Application of Whole Exome-trio Analysis Reveals Rare Variants Associated With Congenital Pouch Colon

Sonal Gupta
Birla Institute of Scientific Research (BISR)
Praveen Mathur
SMS medical college and Hospital, JLN Marg
Ashwani Kumar Mishra
DNA Xperts
Krishna Mohan Medicherla
Birla Institute of Scientific Research (BISR)
Obul Reddy Bandapalli
Hopp Children's Cancer Center, German Cancer Research Center
Prashanth Suravajhala (prash@bioclues.org)
Amrita University

Research Article

Keywords: whole exome sequencing, trio exome, missense variants, congenital pouch colon

Posted Date: January 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1235311/v1

Abstract

Anorectal malformations (ARMs) are individually common, but congenital pouch colon (CPC), a rare anorectal anomaly, causes a dilated pouch in the genitourinary tract. In this work, we attempted to identify *de novo* heterozygous missense variants and further discovered variants of unknown significance (VUS), which could provide insights into CPC manifestation and its etiology. From whole exome sequencing (WES) performed earlier, the trio exomes were analyzed from those who were admitted to J.K. Lon Hospital, SMS Medical College, Jaipur, India between 2011-2017. The proband exomes were compared with the unaffected sibling/family members, and we sought to ask whether any variants of significant interest are associated with the CPC manifestation. The WES data from a total of 64 samples, including 16 affected neonates with their parents and unaffected siblings, were used for the study. Although all samples had unaffected sibling samples and data, we restricted our pool of analyses to all probands (11 male and 5 female) and unaffected parents/siblings. We previously attempted to understand the genetic makeup of CPC and identified genes responsible for the disease using WES. We examined the role of rare allelic variation associated with CPC in a 16 proband/parent trio family comparing the mutations to those of their unaffected parents/siblings. Our study across 16 probands revealed extremely rare variants, *viz.* TAF1B, MUC5B and FRG1. The variants were further validated to reveal disease-causing mutations associated with CPC and genitourinary diseases that could close the gaps of surgery by bringing intervention in therapies.

Introduction

Whole-exome sequencing (WES) has been an invaluable and cost-effective approach to identify genetic variants responsible for both Mendelian and polygenic diseases[1]. In the recent past, it has allowed the detection of clinically relevant genomic regions spanning known unknown regions, disease-associated sites and untranslated regions (UTRs)[2]. In addition to the well-known diseases, prenatal abnormalities, structural anomalies and congenital defects were studied, demonstrating a good diagnostic yield[3,4]. While WES approaches are abundant, they are limited if the disease in question is characteristically rare and medically inconclusive. This could be a deterrent because of the challenges in variant discovery, including rare and low-frequency mutations using next-generation sequencing (NGS) technologies. A recent decrease in the cost of WES and the accuracy of NGS has enabled researchers to study a large number of WES samples, but in the case of rare diseases, studying exome trios (proband/parents) or quads with the addition of siblings to discover single nucleotide variations (SNVs) and indels has proven to be a major landmark in the discovery of rare disease variants. For example, *Boycott et al., 2013; Gahl et al., 2012; Jacob et al., 2013* employed characteristic trio-exome analysis to infer candidate or driver mutations in rare diseases[5–7]. With human disease and genetic variation studies largely driven by NGS, a paramount challenge would be to explore *de novo* mutations, i.e., those not inherited from either father or mother. To check this, parent–child trios/quad WES analysis could be a powerful approach, although biases impede the identification of potential *de novo* mutations. Nevertheless, a majority of mutations may not transcend from parent to offspring, making comprehensive genetic variants to be analyzed to

confirm them as causal[8]. Trio-based exome sequencing has provided a benefit for identifying *de novo* variants in rare diseases, attributing them to largely heterozygous/causal mutations[9]. For rare diseases, although WES analysis often makes assumptions regarding disease inheritance (*de novo* vs. recessive), variant frequency and genetic heterogeneity, it has opened the path toward improved disease management or prognosis and effective therapies. For example, WES trios in schizophrenia patients for recessive genotypes were studied, with rare mutations in voltage gated sodium ion channels contributing to the disorder[10]. In another study, Jin *et al*. (2017) identified pathogenic mutations with an increased rate of *de novo* mutations in early-onset high myopia (EOHM) patients[11]. Recently, Quinlan-Jones *et al*. correlated proband–parent trios to determine the clinical utility of molecular autopsy underlying the etiology of structural anomalies[12].

Congenital pouch colon (CPC) is a rare type of high anorectal malformation wherein a part of or the entire colon becomes dilated in the form of a pouch with a fistula connecting the genitor-urinary tract[13]. Most of the incidences have been reported from India, with cases common to other countries accounted for, although males are prone to be largely affected, i.e., male to female ratio of 4:1[14]. From WES approaches, we previously identified mutations that are causal to CPC and reported candidate missense mutations[15]. In another study, we inferred the role of long noncoding RNAs (lncRNAs) from WES and identified *lnc*-EPB41-1-1, located in the intergenic regions of *EPB41* that are known to interact with *KIF13A*[16]. In this extended WES study of CPC, we examined the role of rare allelic variation associated with CPC in a 16 proband/parent trio family comparing the mutations to those of their unaffected parents/siblings. Keeping in view of understanding the genetic basis of CPC that could possibly delve into variation, an attempt was made to discover variants contributing to phenotypic heterogeneity.

