Whole exome-trio analysis reveals rare variants associated with Congenital Pouch Colon

Sonal Gupta^{1,2}, Praveen Mathur³, Ashwani Kumar Mishra⁴, Krishna Mohan Medicherla¹ and Prashanth Suravajhala^{1*}

1. Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research (BISR), Statue Circle, Jaipur 302021, RJ India
2. Department of Biotechnology, Amity University Rajasthan, Kant Kalwar, Jaipur 303002, RJ India
3. Department of Pediatric surgery, SMS medical college and Hospital, JLN Marg, Jaipur 302004, RJ India
4. DNA Xperts, Noida, India

*Correspondence: prash@biser.res.in

Abstract

Anorectal malformations (ARM) are individually common but Congenital Pouch Colon (CPC), a rare anorectal anomaly causes a dilated pouch in genitourinary tract. We have earlier attempted to understand the clinical genetic makeup of CPC and identified genes responsible for the disease using whole exome sequencing (WES). Here we report our studies of CPC, by identifying *de novo* heterozygous missense mutations in 16 proband-parent trios and further discover variants of unknown significance which could provide insights into CPC manifestation and its etiology. Our study confirms candidate mutations in genes, *viz.* C7orf57, C10orf120, C9orf84 and MUC16, CTC1 particularly emphasizing the role of hypothetical genes or open reading frames causing this developmental disorder. Variant validation revealed disease causing mutations associated with CPC and genitourinary diseases which could close the gaps of surgery in bringing intervention in therapies.

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

Whole Exome Sequencing (WES) has been an invaluable and cost-effective genomic approach to identify genetic variants responsible for both Mendelian and polygenic diseases (Ng *et al.*, 2010). In the recent past, it has allowed detecting clinically relevant genomic regions spanning the known unknowns regions, disease-associated sites and untranslated regions (UTRs) (Shen *et al.*, 2014). In addition to the well known diseases, prenatal abnormalities, structural anomalies and congenital defects were studied demonstrating a good diagnostic yield (Greenbaum *et al.*, 2019; Mone *et al.*, 2018). While WES approaches are abundant, they are limited if the disease in question is characteristically rare and medically inconclusive. This could be deterrent because of the challenges in variant discovery, including rare and low-frequency mutations using next generation sequencing (NGS) technologies. Recent decrease in cost of WES and accuracy of the NGS has enabled the researchers to study trios (proband/parents) or quad, in case of an addition of sibling to discover single nucleotide variations (SNVs) and indels. In the recent past, several research groups reported genetic diagnosis for rare disorders using WES applications while a few used characteristic trio-exome analysis to infer the candidate or driver mutations (Boycott *et al.*, 2013; Gahl *et al.*, 2012; Jacob *et al.*, 2013). With human disease and genetic variation studies largely driven by NGS, a paramount challenge would be to explore *de novo* mutations, i.e. those inherited from either father or mother. To check this, parent-child trios/quad WES analysis could be a powerful approach although biases impede identifying potential *de novo* mutations. Nevertheless, a majority of mutations may not transcend from parent to offspring making comprehensive genetic variants to be analyzed for confirming them as causal (MacArthur *et al.*, 2014). The trio-based exome sequencing has provided a benefit for identifying *de novo* variants in rare diseases attributing them to be heterozygous/causal mutations (Zhu *et al.*, 2015). 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 towards 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 (Rees *et al.*, 2015). In another study, Jin *et al.* (2017) have identified the pathogenic mutations with an increased rate of *de novo* mutations in early-onset high myopia (EOHM) patients (Jin *et al.*, 2017). Recently, Quinlan-Jones *et al.* have

correlated the proband-parent trios to determine the clinical utility of molecular autopsy underlying etiology of structural anomalies (*Quinlan-Jones et al., 2019*).

Congenital Pouch Colon (CPC) is a rare type of high anorectal malformation wherein a part of or entire colon gets dilated in the form of pouch having fistular communication with genitor-urinary tract (*Mathur et al., 2002*). Most of the incidences have been reported from India with cases common to other countries accounted although the males are prone to be largely affected, i.e. male to female ratio of 4:1 (*Gupta and Sharma, 2007*). From WES approaches, we have earlier identified mutations that are causal to CPC and reported candidate missense and nonsense mutations in genes, viz. C7orf57, C10orf120, C9orf84 and MUC16, CTC1 respectively (*Mathur et al., 2018*). In a separate study, we have also inferred the role of long non-coding RNAs (lncRNAs) from a WES and identified *lnc-EPB41-1-1*, located in the intergenic regions of EPB41 that is known to be interacting with KIF13A (*Gupta et al., 2018*). In this extended WES study of CPC, we examine the role of *de novo* mutations in a 16 proband/parent trio family comparing the mutations to that of their unaffected parents. Keeping in view of understanding the genetic basis of CPC that could possibly delve into phenotypic variation, an attempt was made to discover a number of *de novo* variants contributing to the heterogeneity. We identified mutations spanning the orphan open reading frames (ORFs), and a few frame shift mutations, besides mutations in GBA3 and HAVCR1 associated with intestine and urogenital tissues.

Results and Discussion

The downstream analyses leading to the variant calling was done carefully to yield the list of final number of variants. We achieved a mean read depth of 80 in the targeted regions for an average depth of coverage 110x. We generated a norm of 840667 total variants comprising 775262 SNPs and 65405 indels per exome in probands as compared to 736820 SNPs and 63418 indels in unaffected samples from a total of 777216 variants (Figure 1; see methods). Missense variants constituted the large 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% stop lost (Table 1). From the final list of inferring causality, we identified 40 SNPs/ indels with a few non-synonymous

mutations as a result of low quality calls (Table 2) which in turn were mapped to 37 single nucleotide variants (SNVs) (Table 3) and 385 copy number variants (CNVs) (Table 4). Given the rare phenotype, the frequency of the confirmed variants was checked for in agreement with that of 1000 genome and ExAC databases. From the reported familial history, the relatedness tests were not felt necessary as the pedigree confirmed correct parenthood for all affected/unaffected samples with the shared homozygosity (Figure 2 and supplementary table 1). Finally, a set of candidate variants inferred from all samples were checked for validation using Sequenom array/plex (Table 5).

Figure 1: Venn diagram showing variants per exome in (a) probands only and (b) controls versus probands.

