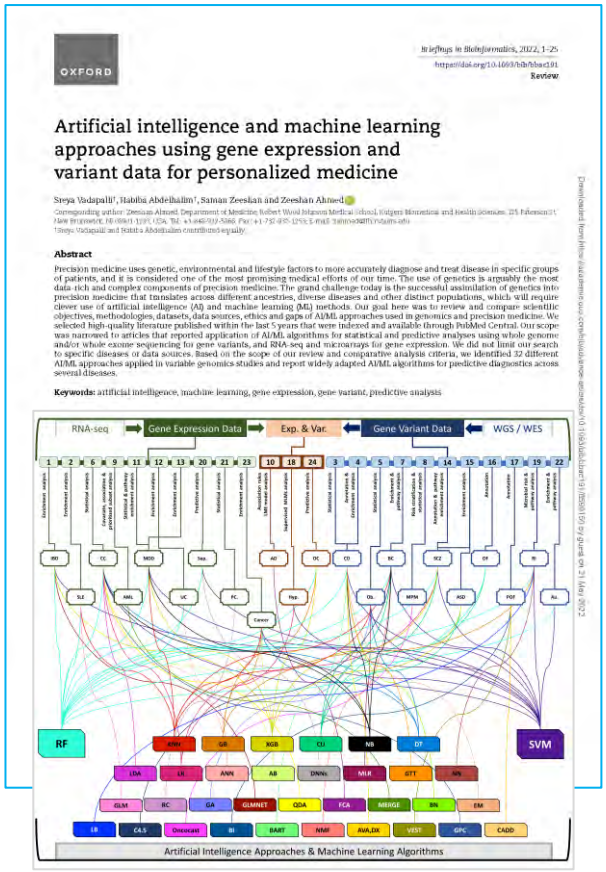
Which AI/ML approach/algorithm is appropriate ?





AI/ML approaches: **Comparative analysis and evaluation. 2022**

Predict CVD with high accuracy with *Hygieia*



OXFORD

Briefings in Bioinformatics, 2022, 1–25

Review

Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine

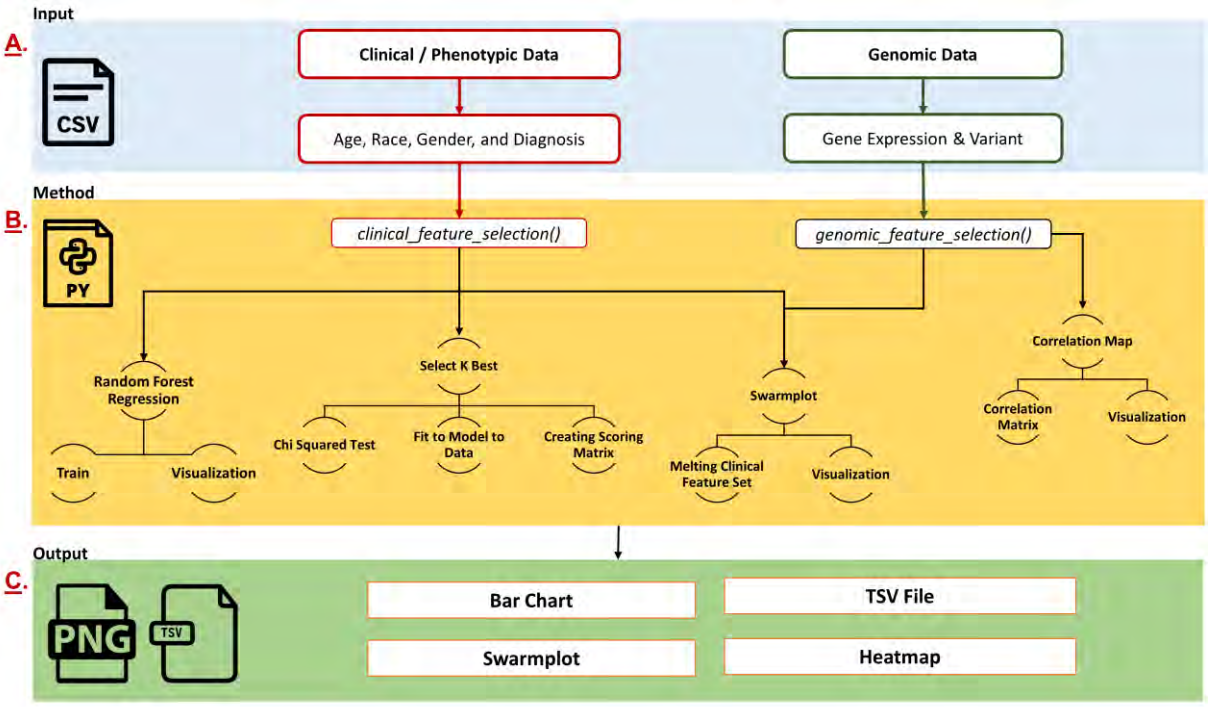
Sreya Vadapalli†, Habiba Abdelhalim†, Saman Zeeshan and Zeeshan Ahmed

Abstract

Precision medicine uses genetic, environmental and lifestyle factors to more accurately diagnose and treat disease in specific groups of patients, and it is considered one of the most promising medical efforts of our time. The use of genetics is arguably the most data-rich and complex components of precision medicine. The grand challenge today is the successful assimilation of genetics into precision medicine that translates across different ancestries, diverse diseases and other distinct populations, which will require clever use of artificial intelligence (AI) and machine learning (ML) methods. Our goal here was to review and compare scientific objectives, methodologies, datasets, data sources, ethics and gaps of AI/ML approaches used in genomics and precision medicine. We selected high-quality literature published within the last 5 years that were indexed and available through PubMed Central. Our scope was narrowed to articles that reported application of AI/ML algorithms for statistical and predictive analyses using whole genome and/or whole exome sequencing for gene variants, and RNA-seq and microarrays for gene expression. We did not limit our search to specific diseases or data sources. Based on the scope of our review and comparative analysis criteria, we identified 32 different AI/ML approaches applied in variable genomics studies and report widely adapted AI/ML algorithms for predictive diagnostics across several diseases.