Methods

Trio selection, sample collection and ethical approval

The CPC subjects were recruited from the J.K. Lon Hospital, SMS Medical College, Jaipur, India, in accordance with a protocol approved by the institutional ethics committee (IEC), SMS Medical College and Hospital, Jaipur, Rajasthan. The informed consent was obtained from all subjects, i.e., from parents for themselves (because blood from parents was also taken along with their children) and also for their children. All methods were performed in accordance with the relevant guidelines and regulations as per IEC, SMS Medical College and Hospital coalescing Helsinki guidelines. Blood samples were collected from all the probands, parents and unaffected siblings if any. The WES data from a total of 64 samples, including 16 affected neonates with their parents and unaffected siblings, were used for the study. Although all samples had unaffected sibling samples and data, we restricted our pool of analyses to all probands (11 male and 5 female) and unaffected parents/siblings. The methods and pipeline leading to family/quad analyses are summarized in Figure 3 (Supplementary information).

Variant annotation, filtering and quality control

The details of sequencing and variant calling in CPC subjects have been previously described[15]. Briefly, WES was performed on an Illumina multiplexed sequencer with paired-end chemistry and 110x effective coverage. Using our in-house developed pipeline[56], all unmapped sequence reads were aligned to the human reference genome (hg38), and variants were called. The mutations from the WES study were manually checked to discover the variants, if any, across the samples (Figure 1). First, all variants with MAF <0.05 were interpreted using 'grep' and *awk* liner commands, if present in probands but absent in their respective parents (Supplementary information) and healthy siblings. After manually checking the variants, we confined the prioritization of variants with filters set to an average depth of 250 and MAF <0.01 and MAF<=0.01% across all the trio samples. Further checking with the dbSNP[17], ClinVar[18], GnomAD[19] and COSMIC[20] databases, we used SNP-Nexus[21] to filter mutations listed in a cohort of databases, *viz.* SIFT[1], PolyPhen-2[22], Ensembl Variant Effect Predictor[23], MutationTaster[24], CADD[25] and GERP[26] prioritize pathogenic mutations if they are deleterious in nature. The CNVs and variants of unknown significance (VUS) were inferred by mapping the final list of variants to the SNPnexus. As a final check in reaching consensus for the variants present in all the probands, we checked the variants with multiple bioinformatics tools to find *bona fide* variants at the union of intersection of these methods, which we construe them to be causal.

Downstream analyses

We used the vcftools package (https://vcftools.github.io), and the mutations in probands were further checked with their relative parents/siblings for heterozygosity transmission. We also looked into the homozygous variants and considered them causal for CPC, where the proband was found to be homozygous and their respective parents and unaffected siblings if any were heterozygous for the specific allele. Enrichment analysis of data was performed to calculate the inclusiveness of parameters such as binomial probability and hypergeometric distribution[27]. After high-throughput screening, we undertook a candidate gene-set analysis based upon significantly enriched sets or rare mutations specifically in colon-related disorders. Seeking novel insights into the disorder, we used pathway analysis based upon gene ontology (GO) derived from PANTHER ontology[28] (http://pantherdb.org/tools) and EnrichR[29] (https://amp.pharm.mssm.edu/Enrichr/) annotation terms. Gene Ontology-based annotations included the biological network gene ontology tool (BinGO)[30], a plug-in for ontology annotation in Cytoscape[31] used for ontological analysis in the form of biological, cellular and metabolic processes.

Identification of transcripts for comparative screening

From RNA-Seq, a quality check was ensued after total RNA was isolated and cDNA double strand synthesis was performed on a pair of CPC type-4 samples. RNA-Seq was performed on an Illumina HiSeq 2000 platform with 2x100 bp paired-end sequencing chemistry, which generated ca. 32 million read pairs. The pair of samples was run through differential gene expression analysis using Cufflinks[32] and DESeq[33] pipelines, and a consensus was reached. To infer the role of lncRNAs, UVA FASTA software (https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml#) (v36. 6.8 version) and NONCODE

FASTA repository[34] were downloaded, and the intergenic regions of the genes from WES samples were queried. The lncRNA NONHSAT002007 was identified based on a query coverage e-value < 0.01. The sequences were carefully checked for bidirectional blast hits, and the lncRNA was visualized using an Ensembl genome browser for *bona fidelity*.

SNP Genotyping

For Sequenom genotyping, a multiplexed iPLEX assay was designed for 20 ng of DNA per sample to determine SNP calls using the Agena biosciences assay design suite. MALDI-TOF-MS analysis was performed using the Agena biosciences massArray analyzer platform on 29 SNPs in 37 DNA samples (16 probands and 21 controls). This method consists of five steps: PCR amplification, shrimp alkaline phosphatase treatment, single base extension, nanodispensing, and matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometry[35]. Data acquisition was automatically performed, and the mass window of analyte peak observation was set at 4500-9000 Da. Call frequencies, expressed as percentages, were calculated for each SNP.

Results And Discussion

We achieved a mean read depth of 80 in the targeted regions for an average depth of coverage 110x. We generated an average of 840667 total variants comprising 775262 SNPs and 65405 indels per exome in probands compared to 736820 SNPs and 63418 indels in unaffected samples from a total of 777216 variants (Figure 2; see methods). Missense variants constituted the largest variation followed by loss-of-function variants in cases when compared to unaffected samples, with an overall 0.35% missense, 0.02% frameshift, 0.004% stop gain and 0.0009% stoplost. The downstream analysis leading to variant calling was performed carefully to yield the list of the final number of variants. This was also checked with gene density and high linkage disequilibrium (LD) regions, even as MAF <=0.01 was sought for with SNPsnap having no candidate matches, thereby confirming that these variants are extremely rare. Given the rare phenotype, the frequency of the prioritized variants was checked for in agreement with that of 1000 Genomes, GnomAD and ExAC databases. From the reported familial history, relatedness tests were not felt necessary, as the pedigree confirmed correct parenthood for all affected/unaffected samples (Supplementary Table 1). From the final list of segregated variants (Table 1), we identified three mutations in the*MUC5B, FRG1*, and*TAF1B* genes, which we deem to be extremely rare variants (Table 2a). Furthermore, we also found an AK9 copy number variant (CNV) that was run through SNPsnap[36]to assess whether rare allelic variation was enriched for particular biological annotations (Supplementary Table 6). Finally, a set of candidate variants inferred from all samples was checked for validation using Sequenom array/plex (Table 3).

