Published in final edited form as:
Hum Mutat. 2021 April ; 42(4): e15–e61. doi:10.1002/humu.24172.

A dataset of variants derived from 1455 clinical and research exomes is efficient in variant prioritization for early-onset monogenic disorders in Indians

Neethukrishna Kausthubham[1], **Anju Shukla**[1], **Neerja Gupta**[2], **Gandham SriLakshmi Bhavani**[1], **Samarth Kulshrestha**[3], **Aneek Das Bhowmik**[4,5], **Amita Moirangthem**[6], **Sunita Bijarnia-Mahay**[3], **Madhulika Kabra**[2], **Ratna Dua Puri**[3], **Kausik Mandal**[6], **Ishwar C Verma**[3], **Stephanie L Bielas**[7], **Shubha R Phadke**[6], **Ashwin Dalal**[4], **Katta M Girisha**[1]

[1]Department of Medical Genetics, Kasturba Medical College, Manipal, Manipal Academy of Higher Education, Manipal, India

[2]Division of Genetics, Department of Pediatrics, All India Institute of Medical Sciences, New Delhi, India

[3]Institute of Medical Genetics and Genomics, Sir Ganga Ram Hospital, New Delhi, India

[4]Diagnostics Division, Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India

[5]Diagnostics Lab (ASPIRE), CSIR-Centre for Cellular & Molecular Biology, CCMB Annexe II, Hyderabad- 500039, India

[6]Department of Medical Genetics, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, India

[7]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Abstract

Given the genomic uniqueness, a local dataset is most desired for Indians, who are underrepresented in existing public databases. We hypothesize patients with rare monogenic

Address for correspondence: Dr Girisha KM, Professor, Department of Medical Genetics, Kasturba Medical College, Manipal, Manipal Academy of Higher Education, Manipal- 576104, Tel No.: 91-820-2923149, girisha.katta@manipal.edu.

AUTHOR CONTRIBUTIONS
Conceptualization: N.K., A.S., S.L.B., K.M.G.; Data curation: N.K., S.K., A.D.B.; Formal Analysis: N.K., A.S., G.S.B., K.M.G.; Funding acquisition: A.S., N.G., M.K., R.D.P., S.L.B., S.R.P., A.D., K.M.G.; Investigation: N.K., A.S., N.G., A.M., S.B.M, M.K., R.D.P., K.M., I.C.V., S.R.P., A.D., K.M.G.: Methodology: N.K., S.L.B., K.M.G.; Resources: A.S., N.G., G.S.B., S.K., A.D.B., A.M., S.B.M, M.K., R.D.P., K.M., I.C.V., S.R.P., A.D., K.M.G.; Supervision: K.M.G.; Writing-original draft: N.K., K.M.G.; Writing-review & editing: N.K., A.S., N.G., G.S.B., S.K., A.D.B., A.M., S.B.M, M.K., R.D.P., K.M., I.C.V., S.R.P., A.D., K.M.G.

Data Repository Information:
http://cdfd.org.in/labpages/diag_datasets.html

CONFLICT OF INTEREST
The authors declare no conflict of interest.

WEB RESOURCES
https://gnomad.broadinstitute.org/
http://simple-clinvar.broadinstitute.org/
https://www.ncbi.nlm.nih.gov/clinvar/
https://browser.genomeasia100k.org/
https://omim.org/

disorders and their family members can provide a reliable source of common variants in the population. Exome sequencing (ES) data from families with rare Mendelian disorders was aggregated from five centers in India. The dataset was refined by excluding related individuals and removing the disease-causing variants (refined cohort). The efficiency of these datasets was assessed in a new set of 50 exomes against gnomAD and GenomeAsia. Our original cohort comprised 1455 individuals from 1207 families. The refined cohort had 836 unrelated individuals that retained 1,251,064 variants with 181125 population specific and 489618 common variants. The allele frequencies from our cohort helped to define 97609 rare variants in gnomAD and 44520 rare variants in GenomeAsia as common variants in our population. Our variant dataset provided additional 1.7% and 0.1% efficiency for prioritizing heterozygous and homozygous variants respectively for rare monogenic disorders. We observed additional 19 genes/human knockouts. We list carrier frequency for 142 recessive disorders. This is a large and useful resource of exonic variants for Indians. Despite limitations, datasets from patients are efficient tools for variant prioritization in a resource limited setting.

Keywords

Variant dataset; Exomes; Indian population; Monogenic disorders

1 INTRODUCTION

The datasets of human genomic variants are becoming increasingly indispensable for genomic medicine. Several large-scale global efforts to sequence individuals of diverse backgrounds have provided a wide range of genomic variations (Summarized in Table 1). The 1000 Genomes (Auton et al., 2015) followed by National Heart, Lung and Blood Institute (NHLBI) sponsored Exome Sequencing Project (ESP) (Fu et al., 2013), ExAC (Monkol Lek et al., 2016) and gnomAD projects (KarczewskiFrancioli, et al., 2020) have added substantially to our understanding of common and rare allelic variations among multiple populations. These large-scale variant databases have demonstrated their utility in deciphering variant pathogenicity, novel disease-gene associations, gene essentiality, drug discovery among several others (M. Lek et al., 2016; Minikel et al., 2020).

In spite of these enormous efforts, several populations remain underrepresented (Popejoy & Fullerton, 2016; Sirugo, Williams, & Tishkoff, 2019). The global and local projects have revealed that rare variants are more likely to be population specific (Auton et al., 2015; KarczewskiFrancioli, et al., 2020). This has led to several regional and population specific efforts (Ameur et al., 2017; Fattahi et al., 2019; John et al., 2018; Le et al., 2019; Lee et al., 2017) (see table 1). The utility of these datasets is immense in exploring the allele frequencies and carrier status for monogenic disorders in the local populations. Even the disease-causing variations are known to be population specific for common and rare diseases (Sirugo et al., 2019). Utility of population specific datasets extend beyond the locals to migrants and other populations.

2 DATA SPECIFICATIONS

Data type	Table, text file, figure
Data acquisition method	Exome sequencing
Data format	Filtered and analyzed
Experimental factors	None
Experimental features	Aggregation of exome sequencing data from families suspected to have rare Mendelian disorders from five different centers across India was performed, followed by data processing and variant calling using in-house pipeline. Curation of this aggregated data was performed by discounting the disease-causing variants and related individuals to create a reference variant dataset. The efficiency of these variant datasets for variant filtering for rare Mendelian disorders was then assessed.
Data source and location	KMC: Kasturba Medical College, Manipal, India SGPGIMS: Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, India CDFD: Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India AIIMS: All India Institute of Medical Sciences, New Delhi, India SGRH: Sir Gangaram Hospital, New Delhi, India
Data accessibility	Variant datasets are available for download from the following link. http://cdfd.org.in/labpages/diag_datasets.html

3 IMPACT OF THE DATA

Indian population is highly diverse and heterogenous (Chaubey, Metspalu, Kivisild, & Villems, 2007; “The Indian Genome Variation database (IGVdb): a project overview,” 2005; Xing et al., 2010). The current representation of the Indians in the available variant databases or datasets is illustrated in table 2. We would like to emphasize the term ‘Asians’ is a geographic descriptor for ethnically diverse Chinese, Indian and South-Eastern population and often the second most populous country is not included in the studies on ‘Asians’. IndiGenomes, (Jain et al., 2020) a recent addition and Singapore Genome Project (Wu et al., 2019) are the currently available large datasets for Indians. These datasets are far from capturing the complete spectrum of genomic variation of the Indian population. The burden of genetic disorders, though not systematically determined, is likely to be enormous due to huge population, complex population architecture, consanguinity and endogamy (Aggarwal & Phadke, 2015; Verma & Bijarnia, 2002). The wider availability of genomic tools like exome sequencing has led to a considerable increment in the number of individuals receiving a genetic diagnosis recently. Genomic variations from a healthy population is ideal to establish a reference variant catalogue of any of the population. As disease causing variants are expected to occur at a very low frequency, we hypothesized that patients with rare monogenic disorders would still provide a source of common variants in a given population. It is expensive to conduct such studies in healthy individuals and otherwise would deprive them of advances in genomic healthcare (incorrect assessment of genetic variants/risk factors and response to drugs). In this study, we delineate the genomic variants obtained from a cohort of families undergoing exome sequencing and assess the population based genomic variability. We also demonstrate the clinical implications of these variants. Finally, we illustrate the utility of such a local cohort in clinical medicine and research.