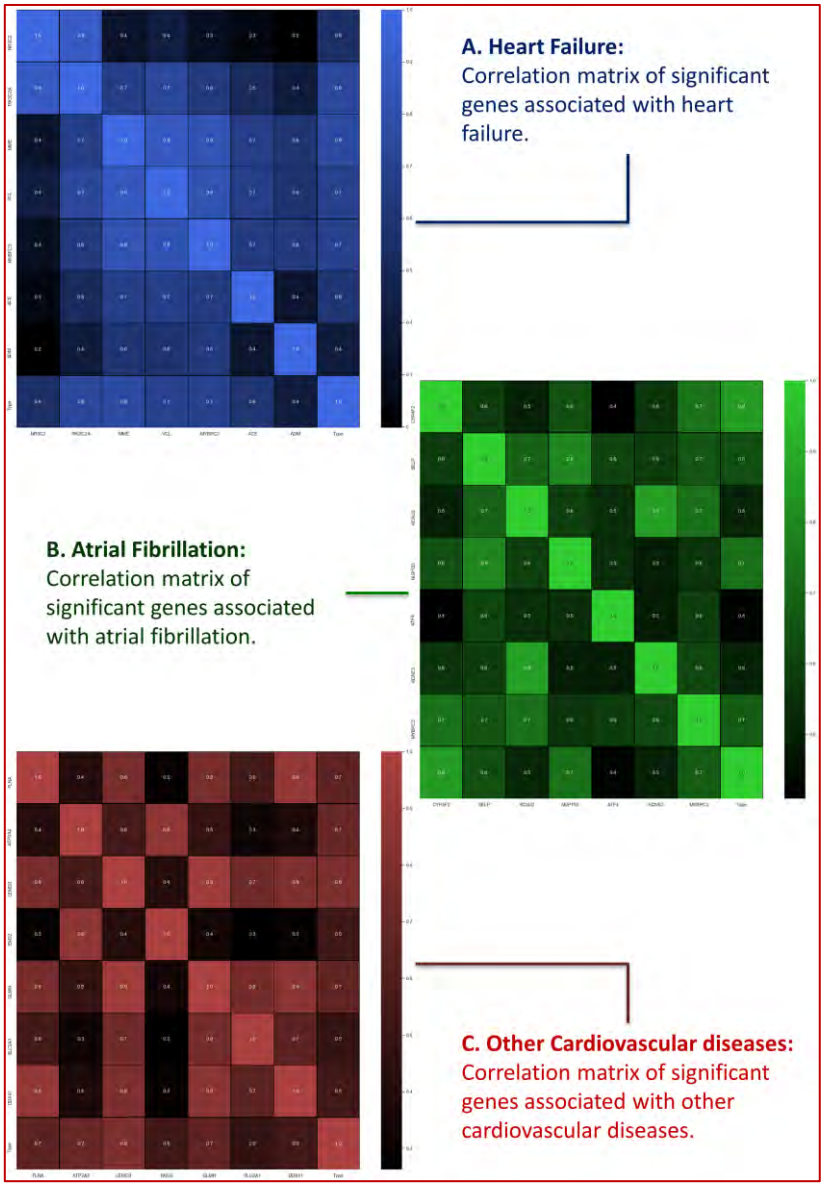
Keywords: artificial intelligence, machine learning, gene expression, gene variant, predictive analysis

AI/ML approaches: **Comparative analysis and evaluation. 2022**



Hygieia: **AI/ML pipeline for predictive analysis.**

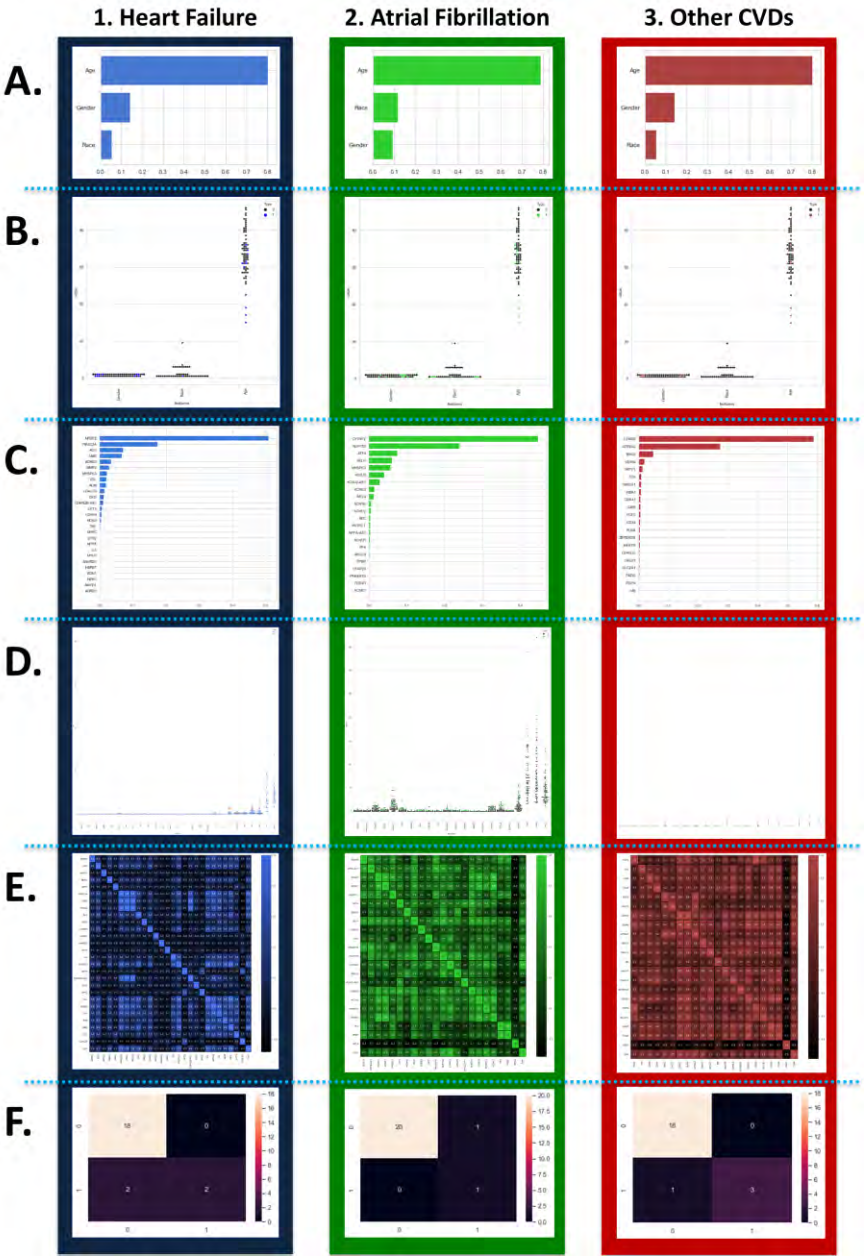
Predict CVD with high accuracy



A. Heart Failure:
Correlation matrix of significant genes associated with heart failure.

B. Atrial Fibrillation:
Correlation matrix of significant genes associated with atrial fibrillation.

C. Other Cardiovascular diseases:
Correlation matrix of significant genes associated with other cardiovascular diseases.



(A) Population distribution based on clinical features; (B) Correlation matrix; (C) Gene ranking; (D) Feature Swarm Plot; (E) Correlation matrix of genes; (F) Confusion matrix of genes.

Summary: Predictive Analysis

- We used our **open-source AI/ML ready pipeline i.e., *Hygieia***, which is based on the Random Forest (RF) for regression analysis and predicting disease without requiring hyperparameter tuning.
- We **trained our model on different cross-sections** of the three different matrices based on HF, AF, and other CVDs.
- We uncovered an interesting correlation between age, gender, race, and diagnosis. During our analysis, it was observed that **age and gender appeared to have a high correlation in HF and other CVDs while age, and race were highly correlated in AF**.
- We observed the most significant genes associated with HF, AF, and other CVDs based on the RF feature importance global variable. **A score was assigned to each gene, which represents the feature importance for the model in stratifying CVD patients**.
- Visible data clusters were observed for the genes highly correlated, downregulated and with altered expression in CVD patients compared to healthy individuals. Our model was able to correctly classify individuals as CVD patients and predict CVD with **95% accuracy**.
- We observed and reported overlapping in **significant results produced in gene expression, variant, phenotypic, and predictive analyses**, which include genes associated with HF, AF, and other CVDs.

