Analysis of a real patient's genetic variants using Nirvana annotation tool and automated integration of CardioDB database

1st Bianca Popa
Department of Computer Engineering
Valencia Polytechnic University
Valencia, Spain
bianca-popa@hotmail.com

***Abstract*—The absence of standardized genomic database nomenclature, coupled with a lack of comprehensive tools to automate genetic variant analysis, contributes to a lengthy and arduous diagnostic process of genetic diseases. It is imperative, therefore, to integrate and develop new tools and methods to automate the task of analyzing genetic variants. In this paper, we utilize an open-source tool called Nirvana to annotate genetic variants in a real clinical case. Additionally, we present an automated technique for incorporating the CardioDB database into Nirvana, leveraging its information during the annotation process. Analysis of our patient's genetic variations reveals a high occurrence of variants associated with cardiac disease in general, specifically a type of cardiomyopathy. The probable genetic diagnosis is therefore related to this specific condition. These results indicate that the use of automated tools and methods for integrating genomic databases and annotating variants has a positive impact on the analysis of a patient's genetic profile, reducing its time and the cost of expertise.**

***Index Terms*—Nirvana, annotation process, database integration, automation, variant analysis, cardiomyopathy, cardiac disease**

I. Introduction

Despite significant advances in sequencing technologies and the accumulation of vast amounts of genomic data, the analysis of genetic variants remains a considerable challenge in clinical practice. The lack of standardized annotation and classification of variants across databases as well as the lack of comprehensive tools to automate the analysis process are barriers to effective variant analysis. Diagnostic procedures still rely heavily on the expertise of geneticists and molecular biologists who manually examine variants one by one. This process, which is both time-consuming and laborious, increases the risk of human error and subjectivity when interpreting variations [5].

In this context, we are interested in an open source tool, Nirvana, which brings together relevant information from different genomic databases to characterize genetic variants [3]. We have developed an automated method to integrate an additional database, CardioDB, into the Nirvana tool for adding custom annotations. This method relies on the execution of a specific Python script. We then applied Nirvana's annotation process enriched with data from CardioDB to the actual genetic variants of a patient provided to us in VCF format. Finally, we analyzed the variants whose clinical significance has been reported as pathogenic to hypothesize about the patient's genetic diagnosis.

We wanted to demonstrate that it is possible, using automatic tools and methods, to speed up the process of analyzing genetic variations for a real patient case with a heart disease profile. Since the analysis was carried out solely from the data obtained using these methods, it is also possible, to a certain extent, to reduce the need for recourse to medical expertise. Obviously, the quality of the analysis is limited by the power and precision of the automatic methods used.

II. Materials and Methods

The Nirvana tool, which provides clinical-grade annotation of genomic variants of different types, was used to annotate one particular VCF format file named `VCF_2.vcf`, which corresponds to the genetic variants of a real patient. This patient, whose information is anonymized, will be identified as patient 2. The patient's VCF format file represents the input to the annotation process, while the output corresponds to a JSON format file generated by the tool itself. This JSON file is the one used for the analysis of the patient's genetic profile in subsequent sections.

In order to enrich the results of the annotation process, an external database, CardioDB, has been integrated into the Nirvana tool. An upstream transformation of the variant data present within CardioDB was necessary to meet the format expected by Nirvana for integration. This transformation takes as input the CSV file of the database containing 1216 variants (only substitutions) with 9 attributes, namely the gene, the nucleotide change, the protein change, the variant consequence, the OMGL class, the LMM class, the phenotype, type and location on the GRCh37 reference genome, and an augmented attribute, the correct reference allele. The output is a TSV file in the format accepted by Nirvana for custom annotations. This file contains 7 attributes describing the variants, namely the chromosome, the position, the reference allele, the alternative allele, the OMGL class, the LMM class and the phenotype.

A script written in Python language made it possible to perform the data transformation automatically. The main parts of the script are a method for writing the TSV file header, a method for writing variants, a method for eliminating duplicate variants and a method for sorting variants according to chromosome and position. These last two methods are necessary to prevent errors during integration. These methods are defined in the `cardiodb_for_nirvana.py` file. Other utility functions for the subsequent analysis of pathogenic genetic variations are defined in the `analyze_variants.py` file. All of the code is available in the project's GitHub repository at https://github.com/BiancaP-hub/variant_analysis.

