Whole-Exome Analysis Reveals X-linked Recessive Mutations underlying Premature Ovarian Failure using Galaxy Platform

Rashid Saif[1*], Tania Mahmood[1], Aniqa Ejaz[1], Saeeda Zia[2], Saqer Sultan Alotaibi[3]

[1]Decode Genomics, 323-D, Punjab University Employees Housing Scheme (II), Lahore, Pakistan
[2]Department of Sciences and Humanities, National University of Computer and Emerging Sciences, Lahore, Pakistan
[3]Department of Biotechnology, College of Science, Taif University, Taif 21944, Saudi Arabia

[*]Corresponding Author: rashid.saif37@gmail.com

Abstract

An in-silico WES approach using the Galaxy platform was adopted in the current study to predict the genetic basis of Premature Ovarian Failure (POF), where three affected patients in a Saudi Arabian family of seven, found associated with X-linked recessive mutations. The current analysis discovered 518,054 variants using FreeBayes variant caller that had 1,461,864 effects on variable sites in the genome revealed by SnpEff software. The causal genetic mutations were filtered and annotated with the ClinVar database using the GEMINI tool. This tool retained 369 pathogenic mutations harboring 130 genes. Among the total, 268 variants positioned on 69 genes are shared with three affected individuals, 61 variants on 23 genes are shared by any two of the affected individuals, and 40 of the variants on 38 genes are present in any one of the affected sample. Two mutations in one of the already POF-associated, *POF1B* gene were also observed e.g. (i) g.84563135T>A; p.M349L and (ii) g.84563194C>T; p.R329Q in the two affected individuals *i.e.* IV-I-C & IV-6 in the current data. This gene consists of 17 exons that span the region of >100 kb. The putative function of this gene in regulating the actin cytoskeleton due to homology with myosin tail and maintains a number of oocytes during fetal ovary development. In a nutshell, this Galaxy pipeline facilitates all-in-one to pinpoint not only the known pathogenic gene mutations for this disorder but few other novel genetic variants as well, whose gene-disease association may be validated by further experimental studies.

Key words: WES analysis, POF, Galaxy platform, GEMINI tool, X-linked recessive, ClinVar

Introduction

An in-silico Next-Generation Sequencing (NGS) approach, the computational whole-exome sequencing (WES) analyses has expedited the identification of genetic etiologies of various complex diseases. The skilled personnel required to utilize the available computational resources for this purpose is a real bottleneck (1). However, the development and easy availability of Galaxy platform, which is an umbrella for plethora of tools, has met the scale-out analysis needs and has also helped in making this task unchallenging. The relevance of in-silico WES technology in analyzing complex genetic traits has led to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 May 2021 doi:10.20944/preprints202105.0519.v1

comprehensive genetic studies on patients with POF disorder. Early cessation of ovarian function prior to age 40 in women, refer to as POF, leads to successive concomitant disorders such as, rise in serum follicle-stimulating hormone (FSH) levels to greater than 40 IU/L and amenorrhea of 4-6 months (2). The occurrence of POF in women above 40 years of age is 1/100 while, this is less frequent 1/1000 in women below age, 30 (3). In many idiopathic POF cases, there is greater 50 to 90% likelihood of genetic involvement. Among them, it is observed that 10-30% are the cases where first degree relative is affected also with 6 times chances of having POF in woman with affected mother (4). The genetic architecture of the POF still remain to be elucidated, however, few clinical exome based diagnostic analyses provide key mutations in *POF1B* gene and their association with POF disorder. A critical region that spans 100kb of genomic DNA and consist of 17 exons on X chromosome, which function for early normal ovarian growth, harbors this gene (5). Expression analysis of this gene in mouse proposed its role in early ovary growth as, this gene also escapes the X chromosome inactivation which suggest its contribution in ovary development (6).

Here, we intend to implement the WES pipeline on Galaxy platform to unveil the causal variants responsible for the POF disorder. For this purpose, we uploaded seven paired-end exome sequencing reads in galaxy history that were retrieved from ENA database. The data was of a Saudi Arabian consanguineous family where the clinical investigation revealed three sisters, IV-1-C, IV-6, and SAPOF with POF disease. Candidate gene mutations were then annotated with ClinVar entries, disease names and its phenotypes. The in-silico analysis on Galaxy software holds valuable place in clinical studies for diagnosing the genetic causes of heterogeneous diseases.

Material and Methods

WES family data retrieval and quality assessment

Whole-exome paired-end fastq datasets of a Saudi Arabia family with POF phenotype were retrieved from European Nucleotide Archive (ENA) under projectID = PRJNA260607 (7). These WES datasets were uploaded in Galaxy history that comprises seven members of a family where parents are immediate cousins. The three patients age 19, 24 and 35 years suffer from primary amenorrhea, hypothyroidism and hypergonadotropic hypogonadism while two daughters remain unaffected from the disorder (4). The sequencing reads were assessed for validating the quality of data using FastQC software (Galaxy v. 0.72+galaxy1) (8). A customized report of all the FastQC results was generated using MultiQC (Galaxy v. 1.7) (9).