Related – Most Recent Peer Revised Publications by Ahmed Lab

Software Impacts 16 (2023) 100493

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

Hygieia: AI/ML pipeline integrating healthcare and genomics data to investigate genes associated with targeted disorders and predict disease

William DeGroat [a,1], Vignesh Venkat [a,1], Widnie Pierre-Louis [a], Habiba Abdelhalim [a], Zeeshan Ahmed [a,b,*]

[a] Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick, NJ, USA
[b] Department of Medicine, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson Street, New Brunswick, NJ, USA

ARTICLE INFO

Keywords:
Hygieia
Artificial Intelligence
Machine learning
Genomics
Healthcare
Predictive analysis

ABSTRACT

Due to the advancements in sequencing technologies, genomics data is developing at an unmatched pace and levels to foster translational research. Over ten million genomics datasets have been produced and publicly shared in the year 2022. Genome-wide association studies (GWAS) have remarkably assisted in understanding the genetic basis of human disease by uncovering millions of loci associated with various complex phenotypes. However, GWAS are unable to predict disease and detect all the heritability explained by single nucleotide polymorphisms (SNPs) and can only target specific variants. The rightful use of the artificial intelligence (AI) and machine learning (ML) techniques can accelerate our ability to leverage and extend the information contained within the original data, and model patient-specific genomics data against publicly available annotation repositories for understanding how coding and non-coding genomic variations are connected to disease mechanisms. The grand challenge here is assimilation of genetics into precision medicine that translates across different ancestries, diverse diseases, and other distinct populations with the implementation of effective AI/ML methods. We present first AI/ML ready pipeline i.e., *Hygieia*., integrating genomics and clinical data to investigate genes associated with the targeted disorders and predict disease with high accuracy. *Hygieia* can utilize broad dataset sizes with heterogeneous levels of granularity and offer a supervised approach to analyze integrated gene expression and multivariate clinical data. It includes the Random Forest based model for regression analysis and predict without hyperparameter tuning. We trained and tested our model across variable disorders and using diverse datasets. *Hygieia* is an open-source and simple to use pipeline, which does not strong require computational background to execute.

Code metadata

Current Code Version	*Hygieia v1.0.2*
Permanent Link to Repository	https://github.com/SoftwareImpacts/SIMPAC-2022-266
Reproducible Capsule	https://codeocean.com/capsule/7964745/tree/v1
Legal License	*GNU General Public License (GPL)*
Code Versioning System	*Git*
Software Code Languages	*Python 3.10.9*
Compilation Requirements, Dependencies	*pandas, sci-kit learn, matplotlib, seaborn*
Support email for questions	zahmed@ifh.rutgers.edu

1. Introduction

Precision and genomics medicine is driven by the paradigm shift of empowering clinicians to predict the most appropriate course of action

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Correspondence to: Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick, 08901, NJ, USA.
E-mail address: zahmed@ifh.rutgers.edu (Z. Ahmed).
[1] Equally contributing first authors.

https://doi.org/10.1016/j.simpa.2023.100493
Received 26 January 2023; Received in revised form 4 March 2023; Accepted 6 March 2023

Genomics 115 (2023) 110564

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

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

Vignesh Venkat [a,1], Habiba Abdelhalim [a,1], William DeGroat [a], Saman Zeeshan [b], Zeeshan Ahmed [a,c,*]

[a] Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson St, New Brunswick, NJ, USA
[b] Rutgers Cancer Institute of New Jersey, Rutgers University, 195 Little Albany St, New Brunswick, NJ, USA
[c] Department of Medicine, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson St, New Brunswick, NJ, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Atrial fibrillation
Cardiovascular diseases
Gene expression
Heart failure
Machine learning
Predictive analysis

ABSTRACT

Cardiovascular disease (CVD) is the leading cause of mortality and loss of disability adjusted life years (DALYs) globally. CVDs like Heart Failure (HF) and Atrial Fibrillation (AF) are associated with physical effects on the heart muscles. As a result of the complex nature, progression, inherent genetic makeup, and heterogeneity of CVDs, personalized treatments are believed to be critical. Rightful application of artificial intelligence (AI) and machine learning (ML) approaches can lead to new insights into CVDs for providing better personalized treatments with predictive analysis and deep phenotyping. In this study we focused on implementing AI/ML techniques on RNA-seq driven gene-expression data to investigate genes associated with HF, AF, and other CVDs, and predict disease with high accuracy. The study involved generating RNA-seq data derived from the serum of consented CVD patients. Next, we processed the sequenced data using our RNA-seq pipeline and applied GVViZ for gene-disease data annotation and expression analysis. To achieve our research objectives, we developed a new Findable, Accessible, Intelligent, and Reproducible (FAIR) approach that includes a five-level biostatistical evaluation, primarily based on the Random Forest (RF) algorithm. During our AI/ML analysis, we have fitted, trained, and implemented our model to classify and distinguish high-risk CVD patients based on their age, gender, and race. With the successful execution of our model, we predicted the association of highly significant HF, AF, and other CVDs genes with demographic variables.

1. Introduction

Cardiovascular disease (CVD) is the leading causes of mortality and loss of disability adjusted life years (DALYs) globally [1–3]. The World Health Organization (WHO) states that over 75% of premature CVDs are preventable with a better understanding of risk factors and gene-disease associations [2]. CVDs like Heart Failure (HF) and Atrial Fibrillation (AF) are associated with physical impacts on the heart muscles [1,2]. HF occurs due to weak heart muscles that impact the efficiency of pumping blood to the body's cells [1]. While AF occurs due to the high-frequency excitation of the atrium, resulting in both dyssynchronous atrial contraction and the irregularity of ventricular excitation [3,4]. Genomics studies done using genome-wide association studies (GWAS) have aided in disease prediction [5,6], discovery of genetic loci and alleles

Abbreviations: Artificial intelligence, (AI); Atrial fibrillation, (AF); Cardiovascular diseases, (CVDs); Computerized tomography, (CT); Differentially expressed genes, (DEGs); Electronic health records, (EHR); Extract, transfer, and load, (ETL); Fragments per kilobase million, (FPKM); Genome-wide association studies, (GWAS); Heart failure, (HF); Institutional review board, (IRB); Machine learning, (ML); Mean expressed transcript lengths, (METL); Next generation sequencing, (NGS); Normalized enrichment score, (NES); Random forest, (RF); Reads per kilobase of transcript per million mapped reads, (RPKM); RNA-sequencing, (RNA-seq); Sci-kit learn, (Sklearn); Support vector machine, (SVM); Transcripts per million, (TPM); Visualizing genes with disease-causing variants, (GVViZ); World Health Organization, (WHO); Whole genome sequencing, (WGS); Whole exome sequencing, (WES).

* Corresponding author at: Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick 08901, NJ, USA.
E-mail address: zahmed@ifh.rutgers.edu (Z. Ahmed).
[1] Equally contributing first authors.

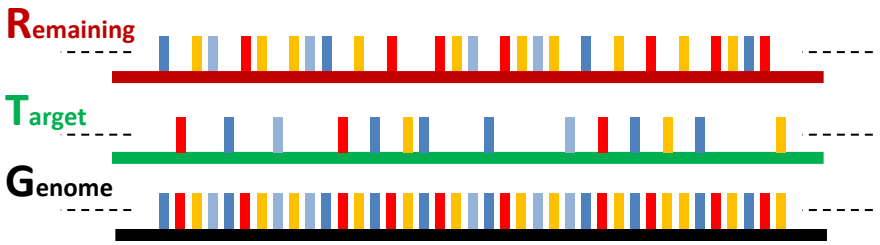
https://doi.org/10.1016/j.ygeno.2023.110564
Received 19 November 2022; Received in revised form 6 February 2023; Accepted 11 February 2023
Available online 20 February 2023
0888-7543/

Biomarkers discovery

Identify new predictive biomarkers

using data-driven approach with a novel nexus of AI/ML algorithms to reveal biological patterns explaining the causal relationship based on genomic, transcriptomic, and phenotypic data of patients with CVD.

