Pak. J. Biochem. Mol. Biol., 2022, 55(3):162-168

Research paper

Identification of rare deleterious genetic variations in a Pakistani obese Individual

Muhammad Shakeel[1], Waqasuddin Khan[1,3], Muhammad Irfan[1], Ishtiaq Ahmad Khan[1*], M. Kamran Azim[2*]

[1]Jamil-ur-Rahman Center for Genome Research, Dr. Panjwani Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi, Karachi-75270, Pakistan.

[2]Department of Biosciences, Mohammad Ali Jinnah University, Karachi, Pakistan.

*Corresponding authors: Dr. Ishtiaq Ahmed (ishtiaq.ahmed@iccs.edu); Prof. M. Kamran Azim (kamran.azim@jinnah.edu)

ABSTRACT

Obesity is one of the major health concerns globally, which is increasing the burden of other lifestyle diseases. This study involves genome-wide finding of potentially deleterious variants in an obese individual using NGS technology. The annotation of the sequencing data with ANNOVAR, CADD, and VEP tools, emphasized multiple genetic risk factors for obesity, Hyperlipidemia, and associated comorbidities. These analyses showed 425 missense SNVs predicted as deleterious by SIFT, Polyphen2, and CADD, including a novel homozygous SNV in obesity-associated 5-methyltetrahydrofolate-homocysteine methyltransferase reductase (MTRR). Two protein truncating SNVs, heterozygous rs328 (p: Ser474X) in lipoprotein lipase (LPL), and homozygous rs885985 (p: Glu37X) in Claudin 5 (CLDN5) were also found. By comparing population allele frequencies, 11 predicted deleterious missense SNVs were found to have higher allele frequency in South Asians compared to global populations of 1000 Genomes Project.

KEYWORDS: Deleterious variants, Prioritization of variants, Obesity, MTRR, LPL, CLDN5

INTRODUCTION

Obesity is a highly prevalent metabolic disorder worldwide [1]. Obesity is associated with many lifestyle disorders including cardiovascular diseases. Large cohort studies have shown a genetic predisposition to obesity e.g., abdominal obesity-metabolic syndrome 3 (OMIM#615812).

Whole genome sequencing by next-generation DNA sequencing (NGS) technology followed by bioinformatics analysis provides an opportunity to identify genomic variants involved in genetic disorders. In this study, we sequenced the genome of an obese Pakistani citizen with coronary heart disease and hyperlipidemia using the NGS technology. We analyzed the generated data for the identification of deleterious genetic variants and their corresponding mutational load analysis related to hyperlipidemia and coronary heart disease.

MATERIALS AND METHODS

For this study, a Pakistani individual with obesity and comorbidity of hyperlipidemia and type-2 diabetes was selected for whole-genome sequencing (WGS) and its subsequent analysis. The individual was obese with a body mass index (BMI) of 38.3. The study was conducted after the informed consent of the subject. WGS was carried out using Applied Biosystems SOLiD® 5500xl next-generation genome analysis sequencer.

The detailed methodology is given below:

Sample Collection and DNA Isolation

About 2 mL blood of the subject was collected in K2-EDTA blood collection tubes. The genomic DNA was isolated from the whole blood immediately after collection, using the QIAamp DNA Mini Kit (Qiagen Inc. USA). The isolated genomic DNA was assessed by 1% agarose gel electrophoresis and quantified with Qubit® Fluorometer (Thermo Fischer, USA). The A260/A280 ratio was determined using a nanodrop instrument (Jenway Inc.) to assess the purity of genomic DNA.

Library Preparation and DNA Sequencing

A mate-paired library with an insert size of 1,300 bps was prepared for whole genome sequencing as per the SOLiD® Mate-Paired library preparation guide. For this, fragmentation of 5 µg genomic DNA was carried out by the Covaris™ ultrasonication machine followed by gel electrophoresis. The quality of the sequencing library was checked by Bioanalyzer 2100. The size selection of the library was carried out on 2% agarose gel using E-Gel® Electrophoresis System.

The emulsion of the template library was performed in SOLiD® EZ Bead™ Emulsifier (Life Technologies Inc., USA) using the SOLiD® EZ Bead™ Emulsifier E80 reagent kit (Life Technologies Inc., USA). The emulsion was subjected to emulsion PCR in the SOLiD® EZ Bead™ Amplifier (Life Technologies Inc., USA), followed by the enrichment of beads with amplified fragments using the SOLiD® EZ Bead™ Enricher (Life Technologies Inc., USA). The beads were loaded onto the flow chip as per the manufacturer's instructions, and 60x2 bp co-forward sequencing of the mate-paired library was carried out using the SOLiD® 5500xl Genetic Analyzer (Life Technologies Inc., USA).