4 MATERIALS AND METHODS

4.1 Design of the study and selection of subjects

The dataset is collated from five centres across India which perform exome sequencing for rare genetic disorders on clinical and research bases viz., Kasturba Medical College (KMC), Manipal, Sanjay Gandhi Postgraduate Institute of Medical Sciences (SGPGIMS), Lucknow, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad, All India Institute of Medical Sciences (AIIMS), New Delhi and Sir Gangaram Hospital (SGRH), New Delhi. The study subjects include probands with or without their family members and couples undergoing exome sequencing for carrier screening. This data is aggregated over a period of three years from January 2017 to December 2019 from the above centres. For the purpose of this study, we use the following descriptions: singleton - only proband, duo - proband and a similarly affected sibling, trio - proband and unaffected parents, carrier testing-unaffected couple carrier for a possible autosomal recessive condition and others - if multiple members were tested and does not fit into any of the previous categories. The aggregated exomes were processed through a single in-house pipeline and the data was anonymised for the purpose of this study. The overall workflow of our study is illustrated in Figure 1.

4.2 Exome sequencing and data analysis

Next generation sequencing was performed at respective centers or outsourced to a service provider. The particulars of capture kits and sequencing platforms at respective centers are elaborated in Table 3. The overall quality control of the raw reads was performed using FastQC toolkit(Andrews, 2010) followed by alignment of paired-end reads of 100-150 bp to the human reference genome (GRCh37) using BWA-MEM (v0.7.15)(Li, 2013). Sorting and indexing of resulting alignment were done using Picard (v.2.5.0)(Picard). The alignments were then post-processed based on Genome Analysis Toolkit (GATK v3.6) (Van der Auwera et al., 2013) best practices for germline SNV and INDEL discovery from exome sequences. Realignment was performed around known Indels and SNVs using GATK RealignerTargetCreator and IndelRealigner followed by base quality score recalibration using GATK BaseRecalibrator. For each sample, the genomic VCF (gvcf) file was generated using GATK HaplotypeCaller with the appropriate exome capture kit bed file. Joint genotyping was performed for the entire cohort, followed by GATK variant quality recalibration (VQSR) and left normalization using BCFTOOLS (v1.3.1)(Li, 2011) to generate a multi-sample VCF file. Allele state (counts of heterozygotes and homozygotes), were derived using customized Perl scripts. Variants that were below the quality trenches and a call rate of <8% were filtered out during downstream analysis. Pairwise kinship coefficient of aggregated samples were calculated using KING (Manichaikul et al., 2010) and the recommended cut-off of 0.34 was considered as described in the KING manual for identifying duplicate samples or monozygotic twins.

4.3 Variant annotation

Multi-sample VCF files generated for the original and refined cohorts were then annotated using ANNOVAR(Wang, Li, & Hakonarson, 2010) against RefGene, gnomad_exome, gnomad_genome, snp138, clinvar_20190305, exac03 and avsnp150. ANNOVAR –xref argument was used to integrate gene-based cross-reference annotations which included various intolerance scores and tissue-specific expressions and counts of homozygous loss of function variants observed in gnomAD. Counts of heterozygotes and homozygotes from the above cohorts and proportion expressed across transcripts (pext)(Cummings et al., 2020) were integrated using ANNOVAR 'genericdbfile'. Downstream to ANNOVAR annotations, in-house Perl scripts were used to integrate disease phenotypes catalogued in OMIM(Online Mendelian Inheritance in)

4.4 Editorial policies and ethical considerations

Ethics committee approvals and informed consents for exome sequencing were obtained at each participating institution for their respective research projects (Kasturba Medical College & Kasturba Hospital Institutional Ethics Committee, Manipal; Ethics Committee, Sir Ganga Ram Hospital, New Delhi; Institutional Ethics Committee, All India Institute of Medical Sciences, New Delhi; Institutional Ethics Committee, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow; Institutional Ethics Committee, Centre for DNA Fingerprinting and Diagnostics, Hyderabad). The entire study was approved by Kasturba Medical College & Kasturba Hospital Institutional Ethics Committee.

4.5 Curation of dataset

The dataset is obtained from a cohort of individuals with genetic disorders and their family members. Hence, it was curated by discounting the disease-causing genomic variants in probands and relatedness amongst the individuals. For singletons and duo, the exome data of one individual was included after removing the disease-causing variants. Customized Perl script and bash commands were used to remove the disease-causing variants from Genomic Variant Call format (GVCF) files. Only one individual was included in this cohort amongst families undergoing duo sequencing. Two individuals (both parents) were selected from couples undergoing carrier-testing (they are considered healthy individuals) and parents in 'trio' exomes. For others, only unaffected and unrelated individuals in the families were selected. Further, if the pairwise kinship coefficient amongst family members was ≥0.0884, only one of them was considered for further analysis. Data of undiagnosed probands were not included in the cohort. We refer to this dataset as 'refined' cohort. New allele frequencies were derived by multi-sample joint genotyping using GATK v3.6 (allele frequencies as well as allele state).

4.6 Delineation of genomic variants

The variants called from both the cohorts were annotated using ANNOVAR (Wang, Li, & Hakonarson, 2010) (detailed annotation protocol is provided in the Supplementary Materials and Methods). Variant profiles were generated using BCFTOOLS (Li, 2011) (v1.3.1), in-house Perl scripts and multiple BASH and AWK commands. Hard filtering was not performed on variant datasets as to avoid missing the disease-causing variants which might reside in poorly covered regions. Variants with an allele frequency (AF) of 1% or more were classified as 'common', lesser than 1% as 'rare' and lesser than 0.01% as 'very rare'. Variants that were absent in gnomAD, AVSNP-150 build and SNP-138 build were termed 'population specific'. The variants obtained from ANNOVAR were converted to HGVS nomenclature using VariantValidator (Freeman, Hart, Gretton, Brookes, & Dalgleish, 2018).

4.7 Comparison with other genomic datasets

Chromosome-wise VCF files from gnomAD (v.2.1) (Karczewski, Francioli, et al., 2020) and GenomeAsia (Wall et al., 2019) were downloaded and concatenated. Left normalization was done using BCFTOOLS (v1.3.1)(Li, 2011) and the data was converted into ANNOVAR (Wang et al., 2010) format using convert2annovar. Customised Perl script was used to extract allele counts and the number of homozygotes from left normalized VCFs. The extracted data was then integrated with annotation protocol and the allele states of variants derived from our cohort was then compared with both the datasets using multiple BASH and AWK commands. Additionally, counts of homozygous loss of function (HLoF) variants observed in gnomAD data were generated using Perl scripts.

4.8 Clinical significance of the variants in our cohort

Human knockouts (non-essential genes) and presence of reported disease-causing variants were investigated in the refined cohort. To delineate human knockouts exclusive to our cohort, predicted deleterious homozygous single nucleotide variants (SNVs) (stop gain, frameshift and canonical splice site) were sought from the refined. Variants which are already observed in homozygous state in gnomAD or GenomeAsia were excluded from further analysis. Variants with genotype quality >60, supporting read numbers ≥10 and visualization in Integrative Genomics Viewer (IGV v.2.8)(Narang et al., 2010) were used to confirm the true calls. Also, variants located in the last exons of genes were excluded.

We obtained a subset of 887 genes with pediatric relevance and actionable adult-onset conditions with high penetrance with definitive and strong gene-disease association from the curated catalogue of genes for reporting results of newborn genomic sequencing (Ceyhan-Birsoy et al., 2017). This set was used to retrieve the reported disease-causing variants in the refined cohort. Clinvar_20190305 annotations were integrated using ANNOVAR and pathogenic and likely pathogenic variants catalogued in ClinVar (Landrum et al., 2014) were inspected. Variants reported with ‘no assertion criteria’ and ‘conflicting interpretations of pathogenicity’ were excluded.

Carrier frequencies for ‘pathogenic’ (P) and ‘likely pathogenic’ (LP) variants catalogued in ClinVar were calculated in the original cohort for a subset of 628 genes associated with recessive monogenic disorders from the above 887 genes (Ceyhan-Birsoy et al., 2017).

