Whole-Exome Data Analysis: Detection of Candidate Gene Mutations for Mitochondrial Encephalohepatopathy

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

Mitochondrial Encephalohepatopathy (MEH) is an autosomal recessive neurodevelopmental disorder usually accompanied by microcephaly, white matter changes, cardiac and hepatic failure. Here, we applied the whole-exome sequencing (WES) framework on a trio family data with unaffected non-consanguineous parents and proband (neonate girl) with this inherited disorder. A total of 2,928,402 variants were observed with 2,613,746 SNPs, 112,336 multiple nucleotide polymorphisms (MNPs), 72,610 insertions, 113,207 deletions and 16,503 mixed variants. These variations are responsible for 82,813,631 effects on various genomic regions. Our pipeline uncovered candidate gene mutations from these variants and retained a handful of 5,277 variants harboring 3,598 genes, out of which, 8 genes codes for non-coding RNA while 178 genes are those with high impact severity. Among these 178 variants, 125 are de-novo variants that are not previously reported in the ClinVar database. Consistent to previous studies, the leftover high impact severity genes are involved in encephalopathy, Leigh syndrome, Charcot–Marie–Tooth disease, global developmental disorder, seizures, spastic paraplegia, premature ovarian failure, mitochondrial myopathy-cerebellar, ataxia-pigmentary, retinopathy syndrome, ocular and retinal degeneration, deafness, intellectual disability, cardiofacioneurodevelopmental syndrome etc. All these clinical features were also observed in the patient studied. The current analysis highlights and expands the genetic architecture of the MEH phenotype. Furthermore, this pipeline on trio family data significantly broadens the concept of its usefulness as a first-tier diagnostic method in the detection of complex multisystem phenotypic disorders.

Keywords: Mitochondrial Encephalohepatopathy, Trio-family, autosomal recessive, GEMINI tool, ClinVar database.

Introduction

Hepatic encephalopathies (HE) resulting from liver complications contribute to neuropsychiatric disorders that are often followed by changes in mitochondrial membrane potential (MMP) (1). Researches describing the linkage between mitochondria with HE suggested its effect on cerebral energy metabolism and levels of ammonia (hyperammonemia) which affects the TCA cycle, electron transport chain, etc. Due to disturbances in the function of mitochondria, reactive oxygen species elevates which interferes with mitochondrial regeneration in the brain (2). Usually, the inherited mitochondrial encephalohepatopathy (MEH) in neonates affects the central nervous system, the liver, and the heart (3). It is inherited in an autosomal recessive pattern. The genetic insult to mitochondria cause mutations in mtDNA or the MEH may be caused by the mutations in those nuclear genes encoding mitochondrial proteins or cofactors (4). Delineating the genetic architecture of MEH is now made possible with the advancements in next-generation sequencing (NGS) technology which is considered a successful diagnostic tool especially the whole-exome sequencing (WES) approach (5). The focus has now shifted towards Trio-based WES methodologies which have higher diagnostic yield. This holds best in the case of rare genetic disorders where the potential candidate mutations have too low a frequency in the human population. It not only plays a role in diagnosing the inheritance pattern and potential de novo candidate mutations but also compares variants between patients and their selected relatives. Despite the advent of NGS-based standard methods and a great deal of research, there is scanty data available on the molecular basis of this disease.

This study is centered to detect candidate gene variants causing Mitochondrial Encephalohepatopathy (MEH) from whole-exome sequence of Trio family data comprising of unaffected non-consanguineous parents and a proband who inherited this disorder. The proband was a baby girl born at 38 weeks of gestational age after an intricate pregnancy due to intrauterine growth restriction. This Trio family WES data is publicly available on ENA.

Materials and methods

Trio data quality control and mapping

Trio paired-end fastq files of father, mother, and proband suffering from MEH were retrieved from ENA source under project ID: PRJNA673368 (6). Before embarking on further steps, we applied FastQC software (7) to check the quality of NGS reads. The quality checks output was then aggregated using MultiQC software (8). Alignment of the Trio fastq files with GRCh37 human genome assembly (9) was carried out using bwa mem (10) with parameter -R “@RG\tID:SampleID\tSM:Samplename”. These mapped reads were then post processed to know the accurate variant spectrum of the sample. We retained only those reads for which both the forward and the reverse reads have been mapped and then deduplicated the mapped reads using samtools view and picard MarkDuplicates feature respectively (11).