Next



Genomics 115 (2023) 110584

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

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

Vignesh Venkat [a,1], Habiba Abdelhalim [a,1], William DeGroat [a], Saman Zeeshan [b], Zeeshan Ahmed [a,c]

[a] Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson St, New Brunswick, NJ, USA
[b] Rutgers Cancer Institute of New Jersey, Rutgers University, 195 Little Albany St, New Brunswick, NJ, USA
[c] Department of Medicine, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson St, New Brunswick, NJ, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Atrial fibrillation
Cardiovascular diseases
Gene expression
Heart failure
Machine learning
Predictive analysis

ABSTRACT

Cardiovascular disease (CVD) is the leading cause of mortality and loss of disability adjusted life years (DALYs) globally. CVDs like Heart Failure (HF) and Atrial Fibrillation (AF) are associated with physical effects on the heart muscles. As a result of the complex nature, progression, inherent genetic makeup, and heterogeneity of CVDs, personalized treatments are believed to be critical. Rightful application of artificial intelligence (AI) and machine learning (ML) approaches can lead to new insights into CVDs for providing better personalized treatments with predictive analysis and deep phenotyping. In this study we focused on implementing AI/ML techniques on RNA-seq driven gene-expression data to investigate genes associated with HF, AF, and other CVDs, and predict disease with high accuracy. The study involved generating RNA-seq data derived from the serum of consented CVD patients. Next, we processed the sequenced data using our RNA-seq pipeline and applied GVViZ for gene-disease data annotation and expression analysis. To achieve our research objectives, we developed a new Findable, Accessible, Intelligent, and Reproducible (FAIR) approach that includes a five-level biostatistical evaluation, primarily based on the Random Forest (RF) algorithm. During our AI/ML analysis, we have fitted, trained, and implemented our model to classify and distinguish high-risk CVD patients based on their age, gender, and race. With the successful execution of our model, we predicted the association of highly significant HF, AF, and other CVDs genes with demographic variables.

1. Introduction

Cardiovascular disease (CVD) is the leading causes of mortality and loss of disability adjusted life years (DALYs) globally [1–3]. The World Health Organization (WHO) states that over 75% of premature CVDs are preventable with a better understanding of risk factors and gene-disease associations [2]. CVDs like Heart Failure (HF) and Atrial Fibrillation (AF) are associated with physical impacts on the heart muscles [1,2]. HF occurs due to weak heart muscles that impact the efficiency of pumping blood to the body's cells [1]. While AF occurs due to the high-frequency excitation of the atrium, resulting in both dyssynchronous atrial contraction and the irregularity of ventricular excitation [3,4]. Genomics studies done using genome-wide association studies (GWAS) have aided in disease prediction [5,6], discovery of genetic loci and alleles

Abbreviations: Artificial intelligence, (AI); Atrial fibrillation, (AF); Cardiovascular diseases, (CVDs); Computerized tomography, (CT); Differentially expressed genes, (DEGs); Electronic health records, (EHR); Extract, transfer, and load, (ETL); Fragments per kilobase million, (FPKM); Genome-wide association studies, (GWAS); Heart failure, (HF); Institutional review board, (IRB); Machine learning, (ML); Mean expressed transcript lengths, (METL); Next generation sequencing, (NGS); Normalized enrichment score, (NES); Random forest, (RF); Reads per kilobase of transcript per million mapped reads, (RPKM); RNA-sequencing, (RNA-seq); Sci-kit learn, (Sklearn); Support vector machine, (SVM); Transcripts per million, (TPM); Visualizing genes with disease-causing variants, (GVViZ); World Health Organization, (WHO); Whole genome sequencing, (WGS); Whole exome sequencing, (WES).

* Corresponding author at: Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick 08901, NJ, USA.
E-mail address: zahmed@ifh.rutgers.edu (Z. Ahmed).
[1] Equally contributing first authors.

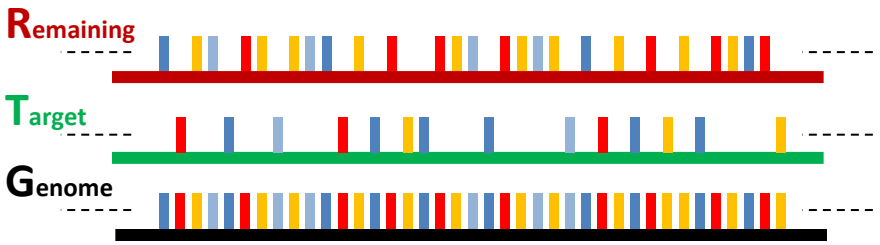
https://doi.org/10.1016/j.ygeno.2023.110584
Received 19 November 2022; Received in revised form 6 February 2023; Accepted 11 February 2023
Available online 20 February 2023

Biomarkers discovery

New AI/ML Pipeline i.e., IntelliGenes

Identify new predictive biomarkers

using data-driven approach with a novel nexus of AI/ML algorithms to reveal biological patterns explaining the causal relationship based on genomic, transcriptomic, and phenotypic data of patients with CVD.



Bioinformatics

OXFORD UNIVERSITY PRESS | Bioinformatics

IntelliGenes: AI/ML pipeline for predictive analyses using multi-genomic profiles

Journal:	*Bioinformatics*
Manuscript ID	Draft
Category:	Applications Note
Date Submitted by the Author:	n/a
Complete List of Authors:	Degroat, William; Rutgers University Institute for Health Health Care Policy and Aging Research Mendhe , Dinesh ; Rutgers Univers Institute for Health Health Care Policy and Aging Research Bhusari , Atharva ; Rutgers Univ titute for Health Health Care Policy and Aging Research Abdelhalim, Habiba; Rutg versity Institute for Health Health Care Policy and Aging Research Zeeshan, Saman; Ru Institute of New Jersey Ahmed, Zeesh Rutgers University Institute for Health Health Care Policy and Aging Research Medicine; Rutgers Robert Wood Johnson Medical School New Brunswick, Department of Medicine
Portal Keywords:	Artificial in e, Machine learning, Genomics, Biomarker, Prediction
Keywords:	Intelli Artificial intelligence, Machine learning, Multi-Genomics, rs, Disease Prediction

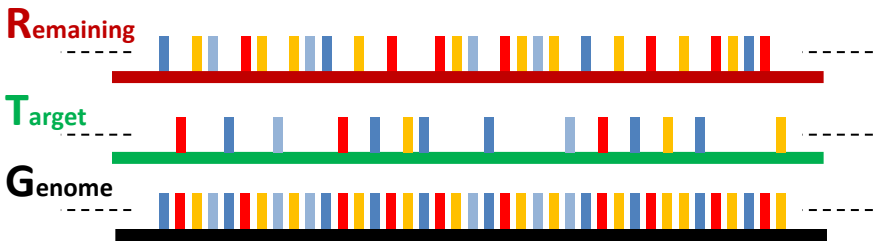
Revised

SCHOLARONE™
Manuscripts

Biomarkers discovery

Identify new predictive biomarkers

using data-driven approach with a novel nexus of AI/ML algorithms to reveal biological patterns explaining the causal relationship based on genomic, transcriptomic, and phenotypic data of patients with CVD.