4.9 Assessment of efficiency of our dataset for variant prioritization

We selected sequencing data of 50 additional singleton exomes that are not part of the original cohort to assess the efficiency of our refined dataset for variant prioritization for rare monogenic disorders. All 50 exomes were captured using Agilent’s SureSelect CREv2 capture kit and sequencing (PE 2X150) was performed based on NovaSeq6000 (Illumina Inc. USA). Annotation of these exome sequencing data was done by integrating allele frequencies and allele states from our datasets in addition to those from gnomAD and GenomeAsia with ANNOVAR. AWK commands were used in this process. The statistical significance of efficiency of filtering for variants with AF<1% was evaluated using R(Team) packages. Shapiro-Wilk test (Royston, 1982) was performed to examine the normality of the data points and Wilcoxon signed-rank test (Forrester & Ury, 1969) was performed to compare the differences between the efficiencies of filtering approaches.

5 DATA

We aggregated 1455 individuals of Indian origin from 1207 families recruited at five different centres in India (Table 3 and Figure 1a). This cohort comprised of individuals with rare monogenic disorders (n=1207, 83%) and their unaffected family members (n=248, 17%). Majority of the patients had a neurocognitive disorder (n=409, 34%) or a skeletal dysplasia (n=267, 22%). Though the cohort comprised individuals of all ages (fetus to 80 years), majority (90.1%) of them belong to pediatric age group (<18 years). Males were higher in the cohort (M/F=1.21). Consanguinity was present in 26.7% while 42.06% families were non-consanguineous and data was unavailable for 31.25% families in the cohort. Most of the families underwent singleton exome sequencing (n=1047, 87%). Diagnostic yield across the cohort was 61% (735/1207 families). Figure 2 and Table 4 provide a detailed demographic summary of the cohort.

5.1 Refined cohort

The refined cohort consisted of 836 unrelated individuals. This dataset consisted of 203 unrelated and apparently healthy individuals and 633 unrelated probands from whom 736 disease causing variants were excluded. These variants included 264 pathogenic (36%), 235 likely pathogenic (32%) and 224 variants of unknown significance (30%)(Richards et al., 2015), but interpreted to explain the disease in the family. Thirteen variants were observed in genes of unknown significance (GUS, n=13, 2%) were also excluded. These GUS are either published or under different experimental stages.

5.2 Spectrum of genomic variants

The total number of variants in the original and refined cohorts were 1844228 and 1449306 respectively. Transitions occurred more frequently than transversions (ratio of 2.2). Application of quality filters (VQSR and variants with a call rate of more than 8%) yielded a subset of 1646560 (SNVs: 92.7%, INDELS: 7.3%) and 1251064 (SNVs: 93.3%, INDELS: 6.7%) variants from original and refined cohort respectively for further downstream analysis (Figure 3a and Table 5). Majority of these variants were observed to be rare with AF <1% (67.2% in original and 60.9% in refined cohort) and more than half of these rare variants were observed with AF ranging from 1 to 0.1% (Figure 3b) and rest of the ~45% of variants were very rare (AF<0.1%). Nearly 42% of the variants were observed only in single individuals. We observed 295,194 (18%) and 181125 (14.6%) population specific variants in our original and refined cohort respectively (Figures 3c and 3d). Seven percent of population specific variants were common and majority (93%) of these variants were rare in our cohort.

In the original cohort, 47% of variants were predicted to reside in the intronic region, 30% of them were exonic and 6% of the variants were found in exon-intron boundaries (Table 6). The distribution of exonic variants included nonsynonymous (16.9%), synonymous (11.2%) and loss-of-function (1.2%) as predicted by their functional impact. Nearly 91% of the loss of function variants were found to be rare in original cohort (Table 7).

Majority of the variants in the original cohort were observed only in heterozygous state (66.1%), 3.8% of the variants were exclusively seen in homozygous state (hemizygous state is included along with homozygous state) and 30.1% of the variants were observed in homozygous as well as in heterozygous states. Comparison of proportion of homozygous variants observed in the cohort against other variant datasets is outlined in Table 8.

We then analysed gnomAD and GenomeAsia for the overlapping alleles. The original cohort consisted of 1242315 (75.4%) variants already catalogued in gnomAD and 770836 (46.8%) variants in GenomeAsia. Among these, 97609 (7.8%) and 44520 (5.7%) variants were found to be rare in gnomAD and GenomeAsia respectively but were common in our cohort, enabling us to classify them as common variants in Indians. Among the shared variants, 704243 (56.6%) variants were found to be rare in our cohort and gnomAD, which increased our confidence of calling these rare variants. Likewise, 368185 (47.7%) variants were observed to be rare in our cohort and GenomeAsia. A similar trend was observed for refined cohort where 9.3% and 6.8% of the shared variants with gnomAD and GenomeAsia were found to be common in our cohort but were rare in these datasets.

5.3 Homozygous loss of function variants and human knockouts

We noted 778 homozygous loss of function variants (homozygous LoF) in 686 genes in the refined cohort. 82% (638/778) homozygous LoF were found in 567 genes which are not yet associated with a human monogenic disease. Among these, 24.9% (159/638) of the variants in 150 genes were unique to our cohort and absent in gnomAD and GenomeAsia. Ninety-two of these were high-quality loss of function variants in 89 genes with at least 10 supporting reads and genotype quality of 60. Seventy-three of these genes were earlier reported to have other homozygous LoF variants in gnomAD. Hence our work enlists additional 19 genes/human knockouts for which a homozygous LoF variant has not been documented in gnomAD (Table 9). However, one of these genes, ADGRF1 (NM_025048.3:c.157C>T, NP_079324.2:p.(Gln53Ter) is listed in GenomeAsia.

We also observed 140 homozygous LoF variants in 122 genes with phenotypic descriptions catalogued in OMIM. Among these, 70 variants were observed in 62 genes that are known to be associated with recessive disorders and 51 of these variants were already noted in gnomAD or GenomeAsia. After applying quality control, only six homozygous LoF variants due to SNVs in genes with phenotypic descriptions catalogued in OMIM were noted in the refined cohort (Table 10).

5.4 Known pathogenic/Likely pathogenic variants

Two hundred and sixteen reported pathogenic or likely pathogenic variants in ClinVar were observed in the refined cohort. We narrowed down the list by considering disease mechanisms, mode of inheritance and allelic state to 13 pathogenic and likely pathogenic variants (Table 11).

5.5 Carrier frequencies for recessive monogenic disorders

Table 12 summarizes the carrier status for recessive disorders (autosomal and X-chromosome) in our original cohort. We list the diseases with at least 10 carriers in table 13.

We observed carrier status for 288 pathogenic and likely pathogenic variants in 161 genes associated with 142 recessive disorders (628 genes were queried). Beta-thalassemia, GJB2 related deafness, Pendred syndrome, cystic fibrosis and Joubert syndrome appeared to have more carriers in our population.

5.6 Utility of the dataset for variant prioritization for monogenic disorders

Assessment of the efficiency of variant filtering for monogenic disorders based on different combination of filters is outlined in Table 14 and demonstrated in Figure 5. The application of allele frequency and homozygous counts from our cohort and those obtained from Gnomad and GenomeAsia led to filtering of an additional 50% homozygous variants (0.2% vs 0.1%) and 37.8% heterozygous variants. The observed differences in the filtering efficiencies for prioritization of heterozygous and homozygous variants were found to be statistically significant (p-value <0.05) based on Wilcoxon signed-rank test (Table 15).

6 DISCUSSION

Indian population is immensely heterogeneous and information on its population structure, variant distribution and their clinical significance is very limited (Indian Genome Variation, 2008; Reich, Thangaraj, Patterson, Price, & Singh, 2009; Xing et al., 2010). In this effort, genomic variants were gathered from a cohort of 1455 individuals of Indian ethnicity. We observed 1.65 million variants with 24.6% new variants that were absent in gnomAD. As this cohort originated from families with suspected monogenic disorders, we derived a refined cohort of unrelated individuals and excluded disease causing variants to make the dataset to represent ‘apparently healthy 836 Indians’. We stress the huge under-representation of Indians in various currently available datasets (Table 2) and this effort has put together the one of the large representation of Indians available till date. This dataset is likely to be useful for genomic healthcare for Indians and Indians living in other countries. The data on allele frequency, gene essentiality and carrier frequency are likely to have wider implication for other populations.