Analysis of Short Reads and Variant Calling

The raw sequencing reads in 'XSQ' (eXtensible SeQuence) format was converted into 'csfasta' (color-space fasta) format using the XSQ Converter tool (Life Technologies, USA). The filtered csfasta formatted reads were aligned with the human reference genome version hg19 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/) using the LifeScope genomic analysis software.

The SAM files were processed by Samtools program [3]. The removal of duplicates, InDels realignment, and base quality recalibration was done by Picard and GATK [4]. The variants calling was done by HaplotypeCaller command of GATK. The false positive variants were checked GATK VariantEval.

Characterization of Single Nucleotide Variants and InDels Associated with Obesity and Hyperlipidemia

We analyzed the SNVs and InDels associated with obesity, hyperlipidemia, and related cardiac disorders according to the pipeline described earlier [5]. The variants were annotated with ANNOAR program.

Three in-silico tools were brought into consideration for determining the possible damaging variants, i.e., SIFT [6], Polyphen2 [7], and combined annotation-dependent depletion CADD [8]. The variants with SIFT score ≤ 0.05, PolyPhen2 HDIV score ≥ 0.957, and CADD Phred score ≥ 15 were considered deleterious, as described previously [5]. To find the variants causing detrimental effects due to stop-lost, stop-gained, start-lost, and splice-site-variant, the Variants Effect Predictor (VEP) tool was used. The data was also filtered by ClinVar [9], OMIM ([10], and GWAS catalog [11] to find already associated variants with these disorders. The ancestral and derived states of the variants were determined from CADD [12].

The derived allele frequencies (DAF) of prioritized variants were compared among the continental populations to determine variants with significantly varying frequencies among populations. The comprehensive workflow for generating sequencing data and its analysis for potentially detrimental variants have been demonstrated in Figure 1.

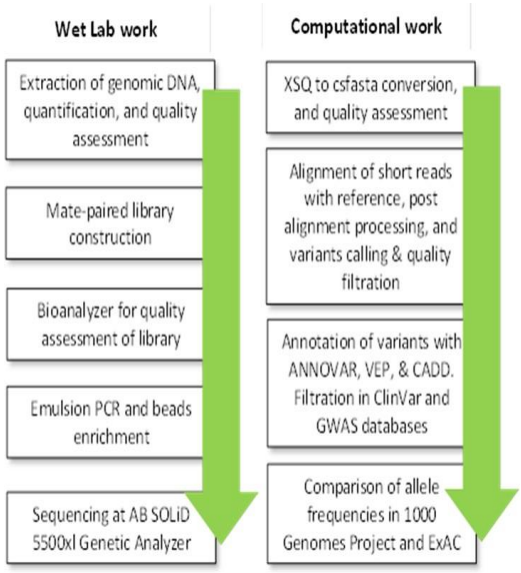

Figure 1: Workflow for next-generation DNA sequencing and variants analysis of the human genome.

RESULTS AND DISCUSSION

After the conversion of XSQ files into csfasta format, a total of 2.065 billion short reads of DNA was obtained. The filtration of low-quality short DNA reads improves the accuracy of alignment, and percentage of coverage thereby facilitating the calling of variants. After removing the reads with low-quality scores and missing calls, about 1.340 billion (51.43%) short DNA reads with their matching mate pairs were retained. There were 312,849,478 short reads which aligned uniquely (after removing duplicates) with the human reference genome along with their mates.

Applying the best practices of PICARD and GATK tools, 2,568,249 variants were called. Filtration of the low-quality variants (variants with quality score (QV) < 20), yielded a total of 2,167,161 variants. This included 2,055,524 single nucleotide variants (SNVs) and 111,664 InDels. The median depth of variants (DP) from the filtered VCF file was found to be 6. The transitions/transversions ratio (Ti/Tv) of the variants for the whole genome was 2.14, an acceptable ratio for the whole genome analysis [13]. There were 41,088 (1.90%) novel variants. The annotation of variants Carried out by the ANNOVAR program represents the number of variants at different genomic regions (Table 1).

Table 1: Number of SNVs in gene and intergenic regions in obese Pakistani individual analyzed during this study.

Gene region	No. of variants
Exonic	14,213
Intergenic	1,190,116
Intronic	771,143
5'UTR	2,794
3'UTR	17,434
Upstream	12,997
Downstream	15,046
Synonymous	7,143
Nonsynonymous	6,434

The nonsynonymous to synonymous ratio was found to be 0.90, which is consistence with the ratio reported for South Asian populations (1000 Genomes Project Consortium, 2015). The annotations with SIFT, PolyPhen2, and CADD prioritized 1,148, 730, and 29,960 variants as deleterious, respectively. Overall, there were 425 variants predicted as deleterious by all three tools (Figure 2) in 385 genes.