Biomarker discovery using IntelliGenes

Title

Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine

Running Head

Discovering biomarkers and predicting CVD using AI/ML

Authors

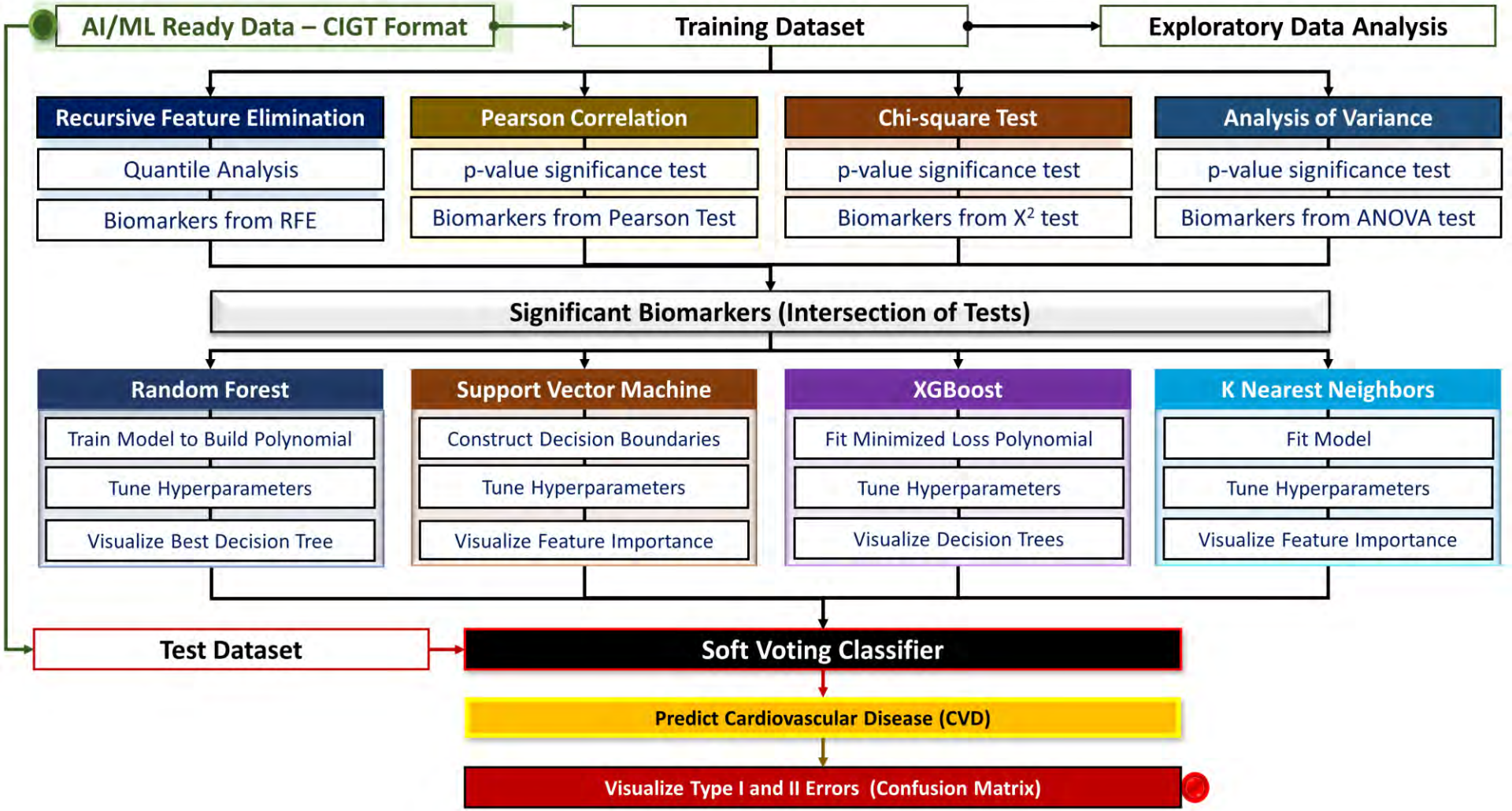
William DeGroat[1], Habiba Abdelhalim[1], Kush Patel[1], Dinesh Mendhe[1], Saman Zeeshan[2], and Zeeshan Ahmed[1, 3, *]

Affiliations

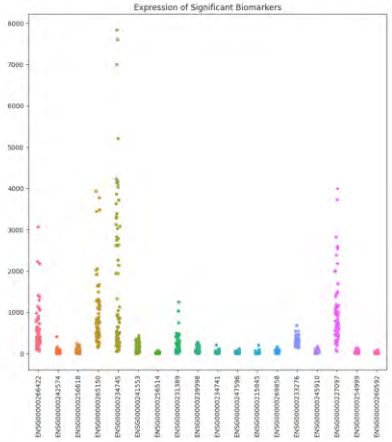
1. Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson St, New Brunswick, NJ, USA.
2. Rutgers Cancer Institute of New Jersey, Rutgers University, 195 Little Albany St, New Brunswick, NJ, USA.
3. Department of Medicine/Cardiovascular Disease and Hypertension, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson St, New Brunswick, NJ, USA.

***Corresponding author:** Zeeshan Ahmed, Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick, 08901, NJ, USA. (zahmed@ifh.rutgers.edu).