Aligning WES datasets and annotating variants

Mapping of sample reads to hg19 reference genome was carried out by specifying a unique read group identifier and sample name to each dataset using BWA-MEM (Galaxy v. 0.7.17.1) (10). We skipped the alignments for which both reads and mates are unmapped by Filter SAM or BAM tool (Galaxy v.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 May 2021 doi:10.20944/preprints202105.0519.v1

1.8+galaxy1). Sorting on coordinate basis was conducted using SortSam (Galaxy v. 2.18.2.1) and MarkDuplicates tool (Galaxy v. 2.18.2.2) (11) was utilized to mark the duplicate reads. Sequence variations were found, normalized by left-aligning, splitting the multiallelic records into biallelic ones using FreeBayes (Galaxy v. 1.1.0.46-0) and bcftools norm (Galaxy v. 1.10) (12) respectively. The filtered polymorphisms were functionally annotated using SnpEff eff (Galaxy v. 4.3+T.galaxy1) (13).

Predicting the putative pathogenic gene variants

Narrowing down our search for deleterious variants that define the patient's POF phenotype, we first loaded the annotated variant information into the GEMINI database framework using GEMINI load tool (Galaxy v. 0.20.1) (14). This GEMINI-specific database dataset along with pedigree file was fed to GEMINI inheritance pattern (Galaxy v. 0.20.1) tool adding the x-linked recessive mode of inheritance constraint to the identified variants. This tool integrated additional annotations from ClinVar database and returned a handful of possible pathogenic mutations.

Results

Alignment and variant annotation of WES POF datasets

The quality statistics of paired-end POF family WES datasets was figured out to mitigate any sequencing errors prior to mapping it. The computed statistics clearly indicated a good overall quality of datasets with average 50% GC content and 100bp average sequence length (Figure 1).

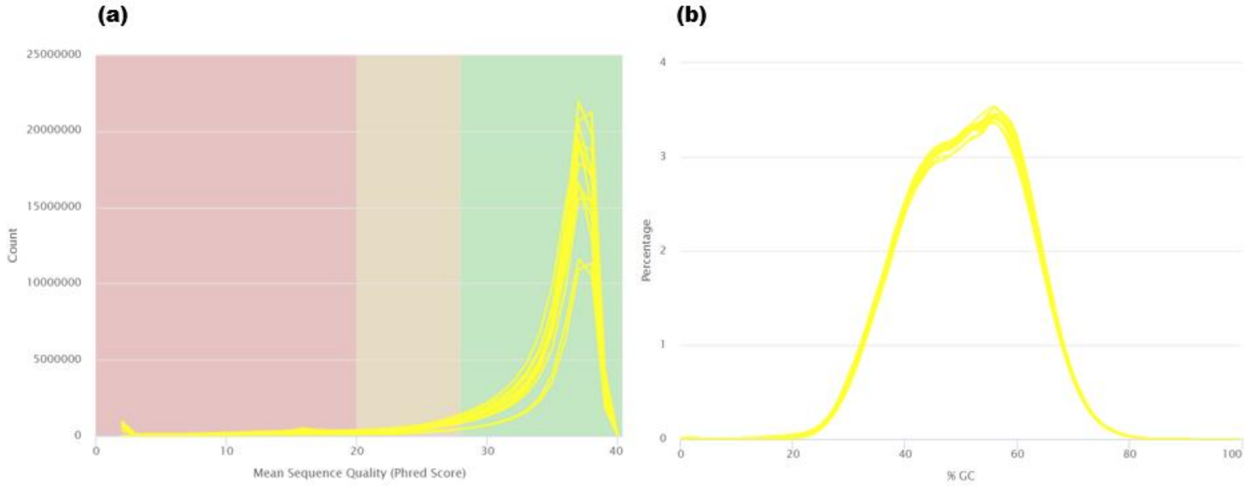

Fig. 1 (a) The number of reads with average quality scores. **(b)** The average GC content of reads.

Paired-end mapping was applied to map the variable sites in these datasets against hg19 reference genome. Overall 99.87% of reads paired properly with the reference genome, the rest were discarded and duplicate sequences were marked. By calling variants across all samples simultaneously, we obtained 518,054 variable sites, each appear after nearly 5,966 bases. The obtained extensively variable sites were

doi:10.20944/preprints202105.0519.v1

categorized which include 469,194 SNPs with 1.5949 transition/transversion ratio of observed events, 8,608 multiple nucleotide polymorphisms (MNPs), 14,540 insertions, 31,126 deletions and 2,238 mixed variants. These polymorphisms produced 1,461,864 effects at different sites in the genome which are outlined in Table 1.