Figure 2: Families with inherited mutations in CPC probands (marked with arrow, please refer supplementary Table 1 for details)

We found two independent lines of evidence from various tools and reached consensus in detecting variants from each samples with all filtering steps. For example, the pathogenic mutations contributing to the each relevant proband were screened first. Altogether, we observed that MUC3A, TAF1B, in addition to the two hypothetical protein-coding genes, viz. C10orf120 and C9orf84, contribute to CPC and these could be interesting as they could be novel genes responsible for pathogenesis of CPC. Although from dbSNP, we checked whether or not the SNPs are missense/frameshift variants, we attempted to validate them according to their clinical/pathogenic significance besides checking the heterozygous mutations for minor allele frequencies (MAF<0.01). Among them, 11 variants were already reported in dbSNP (C7orf57, C10orf120, C7orf31, C9orf84, C11orf96, OR5AR1, TAF1B, MUC3A, DLC1, AK9, FGFR4, KRTAP4-1) whereas AK9 was found to be *de novo* in two probands (Z19 and Z99) with another putative *de novo* variant, viz. FRG1, which was later found to be a false positive post Sequenom validation. In summary, we observed FRG1 and AK9 as *de novo* variants in CPC-XI and CPC-XII, CPC-XVI families. While FRG1 is a protein coding gene involved in mRNA transport and processing linked with epigenetic regulation of muscle differentiation by regulating histone-lysine N-methyltransferase activity (*Sun et al., 2011*), it is associated with disease pathogenesis and its over-expression leads to development of facio-scapulohumeral muscular dystrophy (FSHD1)-like symptoms. On the other hand, AK9 is involved in nucleotide metabolism pathways and maintains homeostasis of cellular nucleotides (*Amiri et al., 2013*).

One of the important challenges for researchers and clinicians nowadays in investigating rare disorders involves predicting pathogenicity of variants of

unknown significance (VUS). With several guidelines mentioned for predicting pathogenicity of variants (*Landrum et al., 2016; MacArthur et al., 2014; Richards et al., 2015*), molecular investigators face a daunting task in considering a rare variant as benign or pathogenic and inferring them to pathogenesis. In rare diseases, where only a minority of population is getting affected and also prevalent to a specific geographical location, identifying and considering VUS will require a thoughtful consideration. In our study, candidate variants such as SQSTM1 (rs10277) , GBA3, FGFR4 and HAVCR1 are shown to be reported in urogenital and colon related disorders and are not reported as deleterious in most of the pathogenicity determination databases (Supplementary figure 1). A notable among them, *viz.* sequestome (SQSTM1 or P62) gene encodes a multifunctional scaffolding protein involved in multiple cellular processes (*Sanz et al., 1999*) besides showing mitochondrial integrity, import and dynamics as a discriminating autophagy receptor (*Seibenhener et al., 2013*). In addition, p62/SQSTM1 is expressed ubiquitously in various cell types such as cytoplasm, nucleus and lysosomes (*Pankiv et al., 2010*) and is known to be over expressed in various human genitourinary diseases including colon cancer (*Mohamed et al., 2015*), hepatocellular carcinoma (*Bao et al., 2014*) and prostate cancer (*Kitamura et al., 2006*) (Supplementary Table 2). Although multiple lines of evidence, suggest that apart from VUS, mutations yielding somatic copy number alterations (SCNA) could not be ruled, we found these missense mutations in C7orf57, C9orf84, ORF5AR1, FGFR4, HLA-DRB5, NOTCH2NLA, MUC5B genes associated with CPC. Interestingly, we observed CNVs in CDC27, HLA-DRB5 and MST1L genes with maximum number of gains and losses (Table 4). Particularly, CDC27, a Cdk-inhibitory kinase of the anaphase-promoting complex is known to be associated with colon inactivation.

To confer the mutations associated with pathogenesis, we observed that C7orf31, OR5AR1, PABPC1, CTBP2 and NOTCH2NLA genes harbor deleterious mutations which were screened through SIFT, PolyPhen2, MutationTaster, Ensembl VEP, Phred CADD score (>18) GERP score and COSMIC prediction tools (Table 2). Interestingly, CTBP2 gene which has a heterozygous missense mutation (rs61870306; c.G1147C; p.N249K) was predicted as lethal with Polyphen score of 0.996 showing high conservation among species. Therefore, we considered to model the PDB structure with these mutations showing unfavorable free energy change ($\Delta\Delta G$) with reduced

stability of -1.82 that may affect the protein stability (Figure 3). The interaction between nearby amino acids and other surrounding residues could ultimately reduce the protein stability.

(a)

(b)

Figure 3: Wild-type and mutant (Asp249Lys) residues are colored in light-green and represented as sticks alongside the surrounding residues.

Dissecting the genetic architecture of a rare disease is certainly an arduous task. Our CPC exome analyses that we have done was intended to fill this gap to detect mutations, SNVs to chromosomal aberrations affecting the focal genes/loci. Assessing these variants in CPC has provided an ample evidence of strata coherent to CPC traits. In our study, at least 4 of the 40 confirmed variants from trio-exome analysis were reported in colon related ailments viz. *DLC1*, *HAVCR1*, *FGFR4* and *GBA3*. While we have aimed to characterize the variants by employing different approaches, a polygenic model is assumed. This could be compounded with two assumptions (a) capturing exomes and identifying deleterious mutations from high depth of coverage exome and (b) identifying large cohort of mutations that fall in low depth of coverage exomes. Though 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 rare variants and adopting a strategy of identifying them in high depth exome, we validated all the 16 trios. Nevertheless, we could not compare the detection yield inherent to these spectrums 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 majority of the diseased phenotypes (*Jin et al., 2017*). Our findings are in agreement with a large number of reports for rare variants, suggesting that cumulative

contribution of *de novo* variants across different genes is associated with distinct phenotypes.

One of the interesting findings that have emerged from our study is the role of the known unknowns or hypothetical genes establishing their roles in various diseases. A notable among them is C10orf120 gene which harbors CTCF binding site even as these mutations remain undefined for most disease types including cancer (*Guo et al., 2018*). We observe that there is significant enrichment of indels associated with intestinal/colon related-specific genes as these mutations are widespread in tissues showing chromosomal instability, co-occurring with neighbouring chromosomal aberrations, and are frequent in colon, rectum and gastrointestinal tumours 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). A careful check in Ensembl regulatory feature reveals that the C/T mutation was experimentally validated in mouse (supplementary figure 2) supporting the aforementioned argument. Another orphan ORF, *viz.* C7orf31 also harbours CTCF binding sites and this is in agreement with the fact that the functional pathway analysis showed significant enrichment for biological processes associated with regulation wherein most of our variants are known to be associated with cellular and metabolic pathways (Table 6). Interestingly, we observed that these two ORFs (C10orf120 and C7orf31) are localized to mitochondria and so we asked if there is a plausible role of hypothetical ORFs harboring candidate mutations inherited from mother (rs2947594 and rs2285738). On the other hand, we argue that these mutations are largely autosomal dominant/ *de novo* with a few x-linked or mitochondrial or maternally inherited genes (Supplementary Table 4). Among them, fibroblast growth factor receptor 4 gene (FGFR4) is known to be associated with genitourinary diseases. Although their maternal association cannot be ruled out, we construe that these candidate genes could be promising biomarkers as precursors of CPC. In addition, we screened our variants from Indian Genome Variant Database (IGVDB) and found that they are already reported in Indian subpopulation (*Brahmachari et al., 2005*) and this stratification allowed us to review the patterns influencing common and rare variants. In principle, the rare variants were found to have stronger patterns when compared to the 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 erstwhile study¹⁶ and reconfirmed whether or not 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 way for driver mutations with the mutations in non-coding genes suppressed for selective pressure. On a different note, we aimed to evaluate exome enrichment to that of transcriptome and anticipated if 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 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 confirm this inferring global gene expression from just a pair of dataset. 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.