We found two independent lines of evidence from various tools and reached consensus in detecting variants from each sample with all filtering steps. For example, the pathogenic mutations contributing to each relevant proband were screened first wherein three variants, *viz.* FRG1 (4-189957414), TAF1B (2-9904885) and MUC5B (11-1238987) were confirmed through ClinVar and GnomAD. While the FRG1

mutation happens to be a CNV-associated missense seen in Z15 and Z99 samples, it is further augmented by the fact that low expression of FRG1 is associated with tumor progression in the colon[55]. The TAF1B mutation is highly associated with colorectal tissues, as we found these stopgain/missense mutations to be highly prolific in some of the aggressive CPC probands, *viz.* Z15, Z19, Z34, Z42, Z46, Z54, Z62, Z66, Z74. TAF1B is known to be the second largest subunit of the TATA box-binding protein (TBP)-containing promoter selectivity factor TIF-1B/SL1 and is connected with ribosomal transcription[51]. We found that the number of detected variants was associated with colon, rectal and genitourinary tissues and the presence of vivid disease occurrence. The density of these pouch-related tissues with missense/stop gain mutations can be attributed to aggressive CPC type IV.

Dissecting the genetic architecture of a rare disease is certainly an arduous task. Our CPC exome analysis was intended to fill this gap by detecting SNVs and CNVs affecting the focal genes/loci. Assessing these variants in CPC has provided ample evidence of strata coherent to CPC traits. In our study, at least three other genes from trio-exome analysis were reported in colon-related ailments, *viz*. DLC1, HAVCR1 and GBA3, but their allele frequencies are not *bona fide* and comparable. While we aimed to characterize the variants by employing different approaches, a polygenic model was assumed. This could be compounded with two assumptions: (a) capturing exomes and identifying deleterious mutations from a high depth of coverage exomes and (b) identifying large cohorts of mutations that fall in a low depth of coverage exomes. Although we observed both classes of genetic variation contributing to the etiology of the disease, inferring proband-parent trios and detecting *de novo* and transmitted genetic variants is quite a challenge. By considering extremely rare variants and adopting a strategy of identifying them in high-depth exomes, we validated all 16 trios. Nevertheless, we could not compare the detection yield inherent to this spectrum of patients owing to lack of CPC phenotype and trio-exome studies of similar design. Although previous studies have shown relatively similar methods, they detected medically relevant variants in the majority of the diseased phenotypes[11]. Our findings are in agreement with a large number of reports for rare variants, suggesting that the cumulative contribution of variants across different genes is associated with distinct phenotypes. In addition, an important challenge for researchers and clinicians currently in investigating rare disorders involves predicting pathogenicity for VUS. With several guidelines mentioned for predicting the pathogenicity of variants[8,18,39], molecular investigators face a daunting task in considering a rare variant as benign or pathogenic and inferring them to be pathogenic. In explaining the germline/heritability of complex variants, based on the rare-variant hypothesis, we argue that the extremely rare variants are associated with phenotype sampling[54]. Next, we showed how we can influence and prioritize these extremely rare variants and further proposed an optimization procedure to check the variants called between MAF <0.01 and MAF 0.01%. To address this, we discarded many variants that have MAF<0.01 and finally expanded the current annotation and prioritization to accommodate the CPC framework.

In rare diseases, where only a minority of the population is affected and prevalent in a specific geographical location, identifying and considering VUS will require thoughtful consideration. A notable among them, *viz.* The sequestome (SQSTM1 or P62) gene encodes a multifunctional scaffolding protein

involved in multiple cellular processes[40–41] in addition to showing mitochondrial integrity, import and dynamics as a discriminating autophagy receptor[42]. In addition, p62/SQSTM1 is ubiquitously expressed in various cell types, such as the cytoplasm, nucleus and lysosomes[43,] and is known to be overexpressed in various human genitourinary diseases, including colon cancer[44], hepatocellular carcinoma[45] and prostate cancer[46] (Supplementary Table 2). While KCNJ12 was also among the *bona fide* variants, our variant classification did not compel us to be considered extremely rare (not shown). It is known to initiate transcription by RNA polymerase I and acts as a channel for regulatory signals, while *KCNJ12* encodes ATP-sensitive inward rectifier potassium channel 12 and is subtly associated with repolarization of channels[52–53]. A list of segregated variants and those that were not extremely rare variants in the form of SQSTM1 and KCNJ12 were also considered for validation. The SQSTM1 mutation is invariably inherited from the father in the case of the Z12 index case and from the mother in the case of Z54. As the aforementioned variants were considered extremely rare variants, we also found AK9 (6-109528998.109528999 C|-) to be consistently seen across all probands. This deletion is invariably associated with nucleotide metabolism pathways and maintains homeostasis of cellular nucleotides[38]. Although AK9 (deletion) was phenomenally seen in all probands, we considered it a rare variant candidate for validation. Although multiple lines of evidence suggest that apart from VUS, mutations yielding somatic copy number alterations (SCNA) could not be ruled, as we found some uncommon missense mutations with MAF <0.05 in C7orf57, C9orf84, ORF5AR1, FGFR4, HLA-DRB5, NOTCH2NLA, and MUC5B genes that could be nonpathogenic/causal to CPC.