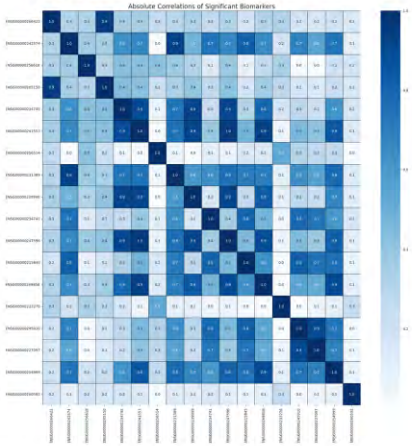
IntelliGenes: Nexus of AI/ML approaches



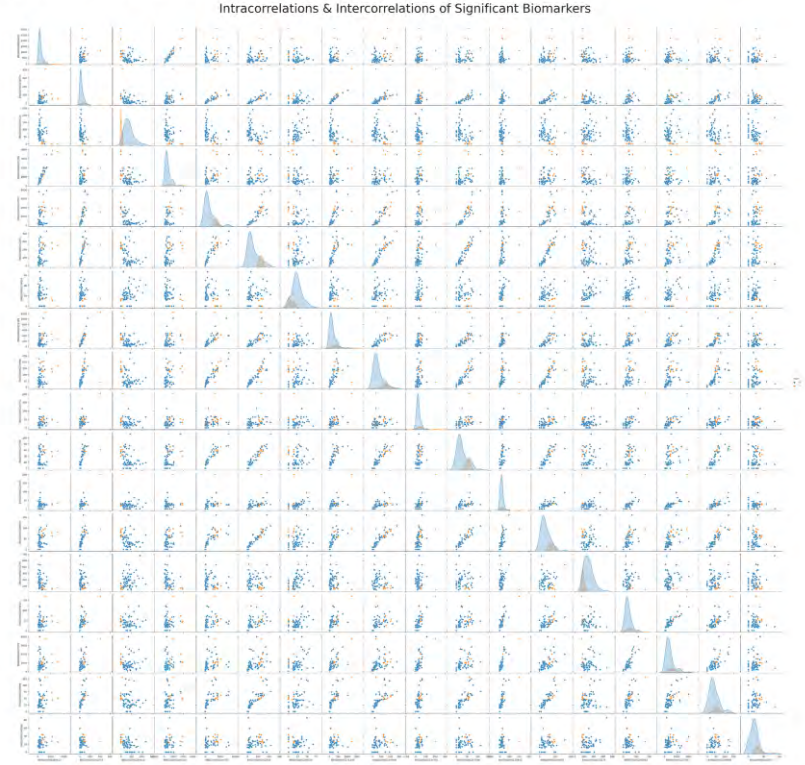
A). Biomarker Expression



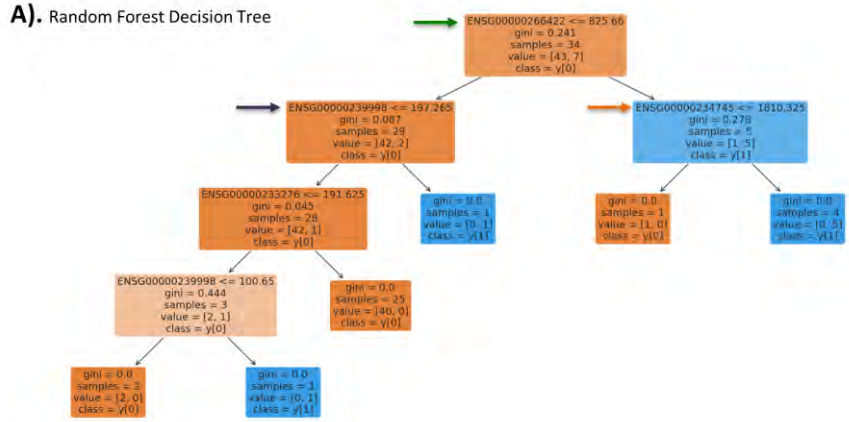
B). Biomarker Correlations



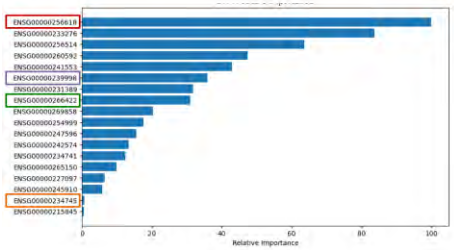
C). Biomarker Pairwise Relationships



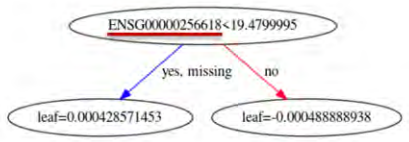
A). Random Forest Decision Tree



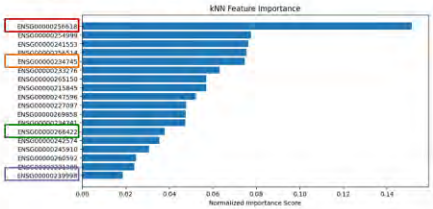
B). Support Vector Machine Feature Importance



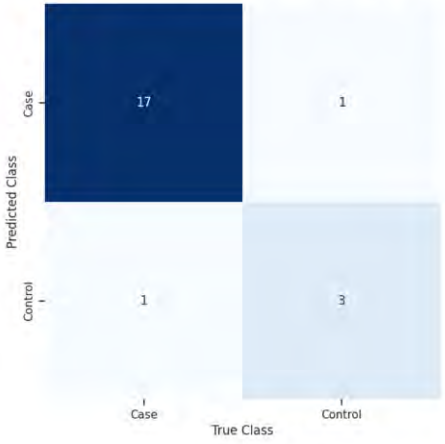
C). XGBoost Decision Tree



D). k-Nearest Neighbors



E). Soft Voting Machine Predictions Confusing Matrix



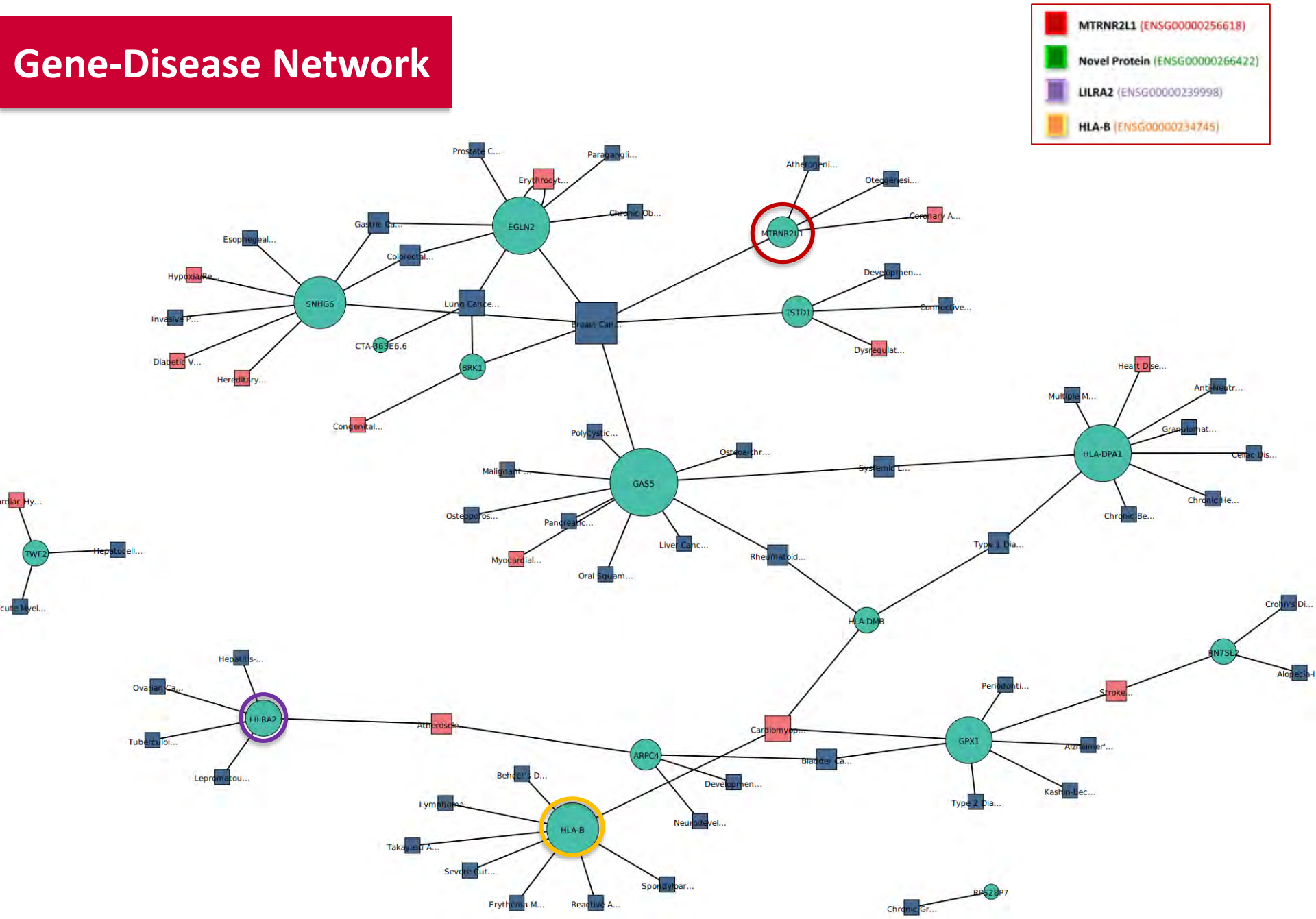
MTRNR2L1 (ENSG00000256618)