In conclusion, we argue that the genetic variability in CPC has had a remarkable significance not only with the parents but also within each proband. The discovery of causal mutations could provide insights into the developmental disorders/anorectal malformation such as CPC and its etiology, which closes the gaps of surgery, taking forward to precision therapy.

Methods

Trio selection

The CPC subjects were recruited from the SPJMC (J.K. Lon Hospital), SMS Medical College, Jaipur in accordance with a protocol approved by the institutional ethics committee (IEC) of the hospital. A written informed consent was provided by the parents on behalf of their children. The 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 probands (11 male and 5 female) and unaffected parents/sibling. The methods and pipeline leading to family/quad analyses is summarized in Figure 4 (Supplementary Methods).

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 Illumina multiplexed sequencer with paired-end chemistry and 110x effective coverage. Using our in-house developed pipeline, all unmapped sequence reads were aligned to the human reference genome (hg38) and variants were called. The mutations from WES study were manually checked to discover the variants, if any across the samples. To start with, *de novo* variants were interpreted using 'grep' and *awk* liner commands, if present in probands but absent in their respective parents (Supplementary information). After manually checking the variants, we confined the prioritization of variants to an average depth of 50 across all the trio samples. Further checking with dbSNP (Sherry *et al.*, 2001), ClinVar (Landrum *et al.*, 2016) and COSMIC (Tate *et al.*, 2019) databases, we used SNP-Nexus (Chelala *et al.*, 2009) to filter mutations listed in a cohort of databases, viz. SIFT (Ng and Henikoff, 2003), PolyPhen-2 (Adzhubei *et al.*, 2010), Ensembl Variant Effect Predictor (McLaren *et al.*, 2016), MutationTaster (Schwarz *et al.*, 2014), CADD (Kircher *et al.*, 2014) and GERP (Cooper *et al.*, 2005), and confirmed the pathogenic mutations, if they are deleterious in nature. The CNVs and VUS were inferred by mapping the final list of variants to SNP-Nexus. As a final check in reaching consensus for the common variants across all the probands, we compared the variants from multiple bioinformatics tools so as to find *bona fide* variants at the union of intersection of these methods which we construe them to be causal.

Downstream analyses

To confirm heterozygous mutations, we used *vcftools* package (<https://vcftools.github.io>) and the mutations in probands were further checked with their relative parents/sibling for heterozygosity transmission. We also looked into the homozygous variants and considered them as causal for CPC where proband is found heterozygous and their respective parents and unaffected sibling if any were

homozygous for the specific allele. Enrichment analysis of data was done to calculate inclusive of parameters like binomial probability, hyper geometric distribution (Huang et al., 2009). After high throughput screening, we undertook a candidate gene-set analysis based upon significantly enriched fsets 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 (Ashburner et al., 2000) (<http://pantherdb.org/tools>), EnrichR (Chen et al., 2013) (<https://amp.pharm.mssm.edu/Enrichr/>) annotation terms. Gene Ontology based annotations included Biological Networks Gene Ontology tool (BinGO) (Maere et al., 2005), a plug-in for ontology annotation in Cytoscape (Shannon et al., 2003) used for ontological analysis in the form of biological, cellular and metabolic processes.

Identification of transcripts for comparative screening

From RNA-Seq, quality check was ensued after total RNA was isolated and cDNA double strand synthesis on a pair of CPC type-4 samples. The RNA-Seq was performed on Illumina HiSeq 2000 platform with 2x100bp paired end sequencing chemistry which generated ca. 16 million reads. The pair of samples was run through differential gene expression analysis using Cufflinks (Trapnell et al., 2012)

and DESeq (Anders *et al.*, 2010) pipelines and a consensus was reached. For inferring 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 (Fang *et al.*, 2017) were downloaded and the intergenic regions of the genes from WES samples were queried. The lncRNA - NONHSAT002007 was identified based on the 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, Multiplexed iPLEX Assay was designed using Assay design suite by Agena Biosciences and MALDI-TOF-MS analysis was performed using 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 (Gabriel *et al.*, 2009). Data acquisition was automatically performed and mass window of analyte peak observation was set at 4500-9000 Da. Call frequencies, expressed as percentages was calculated for each SNP. For Sequenom, 20 ng of DNA with reference was used to determine 40 SNP calls; however 29 SNPs were validated due to the general failure of SNP calls in, viz. rs483929, rs10277, rs4935, rs8002, rs9260191, rs1051454, rs1141701, rs76046078, rs76325618, and rs796106620 which, therefore, were excluded from our analysis.

Competing interests: The authors declare no competing interests.

Authors' contributions: S.G. and P.S. wrote the initial draft and analysed the data. A.K performed Sequenom analysis on discussion with K.M., P.M. and P.S.; K.M. and P.S. conceived the project. P.S. proofread the manuscript before all authors agreed to it.

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

Acknowledgements: 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 towards the support of Bioinformatics Infrastructure Facility at BISR, Jaipur.

References

1. Ng, SB, Buckingham, KJ, Lee, C, Bigham, AW, Tabor, HK, Dent, KM, Huff, CD, et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1), 30-5. doi: 10.1038/ng.499.
2. Shen, Y, Lopez, R & Milos, PM. (2014). The Rapidly Emerging Role for Whole Exome Sequencing in Clinical Genetics. *Curr. Genet. Med. Rep*, 2, 103. doi: 10.1007/s40142-014-0044-8.
3. Greenbaum, L, Pode-Shakked, B, Eisenberg-Barzilai, S, Dicastro-Keidar, M, Bar-Ziv, A, Goldstein, N, Reznik-Wolf, H, et al. (2019). Evaluation of Diagnostic Yield in Fetal Whole-Exome Sequencing: A Report on 45 Consecutive Families. *Front. Genet*, 10, 425. doi: 10.3389/fgene.2019.00425.
4. Mone, F, Quinlan-Jones, E & Kilby, MD. (2018). Clinical utility of exome sequencing in the prenatal diagnosis of congenital anomalies: A Review. *Eur J Obstet Gynecol Reprod Biol*, 231, 19-24 doi: 10.1016/j.ejogrb.2018.10.016.
5. Gahl, WA, Markello, TC, Toro, C, Fajardo, KF, Sincan, M, Gill, F, et al. (2012). The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet Med*, 14(1), 51-59. doi:10.1038/gim.0b013e318232a005.
6. Boycott, KM, Vanstone, MR, Bulman, DE & MacKenzie, AE. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet*, 14, 681-691.
7. Jacob, HJ, Abrams, K, Bick, DP, Brodie, K, Dimmock, DP, Farrell, M, et al. (2013). Genomics in clinical practice: lessons from the front lines. *Sci Transl Med*, 5, 194-195. doi: 10.1126/scitranslmed.3006468.
8. MacArthur, DG, Manolio, TA, Dimmock, DP, Rehm, HL, Shendure, J, Abecasis, GR, et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497), 469-476. doi:10.1038/nature13127