One of the interesting findings that emerged from our study is the role of the known unknowns or hypothetical genes that could be predecessors of noncoding; hence, their establishing roles in diseases such as CPC is limited[47]. Notably, the C10orf120 gene harbors CTCF binding sites, as these mutations remain undefined for most disease types, including cancer[48]. We observed that there was significant enrichment of CNVs and indels associated with intestinal/colon-related-specific genes, as they are widespread in tissues showing chromosomal instability, cooccur with neighboring chromosomal aberrations and are frequent in colon, rectum and gastrointestinal tumors but rare in other diseases. We argue that this mutational disruption associated with CTCF binding sites could be associated with pathogenesis, as it appears to be conserved in a majority of CPC probands (Supplementary Table 3). Another orphan ORF, *viz.* C7orf31 also harbors CTCF binding sites, which in fact showed significant enrichment for biological processes associated with regulatory, cellular and metabolic pathways (Table 4). Another C10orf120 has somatic variants subtly contributing to CPC pathogenesis and the maximum number of gains and losses observed in the form of CNVs in the CDC27, HLA-DRB5 and MST1 L genes (Table 2b). Although some of these ORFs' maternal association and pathogenicity cannot be ruled out, we construe that there are candidate genes that could be promising biomarkers as precursors of CPC, which is beyond the scope of this canonical hypothesis (Supplementary Table 4). In addition, we screened our variants from the Indian Genome Variant Database (IGVDB) and found that they are already reported in the Indian subpopulation[49,] and this stratification allowed us to review the patterns influencing common and rare variants. In principle, rare variants were found to have stronger patterns than common

variants. Thus, there is an inherent need to study the mutations in the known unknown regions, which would possibly delve into understanding rare diseases.

To gain insights into the role of lncRNAs, we revisited our hypothesis from our previous study[16] and reconfirmed whether NONHSAT002007 was inferred in WES samples with predictions from the NONCODE database. While we did not find mutations in lncRNAs from trio-exome analyses, we argue that the mutations in essential genes tend to be causal for rare diseases, paving the way for driver mutations with the mutations in noncoding genes suppressed for selective pressure. On a different note, we aimed to evaluate exome enrichment to that of the transcriptome and anticipated that any of the variants seen from our WES study could be associated with the differentially expressed genes (DEGs). From the transcriptome pair of CPC type-4 (proband and its unaffected parent), we observed several transcripts, alternative splice variants and fusion genes, but none of them could be associated with the causal genes inferred from the exome study (Supplementary Table 5). Although we found RGPD2 and RGPD4 genes known to be significantly associated with bowel/colon as among the top enriched, nevertheless, it is hypothetical to infer global gene expression from just a pair of datasets. This approach, if studied on all samples, we believe, could identify transcripts present at low levels, which in fact could be associated with the pathogenesis of CPC. As CPC cases emerge, it is difficult to classify the clinical significance of pathogenic variants without trio analysis, especially the interpretation of *de novo* variants. The trio exome data could provide insights into whether the variant could be inherited and whether the carrier mutation is transcended to the progeny. In conclusion, we argue that the genetic variability in CPC has had remarkable significance not only with the parents but also within each proband. With genetic diseases leading causes of death in infants, rapid clinical/trio exome sequencing could provide early diagnoses that could impact decision making in critically ill pediatric patients. Although more efficient early diagnosis could be made with intervention using chromosomal microarray (CMA), it was shown that the clinical utility of trio/exome/whole genome sequencing is more than the CMA[50]. The discovery of causal mutations could provide insights into developmental disorders/anorectal malformations, such as CPC, and their etiology, which closes the gaps of surgery, taking forward to precision therapy.

Declarations

Competing interests:

The authors declare no competing interests.

Authors' contributions:

S.G. and P.S. wrote the initial draft and analyzed the data. A.K performed Sequenom analysis on discussion with K.M., P.M. ORB and P.S.; K.M., O.R.B and P.S. conceived the project. O.R.B and P.S. proofread the manuscript before all authors agreed to it.

Funding:

P.M., and K.M. gratefully acknowledge the support of Indian Council of Medical Research toward Grant #5/41/11/2012 RMC.

Acknowledgments:

S.G. acknowledges the support of senior research fellowship from the Council of Scientific and Industrial Research, Government of India (Award #09/892(0002)2015-EMR-I). K.M. and P.S. gratefully acknowledges the support of BTIS-Net, Government of India and Advanced Bioinformatic Centre, Government of Rajasthan toward the support of Bioinformatics Infrastructure Facility at Birla Institute of Scientific Research (BISR), Jaipur.