Novel Protein (ENSG00000266422)

LILRA2 (ENSG00000239998)

HLA-B (ENSG00000234745)

Gene-Disease Network



Validating List of Biomarkers with EHR

Diagnosis	ICD9	ICD10	Gene					
Type 2 or unspecified diabetes mellitus with peripheral circulatory disorder [Type 2 Diabetes]	250	E11.51	GPX1					
Osteoarthritis [Osteoarthritis]	715	M19.90	GAS5					
History of non-Hodgkins lymphoma [Diffuse Large B-cell Lymphoma]**	V10	Z85.72	HLA-B					
Malignant neoplasm of upper-outer quadrant of right female breast. unspecified estrogen receptor status (CMS/HCC) [Breast Cancer]	174	C50.411	MTRNR2L1	GAS5	TSTD1	EGLN2	SNHG6	BRK1
Seronegative arthritis [Rheumatoid Arthritis]**	716	M13.80	HLA-DMB	GAS5				
Mass of upper inner quadrant of right breast [Breast Cancer]	611	N63.12	MTRNR2L1	GAS5	TSTD1	EGLN2	SNHG6	BRK1
Coronary artery disease involving native heart with angina pectoris. unspecified vessel or lesion type (CMS/HCC) [Coronary Artery Disease]	414	I25.119	MTRNR2L1					
Special screening for malignant neoplasms. colon [Colorectal Cancer]	V76	Z12.11	EGLN2	SNHG6				
Seronegative rheumatoid arthritis (CMS/HCC) [Rheumatoid Arthritis]	714	M06.00	HLA-DMB	GAS5				
Family history of ovarian cancer [Ovarian Carcinoma]	V16	Z80.41	LILRA2					
Malignant neoplasm of upper lobe. right bronchus or lung (CMS/HCC) [Lung Cancer]	162	C34.11	EGLN2	BRK1	CTA-363E6.6			
Other malignant lymphoma of extranodal or solid organ sites [Diffuse Large B-cell Lymphoma]**	202	C85.89	HLA-B					
Other diabetic neurological complication associated with other specified diabetes mellitus (CMS/HCC) [Type 1 Diabetes]	249	E13.49	HLA-DMB	HLA-DPA1				
NSTEMI (non-ST elevated myocardial infarction) (CMS/HCC) [Myocardial Infarction]	410	I21.4	GAS5					
Obscure cardiomyopathy of Africa (CMS/HCC) [Cardiomyopathy]	425	I42.8	HLA-DMB	HLA-B	GPX1			
Other atherosclerosis of native artery of extremity [Atherosclerosis]	440	I70.299	ARPC4	LILRA2				
Family history of ischemic heart disease [Coronary Artery Disease]	V17	Z82.49	MTRNR2L1					
Wegeners granulomatosis (CMS/HCC) [Granulomatosis with Polyangiitis]	446	M31.30	HLA-DPA1					
Mantle cell lymphoma (CMS/HCC) [Diffuse Large B-cell Lymphoma]**	200	C83.10	HLA-B					
Viral hepatitis [Chronic Hepatitis B Virus]**	070	B19.9	HLA-DPA1					
Hereditary and idiopathic peripheral neuropathy [Neurodevelopmental Disorders]**	356	G60.9	ARPC4					
Malignant neoplasm of posterior wall of bladder (CMS/HCC) [Bladder Cancer]	188	C67.4	ARPC4	GPX1				
Carcinoma in situ of breast [Breast Cancer]	233	D05.90	MTRNR2L1	GAS5	TSTD1	EGLN2	SNHG6	BRK1
Need for prophylactic vaccination and inoculation against viral hepatitis [Chronic Hepatitis B Virus]**	V05	Z23	HLA-DPA1					
Polycystic ovaries [Polycystic Ovary Syndrome]	256	E28.2	GAS5					
Malignant neoplasm of colon (CMS/HCC) [Colorectal Cancer]	153	C18.9	EGLN2	SNHG6				
Telangiectasia [Hereditary Haemorrhagic Telangiectasia]**	448	I78.1	SNHG6					
Malignant neoplasm of prostate (CMS/HCC) [Prostate Cancer]	185	C61	EGLN2					
Interstitial lung disease (CMS/HCC) [Connective Tissue Disease-Associated Interstitial Lung Disease]**	515	J84.9	TSTD1					
Chronic periodontitis [Periodontitis]	523	K05.30	GPX1					
Secondary malignant neoplasm of lung (CMS/HCC) [Lung Cancer]	197	C78.00	EGLN2	BRK1	CTA-363E6.6			
Squamous cell cancer of epiglottis (CMS/HCC) [Oral Squamous Cell Carcinoma]	161	C32.1	GAS5					
Chronic obstructive pulmonary disease. unspecified COPD type (CMS/HCC) [Chronic Obstructive Pulmonary Disease]	496	J44.9	EGLN2					
Chronic myeloid leukemia (CMS/HCC) [Acute Myeloid Leukemia]**	205	C92.10	TWF2					
Ectopic pregnancy without intrauterine pregnancy [Development of Ectopic Pregnancy]	633	O00.90	TSTD1					
Old myocardial infarction [Myocardial Infarction]	412	I25.2	GAS5					

Summary

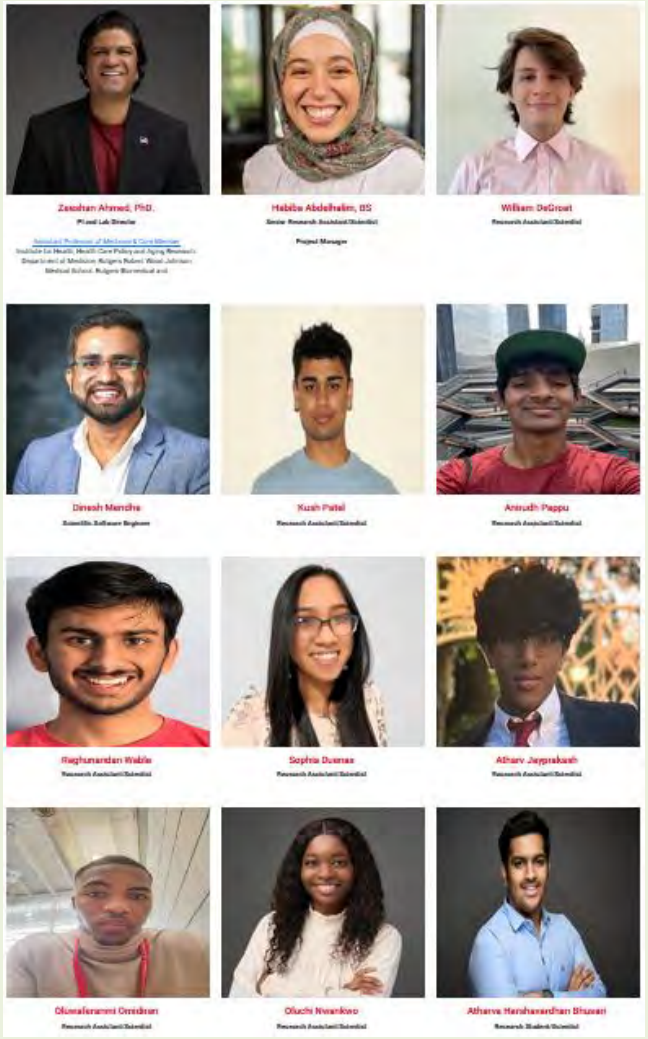
- Personalized interventions are deemed vital given the intricate characteristics, advancement, inherent genetic composition, and diversity of CVD.
  - Implemented orthodox bioinformatics analyses of RNA-seq and WGS data.
  - Investigated genes, known to be associated with CVDs.
- The appropriate utilization of AI and ML methodologies can yield novel understandings of CVDs, enabling improved personalized treatments through predictive analysis and deep phenotyping.
  - Developed AI/ML approaches to
    - predict CVD and identify risk factors
    - discover novel biomarkers
- The synergistic use of multiple AI algorithms provides more accurate results, draws insightful conclusions, and precise predictions about real-world problems compared to single AI algorithm on its own.
- With its successful implementation, our newly developed predictive engine can provide a valuable framework for identifying patients with CVDs based on their biomarker profiles.