9. Zhu, X, Petrovski, S, Xie, P, Ruzzo, EK, Lu, YF, McSweeney, KM, et al. (2015). Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet Med*, 17(10), 774–781. doi:10.1038/gim.2014.191.
10. Rees, E, Kirov, G, Walters, JT, Richards, AL, Howrigan, D, Kavanagh, DH, et al. (2015). Analysis of exome sequence in 604 trios for recessive genotypes in schizophrenia. *Transl. Psychiatry*, 5(7), e607. doi:10.1038/tp.2015.99.
11. Jin, ZB, Wu, J, Huang, XF, Feng, CY, Cai, XB, Mao, JY, et al. (2017). Trio-based exome sequencing arrests de novo mutations in early-onset high myopia. *Proc. Natl. Acad. Sci. U.S.A.*, 114 (16), 4219–4224. doi:10.1073/pnas.1615970114.
12. Quinlan-Jones E, Lord J, Williams D, Hamilton S, Marton T, Eberhardt RY, et al. (2019). Molecular autopsy by trio exome sequencing (ES) and postmortem examination in fetuses and neonates with prenatally identified structural anomalies. *Genet Med*, 21(5):1065-1073. doi: 10.1038/s41436-018-0298-8.
13. Mathur, P, Prabhu, K & Jindal, D. (2002). Unusual presentations of pouch colon. *J Pediatr Surg*, 37:1351–3.
14. Gupta, DK & Sharma, S. (2007). Congenital pouch colon-Then and now. *J Indian Assoc Pediatr Surg*, 12, 5-12.
15. Mathur, P, Medicherla, KM, Chaudhary, S, Patel, M, Bagali, P & Suravajhala, P. (2018). Whole exome sequencing reveals rare variants linked to congenital pouch colon. *Sci. Rep*, 8, 6646. doi:10.1038/s41598-018-24967-y.
16. Gupta, S, Gupta, N, Tiwari, P, Menon, S, Mathur, P & Kothari, SL, et al. (2018). Inc-EPB41-Protein Interactions associated with Congenital Pouch Colon. *Biomolecules*, 8, 95. doi: 10.3390/biom8030095.
17. Sun, CJ, van Koningsbruggen, S, Long, SW, Straasheijm, K, Klooster, R, Jones, TI, et al. (2011). Facioscapulohumeral muscular dystrophy region gene 1 (FRG1) is a dynamic RNA-associated and actin bundling protein. *J Mol Biol*, 411(2): 397– 416. doi:10.1016/j.jmb.2011.06.014.
18. Amiri, M, Conserva, F, Panayiotou, C, Karlsson, A & Solaroli, N. (2013). The human adenylate kinase 9 is a nucleoside mono- and diphosphate kinase. *Int. J. Biochem. Cell Biol*, 45(5), 925-931 (2013).
19. Richards, S, Aziz, N, Bale, S, Bick, D, Das, S & Gastier-Foster, J, et al. (2015). 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. *Genet Med*, 17(5), 405-24.
20. Landrum, MJ, Lee, JM, Benson, M, Brown, G, Chao, C, Chitipiralla, S, et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*, 44(D1), D862-8. Available at: <https://www.ncbi.nlm.nih.gov/clinvar/> (Accessed: 27th September 2019).
 21. Sanz, L, Sanchez, P, Lallena, J, Diaz Meco MT & Moscat, J. (1999). The interaction of p62 with RIP links the atypical PKCs to NFκB activation. *EMBO J*, 18, 3044-3053.
 22. Seibenhener, ML, Du, Y, Diaz Meco, MT, Moscat, J, Wooten, MC & Wooten, MW. (2013). A role for sequestosome 1/p62 in mitochondrial dynamics, import and genome integrity. *Biochim Biophys Acta*, 1833, 452-459.
 23. Pankiv, S, Lamark, T, Bruun, JA, Overvatn, A, Bjorkoy, G & Johansen, T. (2010). Nucleocytoplasmic shuttling of p62/SQSTM1 and its role in recruitment of nuclear polyubiquitinated proteins to promyelocytic leukemia bodies. *J Biol Chem*, 285, 5941-5953.
 24. Mohamed, A, Ayman, A, Deniece, J, Wang, T, Kovach, C, Siddiqui, MT, et al. (2015). P62/ubiquitin IHC expression correlated with clinicopathologic parameters and outcome in gastrointestinal carcinomas. *Front Oncol*, 5, 70.
 25. Bao, L, Chandra, PK, Moroz, K, Zhang, X, Thung, SN, Wu, T, et al. (2014). Impaired autophagy response in human hepatocellular carcinoma. *Exp Mol Pathol*, 96, 149-154.
 26. Kitamura, H, Torigoe, T, Asanuma, H, Hisasue, SI, Suzuki, K, Tsukamoto, T, et al. (2006). Cytosolic overexpression of p62 sequestosome 1 in neoplastic prostate tissue. *Histopathology*, 48, 157- 161.
 27. Guo, YA, Chang, MM, Huang, W, Ooi, WF, Xing, M, Tan, P, et al. (2018). Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat Commun*, 9(1), 1520. doi: 10.1038/s41467-018-03828-2.
 28. Brahmachari, S, Singh, L, Sharma, A, Mukerji, M, Ray, K, Roychoudhury, S, et al. (2005). The Indian Genome Variation Consortium: The Indian Genome Variation database (IGVdb): a project overview. *Hum Genet*, 118, 1-11.
 29. Sherry, ST, Ward, MH, Kholodov, M, Baker, J, Phan, L, Smigielski, EM, et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*,

- 29(1), 308–311. doi:10.1093/nar/29.1.308. Available at: <https://www.ncbi.nlm.nih.gov/snp/>. (Accessed: 22nd September 2019).
30. Tate, JG, Bamford, S, Jubb, HC, Sondka, Z, Beare, DM, Bindal, N, et al. (2019). COSMIC: the Catalogue of Somatic Mutations in Cancer, *Nucleic Acids Res*, 47(D1), D941–D947. doi:10.1093/nar/gky1015. Available at: <https://cancer.sanger.ac.uk/cosmic>. (Accessed: 16th September 2019).
31. Chelala, C, Khan, A & Lemoine, NR. (2009). SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25(5), 655–661. doi:10.1093/bioinformatics/btn653. Available at: <https://www.snp-nexus.org/index.html>. (Accessed: 20th October 2019).
32. Ng, PC & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13), 3812–3814. doi:10.1093/nar/gkg509. Available at: <https://sift.bii.a-star.edu.sg/>. (Accessed: 12th October 2019).
33. Adzhubei, IA, Schmidt, S, Peshkin, L, Ramensky, VE, Gerasimova, A, Bork, P, et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. Available at: <http://genetics.bwh.harvard.edu/pph2/>. (Accessed: 21st September 2019).
34. McLaren, W, Gil, L, Hunt, SE, Riat, HS, Ritchie, GR, Thormann, A, et al. (2016). The Ensembl Variant Effect Predictor. *Genome biology*, 17(1), 122. doi:10.1186/s13059-016-0974-4. Available at: https://asia.ensembl.org/Homo_sapiens/Tools/VEP. (Accessed: 14th October 2019).
35. Schwarz, JM, Cooper, DN, Schuelke, M, Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*, 11(4):361–2. Available at: <http://www.mutationtaster.org/>. (Accessed: 19th September 2019).
36. Kircher, M, Witten, DM, Jain, P, O’Roak, BJ, Cooper, GM & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46 (3), 310–315. doi:10.1038/ng.2892.
37. Cooper, GM, Stone, EA, Asimenos, G, NISC Comparative Sequencing Program, Green, ED, Batzoglou, S, et al. (2005). Distribution and intensity of constraint