References

1. Ng PC & Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research.* 2003. doi:10.1093/nar/gkg509. Available At: Https://Sift.Bii.A-star.Edu.Sg/. (Accessed: 12th October 2021).
2. Shen Y, Lopez R & Milos PM. The Rapidly Emerging Role for Whole Exome Sequencing in Clinical Genetics. *Current Genetics Medicine Reports*. 2014. doi:10.1007/s40142-014-0044-8.
3. Greenbaum L. *et al.* Evaluation of Diagnostic Yield in Fetal Whole-Exome Sequencing: A Report on 45 Consecutive Families. *Frontier in Genetics.* 2019. doi:10.3389/fgene.2019.00425.
4. Mone F, Quinlan-Jones E & Kilby MD. Clinical utility of exome sequencing in the prenatal diagnosis of congenital anomalies: A Review. *European Journal of Obstetrics and Gynecology and Reproductive Biology*. 2018. doi:10.1016/j.ejogrb.2018.10.016.
5. Boycott KM, Vanstone MR, Bulman DE & MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nature Reviews Genetics.* 2013. doi:10.1038/nrg3555.
6. Gahl WA. *et al.* The national institutes of health undiagnosed diseases program: Insights into rare diseases. *Genetics in Medicine.* 2012. doi:10.1038/gim.0b013e318232a005.
7. Jacob HJ. *et al.* Genomics in clinical practice: Lessons from the front lines. *Science Translational Medicine.* 2013. doi:10.1126/scitranslmed.3006468.
8. MacArthur DG. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature.* 2014. doi:10.1038/nature13127.
9. Zhu X. *et al.* Whole-exome sequencing in undiagnosed genetic diseases: Interpreting 119 trios. *Genetics in Medicine.* 2015. doi:10.1038/gim.2014.191.
10. Rees E. *et al.* Analysis of exome sequence in 604 trios for recessive genotypes in schizophrenia. *Translational Psychiatry.* 2015. doi:10.1038/tp.2015.99.
11. Jin ZB. *et al.* Trio-based exome sequencing arrests de novo mutations in early-onset high myopia. *Proceedings of the National Academy of Sciences of the United States of America*. 2017. doi:10.1073/pnas.1615970114.

12. Quinlan-Jones E. *et al.* Molecular autopsy by trio exome sequencing (ES) and postmortem examination in fetuses and neonates with prenatally identified structural anomalies. *Genetics in Medicine.* 2019. doi:10.1038/s41436-018-0298-8.
13. Mathur P, Prabhu K & Jindal D. Unusual presentations of pouch colon. *Journal of Pediatric Surgery.* 2002. doi:10.1053/jpsu.2002.35007.
14. Gupta D & Sharma S. Congenital pouch colon - Then, and now. *Journal of Indian Association of Pediatric Surgeons.* 2007. doi:10.4103/0971-9261.31081.
15. Mathur P. *et al.* Whole exome sequencing reveals rare variants linked to congenital pouch colon. *Scientific Reports.* 2018. doi:10.1038/s41598-018-24967-y.
16. Gupta S. *et al.* Lnc-EPB41-protein interactions associated with congenital pouch colon. *Biomolecules.* 2018: 8.
17. Sherry ST. *et al.* DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research.* 2001. doi:10.1093/nar/29.1.308. Available At: Https://Www.Ncbi.Nlm.Nih.Gov/Snp/. (Accessed: 22nd September 2021).
18. Landrum MJ. *et al.* ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research.* 2016. doi:10.1093/nar/gkv1222. Available At: Https://Www.Ncbi.Nlm.Nih.Gov/Clinvar/(Accessed: 27th September 2021).
19. Karczewski KJ. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020. doi:10.1038/s41586-020-2308-7.
20. Tate JG. *et al.* COSMIC: The Catalog Of Somatic Mutations In Cancer. *Nucleic Acids Research.* 2019. doi:10.1093/nar/gky1015. Available At: Https://Cancer.Sanger.Ac.Uk/Cosmic. (Accessed: 16th September 2021).
21. Chelala C, Khan A & Lemoine NR. SNPnexus: A web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics.* 2009. doi:10.1093/bioinformatics/btn653. Available At: Https://Www.Snp-nexus.Org/Index.Html. (Accessed: 20th October 2021).
22. Adzhubei IA. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods*. 2010. doi:10.1038/nmeth0410-248. Available At: Http://Genetics.Bwh.Harvard.Edu/Pph2/. (Accessed: 21st September 2021).
23. McLaren W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology.* 2016. doi:10.1186/s13059-016-0974-4. Available At: Https://Asia.Ensembl.Org/Homo_sapiens/Tools/Vep. (Accessed: 14th October 2021).
24. Schwarz JM, Cooper DN, Schuelke M & Seelow D. Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nature Methods.* 2014. doi:10.1038/nmeth.2890. Available At: Http://Www.Mutationtaster.Org/ (Accessed: 19th September 2021).
25. Kircher M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics.* 2014. doi:10.1038/ng.2892.

26. Cooper GM. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*. 2005. doi:10.1101/gr.3577405.
27. Huang DW, Sherman BT & Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*. 2009. doi:10.1093/nar/gkn923.
28. Ashburner M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics.* 2000. doi:10.1038/75556.
29. Chen EY. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics.* 2013. doi:10.1186/1471-2105-14-128. Available At: Https://Amp.Pharm.Mssm.Edu/Enrichr/. (Accessed: 21st September 2021).
30. Maere S, Heymans K & Kuiper M. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics.* 2005. doi:10.1093/bioinformatics/bti551. Available At: Https://Doi.Org/10.1093/Bioinformatics/Bti551. (Accessed: 19th September 2021).
31. Shannon P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003. doi:10.1101/gr.1239303.
32. Trapnell C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocol.* 2012. doi:10.1038/nprot.2012.016.
33. Anders S. & Huber W. Differential expression analysis for sequence count data. *Genome Biology.* 2010. doi:10.1186/gb-2010-11-10-r106.
34. Fang S. *et al.* NONCODEV5: A comprehensive annotation database for long noncoding RNAs. *Nucleic Acids Research.* 2018. doi:10.1093/nar/gkx1107. Noncode. Available At: Http://Www.Noncode.Org/Show_rna.Php?Id=Nonhsat002007&Version=2&Utd=1#. Www.Noncode.Org (Accessed: 9 November 2021).
35. Gabriel S, Ziaugra L & Tabbaa D. SNP genotyping using the sequenom massARRAY iPLEX Platform. *Current Protocols in Human Genetics.* 2009. doi:10.1002/0471142905.hg0212s60.
36. Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: A Web-based tool for identification and annotation of matched SNPs. *Bioinformatics.* 2015. doi:10.1093/bioinformatics/btu655.
37. Sun CYJ. *et al.* Facioscapulohumeral muscular dystrophy region gene 1 Is a dynamic RNA-associated and actin-bundling protein. *Journal of Molecular Biology.* 2011. doi:10.1016/j.jmb.2011.06.014.
38. Amiri, M., Conserva, F., Panayiotou, C., Karlsson, A. & Solaroli, N. The human adenylate kinase 9 is a nucleoside mono- and diphosphate kinase. *The International Journal of Biochemistry & Cell Biology.* 2013. doi:10.1016/j.biocel.2013.02.004.
39. Richards S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine.* 2015. doi:10.1038/gim.2015.30.