Lab Members (2020 – 23)

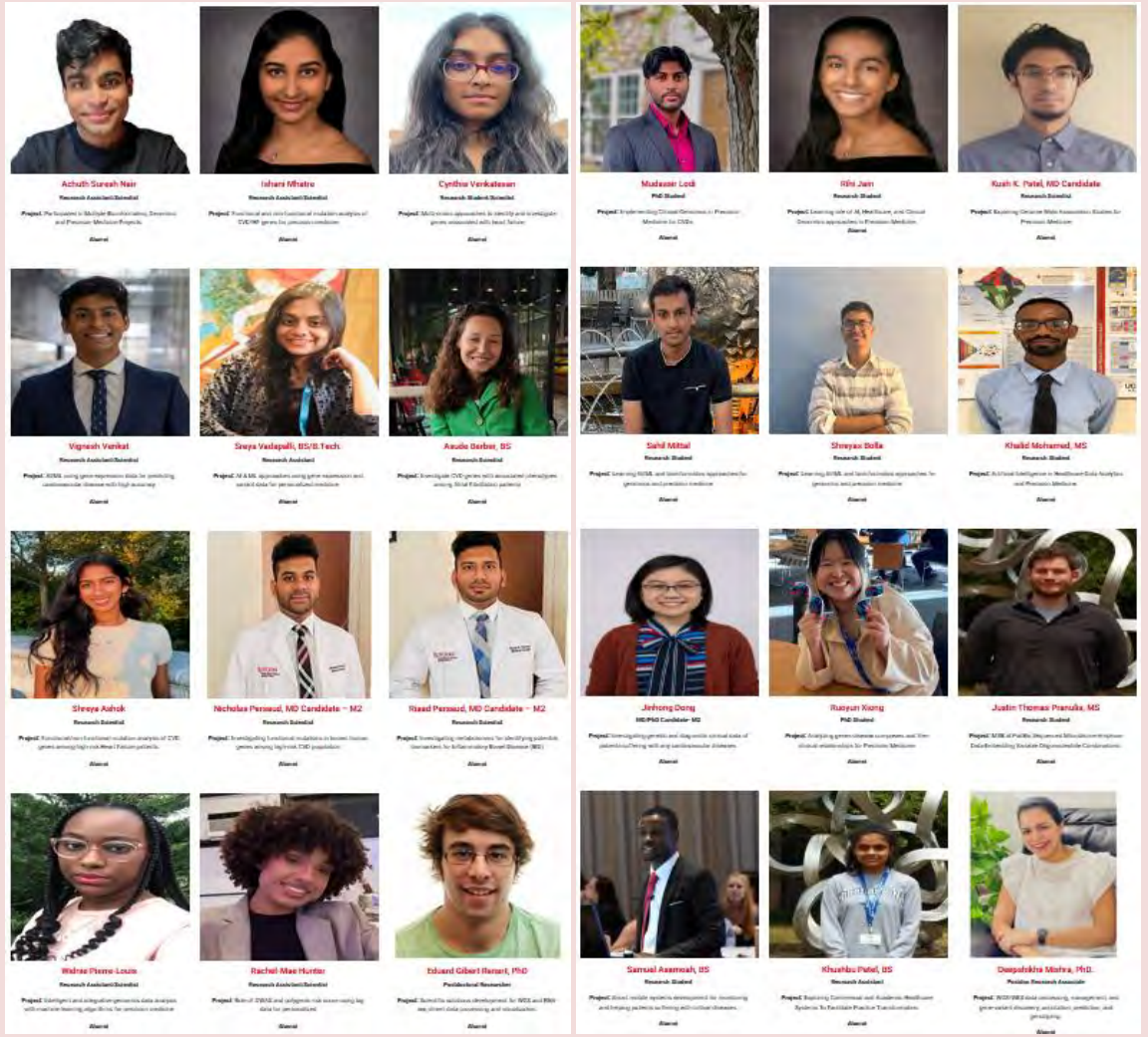
Ahmed Lab

RUTGERS
Institute for Health, Health Care Policy and Aging Research
Rutgers Robert Wood Johnson Medical School
Rutgers Biomedical and Health Sciences

Active



Alumni



Collaborations

Acknowledgement

Bruce T. Liang, M.D., F.A.C.C.
Dean, UConn School of Medicine, UConn Health
Director, Pat and Jim Calhoun Cardiology Center

Partho P. Sengupta, M.D.
Chief of the Division of Cardiovascular Disease & Hypertension.
Chief of Cardiology, Robert Wood Johnson Medical School.

Daniel Fine, D.M.D
Chair, Department of Oral Biology
Rutgers School of Dental Medicine

Danielle Dick, Ph.D.
Director, Rutgers Addiction Research Center (RARC)

Suvi Linna-Kuosmanen, PhD.
Massachusetts Institute of Technology (MIT).

Yuichiro Arima, M.D., PhD.
Kumamoto University, Japan.

Naveena Yanamala, Ph.D.
Director of Data Science and Machine Learning Research, RWJMS

Olga F. Jarrín Montaner, Ph.D.
Director, Community Health and Aging Outcomes Lab

Tammy Chung, Ph.D.
Director, Center for Population Behavioral Health, Rutgers IFH

Saman Zeeshan, Ph.D.
Rutgers Cancer Institute of New Jersey

Ahmed Lab

Grants/Funding

The New York Academy of Sciences

AMED 国立研究開発法人 日本医療研究開発機構 Japan Agency for Medical Research and Development

NIH National Institute on Aging

Rutgers Brain Health Institute

Rutgers Biomedical and Health Sciences

UConn School of Medicine

Google Research

CMS TCPi Transforming Clinical Practice Initiative

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Ahmed Lab

Zeeshan Ahmed, Ph.D.

Email: < zahmed@ifh.rutgers.edu **>**

Ahmed Lab: < https://promis.rutgers.edu/ **>**

Precision Medicine Project: < https://sites.rutgers.edu/precision-medicine/ **>**

Institute for Health, Health Care Policy and Aging Research
Rutgers University-New Brunswick

Robert Wood Johnson Medical School
Rutgers Biomedical and Health Sciences