- in mammalian genomic sequence. *Genome Res*, 15(7), 901-13. doi: 10.1101/gr.3577405.
38. Huang, D, Sherman, BT, Lempicki, RA. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids Res*, 37(1), 1-13. doi:10.1093/nar/gkn923.
 39. Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, J M, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1), 25-29. doi:10.1038/75556.
 40. Chen, EY, Tan, CM, Kou, Y, Duan, Q, Wang, Z, Meirelles, GV, et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, [128\(14\)](#). Available at: <https://amp.pharm.mssm.edu/Enrichr/>. (Accessed: 21st September 2019).
 41. Maere, S, Heymans, K, Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks, *Bioinformatics*, 21(16), 3448-3449. Available at: <https://doi.org/10.1093/bioinformatics/bti551>. (Accessed: 19th September 2019).
 42. Shannon, P, Markiel, A, Ozier, O, Baliga, NS, Wang, JT, Ramage, D, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11), 2498-2504. doi:10.1101/gr.1239303.
 43. Trapnell, C, Roberts, A, Goff, L, Pertea, G, Kim, D, Kelley, DR, et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3), 562-78. doi: 10.1038/nprot.2012.016.
 44. Anders, S & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, 11, R106. doi: 10.1186/gb-2010-11-10-r106.
 45. The UVA FASTA Available at: https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml#. (Accessed: 1 November 2019).
 46. Fang, S, Zhang, L, Guo, J, et al. (2017). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res*, NONCODE. Available at: http://www.noncode.org/show_rna.php?

It is made available under a [CC-BY-ND 4.0 International license](#).

[id=NONHSAT002007&version=2&utd=1#](#). www.noncode.org (Accessed: 9 November 2019).

47. Gabriel, S, Ziaugra, L & Tabbaa, D. (2009). SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr Protoc Hum Genet*, 60(1), 2.12.1-2.12.18. doi: 10.1002/0471142905.hg0212s60.

Tables

Table 1: Total variants in cases versus controls							
	Probands (N=16)	Average variants for Proband	Controls(N=46)	Average variants for Control	Total Cases	% Probands	% Controls
Total variants	13450685	840667	3,57,51,971	777216.76	4,92,02,656		
SNPs	12404198	775262	3,38,93,726	736820.13	4,62,97,924	25.2104236	68.88596827
Indels	1046487	65405.4	29,04,835	63148.587	39,51,322	2.12689128	5.903817469
Missense	172232	10764.5	4,83,790	10517.174	6,56,022	0.350046144	0.983259928
Frameshift	10572	660.75	30183	656.15217	40755	0.021486645	0.061344249
Stop gain	2460	153.75	6778	147.34783	9238	0.00499973	0.013775679
Stop lost	469	29.3125	1280	27.826087	1749	0.000953201	0.002601486