The overall proportion of homozygous variants in our cohort is high (34%). The proportion of homozygous variants in GnomAD is 4.1% and 10.19% in exome and genome data respectively whereas 18.6% of the variants from GenomeAsia and 20% of the variants from Kuwaiti exomes are in homozygous state. A higher proportion of homozygous variants in our cohort indicates the higher rates of consanguinity and inbreeding among Asian Indians (Table 8) as compared to other populations such as African/African American, Latino and Non-Finnish European (Bittles & Black, 2010; Karczewski, Francioli, et al., 2020). Higher number of homozygous variants and increased burden of runs of homozygosity reported by Iranian and Kuwaiti population with higher inbreeding levels are in-line with this observation (Fattahi et al., 2019; John et al., 2018). We have also noted remote inbreeding and higher proportion of disease-causing variants in our earlier studies (Bhavani et al., 2015; Bidchol et al., 2014; K. M. Girisha et al., 2019; Shukla et al., 2018).

Nearly 18% (295194/1646560) of the variants observed in the original cohort were population specific. These variants when added to the existing pool of human genomic variation catalogues, increase the diversity and improve the representation of Indian population. Our dataset catalogued 7% of the population specific as common variants. However, a larger portion (93%) of these variants were noted to be rare due to small cohort size, indicating the need for large-scale sequencing efforts in Indians. Additionally, this dataset helps to redefine 7.8% and 9.3% of the rare variants observed in gnomAD as common variants and 5.7% and 6.8% of the rare variants observed in GenomeAsia as common variants based on the observed allele frequency in original and refined cohort respectively. Overall, the differences in allele frequencies were consistent with the previously conducted ethnic specific studies and highlights the lack of representation of these ethnicities in the available large-scale datasets (Fattahi et al., 2019; John et al., 2018).

Ethnic specific datasets have provided several insights into the clinical significance of genomic variants (Fattahi et al., 2019; John et al., 2018; Le et al., 2019). We evaluated this in terms of presence of human knockouts, reported disease-causing variants and carrier frequencies of recessive monogenic disorders in our cohort. The availability of large-scale datasets including population specific datasets are contributing to identification of biallelic LoFs or human knockouts (Alkuraya, 2015; M. Lek et al., 2016; Sulem et al., 2015; Wall et al., 2019). These variants in known disease-causing genes have often led to recognition of distinct but different phenotype from those reported for other class of variants (Shamseldin et al., 2015). Also, knockouts in healthy individuals in, previously reported disease-causing genes can raise questions against the reported disease mechanism (Alsalem, Halees, Anazi, Alshamekh, & Alkuraya, 2013). Further, these variants in genes not known to cause any human disease can add to the existing knowledge of non-essential genes (Monkol Lek et al., 2016). We list 19 novel HLoFs/human knockouts in genes that are not yet known to be associated with human disease. The truncating variants identified in *ETV7*, *HOPX* and *FOXM1* were observed in unaffected family members and the rest of the variants were observed in affected individuals with an identified genetic cause. The importance of HLoF variants in *FOXM1* and *HOPX* are uncertain as they are present in those exons which are barely expressed in most human tissues. Also, HLoF variants in *SCYL2* and *FOXM1* with high pLi scores seem to suggest that more ethnic specific sequencing may redefine the existing catalogue of essential genes. Details of gene, variant and genetic diagnosis of the individuals are given in Table 9. As most of the HLoF variants are observed in affected individuals, we advise cautious approach until further evidences are available as blended phenotypes or dual diagnoses are possible in our subjects. We also would like to consider absence of phenotype due to late onset diseases and variable expression in these 19 individuals.

Additionally, six HLoF variants were observed in known disease-causing genes in the refined cohort (Table 10). However, truncating variants observed in *ALMS1* and *PKD1L1* were found to be multi-nucleotide variants leading to possible rescue of these HLoFs. In one family with *UMOD* truncating variant, we could not rule out the possibility of a blended phenotype of *UMOD* related kidney disease and osteogenesis imperfecta, as the child was aged one year and could not be assessed for renal phenotype. A HLoF in *PRPH* known as a susceptibility gene for amyotrophic lateral sclerosis was observed in a one-year-old. The clinical implication of this variant would be difficult to interpret (Ahmeti et al., 2013). The HLoF noted in *MOCOS* can be explained by later age of onset and report of several asymptomatic individuals with xanthinuria, type II (Akıncı, Çakıl, & Öner, 2013).

In one family with HLoF in *PLA2G6*, possibility of blended phenotype of *PLA2G6* related neurodegeneration and Omenn syndrome could not be ruled out as the proband succumbed at the age of 4 months.

Thirteen reported pathogenic/likely pathogenic variants in *GJB2, TSC2, G6PD, BRCA1, TTR, F11, GLA, PKLR, MYH7, LDLR, DMD* were observed in the refined cohort (Table 11). Pathogenic variants in *BRCA1* and *TTR* that are exclusively adult-onset diseases are expected to be observed in a predominantly pediatric cohort like ours. The age of individuals with these variants were five months and five years respectively. Variable severity and age of onset is reported for *GJB2* related palmoplantar keratoderma with deafness whereas variable severity and reduced penetrance are known for tuberous sclerosis-2 and G6PD related anemia. Pathogenic variants in *F11* are usually known to result in excessive bleeding only after surgery and may go unrecognised. Significant phenotypic heterogeneity, later onset disease forms and asymptomatic individuals are reported for Fabry disease (Eng & Desnick, 1994) and the unaffected family member observed with the disease-causing variant in *GLA* may be exhibiting the same phenomena. Three of the variants, observed in *PKLR, MYH7 and LDLR* were earlier classified as pathogenic or likely pathogenic (P/LP). However, the current version of ClinVar (accessed on 18-08-2020) has re-classified these as variants with conflicting interpretation of the pathogenicity. The pathogenic variant in *DMD* too was reclassified as benign. Eleven of these variants were also observed in gnomAD in corresponding allele states. Hence, these findings provide further evidence for non-penetrance and clinical variability in these conditions and highlight the significance of updating the resources utilized for variant interpretation periodically for possibility of re-classification of these variants.

High burden of monogenic disorders is well described in the Indian population (Kaur & Singh, 2010; Sachdeva et al., 2012; Singh et al., 2010; Sivasubbu & Scaria, 2019; Venugopal et al., 2018). However, the incidence and prevalence and consequently the carrier frequencies for most disorders remain unknown. Prenatal and/or pre-conceptional expanded carrier screening as in several other nations is not yet practiced widely in India. Carrier screening was mostly carried out in couples with a previous history of putative recessive disease in the deceased offspring and non-availability of samples from the proband/s. Globally, thalassemia and structural hemoglobinopathies are the commonest monogenic disorders and India too has a huge burden of these conditions (Colah, Italia, & Gorakshakar, 2017; Sivasubbu & Scaria, 2019; Williams & Weatherall, 2012). Patients with β thalassemia and sickle cell disease in India are estimated to be 100,000 and 150,000 respectively and the reported average prevalence of carriers for β thalassemia is 3-4% (Colah et al., 2017) whereas sickle cell disease was observed with a carrier frequency ranging from ~1 to 40% in specific subpopulations in India (Hockham et al., 2018). Highest number of carriers (n=44, 3.02%) were observed for beta-thalassemia in our study population. Among these, NM_000518.5:c.92+5G>C, one of the most common pathogenic variant in HBB in India was observed with highest frequency (n=24) in our cohort (Grow, Vashist, Abrol, Sharma, & Yadav, 2014). Among the carriers observed for GJB2, NM_004004.5:c.71G>A was the most commonly observed pathogenic variant (31 carriers). GnomAD has reported 151 and 134 allele counts for the corresponding disease-causing variants in *HBB* and *GJB2* respectively and interestingly more than 90% of these variants were from South

Asian populations. GenomeAsia has reported 22 carriers for the *HBB* variant and more than 90% of the carriers were South Asians. Similarly, among the 13 carriers observed in GenomeAsia for the *GJB2* variant, 85% were South Asians. Cystic fibrosis is reported to be the commonest recessive disorder in Caucasian population with an observed carrier rate of 3.7% (Goldstein & Prystowsky, 2017; Zvereff, Faruki, Edwards, & Friedman, 2014). It is also found to be more common in Indian population with a carrier rate of 0.4% for one of the most common variant NP_000483.3:p.(Phe508del) in *CFTR* (Kapoor et al., 2006; Prasad, Sharma, & Kaur, 2010). However, carrier rate for all the disease-causing variants in *CFTR* in Indian population is not yet available. A total of 20 carriers were observed for seven P/LP variants in *CFTR* yielding a net carrier rate of 1.37% and the carrier rate for NP_000483.3:p.(Phe508del) was 0.27%. Nearly 5% of the recessive deafness in South Asians is due to the disease causing variants observed in *SLC26A4* (Park et al., 2003) and 21 carriers were observed for 7 disease causing in variants in *SLC26A4* in our cohort. A high frequency of few rare monogenic disorders like Joubert syndrome (16 carriers), ataxia-ocular apraxia 2 (13 carriers) and Mucolipidosis II (12 carriers) was observed in the cohort. However, this data is unlikely to represent a high carrier frequency of these disorders as these data is derived from a biased cohort of families with predominantly neurodevelopmental and skeletal disease phenotypes.