40. Kim, S., Eun, H. & Jo, E.-K. Roles of Autophagy-Related Genes in the Pathogenesis of Inflammatory Bowel Disease. *Cells.* 2019. doi:10.3390/cells8010077.
41. Sanz, L., Sanchez, P., Lallena, M. J., Diaz-Meco, M. T. & Moscat, J. The interaction of p62 with RIP links the atypical PKCs to NF-κB activation. *EMBO J.* 1999. doi:10.1093/emboj/18.11.3044.
42. Seibenhener, M. L. *et al.* A role for sequestosome 1/p62 in mitochondrial dynamics, Import and genome integrity. *Biochimica et Biophysica Acta - Molecular Cell Research.* 2013. doi:10.1016/j.bbamcr.2012.11.004.
43. Pankiv S. *et al.* Nucleocytoplasmic shuttling of p62/SQSTM1 and its role in recruitment of nuclear polyubiquitinated proteins to promyelocytic leukemia bodies. *Journal of Biological Chemistry.* 2010. doi:10.1074/jbc.M109.039925.
44. Mohamed A. *et al.* P62/Ubiquitin IHC Expression Correlated with Clinicopathologic Parameters and Outcome in Gastrointestinal Carcinomas. *Frontiers in Oncology.* 2015. doi:10.3389/fonc.2015.00070.
45. Bao L. *et al.* Impaired autophagy response in human hepatocellular carcinoma. *Experimental and Molecular Pathology.* 2014. doi:10.1016/j.yexmp.2013.12.002.
46. Kitamura H. *et al.* Cytosolic overexpression of p62 sequestosome 1 in neoplastic prostate tissue. *Histopathology.* 2006. doi:10.1111/j.1365-2559.2005.02313.x.
47. Malik G. *et al.* Hypothetical Proteins as Predecessors of Long Noncoding RNAs. *Current Genomics*. 2020. doi:10.2174/1389202921999200611155418.
48. Guo YA. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nature Communications*. 2018. doi:10.1038/s41467-018-03828-2.
49. Brahmachari SK. *et al.* The Indian Genome Variation database (IGVdb): A project overview. *Human Genetics.* 2005. doi:10.1007/s00439-005-0009-9.
50. Clark MM. et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. npj Genomic Medicine 2018; 3:16.
51. Comai L, Zomerdijk JCBM, Beckmann H, Zhou S, Admon A, Tjian R. Reconstitution of transcription factor SL1: exclusive binding of TBP by SL1 or TFIID subunits. *Science.* 1994; 266:1966-1972
52. Koch M, Beckhove P, Op den Winkel J, et al. Tumor infiltrating T lymphocytes in colorectal cancer: Tumor-selective activation and cytotoxic activity in situ. Annals of Surgery. 2006;244(6):986-993. doi:10.1097/01.sla.0000247058.43243.7b
53. Yuan HX, Yan K, Hou DY, et al. Whole exome sequencing identifies a KCNJ12 mutation as a cause of familial dilated cardiomyopathy. Medicine (Baltimore). 2017;96(33):e7727. doi:10.1097/MD.0000000000007727.
54. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genetic Epidemiology. 2011;35(8):790-9. doi: 10.1002/gepi.20628.

55. Tiwari A, Pattnaik N, Mohanty Jaiswal A, Dixit M. Increased FSHD region gene1 expression reduces *in vitro* cell migration, invasion, and angiogenesis, *ex vivo* supported by reduced expression in tumors. Bioscience Reports. 2017;37(5):BSR20171062. doi: 10.1042/BSR20171062.
56. Meena N, Mathur P, Medicherla KM and Suravajhala P. A Bioinformatics Pipeline for Whole Exome Sequencing: Overview of the Processing and Steps from Raw Data to Downstream Analysis. *Bio101*, 2018: e2805. doi: 10.21769/BioProtoc.2805.

Tables

Due to technical limitations, tables xlsx is only available as a download in the Supplemental Files section.