Table 2: Variants inferred to be causal for CPC from trio-exome analysis													
Sno.	Variants	Variant_annotation	Location (hg38)	rsIDs	cDNA_position	AA_position	SIFT annotation	PolyPhen-2	Ensembl_VEP	MutationTaster	Phred_CADD	GERP_score	COSMIC
1	C7orf57	missense	48046564	rs10233232	c.644T>C	p.152M>T	tolerated(0.27)	benign(0.001)	MODERATE	disease causing	14.82	.	COSM4004495
2	C10orf120	missense	122697936	rs2947594	c.837C>T	p.269A>T	tolerated(1)	benign(0)	MODERATE	polymorphism	0.197	4.57	.
3	C7orf31	missense	25142293	rs2285738	c.1616T>C	p.300H>R	deleterious(0.04)	benign(0.031)	MODERATE	polymorphism	22.7	5.78	COSM7221306
4	C9orf84	missense	111700042	rs6477845	c.3031A>G	p.929L>P	tolerated(0.27)	benign(0)	MODERATE	polymorphism	7.633	6.08	COSM5618301
5	C11orf96	missense	43925417	rs1973717	c.76A>G	p.26T>A	tolerated_low_confidence(0.52)	unknown(0)	MODERATE	polymorphism	0.35	2.56	.
6	OR5AR1	stop_gain	56663740	rs11228710	c.79C>T	p.19Q>*	Deleterious	Probably Damaging	HIGH	polymorphism	35	5.33	COSM4145892
7	TAF1B	stop_gain	9904885	rs400917	c.902C>A	p.278Y>*	.	.	HIGH	polymorphism	40	5.67	.
8	MUC3A	frameshift	100960609	rs79874934	c.8921C>T	p.2944Q>*	.	.	HIGH	.	33	1.71	.
9	DLC1	frameshift	13499309	rs11203494	c.1138T>C	p.255N>D	tolerated_low_confidence(1)	benign(0)	MODERATE	polymorphism	6.304	4.98	.
10	AK9	frameshift/de novo	109528998..109528999	rs577355457	c.1130C>T	p.295V>F	.	.	HIGH	.	.	.	COSM1180171
11	FGFR4	frameshift	177090460	rs446382	c.378T>G	p.54R>R	.	.	LOW	.	9.481	.	COSM3761131
12	KRTAP4-1	frameshift	41184560	rs483929	c.783T>C	p.99M>V	.	.	.	.	.	.	COSM4130020
13	HAVCR1	frameshift	157055448	rs1809941	c.665G>T	p.44S>*	.	.	LOW	.	2.149	.	.
14	GBA3	stop_lost	22818881	rs358231	c.538A>C	p.149*->Y	.	.	HIGH	.	0.001	5.71	COSM7243285
15	SQSTM1	missense	179833099	rs55793208	c.1090G>C	p.190E>D	tolerated(0.53)	benign(0)	MODERATE	polymorphism	2.324	.	COSM5019342
16	SQSTM1	missense variant	179833153	rs4935	c.1144C>G	p.208D>E	tolerated(0.8)	benign(0)	MODERATE	polymorphism	0.013	4.79	COSM4159814
17	SQSTM1	3 Prime UTR Variant	179837704	rs199887787	c.2466G>C	.	.	.	MODIFIER	.	35	.	.
18	SQSTM1	3 Prime UTR Variant	179837731	rs10277	c.2493T>C	.	.	.	MODIFIER	.	0.016	.	.
19	SQSTM1	3 Prime UTR Variant	179837132	rs1060271	c.2130T>G	.	.	.	MODIFIER	.	5.181	.	.
20	BBS2	missense	56514589	rs4784677	c.371C>T	p.70S>N	tolerated(1)	benign(0)	MODERATE	polymorphism	20.8	5.54	.
21	PABPC1	missense	100709499	rs139094790	c.276G>A	p.93P>L	deleterious(0.05)	possibly_damaging(0.743)	MODERATE	disease causing	26.1	5.37	COSM4162560
22	HLA-DRB5	splice donor	32530124	rs79192142	c.100G>C	.	.	.	HIGH	.	18.53	.	COSM4003987
23	CTBP2	missense	124994502	rs61870306	c.1147G>C	p.249N>K	deleterious(0)	probably_damaging(0.99)	MODERATE	disease causing	27.2	5.21	COSM3930951
24	CTDSP2	missense	57823987	rs796106620	c.911G>Y	p.203L>I/V	deleterious	Probably Damaging	.	.	.	.	.
25	PABPC3	missense	25096815	rs74040928	c.948G>A	p.206R>H	tolerated(0.24)	benign(0.043)	MODERATE	disease causing	18.87	1.6	COSM4591481
26	MUC16	missense	8888842	rs77650227	c.40861A>G	p.13553Y>H	tolerated_low_confidence(0.36)	possibly_damaging(0.523)	MODERATE	polymorphism	8.253	3.48	COSM5957247
27	OR8U1	missense	56376308	rs76325618	c.685C>A	p.229H>N	tolerated(0.46)	benign(0.04)	MODERATE	polymorphism	2.357	5.78	COSM4145879
28	ZNF717	missense	75736909	rs76046078	c.2841A>G	p.898V>A	tolerated(1)	benign(0.009)	MODERATE	polymorphism	7.176	1.73	COSM4327424
29	MUC6	missense	1016800	rs199626069	c.6064G>A	p.2001P>S	tolerated_low_confidence(0.09)	benign(0.014)	MODERATE	polymorphism	11.6	2.76	COSM4593299
30	CDC27	missense	47143907	rs1141701	c.1198A>C	p.382F>L	tolerated(0.13)	possibly_damaging(0.548)	MODERATE	disease causing	23.3	5.59	COSM4130246
31	KIR2DL4	missense	54806001	rs1051454	c.424G>A	p.138A>T	tolerated(0.14)	benign(0.013)	MODERATE	polymorphism	0.008	1.05	COSM3835746
32	KCNJ12	missense	21416455	rs1612176	c.1483C>G	p.371S>R	tolerated(0.1)	possibly_damaging(0.734)	MODERATE	disease causing	19.55	5.65	COSM4600498
33	FRG1	missense/de novo	189957414	NA	c.449T>C	p.L150S	.	.	.	.	.	.	COSM1131141
34	NOTCH2NL	missense	146156641	rs8002	c.835G>A	p.158T>I	deleterious(0.02)	probably_damaging(0.96)	MODERATE	.	23.6	2.87	COSM4142513
35	ANKRD36	missense	97152524	rs139806872	c.1409G>T	p.395A>S	tolerated_low_confidence(0.16)	probably_damaging(0.91)	MODERATE	polymorphism	0.013	1.34	COSM3727828
36	IGSF3	missense	116603970	rs532709767	c.1278G>C	p.E426D	.	.	.	.	.	.	COSM4142207
37	OR9G1	missense	56700964	rs12421330	c.577G>A	p.193G>S	tolerated(0.05)	possibly_damaging(0.899)	MODERATE	polymorphism	1.877	4.52	COSM4591451
38	HLA-A	non coding	29944621	rs9260191	c.1326*>G	p.339G>*	.	.	LOW	.	11.68	.	COSM5019030
39	MUC5B	missense	1238987	rs79638064	c.3472C>A	p.1138D>E	tolerated(0.13)	benign(0.229)	MODERATE	polymorphism	1.119	3.78	COSM5957711
40	MST1L	non coding transcript exon variant	16758917	rs872964	c.1279C>T	p.P384P	.	.	MODIFIER	.	8.031	.	COSM3747587

* red ones are predicted as deleterious in all databases

Sno.	Position	Ref_Allele	Alt_Allele frequency	SNP_id
1	chr1:16758917	C	T (2.3170%)	rs872964
2	chr1:116603970	C	G (0.0019%)	rs532709767
3	chr1:146156641	G	T (0.5165%)	rs8002
4	chr2:9904885	C	T (82.6419%)	rs400917
5	chr2:97152524	G	T (0.0148%)	rs139806872
6	chr4:22818881	A	T (82.1699%)	rs358231
7	chr5:157055448	G	T (83.0161%)	rs1809941
8	chr5:177090460	T	G (70.9025%)	rs446382
9	chr5:179833099	G	C (1.7278%)	rs55793208
10	chr5:179833153	C	T (58.5686%)	rs4935
11	chr5:179837132	T	G (0.9383%)	rs1060271
12	chr5:179837704	G	C (0.1700%)	rs199887787
13	chr5:179837731	T	C (59.4508%)	rs10277
14	chr6:29944621	A	G (39.2973%)	rs9260191
15	chr6:32530124	C	T (0.1694%)	rs79192142
16	chr7:25142293	T	C (61.7639%)	rs2285738
17	chr7:48046564	T	C (83.3657%)	rs10233232
18	chr7:100960609	C	T (1.1563%)	rs79874934
19	chr8:13499309	T	C (9.5043%)	rs11203494
20	chr8:100709499	G	A (8.3104%)	rs139094790
21	chr9:111700042	A	G (77.7298%)	rs6477845
22	chr10:12269793	C	T (74.6677%)	rs2947594
23	chr10:12499450	G	C (0.0039%)	rs61870306
24	chr11:1016800	G	A (0.5667%)	rs199626069
25	chr11:1238987	C	A (0.1410%)	rs79638064
26	chr11:43925417	A	G (15.3193%)	rs1973717
27	chr11:56376308	C	T (0.0483%)	rs76325618
28	chr11:56663740	C	T (65.5350%)	rs11228710
29	chr11:56700964	G	T (0.2808%)	rs12421330
30	chr12:57823987	G	C (0.0006%)	rs796106620
31	chr13:25096815	G	A (0.2769%)	rs74040928
32	chr16:56514589	C	T (94.0485%)	rs4784677
33	chr17:21416455	C	T (23.5079%)	rs1612176
34	chr17:41184560	T	C (6.6212%)	rs483929
35	chr17:47143907	A	C (0.0090%)	rs1141701
36	chr19:8888842	A	G (0.1488%)	rs77650227
37	chr19:54806001	G	A (26.8694%)	rs1051454

It is made available under a [CC-BY-ND 4.0 International license](#).