Exome sequencing (ES) has emerged as a highly efficient tool for clinical diagnosis and research on monogenic diseases (Rabbani, Tekin, & Mahdieh, 2014). The widespread availability of ES had reduced the discrepancy of patients receiving a genetic diagnosis across the globe. However, the challenges remain in terms of broad testing and prioritization of appropriate candidate variant/s (MacArthur et al., 2014) The availability of data on common variants has eased this process. This has further been facilitated by addition of population specific variants in the recent past. We demonstrate the utility of our dataset collated from families undergoing ES without incurring any additional costs and its efficiency in variant prioritization in a test set of 50 individuals. After filtering the exonic/splicing heterozygous variants with allele frequencies and counts of homozygotes from gnomAD, 5.2% variants remained. Remarkably, with the use of these same filters from the refined cohort, an almost similar number of variants (5.7%) remained. Interestingly, filtering for presumably *de novo* variants based on refined cohort improved variant prioritization by filtering out 97.45% of the heterozygous variants observed in the test set. Similar results were obtained for homozygous variants, where gnomAD and refined cohort resulted in prioritization of ~0.4% of variants for further analysis.

The allele states and their counts derived from the original cohort which comprises of diseased as well as healthy individuals was efficiently used for prioritization of disease-causing variants. As the phenotypic information is available for the complete cohort, very low cut-off values could be used, thus leading to a very small number of candidates. Efficiency of these filters is evident from a subset of our cohort of 115 families with a diagnosis of inherited white matter disorders, a diagnostic yield of 71.28% (72/101) was achieved for singleton exome sequencing (unpublished data). Identification of recurrent disease-causing variants in unrelated families in the original cohort resulted in identification of novel disease-gene associations too (Katta M. Girisha et al., 2019; Shukla et al., 2017).

There are several limitations of the current dataset derived from families with rare monogenic disorders. First, it does not capture all genomic variants and is restricted to only exonic and flanking intronic variants. Hence the utility of the dataset is to a large extent limited to evaluation of monogenic disorders in children. Second, though the dataset was refined to make it as close as possible to a healthy cohort, it is likely to harbour disease causing variants with incomplete penetrance, variable expressivity and blended phenotypes. In such situations, we will be happy to share the phenotypes of the individual carrying such variant for the benefit of community. Third, the sample size is too small to represent the extremely huge and heterogeneous Indian population. Fourth, estimation of human knockouts from the cohort is incomplete as we have not considered LoF variants observed in compound heterozygous state. Fifth, the observed carrier status is biased as the cohort consists of families with monogenic disorders and not the general population. Sixth, the pathogenic variants were queried only from the ClinVar dataset and it does not capture all published disease-causing variants.

Despite these limitations, our dataset is a significant step to understand the genomic architecture and the distribution of alleles in Indian population. The most useful aspect of this dataset is its impact on variant filtering. We demonstrate the utility of combining clinical and research samples in a resource-limited setting and encourage genomic data sharing. Though limited in numbers, we also provide insights into human knockouts, carrier status and other clinically significant variants in our dataset that are not yet available for Indian population.

ACKNOWLEDGMENTS

We express our gratitude to all the patients and their family members who consented to sequencing their exomes at participating centres. We are grateful to the following funding agencies for supporting the respective centres to recruit the families with monogenic disorders and performed exome sequencing.

GRANT NUMBERS

1) Department of Health Research, Government of India. Grant numbers: V.25011/379/2015-GIA/HR and R.11012/02/2018-HR 2) Indian Council of Medical Research, Government of India. Grant numbers: 54/2/2013-HUM-BMS, No.4/13/58/2015/NCD-III, 5/7/1508/2016-CH, ISRM/12(40)/2019, 63/8/2010-BMS and F.No. 63/01/2019-Genomics/BMS, 3) Science and Engineering Research Board, Government of India. Grant numbers: SB/SO/HS/005/2014, YSS/2015/002009 and YSS/2015/001681, 4) Department of Biotechnology, Government of India. Grant numbers: BT/Bio-CARe/07/9889/2013-14, BT/PR9635/MED/97/198/2013, BT/PR13921/MED/12/704/2015 and BT/PR3193/MED/12/521/2011, 5) National Institutes of Health, United states. Grant number: 1R21NS094047–01 and R01 HD093570 01 A1.

DATA AVAILABILITY

Ready to integrate allele frequencies and counts of heterozygous and homozygous alleles from original and refined cohort are available in the following location. The de-identified variant data and phenotypic features of the individuals are available at respective centers and available on reasonable request.

http://cdfd.org.in/labpages/diag_datasets.html

REFERENCES

Aggarwal S, & Phadke SR (2015). Medical genetics and genomic medicine in India: current status and opportunities ahead. Mol Genet Genomic Med, 3(3), 160–171. doi:10.1002/mgg3.150 [PubMed: 26029702]

Ahmeti KB, Ajroud-Driss S, Al-Chalabi A, Andersen PM, Armstrong J, Birve A, … Consortium, A. (2013). Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. Neurobiology of aging, 34(1), 357.e357. doi:10.1016/j.neurobiolaging.2012.07.017

Akıncı N, Çakıl A, & Öner A (2013). Classical xanthinuria: a rare cause of pediatric urolithiasis. Turkish journal of urology, 39(4), 274–276. doi:10.5152/tud.2013.066 [PubMed: 26328123]

Alkuraya FS (2015). Natural human knockouts and the era of genotype to phenotype. Genome Medicine, 7(1), 48. doi:10.1186/s13073-015-0173-z [PubMed: 26029266]

Alsalem AB, Halees AS, Anazi S, Alshamekh S, & Alkuraya FS (2013). Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation. PLoS Genet, 9(12), e1004030. doi:10.1371/journal.pgen.1004030 [PubMed: 24367280]

Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, … Gyllensten U (2017). SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. European journal of human genetics : EJHG, 25(11), 1253–1260. doi:10.1038/ejhg.2017.130 [PubMed: 28832569]

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, … Abecasis GR (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. doi:10.1038/nature15393 [PubMed: 26432245]

Bhavani GS, Shah H, Dalal AB, Shukla A, Danda S, Aggarwal S, … Girisha KM (2015). Novel and recurrent mutations in WISP3 and an atypical phenotype. Am J Med Genet A, 167a(10), 2481–2484. doi:10.1002/ajmg.a.37164 [PubMed: 25988854]

Bidchol AM, Dalal A, Shah H, S S, Nampoothiri S, Kabra M, … Girisha KM (2014). GALNS mutations in Indian patients with mucopolysaccharidosis IVA. Am J Med Genet A, 164a(11), 2793–2801. doi:10.1002/ajmg.a.36735 [PubMed: 25252036]

Bittles AH, & Black ML (2010). The impact of consanguinity on neonatal and infant health. Early Hum Dev, 86(11), 737–741. doi:10.1016/j.earlhumdev.2010.08.003 [PubMed: 20832202]

Ceyhan-Birsoy O, Machini K, Lebo MS, Yu TW, Agrawal PB, Parad RB, … Rehm HL (2017). A curated gene list for reporting results of newborn genomic sequencing. Genetics in medicine : official journal of the American College of Medical Genetics, 19(7), 809–818. doi:10.1038/gim.2016.193 [PubMed: 28079900]