Candidate gene variant detection and annotation

With the samples mapped and postprocessed, we created a multisample VCF file using FreeBayes software (12) to find SNPs, MNPs, indels, etc.VCF normalization was done using bcftools norm (13) with settings -c W -m- both v to split multiallelic variant records and to left-align and normalize indels. Functional genomic effects were added to the normalized VCF using SnpEff eff tool (14). Variants in the VCF file were then prioritized based on the relationship between samples and their biological phenotype by generating GEMINI-specific database dataset using GEMINI load --skip-pls --save-info-string -p PED file (15). The candidate gene variants that have the potential to explain the girl’s MEH phenotype were reported using GEMINI autosomal_recessive inheritance pattern (15) adding --filter “impact_severity != ’LOW’”.

Results

Trio data sequencing and quality checks

The quality checks applied on Trio fastq files returned the quality graphs which are shown in Figure 1.

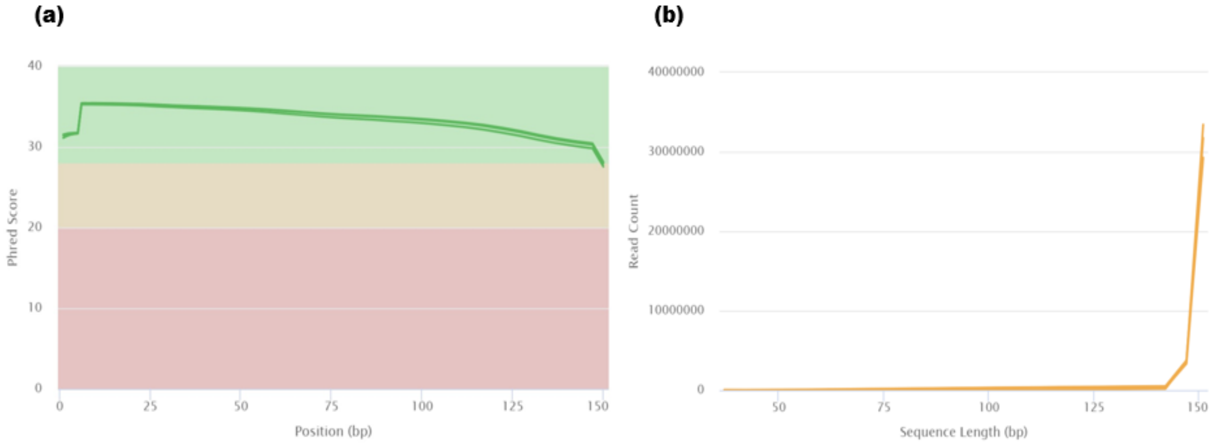

Fig. 1. Graphical illustration of aggregated results of FastQC output. (a) The mean quality value across each base position in the read is shown. The green region highlights the good quality of bases while the red region displays bad quality bases (b) The distribution of fragment sizes (read lengths).

The aggregated statistics of FastQC report generated are summarized in Table 1. Mapping of Trio datasets with reference hg19 genome revealed the average percentage 94.65 of total reads that properly paired which include father = 77623548 reads (94.32), mother = 71653302 reads (94.10), and proband = 65906818 reads (95.53).

Table 1 General statistics of Trio family datasets.