III. Results

A total of 1564 variants of patient 2 were annotated by the Nirvana tool. Of these 1564 variants, 1403 appear in ClinVar's database, while none appear in CardioDB's. Therefore, no variant appears in both databases at the same time. In this context, it is not possible to analyze the concordance between the information provided by ClinVar and that provided by CardioDB. Of the 1403 variants annotated using the ClinVar database, 22 are found to be pathogenic. Table I presents the list of pathogenic variants according to the position where they are found on the corresponding chromosome, as well as their type and their dbSNP identifier (if available).

For the analysis of the effect of these variants on the patient's condition, additional information is necessary on the phenotypes associated with the variants, as given by the ClinVar database, as well as on the genes targeted by the variants and their role. Table I presents the phenotypes associated with each of the pathogenic variants identified, if the information was obtained following the annotation of the patient's VCF file. This table also shows the correspondence between the variants and the targeted genes.

Table II presents the function of all the genes targeted by the pathogenic variants of patient 2. Note that all of those genes have the function of coding a protein playing a specific role in the cell.

The following graph shows the proportion of times the physical characteristics (phenotypes) mentioned in Table I are associated with heart disease or defects, compared to their association with other types of conditions. Among the associations with heart disease, we find in particular those with cardiomyopathies, the proportion of which is also shown. This graphic representation makes it possible to visualize the frequency of appearance of these physical characteristics in the patient.

On the left, the number of variants whose phenotype is associated with heart disease is 9 out of 16 variants with at least one associated phenotype (56.25%), while it is 7 for the other types of conditions. On the right, the number of variants whose phenotype is associated with a type of cardiomyopathy is 5 out of 9 variants associated with a heart condition in general (55.55%).

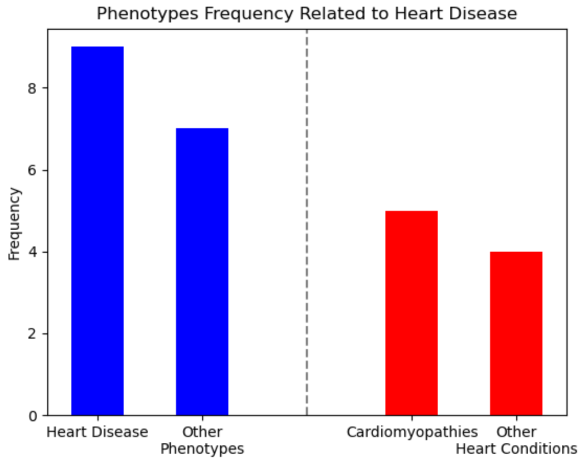

Fig. 1. Frequency of heart disease-related phenotypes compared to other phenotypes

IV. Discussion

First, the inclusion of the CardioDB database in the annotation process was not at all useful for the analysis. Indeed, no patient variant was included in this database. A manual check revealed a variant located at the same position on chromosome 14 in CardioDB as an annotated variant. That variant expresses a similar phenotype and has the same clinical significance (i.e. pathogenicity). However, it does not carry the same alternative allele. It is therefore not the same variant.

To make hypotheses on the genetic diagnosis of the patient, the most important thing is to consider the phenotypes associated with each of the pathogenic variants identified. Remember that a pathogenic variant is a genetic alteration that causes a disease or increases the susceptibility or predisposition to developing it [6]. As can be seen in Fig. 1, the majority of patient physical characteristics are associated with heart disease, particularly disease that affects the muscle of the heart. So the most likely genetic diagnosis for patient 2 is (a predisposition to) some type of cardiomyopathy, a condition that weakens the heart and makes it difficult for it to pump blood to the rest of the body. The symptoms that could be experienced by the patient are difficulty breathing, pain felt in the chest, dizziness or fainting and even, in the worst case, heart failure [1].

Among the different possible types of cardiomyopathies, we find dilated cardiomyopathy (DCM) and hypertrophic cardiomyopathy (HCM). In either case, the heart grows and weakens, making it difficult to pump blood efficiently [1]. The variant at position 23895180 of chromosome 14 is the most significant for making this diagnosis, since it is associated with the greatest number of myopathy-related phenotypes. It is also a variant that has been reviewed by a panel of experts recognized by the United States Food and Drug Administration (FDA), confirming its pathogenicity [2].