Chaubey G, Metspalu M, Kivisild T, & Villems R (2007). Peopling of South Asia: investigating the caste–tribe continuum in India. BioEssays, 29(1), 91–100. doi:10.1002/bies.20525 [PubMed: 17187379]

Colah R, Italia K, & Gorakshakar A (2017). Burden of thalassemia in India: The road map for control. Pediatric Hematology Oncology Journal, 2(4), 79–84. doi:10.1016/j.phoj.2017.10.002

Eng CM, & Desnick RJ (1994). Molecular basis of fabry disease: Mutations and polymorphisms in the human α-galactosidase A gene. Human Mutation, 3(2), 103–111. doi:10.1002/humu.1380030204 [PubMed: 7911050]

Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, … Najmabadi H (2019). Iranome: A catalog of genomic variations in the Iranian population. Hum Mutat, 40(11), 1968–1984. doi:10.1002/humu.23880 [PubMed: 31343797]

Forrester JC, & Ury HK (1969). The Signed-Rank (Wilcoxon) test in the rapid analysis of biological data. Lancet, 1(7588), 239–241. doi:10.1016/s0140-6736(69)91245-8 [PubMed: 4178594]

Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, & Dalgleish R (2018). VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. Human Mutation, 39(1), 61–68. doi:10.1002/humu.23348 [PubMed: 28967166]

Fu W, O’Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, … Akey JM (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature, 493(7431), 216–220. doi:10.1038/nature11690 [PubMed: 23201682]

Girisha KM, von Elsner L, Neethukrishna K, Muranjan M, Shukla A, Bhavani GS, … Mortier G (2019). The homozygous variant c.797G>A/p.(Cys266Tyr) in PISD is associated with a Spondyloepimetaphyseal dysplasia with large epiphyses and disturbed mitochondrial function. Human Mutation, 40(3), 299–309. doi:10.1002/humu.23693 [PubMed: 30488656]

Girisha KM, von Elsner L, Neethukrishna K, Muranjan M, Shukla A, Bhavani GS, … Mortier G (2019). The homozygous variant c.797G>A/p.(Cys266Tyr) in PISD is associated with a Spondyloepimetaphyseal dysplasia with large epiphyses and disturbed mitochondrial function. Hum Mutat, 40(3), 299–309. doi:10.1002/humu.23693 [PubMed: 30488656]

Goldstein DY, & Prystowsky M (2017). Autosomal Recessive Inheritance: Cystic Fibrosis. Academic pathology, 4, 2374289517691769–2374289517691769. doi:10.1177/2374289517691769 [PubMed: 28815197]

Grow K, Vashist M, Abrol P, Sharma S, & Yadav R (2014). Beta thalassemia in india: Current status and the challenges ahead. International Journal of Pharmacy and Pharmaceutical Sciences, 6, 28–33.

Hockham C, Bhatt S, Colah R, Mukherjee MB, Penman BS, Gupta S, & Piel FB (2018). The spatial epidemiology of sickle-cell anaemia in India. Scientific Reports, 8(1), 17685. doi:10.1038/s41598-018-36077-w [PubMed: 30523337]

Indian Genome Variation C (2008). Genetic landscape of the people of India: a canvas for disease gene exploration. Journal of Genetics, 87(1), 3–20. doi:10.1007/s12041-008-0002-x [PubMed: 18560169]

The Indian Genome Variation database (IGVdb): a project overview. (2005). Hum Genet, 118(1), 1–11. doi:10.1007/s00439-005-0009-9 [PubMed: 16133172]

Jain A, Bhoyar RC, Pandhare K, Mishra A, Sharma D, Imran M, … Sivasubbu S (2020). IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. Nucleic Acids Res. doi:10.1093/nar/gkaa923

John SE, Antony D, Eaaswarkhanth M, Hebbar P, Channanath AM, Thomas D, … Thanaraj TA (2018). Assessment of coding region variants in Kuwaiti population: implications for medical genetics and population genomics. Sci Rep, 8(1), 16583. doi:10.1038/s41598-018-34815-8 [PubMed: 30409984]

Kapoor V, Shastri SS, Kabra M, Kabra SK, Ramachandran V, Arora S, … Paul VK (2006). Carrier frequency of F508del mutation of cystic fibrosis in Indian population. J Cyst Fibros, 5(1), 43–46. doi:10.1016/j.jcf.2005.10.002 [PubMed: 16311077]

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, … MacArthur DG (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. bioRxiv, 531210. doi:10.1101/531210

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, … Genome Aggregation Database, C. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809), 434–443. doi:10.1038/s41586-020-2308-7 [PubMed: 32461654]

Kaur A, & Singh JR (2010). Chromosomal Abnormalities: Genetic Disease Burden in India. International Journal of Human Genetics, 10(1–3), 1–14. doi:10.1080/09723757.2010.11886079

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, & Maglott DR (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, 42(Database issue), D980–D985. doi:10.1093/nar/gkt1113 [PubMed: 24234437]

Le VS, Tran KT, Bui HTP, Le HTT, Nguyen CD, Do DH, … Nguyen LT (2019). A Vietnamese human genetic variation database. Hum Mutat, 40(10), 1664–1675. doi:10.1002/humu.23835 [PubMed: 31180159]

Lee S, Seo J, Park J, Nam JY, Choi A, Ignatius JS, … Choi M (2017). Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. Sci Rep, 7(1), 4287. doi:10.1038/s41598-017-04642-4 [PubMed: 28655895]

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, … Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536(7616), 285–291. doi:10.1038/nature19057 [PubMed: 27535533]

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, … Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536(7616), 285–291. doi:10.1038/nature19057 [PubMed: 27535533]

Li H (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England), 27(21), 2987–2993. doi:10.1093/bioinformatics/btr509 [PubMed: 21903627]

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, … Gunter C (2014). Guidelines for investigating causality of sequence variants in human disease. Nature, 508(7497), 469–476. doi:10.1038/nature13127 [PubMed: 24759409]

Minikel EV, Karczewski KJ, Martin HC, Cummings BB, Whiffin N, Rhodes D, … MacArthur DG (2020). Evaluating drug targets through human loss-of-function genetic variation. Nature, 581(7809), 459–464. doi:10.1038/s41586-020-2267-z [PubMed: 32461653]

Narang A, Roy RD, Chaurasia A, Mukhopadhyay A, Mukerji M, & Dash D (2010). IGVBrowser--a genomic variation resource from diverse Indian populations. Database: The journal of biological databases and curation, 2010, baq022. doi:10.1093/database/baq022 [PubMed: 20843867]

Park HJ, Shaukat S, Liu XZ, Hahn SH, Naz S, Ghosh M, … Griffith AJ (2003). Origins and frequencies of SLC26A4 (PDS) mutations in east and south Asians: global implications for the epidemiology of deafness. Journal of Medical Genetics, 40(4), 242. doi:10.1136/jmg.40.4.242 [PubMed: 12676893]

Popejoy AB, & Fullerton SM (2016). Genomics is failing on diversity. Nature, 538(7624), 161–164. doi:10.1038/538161a [PubMed: 27734877]

Prasad R, Sharma H, & Kaur G (2010). Molecular basis of cystic fibrosis disease: an Indian perspective. Indian journal of clinical biochemistry : IJCB, 25(4), 335–341. doi:10.1007/s12291-010-0091-1 [PubMed: 21966101]

Rabbani B, Tekin M, & Mahdieh N (2014). The promise of whole-exome sequencing in medical genetics. Journal of Human Genetics, 59(1), 5–15. doi:10.1038/jhg.2013.114 [PubMed: 24196381]

Reich D, Thangaraj K, Patterson N, Price AL, & Singh L (2009). Reconstructing Indian population history. Nature, 461(7263), 489–494. doi:10.1038/nature08365 [PubMed: 19779445]

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, … Committee, A. L. Q. A. (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. Genetics in medicine : official journal of the American College of Medical Genetics, 17(5), 405–424. doi:10.1038/gim.2015.30 [PubMed: 25741868]

Royston JP (1982). An Extension of Shapiro and Wilk’s W Test for Normality to Large Samples. Applied Statistics, 31(2), 115. doi:10.2307/2347973