Sample Name	Duplicates	GC%	Length	Failed	M Seqs
Father	28.2	47	145 bp	0	41.3
Mother	26.3	47	150 bp	0	38.2
Proband	28.0	47	149 bp	20	34.6

***M Seqs =** Total Sequences (millions),
Failed = Percentage of modules failed in FastQC report

Mutation in Trio samples and their functional effects

In search for the evidence of sequence deviation from the reference genome, we identified 2,928,402 variants in total Table S1 with 1 variant occurring after every 1,101 bases. From the identified variants, we obtained 2,613,746 SNPs, 112,336 multiple nucleotide polymorphisms (MNPs), 72,610 insertions, 113,207 deletions and 16,503 mixed variants. Out of all these variants in father, mother, and proband, 339618, 291084, and 281312 variants are heterozygous while 576851, 430298 and 482985 variants are homozygous respectively. The overall transition (Ts) rate of SNPs is 2,031,179 (781007 Ts in father, 598680 Ts in mother, and 651492 Ts in proband) and SNPs transversion (Tv) rate is 1,301,190 (496180 Tv in father, 391705 Tv in mother and 413305 Tv in the proband. While prioritizing only the fraction of detected variants that have very clear biological relevance the functional effects of these variants on genomic regions were considered by their type, and region which accounts for a total of 82,813,631 number of effects as detailed in Table 2.

Table 2 List of variant's effect by type and region

TYPE			REGION		
Variant Type	**Count**	**Percent**	**Type**	**Count**	**Percent**
3` UTR	221741	0.267	Downstream	1659379	2.004
5` UTR premature start codon gain	8612	0.01	Exon	1056083	1.275
5` UTR	148763	0.179	Gene	1155	0.001
Bidirectional gene fusion	269	0	Intergenic	219141	0.265
Conservative inframe deletion	567	0.001	Intron	5814504	7.021
Conservative inframe insertion	231	0	Splice site acceptor	14270	0.017
Disruptive inframe deletion	1553	0.002	Splice site donor	24902	0.03
Disruptive inframe insertion	238	0	Splice site region	108038	0.13
Downstream gene	1659379	1.999	Transcript	71941814	86.872
Frame shift	4740	0.006	Upstream	1595360	1.926
Gene fusion	886	0.001	UTR 3`	221741	0.268
Initiator codon	943	0.001	UTR 5`	157244	0.19
Intergenic region	219141	0.264			
Intragenic	64751307	78.02			
Intron	5947187	7.166			
Missense	626578	0.755			
Noncoding transcript exon	202746	0.244			
Noncoding transcript	7190507	8.664			
Splice acceptor	14358	0.017			
Splice donor	25107	0.03			
Splice region	140492	0.169			
Start lost	1507	0.002			
Stop gained	24785	0.03			
Stop lost	2290	0.003			
Stop retained	68	0			
Synonymous	203397	0.245			
Upstream gene	1595360	1.922			

The distribution of different types of variants are graphically displayed in Figure 2 which represents the relevant fraction of intronic variants of all the detected ones.

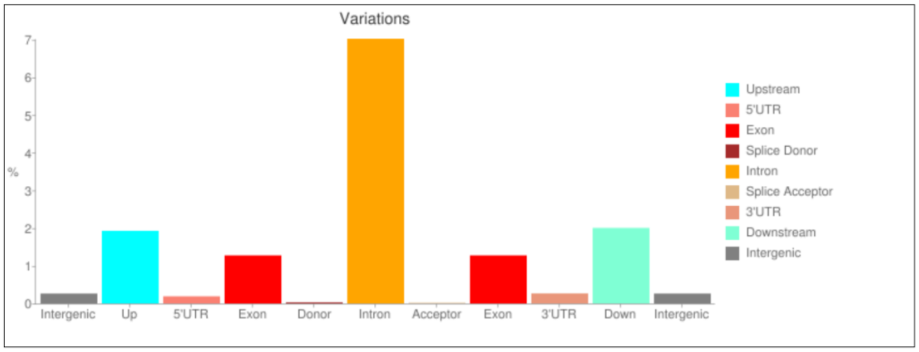

Fig. 2. Bar chart demonstrating the distribution of variants across gene features.

Detection of potential variants responsible for MEH

By tailoring our WES framework to capture autosomal recessive candidate mutations responsible for the girl's MEH phenotype, the analysis retained 5,277 variants residing on 3,598 genes out of which 8 genes codes for non-coding RNA while 178 genes are those with high impact severity. The candidate genes and variants, annotated with ClinVar database are very precisely shown in Table 3, however, the complete list is given in Table S2. None of the variants observed have been clinically reported before in ClinVar database.