The genetic diagnosis of the patient is supported not only

TABLE I
PATIENT 2 VARIANTS CLASSIFIED AS PATHOGENS, ASSOCIATED PHENOTYPES AND TARGETED GENES

Chromosome	Position	Type	dbSNP	Phenotype	Gene
chr2	27535451	deletion	rs766160589	Navajo neurohepatopathy	MPV17
chr2	73716810	SNV	rs3820700	Alstrom syndrome	ALMS1
chr2	220283699	deletion	rs769096434	Desmin-related myofibrillar myopathy	DES
chr2	241808314	SNV	rs34116584	—	AGXT
chr3	8775660	insertion	rs1008642	Distal myopathy, Tateyama type Long QT syndrome	CAV3
chr3	38645420	SNV	rs1805124	Progressive familial heart block, type 1A	SCN5A
chr3	193382703	deletion	—	—	OPA1
chr5	226160 251541	SNV	rs6555055 rs13070	Mitochondrial complex II deficiency, nuclear type 1 Paragangliomas 5	SDHA
chr6	44269191	deletion	rs771808127	—	POLR1C
chr6	152658142	SNV	rs9479297	—	SYNE1
chr7	150648198	SNV	rs1137617	—	KCNH2
chr8	11606312 11612698	SNV	rs3735819 rs804280	Congenital heart disease	GATA4
chr8	30924557	SNV	rs1800389	Werner syndrome	WRN
chr10	121436362	SNV	rs196295	Myofibrillar myopathy 6 Dilated cardiomyopathy 1HH	BAG3
chr11	6411935	SNV	rs1050228	Niemann-Pick disease, type A	SMPD1
chr14	23895180	SNV	rs121913637	Cardiomyopathy Cardiovascular phenotype Hypertrophic cardiomyopathy 1 Myopathy, myosin storage, autosomal recessive Congenital myopathy with fiber type disproportion Dilated cardiomyopathy 1S MYH7-related skeletal myopathy MYH7-related late-onset scapuloperoneal muscular dystrophy Primary familial hypertrophic cardiomyopathy	MYH7
chr14	73664813	deletion	—	Alzheimer disease 3 Frontotemporal dementia Pick disease Acne inversa, familial, 3	PSEN1
chr16	16271357	SNV	rs6416668	—	ABCC6
chr17	29508775	SNV	rs1801052	Neurofibromatosis, type 1	NF1
chr17	78083791	SNV	rs1800305	Glycogen storage disease, type II	GAA

by variants whose phenotype is directly associated with a type of cardiomyopathy, but also by other variants whose associated phenotype's consequences also affect the heart. For example, the variant at position 78083791 of chromosome 17 is responsible for a defect in glycogen storage, which can have severe consequences on the heart. Without adequate glycogen breakdown, excessive amounts of glycogen can build up in heart muscle cells, leading to enlargement of the heart [5]. We thus see how different variants collaborate for the expression of the characteristics associated with cardiomyopathy.

Among the genes targeted by the variants, some have a demonstrated link with cardiomyopathies. Notably, the MYH7 gene was the first gene known to cause hereditary hypertrophic cardiomyopathy. This gene encodes the beta heavy chain of myosin, a protein found in heart muscle [7, p. 4]. This protein plays a fundamental role in the mechanisms of muscle contraction. A mutation in the gene encoding this protein can therefore prevent the heart muscle from contracting enough for the proper transport of blood [4]. Mutations of the MYH7 gene, combined with those of the MYBPC3 gene, account for 40 to 50% of clinically proven cases of HCM [7, p. 4]. A second gene relevant to the analysis is the KCNH2 gene, targeted by the rs1137617 variant. This gene encodes a component of a voltage-activated potassium channel found in cardiac muscle. This channel controls heart repolarization, the process by which cardiac muscle cells return to their pre-contraction electrical state in preparation for the next one [2]. A defect in this channel linked to a mutation in the KCNH2 gene can disrupt this repolarization phase, leading to irregular contractions.

It is important to emphasize that, in the absence of a thorough evaluation of other medical data, such as patient records, medical history and results of additional genetic tests, the proposed diagnosis remains a probabilistic hypothesis.

V. CONCLUSION

Analysis of patient 2's genetic variants from information obtained through Nirvana's annotation process revealed a high frequency of occurrence of phenotypes related to heart disease, and in particular cardiomyopathy. This suggests a probable diagnosis related to this condition in the patient. In the case under study, the integration of the CardioDB database had no impact on the analysis of genetic variants, since none of them were part of it. However, in other patient cases, it is possible that the additional information from this database may reveal other pathogenic variants or rule out some of them.