Table 4: CNVs associated with variants identified from trio-exomes*											
SNP	name	Chr	chrom	Star	chrom	End	Type	Reference	Pubmed	Method	Sample Gain Loss
rs1141701	CDC27	chr17	46006547	47199967	gain+loss	Redon et al 2006	###	BAC aCGH, SNP array	270	171	43
rs79192142	HLA-DRB5	chr6	32403975	32737657	gain+loss	Redon et al 2006	###	BAC aCGH, SNP array	270	165	205
rs79192142	HLA-DRB5	chr6	32480351	32562509	gain+loss	Coe et al 2014	###	Oligo aCGH, SNP arra	29084	264	38
rs872964	MST1L	chr1	16437837	17157486	gain+loss	Redon et al 2006	###	BAC aCGH, SNP array	270	182	33
rs872964	MST1L	chr1	16487425	16935752	gain+loss	Vogler et al 2010	###	Merging, SNP array	1109	115	105
rs872964	MST1L	chr1	16684942	16949734	gain+loss	Coe et al 2014	###	Oligo aCGH, SNP arra	29084	163	38

*The Table is filtered based on the Gain+loss scores. Please unfilter to see the rest. However, the six candidate CNVs are considered the best based on gain+loss filters

Table 5: Variant Validation through Sequenom Mass Array plex

Rs id	nAllele	Coverage	NoCall	Total	COMMON	HET	RARE
rs199626069	2	100%	0	37	16	5	5
rs11228710	2	100%	0	37	16	11	11
rs77650227	2	100%	0	37	13	4	4
rs10233232	2	100%	0	37	11	9	9
rs358231	2	100%	0	37	15	7	7
rs55793208	2	100%	0	37	9	12	12
rs2947594	2	100%	0	37	5	11	11
rs139094790*	2	95%	2	37	12	4	4
rs199887787	2	100%	0	37	12	4	4
rs2285738	2	100%	0	37	9	12	12
rs1809941	2	100%	0	37	10	12	12
rs446382	2	100%	0	37	12	9	9
rs79638064	2	100%	0	37	7	10	10
rs139806872	2	100%	0	37	11	8	8
rs74040928	2	100%	0	37	16	5	5
rs1060271	2	100%	0	37	14	3	3
rs577355457	2	100%	0	37	25	12	12
rs11203494	2	100%	0	37	11	7	7
rs1973717	2	100%	0	37	12	4	4
rs532709767	2	100%	0	37	13	4	4
rs872964	2	100%	0	37	9	13	13
rs400917	2	100%	0	37	10	7	7
rs4784677	2	100%	0	37	9	10	10
rs79874934	2	100%	0	37	8	12	12
rs79192142	2	100%	0	37	5	13	13
rs61870306	2	100%	0	37	4	16	16
rs12421330	2	100%	0	37	13	4	4
rs1612176	2	100%	0	37	15	5	5
rs6477845	2	100%	0	37	12	12	12