Figures

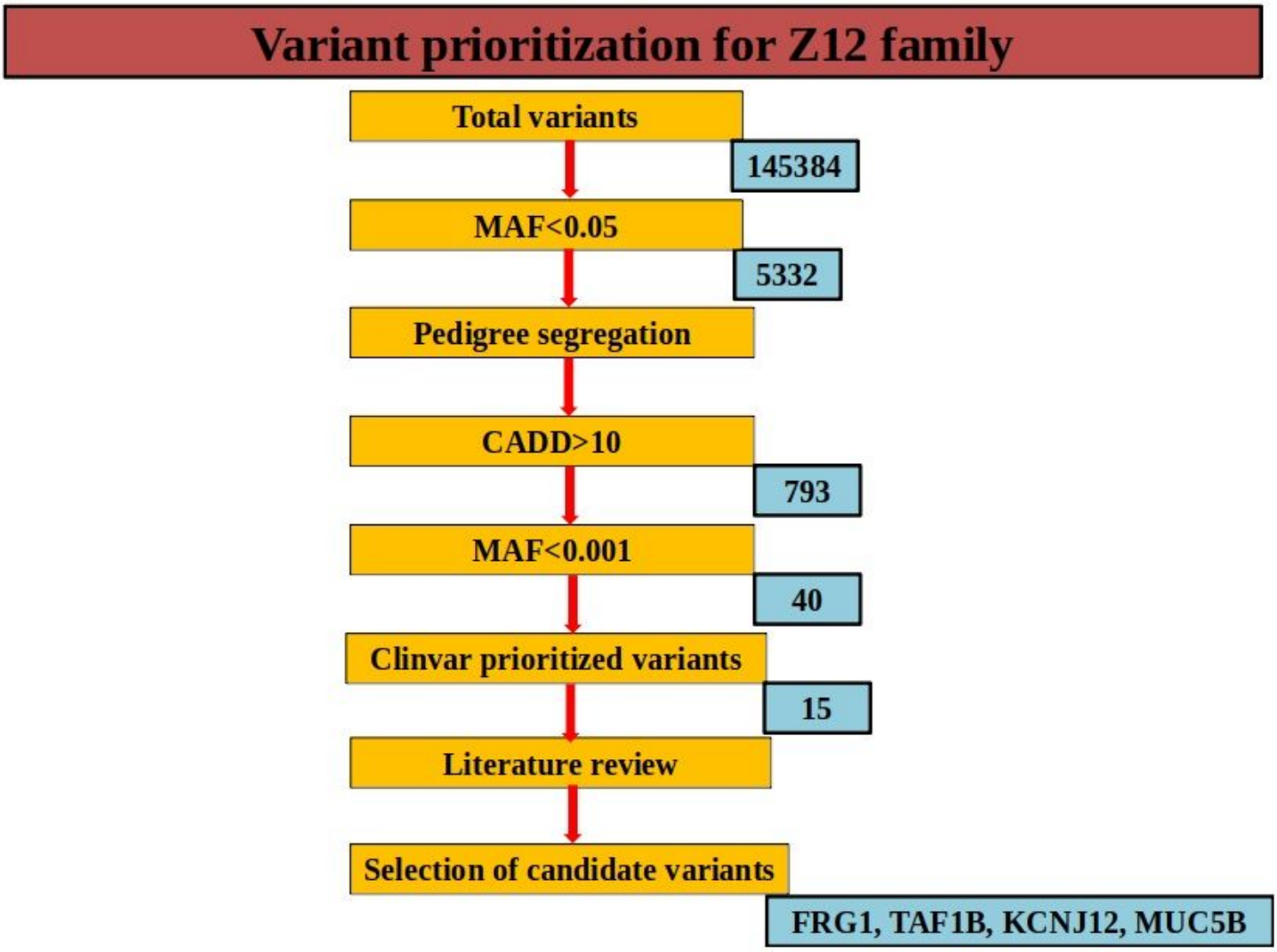

Figure 1

A schematic diagram for the prediction of causal variants in the Z12 family.