Table 1 Characterization of variant types based on their type and region

TYPE			REGION		
Variant Type	**Count**	**Percent**	**Type**	**Count**	**Percent**
3' UTR	25,273	16790	Downstream	127,363	8.71
5' UTR premature start codon gain	3,022	2010	Exon	432,664	29.60
5' UTR	18,503	12300	Gene	1	0
Conservative inframe deletion	116	80	Intergenic	112,227	7.68
Conservative inframe insertion	134	90	Intron	591,797	40.48
Disruptive inframe deletion	265	180	Splice site acceptor	2,149	0.15
Disruptive inframe insertion	92	60	Splice site donor	2,947	0.20
Downstream gene	127,363	84630	Splice site region	29,513	2.02
Frameshift	1,162	770	Transcript	14,216	0.97
Gene fusion	1	0	Upstream	102,191	6.99
Initiator codon	14	10	UTR 3`	25,273	1.73
Intergenic region	112,227	74570	UTR 5`	21,523	1.47
Intragenic	4	0			
Intron	622,123	413390			
missense	227,634	151260			
Non coding transcript exon	34,194	22720			
Non coding transcript	7	0			
Protein-protein contact	1,729	1150			
Sequence feature	14,205	9440			
Splice acceptor	2,152	1430			
Splice donor	2,984	1980			
Splice region	38,242	25410			
Start lost	323	210			
Stop gained	18,643	12390			
Stop lost	193	130			
Stop retained	124	80			
Structural interaction	41,903	27840			
Synonymous	110,095	73160			
Upstream gene	102,191	67900			

The significant fraction of variations appeared on intronic region and on exons as pictured in Figure 2.

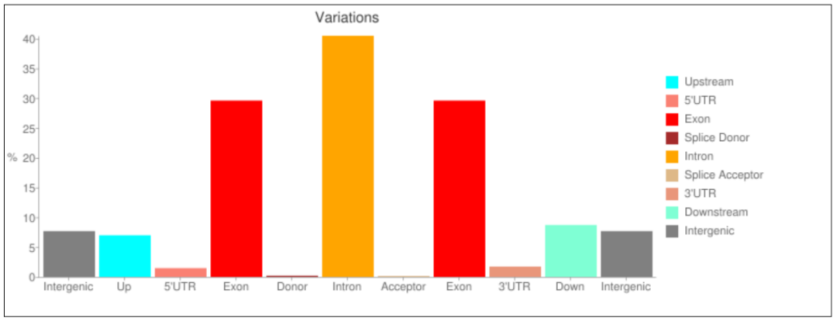

Fig. 2 Graphical illustration of the frequency of different types of variations

Detection of pathogenic and novel mutations associated with POF

To determine the mutational effect of the variants on patients, we screened the genes harboring variants that holds for x-linked recessive inheritance pattern type rather than autosomal recessive. This candidate gene mutation detection pipeline retained 369 putative mutations harboring 130 genes. This tool retained 369 pathogenic mutations harboring 130 genes. Of these, 268 variants positioned on 69 genes are spotted in all three affected individuals, other 61 variants on 23 genes are present in any two affecteds, while 40 of the variants on 38 genes are observed in any one of the sample. Only 1 gene harboring two mutations in IV-I-C and IV-6 affecteds was found previously associated with POF disease Table 2. While 247 of 369 candidate mutations positioned on 87 genes are not annotated before in ClinVar database. Remaining genes annotation shows the variants association with cardiac, neurological, muscular, ocular disorders etc. Table S1.

Table 2 Details of putative gene found on Chr.X associated with POF in affected daughters (IV-I-C and IV-5-D).

Genes	Start	Ref	Alt	Impact	ClinVar disease name / phenotype	rs_id	Genotypes
POF1B	84563134	T	A	Missense	non-syndromic X-linked intellectual disability, premature ovarian failure 2b	rs363774	A/A
POF1B	84563193	C	T	Missense	non-syndromic X-linked intellectual disability, premature ovarian failure 2b	rs75398746	T/T

Discussion

In current computational investigation, we recruited WES datasets on Galaxy software of a Saudi Arabian family where three sisters IV-1-C, IV-6, and SAPOF were clinically diagnosed with POF disorder. They were found suffering from idiopathic hypergonadotropic primary amenorrhea with hypothyroidism, atrophic ovaries having normal 46, XX karyotype (4). Prior SNP analysis and functional study on this family data by (4), identified an autosomal recessive mutation on *MCM8* gene c.446C>G; p.P149R that manifests POF, endocrine dysfunction and chromosomal instability. Consequently, we searched for the causative pathogenic mutation that met the X-linked recessive inheritance filter criteria by executing Galaxy software WES pipeline. Genetic predispositions for POF often comply with an X-chromosomal inheritance pattern and these families usually have an early onset of this disorder before age 31 (15). Our WES framework divulged two known homozygous mutations on *POF1B* gene, (i) g.84563135T>A; p.M349L and, (ii) g.84563194C>T; p.R329Q. There are no evident studies on former mutation that can best describe its pathogenicity, however, analyses on later mutation describing the role, functioning and alteration are present. Moreover, we discovered some novel candidate mutations from 268 that are not reported before in ClinVar, but might have role in POF due to its incidence in all patients.