This included a novel homozygous SNV in the gene MTRR (chr5:7897285A>T; p.Q626L). MTRR encodes 5-methyltetrahydrofolate-homocysteine methyltransferase reductase is involved in the synthesis of methionine. A missense mutation in this gene has been linked to overweight and obesity in the Han Chinese population.

We found several deleterious variants in three genes i.e., KCNJ12, CDC27, and HYDIN (6 deleterious nonsynonymous variants in each gene).

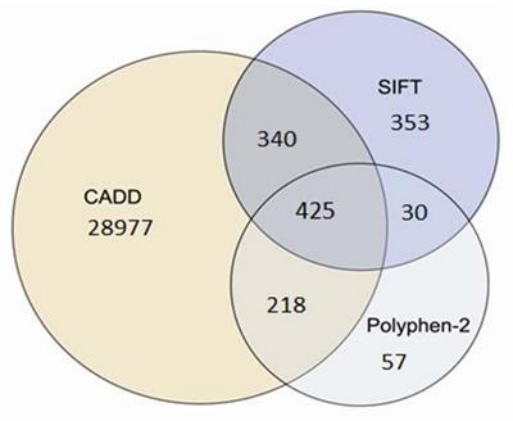

Figure 2: Venn diagram of deleterious variants showing deleterious variants prioritize by Polyphen2, SIFT, and CADD tools.

==========

The gene KCNJ12 encodes a potassium voltage-gated channel subfamily J member 12 which plays an important role in controlling rectifying current in cardiac cells whereby taking part in cardiac conduction. In a Chinese family, the gene was found to be associated with dilated cardiomyopathy [14]. The CDC27 encodes cell division cycle 27 protein which is a component of the anaphase-promoting complex (APC), and HYDIN encodes axonemal central pair apparatus protein which functions in cilia motility.

Both these genes have not been previously reported in cardiac disorders. For comparison, the deleterious missense variations in these three genes were determined using 5 randomly selected male individuals from the PJL dataset of the 1000 Genomes Project. The KCNJ12 and CDC27 genes contained, on average, 2 deleterious missense SNVs, while HYDIN contains 8 deleterious missense SNVs in the selected male individuals from the PJL dataset. This gave a clue that HYDIN normally contains a high number of deleterious missense SNVs, while KCNJ12 and CDC27 contained more deleterious missense SNVs than in normal individuals of the same population.

These two genes can further be assessed in a large cohort of hyperlipidemic, obese individuals with cardiac disorders.

Bioinformatics analysis revealed a homozygous stop-gained SNV rs885985 (G>A, p.Q37X; ENST00000403084.1, ENST00000406028.1, ENST00000413119.2) in the CLDN5 gene.

This gene encodes Claudin-5, a membrane protein that forms strands of tight junctions. The CLDN5 is expressed in fat tissues [15] and the decreased level of claudin-5 has been found to be associated with heart failure [16]. A heterozygous stop-gained SNV rs328 (C>G, p.S474X; ENST00000311322.1, ENST00000650287.1) was observed in LPL (lipoprotein lipase) gene which is implicated in hyperlipidemia ([17] and risk of CVD [18]. In addition, a single nucleotide variant rs2516839 (C>T) that is homozygous was identified in the 5' untranslated region of USF1. This variant has been linked to a doubled risk of sudden cardiac death. USF1 is responsible for coding the upstream transcription factor 1 from the leucine zipper family and has been linked to the development of atherosclerosis and hyperlipidemia.

The frequency of this variant is 0.53 in Pakistani dataset (PJL) of 1000 Genomes database, 0.44 and 0.45 in two neighboring populations Indian Telugu (ITU) and Gujrati Indian (GIH) respectively, 0.15 in African, 0.46 in American, and 0.63 in European populations. A homozygous single nucleotide variant rs71457130 (C>T) was found in the 3´-UTR of LRP6

gene which encodes LDL receptor-related protein 6.

This protein functions in receptor-mediated endocytosis of lipoproteins. Though several SNPs in LRP6 have been found to be associated with CVDs, this SNV has not been illustrated previously with the consequence of cardiac disorder.

A comparison was made between South Asian (SAS) and global populations of the 1000 Genomes Project to determine the derived allele frequencies (DAF) of predicted deleterious variants. The results revealed that 11 single nucleotide variants (SNVs) had higher DAF in south Asians than in global populations (Figure 3).