Sachdeva K, Saxena R, Puri R, Bijarnia S, Kohli S, & Verma IC (2012). Mutation analysis of the CFTR gene in 225 children: identification of five novel severe and seven reported severe mutations. Genet Test Mol Biomarkers, 16(7), 798–801. doi:10.1089/gtmb.2011.0283 [PubMed: 22299590]

Shamseldin HE, Tulbah M, Kurdi W, Nemer M, Alsahan N, Al Mardawi E, … Alkuraya FS (2015). Identification of embryonic lethal genes in humans by autozygosity mapping and exome sequencing in consanguineous families. Genome Biology, 16(1), 116. doi:10.1186/s13059-015-0681-6 [PubMed: 26036949]

Shukla A, Das Bhowmik A, Hebbar M, Rajagopal KV, Girisha KM, Gupta N, & Dalal A (2018). Homozygosity for a nonsense variant in AIMP2 is associated with a progressive neurodevelopmental disorder with microcephaly, seizures, and spastic quadriparesis. J Hum Genet, 63(1), 19–25. doi:10.1038/s10038-017-0363-1 [PubMed: 29215095]

Shukla A, Hebbar M, Srivastava A, Kadavigere R, Upadhyai P, Kanthi A, … Girisha KM (2017). Homozygous p.(Glu87Lys) variant in ISCA1 is associated with a multiple mitochondrial dysfunctions syndrome. Journal of Human Genetics, 62(7), 723–727. doi:10.1038/jhg.2017.35 [PubMed: 28356563]

Singh I, Faruq M, Mukherjee O, Jain S, Pal PK, Srivastav MV, … Mukerji M (2010). North and South Indian populations share a common ancestral origin of Friedreich’s ataxia but vary in age of

GAA repeat expansion. Ann Hum Genet, 74(3), 202–210. doi:10.1111/j.1469-1809.2010.00569.x [PubMed: 20374234]

Sirugo G, Williams SM, & Tishkoff SA (2019). The Missing Diversity in Human Genetic Studies. Cell, 177(1), 26–31. doi:10.1016/j.cell.2019.02.048 [PubMed: 30901543]

Sivasubbu S, & Scaria V (2019). Genomics of rare genetic diseases-experiences from India. Hum Genomics, 14(1), 52. doi:10.1186/s40246-019-0215-5 [PubMed: 31554517]

Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, … Stefansson K (2015). Identification of a large set of rare complete human knockouts. Nat Genet, 47(5), 448–452. doi:10.1038/ng.3243 [PubMed: 25807282]

Team, R. C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. Accessed 13 March 2020.

Venugopal A, Chandran M, Eruppakotte N, Kizhakkillach S, Breezevilla SC, & Vellingiri B (2018). Monogenic diseases in India. Mutat Res, 776, 23–31. doi:10.1016/j.mrrev.2018.03.003

Verma IC, & Bijarnia S (2002). The burden of genetic disorders in India and a framework for community control. Community Genet, 5(3), 192–196. doi:10.1159/000066335 [PubMed: 14960891]

Wall JD, Stawiski EW, Ratan A, Kim HL, Kim C, Gupta R, … GenomeAsia, K. C. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature, 576(7785), 106–111. doi:10.1038/s41586-019-1793-z [PubMed: 31802016]

Wang K, Li M, & Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38(16), e164–e164. doi:10.1093/nar/gkq603 [PubMed: 20601685]

Williams TN, & Weatherall DJ (2012). World distribution, population genetics, and health burden of the hemoglobinopathies. Cold Spring Harb Perspect Med, 2(9), a011692. doi:10.1101/cshperspect.a011692 [PubMed: 22951448]

Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, … Wang C (2019). Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. Cell, 179(3), 736–749.e715. doi:10.1016/j.cell.2019.09.019 [PubMed: 31626772]

Xing J, Watkins WS, Hu Y, Huff CD, Sabo A, Muzny DM, … Yu F (2010). Genetic diversity in India and the inference of Eurasian population expansion. Genome Biology, 11(11), R113. doi:10.1186/gb-2010-11-11-r113 [PubMed: 21106085]

Zvereff VV, Faruki H, Edwards M, & Friedman KJ (2014). Cystic fibrosis carrier screening in a North American population. Genetics in medicine : official journal of the American College of Medical Genetics, 16(7), 539–546. doi:10.1038/gim.2013.188 [PubMed: 24357848]

References cited in the tables

Ahmed PH, V V, More RP, Viswanath B, Jain S, Rao MS, & Mukherjee O (2019). INDEX-db: The Indian Exome Reference Database (Phase I). J Comput Biol, 26(3), 225–234. doi:10.1089/cmb.2018.0199 [PubMed: 30615482]

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, … National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. doi:10.1038/nature15393 [PubMed: 26432245]

Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, … van Duijn CM (2014). The Genome of the Netherlands: design, and project goals. Eur J Hum Genet, 22(2), 221–227. doi:10.1038/ejhg.2013.118 [PubMed: 23714750]

Cocca M, Barbieri C, Concas MP, Robino A, Brumat M, Gandin I, … Mezzavilla M (2020). A bird's-eye view of Italian genomic variation through whole-genome sequencing. Eur J Hum Genet, 28(4), 435–444. doi:10.1038/s41431-019-0551-x [PubMed: 31784700]

Dopazo J, Amadoz A, Bleda M, Garcia-Alonso L, Aleman A, Garcia-Garcia F, … Antinolo G (2016). 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. Mol Biol Evol, 33(5), 1205–1218. doi:10.1093/molbev/msw005 [PubMed: 26764160]

Edwards N, Olinger E, Adam J, Kelly M, Schiano G, Ramsbottom SA, … Sayer JA (2017). A novel homozygous UMOD mutation reveals gene dosage effects on uromodulin processing and urinary excretion. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association, 32(12), 1994–1999. doi:10.1093/ndt/gfx066 [PubMed: 28605509]

Girisha KM, Bhavani GS, Shah H, Moirangthem A, Shukla A, Kim OH, … Mortier GR (2020). Biallelic variants p.Arg1133Cys and p.Arg1379Cys in COL2A1: Further delineation of phenotypic spectrum of recessive Type 2 collagenopathies. Am J Med Genet A, 182(2), 338–347. doi:10.1002/ajmg.a.61414 [PubMed: 31755234]

Hariprakash JM, Vellarikkal SK, Verma A, Ranawat AS, Jayarajan R, Ravi R, … Sivasubbu S (2018). SAGE: a comprehensive resource of genetic variants integrating South Asian whole genomes and exomes. Database (Oxford), 2018, 1–10. doi:10.1093/database/bay080

Higasa K, Miyake N, Yoshimura J, Okamura K, Niihori T, Saitsu H, … Matsuda F (2016). Human genetic variation database, a reference database of genetic variations in the Japanese population. J Hum Genet, 61(6), 547–553. doi:10.1038/jhg.2016.12 [PubMed: 26911352]

Lan T, Lin H, Zhu W, Laurent T, Yang M, Liu X, … Guo X (2017). Deep whole-genome sequencing of 90 Han Chinese genomes. Gigascience, 6(9), 1–7. doi:10.1093/gigascience/gix067

Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, … MacArthur D. (2016). Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. doi:10.1101/030338

Locke AE, Steinberg KM, Chiang CWK, Service SK, Havulinna AS, Stell L, … Freimer NB (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. Nature, 572(7769), 323–328. doi:10.1038/s41586-019-1457-z [PubMed: 31367044]

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, … Reich D (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature, 538(7624), 201–206. doi:10.1038/nature18964 [PubMed: 27654912]

Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, … Yamamoto M (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun, 6, 8018. doi:10.1038/ncomms9018 [PubMed: 26292667]

Narang A, Uppilli B, Vivekanand A, Naushin S, Yadav A, Singhal K, … Faruq M (2020). Frequency spectrum of rare and clinically relevant markers in multiethnic Indian populations (ClinIndb): A resource for genomic medicine in India. Hum Mutat. doi:10.1002/humu.24102

Naslavsky MS, Yamamoto GL, de Almeida TF, Ezquina SAM, Sunaga DY, Pho N, … Zatz M (2017). Exomic variants of an elderly cohort of Brazilians in the ABraOM database. Hum Mutat, 38(7), 751–763. doi:10.1002/humu.23220 [PubMed: 28332257]