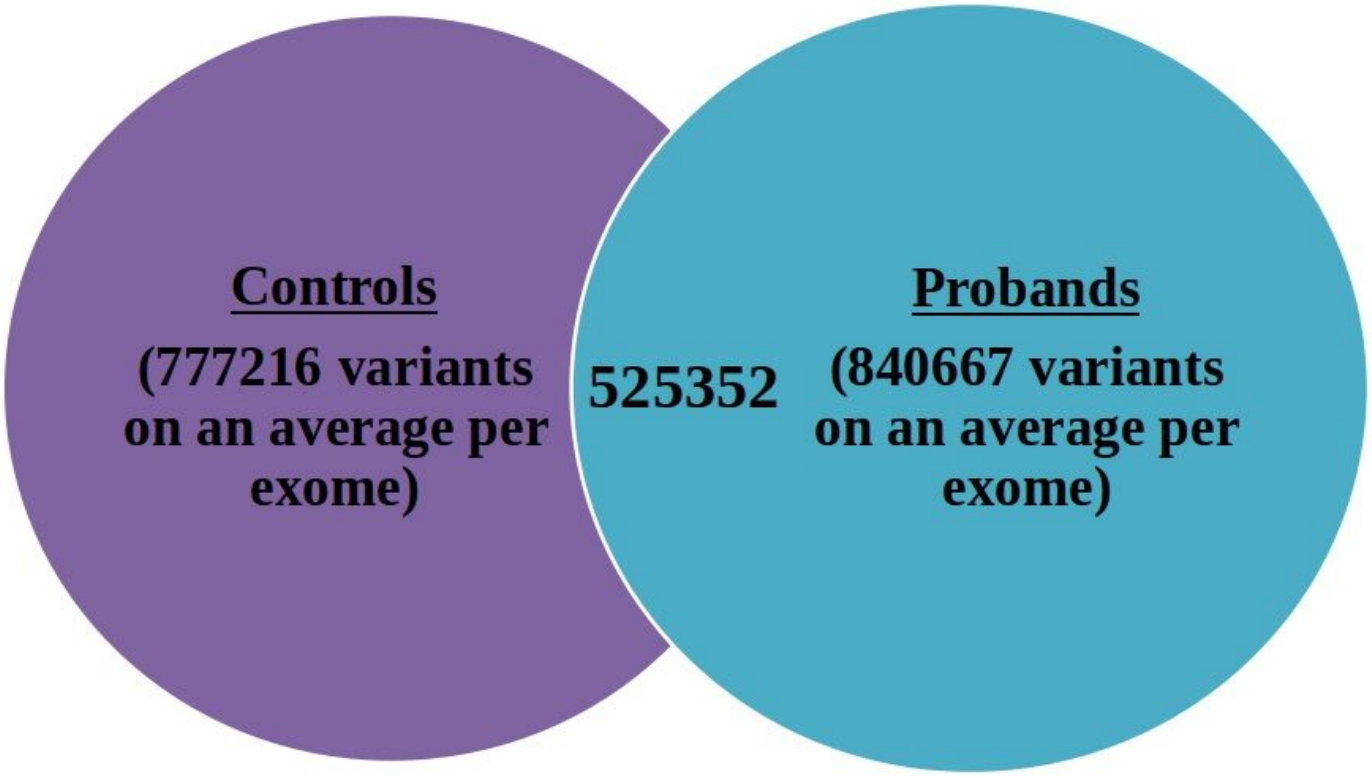

Figure 2

Venn diagram showing variants per exome in controls versus probands. The 525352 are the common variants between them.

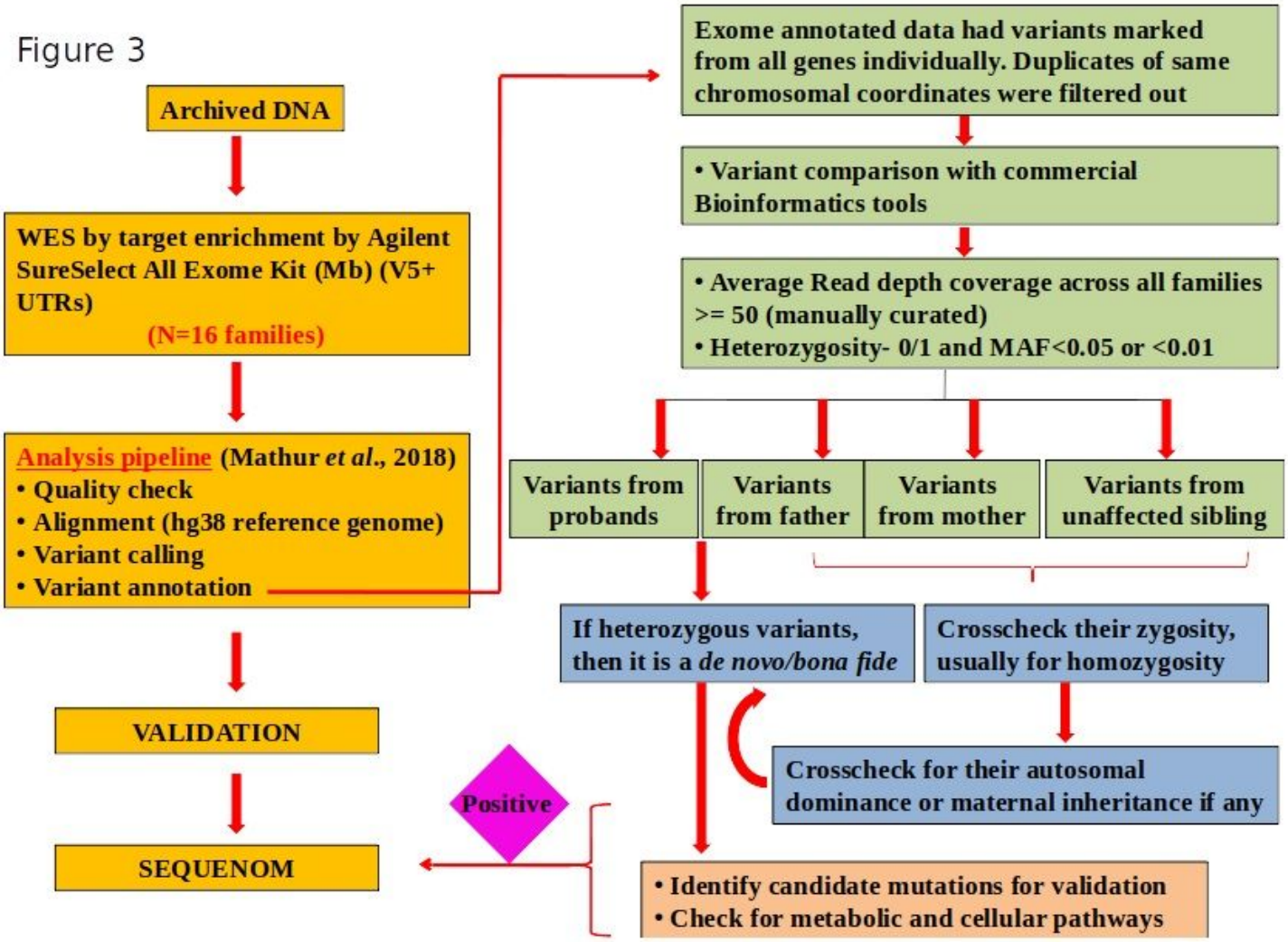

Figure 3

Detailed pipeline for identifying candidate mutations and further screening the data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TrioSupplementarytables.xlsx
- TablesTrio.xlsx