Table 3 Overview of candidate gene mutations annotated with ClinVar database

Chr.	Gene	Start	Ref	Alt	Impact	Impact severity	ClinVar			rs_id	Varia nt ids	Geno.
							Sig	Disease	Pheno.			
NC_000001	NM_001385 640	877830	T	C	missense	MED	None	None	None	None	864	C/C
NC_000001	NM_015658	888658	T	C	missense	MED	None	None	None	None	1110	C/C
NC_000001	NM_015658	889157	GA	CC	Splice region	MED	None	None	None	None	1125	CC/CC
NC_000001	NM_001160 184	909237	G	C	missense	MED	None	None	None	None	1481	C/C
NC_000001	NM_001160 184	909308	T	C	missense	MED	None	None	None	None	1482	C/C
NC_000001	NM_001369 898	914875	T	C	missense	MED	None	None	None	None	1575	C/C
-	-	-	-	-	-	-	-	-	-	-	-	-

Discussion

Herein, Trio pipeline was implemented on a family WES data where the proband inherited the MEH disease who died of it at 30 months of life. Mother had seriously complicated pregnancy due to restricted intrauterine growth. The neurological investigation found that patient had serious global hypotonia just after the birth, white matter lesions and progression of cortical and subcortical atrophy, delayed development and growth, epileptic encephalopathy with spasms, myoclonic and focal seizures at 21 months of age followed by microcephaly, hearing loss, retinal degeneration, portal hypertension and pancytopenia (16). The current findings revealed some key variants of high impact severity in the proband, the genes of which had previously been described in ClinVar database and are associated with the aforementioned clinical features.

For instance, we found frameshift mutations in NM_001037333 (g.156721862T>TC), NM_001291412 (g. 34948683G>GA), NM_001256743 (g. 122336599T/TG) that are reported to be mutated in patients with epileptic encephalopathy, global developmental delay, and mental retardation (17). Splice donor variants were detected in NM_001300908 (g. 67980942A>C), NM_001012759 (g.88780640AGGTGTG>A), the defects of which causes KMT5B-related neurodevelopmental disorder, intellectual disability, microcephaly, facial dysmorphism, renal agenesis, and ambiguous genitalia syndrome (18).

Two putative mutations, one splice acceptor variant NM_001134367 (g. 14444242CT>C) and other stop gained variant NM_181661 (g.100133705T>G), are responsible for Cohen syndrome and retinal degeneration, were found in the proband (19). Moreover, of the variants identified four were located on genes that are involved in muscle disorders such as autosomal recessive spastic paraplegia, Nemaline myopathy 2, and fetal akinesia sequence (TRBV21-2, NM_001164508, NM_022140, NM_001330353) (20), two are related to premature ovarian failure (NM_207421 and NM_001317056) (21), infertility disorders (NM_030930, TRS-GCT6, NM_000348), gastrointestinal related features (IGHV3-79, NM_006249), cardiomyopathy (NM_001278344, NM_001289132) and one gene was identified responsible for deafness (NM_001079812). Additional 125 rare variants were also considered that are not provided in ClinVar database and are of largely unknown function while three variants (NM_016339, NM_00117243 and NM_001291745) observed are associated with various cancers (22).

Conclusion

In this study, we intend to provide the novel candidate gene variants that are responsible for the MEH phenotype of the patient studied by applying Trio approach. To the best of our knowledge, very rare studies have been performed on whole-exome level exploring the candidate gene variants for this disease. Although the study generated a comprehensive list of putative variants which is analytically less challenging still further extensive functional confirmatory studies are to be conducted for more certainty and a better understanding of its association with the disease. However, our findings are useful in contributing rich data set providing novel variants and genes.

Acknowledgement

Authors are thankful to the Meyer Children’s Hospital, Florence, Italy for providing this data publically available for the researchers to analyze it further, to use it as reference and in general to get the community more benefited.