A mutational study of *POF1B* gene performed on a Lebanese family WES data of 5 affected sisters identified the homozygous mutation R329Q by whole-genome SNP typing and homozygosity-by-descent mapping (16). They hypothesized that POF1B shares homology with myosin tail and thus it plays a function in actin-filament interaction. In vitro examination conducted on mutant and wild-type proteins showed hindrance in the interaction of mutant with actin four times than the wild-type POFIB. They speculated that the loss of function of mutant type is probably due to lack of phosphorylation at serine-leucine-arginine site (17). The expression investigation in mouse suggested POFIB functioning in meiotic chromosomal pairing which gets altered after R329Q mutation, leading to cell death and extreme decrease of the number of oocytes during ovary development. This pathogenic variant is responsible for exaggerated germ cell apoptosis and POF. Another finding on similar mutation in Italian population suggested no relation between R329Q mutation and POF due to heterozygous case both in patients and healthy subjects (6). This situation is probable because most of the Italian patients had secondary amenorrhea which contradicts with our patient's condition who have primary amenorrhea. Thus far, the future inclusive genetic studies can be vital to outline the spectrum of human phenotypes linked with *POF1B* gene variants.

Conclusion

Detecting and predicting the pathogenicity of genetic variants using omics data and associating its causality with any disorder is an expedient in-silico approach. We retrieved a WES data of one Saudi Arabian family suffering in POF and firstly, developed a Galaxy pipeline using freely available tools and software to scan candidate gene variants in this disorder. Secondly, two known candidate mutations were found in coding region of *POF1B* gene that are already reported with this disorder along with some other novel variants as well. We found this bioinformatics pipeline practically robust that offers the theoretical basis to find the genetic variants using WES data and its initial detection. Further, these detected variant's association can be validated by wet-lab experimentation for better understandings and more confidence.

Acknowledgement

Authors are thankful to the Magee-Womens Research Institute & Foundation, Pittsburgh, PA, USA for providing this data publically available for the researchers to explore and understand the genetic basis of this disorder for the awareness of affected community and overall humanity in general.

References

1. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology. 2010;11(8):1-13.
2. Chapman C, Cree L, Shelling AN. The genetics of premature ovarian failure: current perspectives. International journal of women's health. 2015;7:799.
3. Jankowska K. Premature ovarian failure. Przeglad menopauzalny= Menopause review. 2017;16(2):51.
4. AlAsiri S, Basit S, Wood-Trageser MA, Yatsenko SA, Jeffries EP, Surti U, et al. Exome sequencing reveals MCM8 mutation underlies ovarian failure and chromosomal instability. The Journal of clinical investigation. 2015;125(1):258-62.
5. GeneCards: The Human Gene Database [Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=POF1B#:~:text=UniProtKB%2FSwiss%2DProt%20Summary%20for,be%20involved%20in%20ovary%20development.
6. Bione S, Rizzolio F, Sala C, Ricotti R, Goegan M, Manzini M, et al. Mutation analysis of two candidate genes for premature ovarian failure, DACH2 and POF1B. Human reproduction. 2004;19(12):2759-66.
7. ENA: European Nucleotide Archive [Available from: https://www.ebi.ac.uk/ena/browser/view/PRJNA260607.
8. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
9. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8.
10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
11. Picard [Available from: http://broadinstitute.github.io/picard/.
12. Porter J, Berkhahn J, Zhang L, editors. A comparative analysis of computational indel calling pipelines for next generation sequencing data. Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP); 2014: The Steering Committee of The World Congress in Computer Science
13. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80-92.

14. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol. 2013;9(7):e1003153.
15. Fassnacht W, Mempel A, Strowitzki T, Vogt P. Premature ovarian failure (POF) syndrome: towards the molecular clinical analysis of its genetic complexity. Current medicinal chemistry. 2006;13(12):1397-410.
16. Lacombe A, Lee H, Zahed L, Choucair M, Muller J-M, Nelson SF, et al. Disruption of POF1B binding to nonmuscle actin filaments is associated with premature ovarian failure. The American Journal of Human Genetics. 2006;79(1):113-9.
17. POF1B ACTIN-BINDING PROTEIN; POF1B https://omim.org/entry/300603?search=prefix%3A%2A%20AND%20chromosome%3AX&highlight=%2A%20%2BX.