CONCLUSION

This study suggests that single nucleotide variants (SNVs) with derived allele frequencies (DAF) in South Asian (SAS) population compared to global populations could potentially be a risk factor for obesity and comorbidities specific to the SAS population.

However, the presence of these variants alone does not necessarily mean that an individual with these variants will develop obesity or any comorbidities associated with it. However, the allele frequency in the SAS population suggests that this population may have a greater predisposition to these conditions.

Further research is necessary to understand the functional implications of these variants and their role in the development of obesity and related comorbidities in the SAS population. This information could be useful in developing targeted interventions to prevent or manage these conditions in this population.

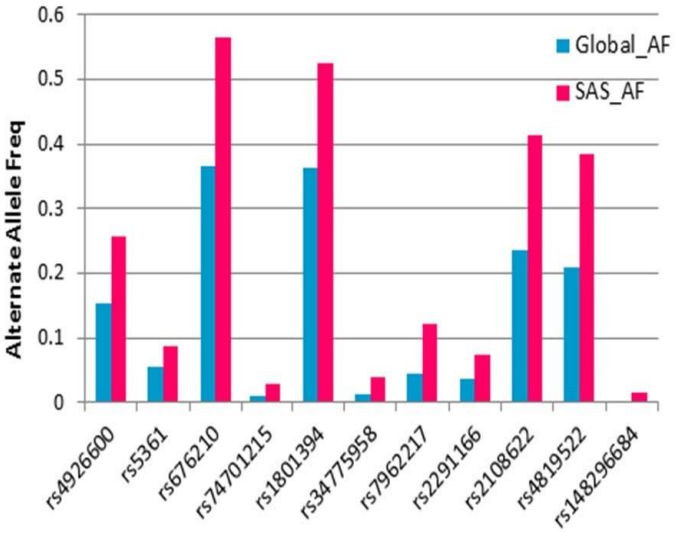

Figure 3: Deleterious SNVs with comparably higher allele frequency in South Asian (SAS) population than in global populations.

==========

REFERENCES

1.Reilly, J. J., El-Hamdouchi, A., Diouf, A., Monyeki, A., & Somda, S. A. (2018). Determining the worldwide prevalence of obesity. The Lancet, 391(10132), 1773-1774.

2. Sasson, A., & Michael, T. P. (2010). Filtering error from SOLiD output. Bioinformatics, 26(6), 303 849-850.

3. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

4. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Daly, M.(2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297-1303.

5. Shakeel, M., Irfan, M., & Khan, I. A. (2018). Estimating the mutational load for cardiovascular diseases in Pakistani population. PLOS ONE, 13(2), e0192446.

6. Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous
variants on protein function using the SIFT algorithm. Nature Protocols, 4(7), 1073-285 1081.

7. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . .Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature methods, 7(4), 248.

8. Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. (2014). Ageneral framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics, 46(3), 310-315.

9. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott,D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, 42(D1), D980-D985.

10. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research, 33(suppl 1), D514-D517.

11. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H.,Hindorff, L. (2013).The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research, 42(D1), D1001-D1006. World Health Organization. (2017b). Obesity and Overweight Fact Sheet October 2017. Retrieved from World Health Organization website: http://www.who.int/mediacentre/factsheets /fs311/en/.

12. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M.,Brent, S. (2016).Ensembl comparative genomics resources. Database, 2016, bav096.

13. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C.,Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491-498.

14. Yuan, H. X., Yan, K., Hou, D. Y., Zhang, Z. Y., Wang, H., Wang, X.,Zhang, L. (2017). Whole exome sequencing identifies a KCNJ12 mutation as a cause of familial dilated cardiomyopathy. Medicine (Baltimore), 96(33), e7727. 337

15. Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., .Edlund, K. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Molecular and Cellular Proteomics, 13(2), 397-406.

16. Swager, S. A., Delfín, D. A., Rastogi, N., Wang, H., Canan, B. D., Fedorov, V. V.,Ziolo, M. T. (2015). Claudin-5 levels are reduced from multiple cell types in human failing hearts and are associated with mislocalization of ephrin-B1. Cardiovascular Pathology, 24(3), 160-167.

17. Shatwan, I. M., Minihane, A.-M., Williams, C. M., Lovegrove, J. A., Jackson, K. G., & Vimaleswaran, K. S. (2016). Impact of lipoprotein lipase gene polymorphism, S447X, on postprandial triacylglycerol and glucose response to sequential meal ingestion. International journal of molecular sciences, 17(397), 1-9.

18. Xie, L., & Li, Y.-M. (2017). Lipoprotein Lipase (LPL) Polymorphism and the Risk of Coronary Artery Disease: A Meta-Analysis. International journal of environmental research and public health, 14(84), 1-7.