References

1. Bai Y, Wang Y, Yang Y. Hepatic encephalopathy changes mitochondrial dynamics and autophagy in the substantia nigra. Metabolic brain disease. 2018;33(5):1669-78.
2. Ruszkiewicz J, Albrecht J. Changes in the mitochondrial antioxidant systems in neurodegenerative diseases and acute brain disorders. Neurochemistry international. 2015;88:66-72.
3. Zeharia A, Friedman JR, Tobar A, Saada A, Konen O, Fellig Y, et al. Mitochondrial hepato-encephalopathy due to deficiency of QIL1/MIC13 (C19orf70), a MICOS complex subunit. European Journal of Human Genetics. 2016;24(12):1778-82.
4. Lee WS, Sokol RJ. Mitochondrial hepatopathies: advances in genetics and pathogenesis. Hepatology. 2007;45(6):1555-65.
5. Gao C, Wang X, Mei S, Li D, Duan J, Zhang P, et al. Diagnostic yields of Trio accompanied by CNVseq for rare neurodevelopmental disorders. Frontiers in genetics. 2019;10:485.
6. ENA-European Nucleotide Archive
[Available from: https://www.ebi.ac.uk/ena/browser/view/PRJNA673368.

doi:10.20944/preprints202105.0068.v1

7. Pereira R, Oliveira J, Sousa M. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. Journal of clinical medicine. 2020;9(1):132.
8. 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.
9. Human Genome Resources at NCBI
[Available from: https://www.ncbi.nlm.nih.gov/genome/guide/human/.
10. Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV. Benchmarking short sequence mapping tools. BMC bioinformatics. 2013;14(1):1-25.
11. Zhao Q, editor A Study on Optimizing MarkDuplicate in Genome Sequencing Pipeline. Proceedings of the 2018 5th International Conference on Bioinformatics Research and Applications; 2018.
12. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. Nature biotechnology. 2020:1-9.
13. Diab NS, King S, Dong W, Allington G, Sheth A, Peters ST, et al. Analysis workflow to assess de novo genetic variants from human whole-exome sequencing. STAR protocols. 2021;2(1):100383.
14. Choi J, Tantisira KG, Duan QL. Whole genome sequencing identifies high-impact variants in well-known pharmacogenomic genes. The pharmacogenomics journal. 2019;19(2):127-35.
15. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol. 2013;9(7):e1003153.
16. Peluso F, Palazzo V, Indolfi G, Mari F, Pasqualetti R, Procopio E, et al. Leopard-like retinopathy and severe early-onset portal hypertension expand the phenotype of KARS1-related syndrome: a case report. BMC medical genomics. 2021;14(1):1-12.
17. Nakashima M, Kato M, Aoto K, Shiina M, Belal H, Mukaida S, et al. De novo hotspot variants in CYFIP2 cause early-onset epileptic encephalopathy. Annals of neurology. 2018;83(4):794-806.
18. Stessman HA, Xiong B, Coe BP, Wang T, Hoekzema K, Fenckova M, et al. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. Nature genetics. 2017;49(4):515-26.
19. Buratti E, Chivers M, Královičová J, Romano M, Baralle M, Krainer AR, et al. Aberrant 5′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. Nucleic acids research. 2007;35(13):4250-63.
20. Słabicki M, Theis M, Krastev DB, Samsonov S, Mundwiller E, Junqueira M, et al. A genome-scale DNA repair RNAi screen identifies SPG48 as a novel gene associated with hereditary spastic paraplegia. PLoS Biol. 2010;8(6):e1000408.
21. Jolly A, Bayram Y, Turan S, Aycan Z, Tos T, Abali ZY, et al. Exome sequencing of a primary ovarian insufficiency cohort reveals common molecular etiologies for a spectrum of disease. The Journal of Clinical Endocrinology & Metabolism. 2019;104(8):3049-67.
22. ClinVar Genomic variation as it relates to human health
[Available from: https://www.ncbi.nlm.nih.gov/clinvar/variation/302141/.