*This SNP failed in validation

Table S: De novo gene set enrichment analyses

GO term	Count	Log2(OR)	Adjusted P-value	Count
pyrimidine-containing compound transmembrane transport (GO:0072531)	0.03495	0.05015	-5.67	2077
regulation of nuclease activity (GO:0032069)	0.003495	0.05015	-2.86	16.16
adenine nucleoside transport (GO:0005504)	0.003495	0.05015	-2.77	15.32
purine ribonucleotide transport (GO:0015868)	0.003994	0.05015	-3.08	17.01
vitamin transmembrane transport (GO:0035461)	0.00499	0.05015	-2.7	14.32
positive regulation of execution phase of apoptosis (GO:1900119)	0.005488	0.05015	-3.3	17.19
negative regulation of metabolic process (GO:0009892)	0.006482	0.05015	-1.67	8.4
negative regulation of oxidoreductase activity (GO:0051354)	0.006482	0.05015	-2.52	12.7
DNA-templated transcriptional preinitiation complex assembly (GO:0070897)	0.00698	0.05015	-1.87	9.28
regulation of execution phase of apoptosis (GO:1900117)	0.00698	0.05015	-2.52	12.52
regulation of oxidoreductase activity (GO:0051341)	0.00698	0.05015	-2.24	11.14
sensory perception of taste (GO:0050909)	0.00698	0.05015	-2.12	10.5
negative regulation of focal adhesion assembly (GO:0051895)	0.007476	0.05015	-2.66	13.04
negative regulation of Rho protein signal transduction (GO:0035024)	0.007476	0.05015	-2.58	12.63
regulation of secretion (GO:0051046)	0.007476	0.05015	-2.06	10.08
negative regulation of adherens junction organization (GO:1903392)	0.007973	0.05015	-2.57	12.43
RNA transcription (GO:0009303)	0.008469	0.05015	-1.91	9.09
negative regulation of stress fiber assembly (GO:0051497)	0.008966	0.05015	-2.22	10.44
cell-substrate adherens junction assembly (GO:0007045)	0.009462	0.05015	-2.18	10.15
focal adhesion assembly (GO:0048041)	0.009462	0.05015	-2.19	10.23
negative regulation of cell junction assembly (GO:1901889)	0.009957	0.05026	-2.36	10.87
negative regulation of actin filament bundle assembly (GO:0032232)	0.01045	0.05036	-1.99	9.07
positive regulation of dephosphorylation (GO:0035306)	0.01243	0.05111	-1.53	6.72
regulation of transmembrane transport (GO:0034762)	0.01243	0.05111	-1.95	8.57
negative regulation of catalytic activity (GO:0043086)	0.01342	0.05111	-1.52	6.54
negative regulation of cell-matrix adhesion (GO:0001953)	0.01392	0.05111	-1.73	7.39
acute inflammatory response (GO:0002526)	0.01441	0.05111	-1.63	6.92
primary neural tube formation (GO:0014020)	0.01441	0.05111	-2.03	8.59
transcription elongation from RNA polymerase I promoter (GO:0006362)	0.01441	0.05111	-2.13	9.05
tube closure (GO:0060606)	0.0149	0.05111	-2.29	9.62
positive regulation of protein dephosphorylation (GO:0035307)	0.01589	0.05111	-1.83	7.59
termination of RNA polymerase I transcription (GO:0006363)	0.01589	0.05111	-1.43	5.9
neural tube closure (GO:0001843)	0.01687	0.05111	-1.81	7.4
transcription from RNA polymerase I promoter (GO:0006360)	0.01687	0.05111	-1.65	6.73
transcription initiation from RNA polymerase I promoter (GO:0006361)	0.01687	0.05111	-1.7	6.93
regulation of ion transmembrane transport (GO:0034765)	0.01786	0.05258	-1.59	6.39
negative regulation of Ras protein signal transduction (GO:0046580)	0.01884	0.05398	-1.39	5.5
regulation of RNA metabolic process (GO:0051252)	0.01983	0.05492	-1.37	5.36
regulation of protein dephosphorylation (GO:0035304)	0.02032	0.05492	-1.94	7.55
sensory perception of bitter taste (GO:0050913)	0.02081	0.05492	-1.38	5.35
regulation of multicellular organismal development (GO:2000026)	0.02179	0.05492	-1.95	7.47
response to hydrogen peroxide (GO:0042542)	0.02179	0.05492	-1.59	6.08
regulation of focal adhesion assembly (GO:0051893)	0.02228	0.05492	-1.9	7.23
positive regulation of gene expression, epigenetic (GO:0045815)	0.02375	0.05687	-1.58	5.9
heart morphogenesis (GO:0003007)	0.02521	0.05687	-1.48	5.45
negative regulation of viral genome replication (GO:0045071)	0.02521	0.05687	-1.45	5.33
regulation of ion transport (GO:0043269)	0.02521	0.05687	-1.88	6.92
negative regulation of cellular catabolic process (GO:0031330)	0.02814	0.05944	-1.2	4.28
positive regulation of cell death (GO:0010942)	0.02863	0.05944	-1.98	7.02
regulation of Rho protein signal transduction (GO:0035023)	0.02863	0.05944	-2	7.09
regulation of stress fiber assembly (GO:0051492)	0.03009	0.05944	-1.28	4.47
response to reactive oxygen species (GO:0000302)	0.03009	0.05944	-1.26	4.4
negative regulation of viral life cycle (GO:1903901)	0.03058	0.05944	-1.39	4.86
regulation of actin filament-based process (GO:0032970)	0.03106	0.05944	-1.25	4.34
regulation of viral genome replication (GO:0045069)	0.03155	0.05944	-1.32	4.56
cellular response to type I interferon (GO:0071357)	0.03252	0.05944	-1.35	4.63
O-glycan processing (GO:0016266)	0.03252	0.05944	-1.67	5.71
type I interferon signaling pathway (GO:0060337)	0.03252	0.05944	-2.23	7.62
interferon-gamma-mediated signaling pathway (GO:0060333)	0.03495	0.06278	-1.28	4.28
regulation of gene expression, epigenetic (GO:0040029)	0.03688	0.06471	-1.34	4.41
regulation of actin cytoskeleton organization (GO:0032956)	0.03737	0.06471	-1.38	4.53
nucleic acid-templated transcription (GO:0097659)	0.03785	0.06471	-1.75	5.74
activation of cysteine-type endopeptidase activity involved in apoptotic pro	0.0393	0.06508	-1.33	4.3
ion transport (GO:0006812)	0.0393	0.06508	-1.25	4.06
cell-matrix adhesion (GO:0007160)	0.04459	0.07272	-1.44	4.47
regulation of cytoskeleton organization (GO:0051493)	0.04699	0.07474	-1.6	4.91
regulation of cell death (GO:0010941)	0.04747	0.07474	-1.31	4.01
negative regulation of cell motility (GO:2000146)	0.04794	0.07474	-1.2	3.64
DNA-templated transcription, termination (GO:0006353)	0.05224	0.08013	-1.44	4.26
DNA-templated transcription, elongation (GO:0006354)	0.05319	0.08013	-1.68	4.92
positive regulation of cysteine-type endopeptidase activity involved in apop	0.05367	0.08013	-1.34	3.93
protein O-linked glycosylation (GO:0006493)	0.05652	0.08276	-1.24	3.57
cellular response to interferon-gamma (GO:0071346)	0.057	0.08276	-1.88	5.39
negative regulation of cell migration (GO:0030336)	0.05937	0.08368	-1.22	3.45
stimulatory C-type lectin receptor signaling pathway (GO:0002223)	0.05937	0.08368	-1.29	3.64
innate immune response activating cell surface receptor signaling pathway	0.06078	0.08368	-1.4	3.92
organic anion transport (GO:0015711)	0.06078	0.08368	-1.12	3.13
regulation of small GTPase mediated signal transduction (GO:0051056)	0.06832	0.09284	-1.25	3.37
heart development (GO:0007507)	0.07253	0.09732	-1.73	4.54
positive regulation of protein modification process (GO:0031401)	0.07906	0.1047	-1.02	2.59
DNA-templated transcription, initiation (GO:0006352)	0.08968	0.1169	-1.28	3.08
receptor-mediated endocytosis (GO:0006898)	0.0906	0.1169	-1.32	3.18
protein homooligomerization (GO:0051260)	0.09152	0.1169	-1.41	3.38
protein oligomerization (GO:0051259)	0.1038	0.131	-1.63	3.68
apoptotic process (GO:0006915)	0.1102	0.1374	-1.9	4.2
sensory perception of smell (GO:0007608)	0.1146	0.1413	-1.57	3.4
inflammatory response (GO:0006954)	0.1196	0.1457	-1.67	3.54
endocytosis (GO:0006897)	0.1245	0.1499	-1.59	3.32
ion transport (GO:0006811)	0.1359	0.1619	-1.09	2.18
positive regulation of apoptotic process (GO:0043065)	0.1438	0.1693	-1.58	3.06
regulation of cell migration (GO:0030334)	0.1477	0.172	-1.19	2.27
sensory perception of chemical stimulus (GO:0007606)	0.1524	0.1756	-1.51	2.83
transcription, DNA-templated (GO:0006351)	0.1649	0.1879	-1.73	3.12
negative regulation of cell proliferation (GO:0008285)	0.1678	0.1892	-1.72	3.07
cellular macromolecule biosynthetic process (GO:0034645)	0.1695	0.1892	-1.63	2.9
transmembrane transport (GO:0055085)	0.1742	0.1923	-1.28	2.24
gene expression (GO:0010467)	0.188	0.2054	-1.17	1.96
regulation of intracellular signal transduction (GO:1902531)	0.1925	0.2082	-1.82	2.99
neutrophil degranulation (GO:0043312)	0.2157	0.2298	-1.85	2.85
neutrophil activation involved in immune response (GO:0002283)	0.2173	0.2298	-1.17	1.78
neutrophil mediated immunity (GO:0002446)	0.2189	0.2298	-1.79	2.72
positive regulation of cellular process (GO:0048522)	0.2316	0.2407	-1.26	1.84
negative regulation of cellular process (GO:0048523)	0.2375	0.2445	-1.29	1.85
cytokine-mediated signaling pathway (GO:0019221)	0.2755	0.2807	-1.25	1.61
regulation of cell proliferation (GO:0042127)	0.3145	0.3175	-1.05	1.21
positive regulation of gene expression (GO:0010628)	0.3255	0.3255	-1.52	1.71

Control

(7710 variants
on an average per
cluster)

13312

Proband

(4667 variants
on an average per
cluster)

750000 CWTs

400,000 PWTs
to be recovered

300,000 CWTs used
in CWTs

500,000
Total
recovered

750,000 CWTs