TABLE II
FUNCTION OF THE GENES TARGETED BY IDENTIFIED PATHOGENIC VARIANTS

Gene	Gene function
MPV17	Encodes a mitochondrial inner membrane protein that is involved in mitochondrial deoxynucleotide homeostasis and maintenance of mtDNA.
ALMS1	Encodes a protein containing a large tandem-repeat domain as well as additional low complexity regions.
DES	Encodes a muscle-specific class III intermediate filament.
AGXT	Encodes alanine:glyoxylate aminotransferase (AGT; EC 2.6.1.44), whose activity is largely confined to peroxisomes in the liver.
CAV3	Encodes a caveolin family member, which functions as a component of the caveolae plasma membranes found in most cell types.
SCN5A	Encodes an integral membrane protein and tetrodotoxin-resistant voltage-gated sodium channel subunit. This protein is found primarily in cardiac muscle and is responsible for the initial upstroke of the action potential in an electrocardiogram.
OPA1	Encodes a protein that localizes to the inner mitochondrial membrane and regulates several important cellular processes including stability of the mitochondrial network.
SDHA	Encodes a major catalytic subunit of succinate-ubiquinone oxidoreductase, a complex of the mitochondrial respiratory chain.
POLR1C	Encodes a subunit of both RNA polymerase I and RNA polymerase III complexes.
SYNE1	Encodes nesprin-1, a member of the spectrin family of structural proteins that link the nuclear plasma membrane to the actin cytoskeleton.
KCNH2	Encodes the pore-forming subunit of a rapidly activating-delayed rectifier potassium channel that plays an essential role in the final repolarization of the ventricular action potential.
GATA4	Encodes a member of the GATA family of zinc-finger transcription factors. This protein is thought to regulate genes involved in embryogenesis and in myocardial differentiation and function.
WRN	Encodes a member of the RecQ subfamily of DNA helicase proteins. The encoded nuclear protein is important in the maintenance of genome stability.
BAG3	Encodes one of the cytoprotective proteins that bind to and regulate Hsp70 family molecular chaperones.
SMPD1	Encodes a lysosomal acid sphingomyelinase that converts sphingomyelin to ceramide.
MYH7	Encodes the beta-cardiac/slow skeletal myosin heavy chain (MyHC-slow), expressed predominantly in the cardiac ventricles and slow skeletal (type 1) myofibers.
PSEN1	Encodes presenilin-1, which forms the catalytic component of gamma-secretase.
ABCC6	Encodes a member of the superfamily of ATP-binding cassette (ABC) transporters. The encoded protein, a member of the MRP subfamily, is involved in multi-drug resistance.
NF1	Encodes neurofibromin, a cytoplasmic protein that is predominantly expressed in neurons, Schwann cells, oligodendrocytes, and leukocytes.
GAA	Encodes lysosomal alpha-glucosidase, which is essential for the degradation of glycogen to glucose in lysosomes.

We successfully utilized automated techniques to analyze a comprehensive set of 1564 variants in a patient with a heart disease profile. Through this process, we were able to identify pathogenic variants and make a probable genetic diagnosis without the need to individually examine each variant and with limited prior domain knowledge. In the future, it would be valuable to develop a knowledge model (conceptual model) that formalizes the analysis of genetic variations, thereby enhancing the overall understanding and interpretation of such data.

ACKNOWLEDGMENT

I sincerely thank my Genomic Database Design (14104) professors at the Polytechnic University of Valencia for providing me with the data used for the genetic variant analysis and for their support in resolving technical issues encountered during the completion of this project. Their guidance and assistance have played a pivotal role in the successful completion of this scientific paper.

REFERENCES

[1] “American Heart Association,” www.heart.org. https://www.heart.org (accessed May 12, 2023).

[2] ClinVar, “ClinVar.” https://www.ncbi.nlm.nih.gov/clinvar/ (accessed May 12, 2023).

[3] Illumina, “GitHub - Illumina/Nirvana: The nimble & robust variant annotator,” GitHub. https://github.com/Illumina/Nirvana (accessed May 12, 2023).

[4] M. Hesaraki et al., “A Novel Missense Variant in Actin Binding Domain of MYH7 Is Associated With Left Ventricular Noncompaction,” Frontiers in Cardiovascular Medicine, vol. 9, Apr. 2022, doi: 10.3389/fcvm.2022.839862.

[5] “National Center for Biotechnology Information.” https://www.ncbi.nlm.nih.gov/ (accessed May 12, 2023).

[6] “NCI Dictionary of Genetics Terms,” National Cancer Institute. https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/pathogenic-variant (accessed May 12, 2023).

[7] P. Richard, F. Ader, and P. Charron, “Génétique des cardiomyopathies héréditaires,” EMC - Cardiologie, vol. 13, no. 3, pp. 1–19, Aug. 2018, doi: 10.1016/S1166-4568(18)53103-X.