Rezende-Lima W, Parreira KS, García-González M, Riveira E, Banet JF, & Lens XM (2004). Homozygosity for uromodulin disorders: FJHN and MCKD-type 2. Kidney Int, 66(2), 558–563. doi:10.1111/j.1523-1755.2004.00774.x [PubMed: 15253706]

Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, … Gleeson JG (2016). Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet, 48(9), 1071–1076. doi:10.1038/ng.3592 [PubMed: 27428751]

Tang D, Anderson D, Francis RW, Syn G, Jamieson SE, Lassmann T, & Blackwell JM (2016). Reference genotype and exome data from an Australian Aboriginal population for health-based research. Sci Data, 3, 160023. doi:10.1038/sdata.2016.23 [PubMed: 27070114]

Upadhyay P, Gardi N, Desai S, Sahoo B, Singh A, Togar T, … Dutt A (2016). TMC-SNPdb: an Indian germline variant database derived from whole exome sequences. Database: The journal of biological databases and curation. doi:10.1093/database/baw104

Zhang W, Meehan J, Su Z, Ng HW, Shu M, Luo H, … Hong H (2014). Whole genome sequencing of 35 individuals provides insights into the genetic architecture of Korean population. BMC Bioinformatics, 15 Suppl 11, S6. doi:10.1186/1471-2105-15-s11-s6

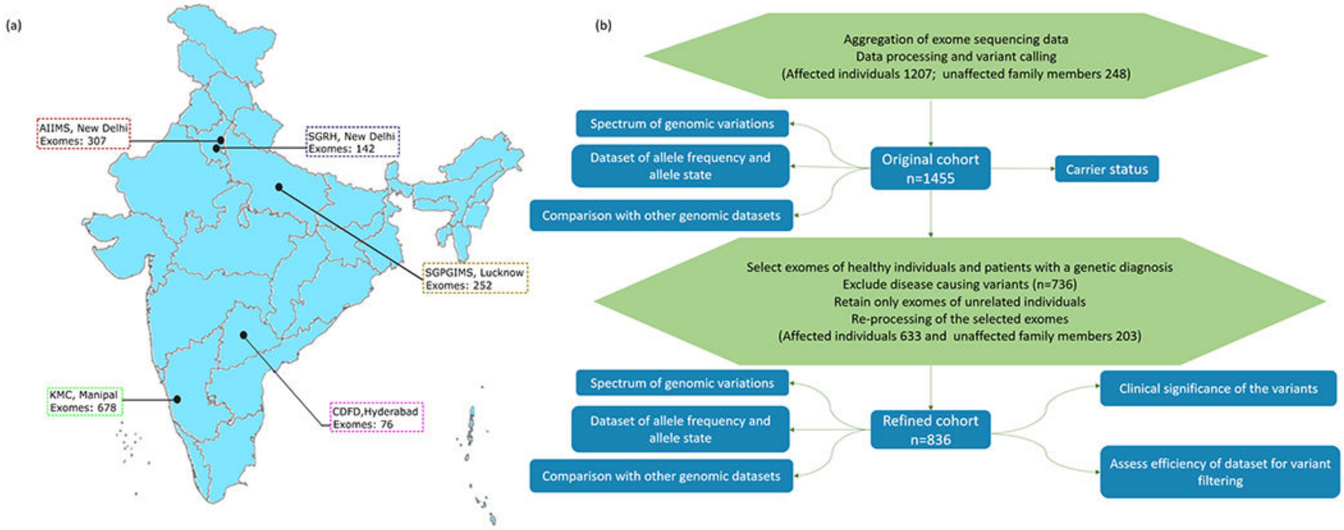
(a)
AIIMS, New Delhi
Exomes: 307
SGRH, New Delhi
Exomes: 142
SGPGIMS, Lucknow
Exomes: 252
KMC, Manipal
Exomes: 678
CDFD,Hyderabad
Exomes: 76
(b)
Aggregation of exome sequencing data
Data processing and variant calling
(Affected individuals 1207; unaffected family members 248)
Spectrum of genomic variations
Dataset of allele frequency and allele state
Comparison with other genomic datasets
Original cohort n=1455
Carrier status
Select exomes of healthy individuals and patients with a genetic diagnosis
Exclude disease causing variants (n=736)
Retain only exomes of unrelated individuals
Re-processing of the selected exomes
(Affected individuals 633 and unaffected family members 203)
Spectrum of genomic variations
Dataset of allele frequency and allele state
Comparison with other genomic datasets
Refined cohort n=836
Clinical significance of the variants
Assess efficiency of dataset for variant filtering

Figure 1:

(a) Location of collaborating centres and **(b)** flowchart representing the overview of our study design.

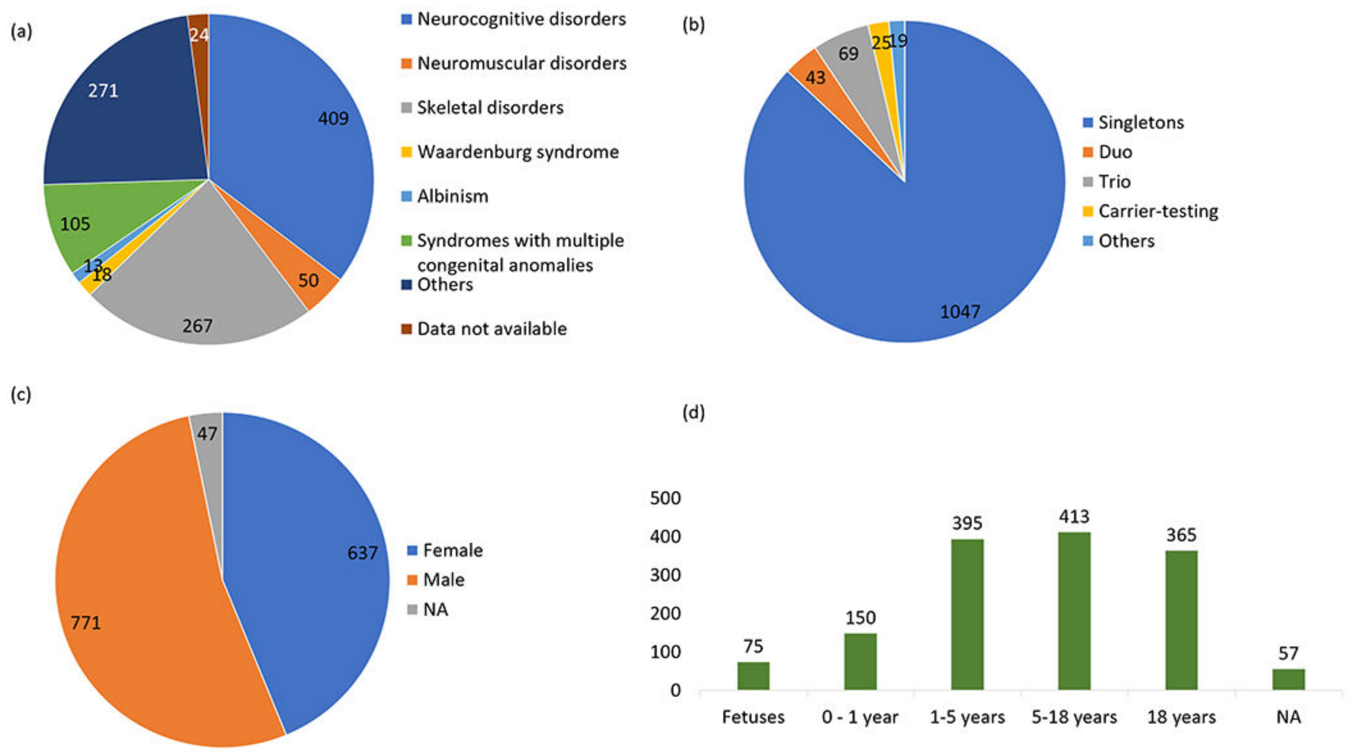

Figure 2:

(a) Referral patterns in the cohort for exome sequencing. **(b)** Exome sequencing approaches in this study. **(c)** Our cohort comprised predominantly of males. **(d)** Age distribution of the individuals included in the cohort. NA: data not available.

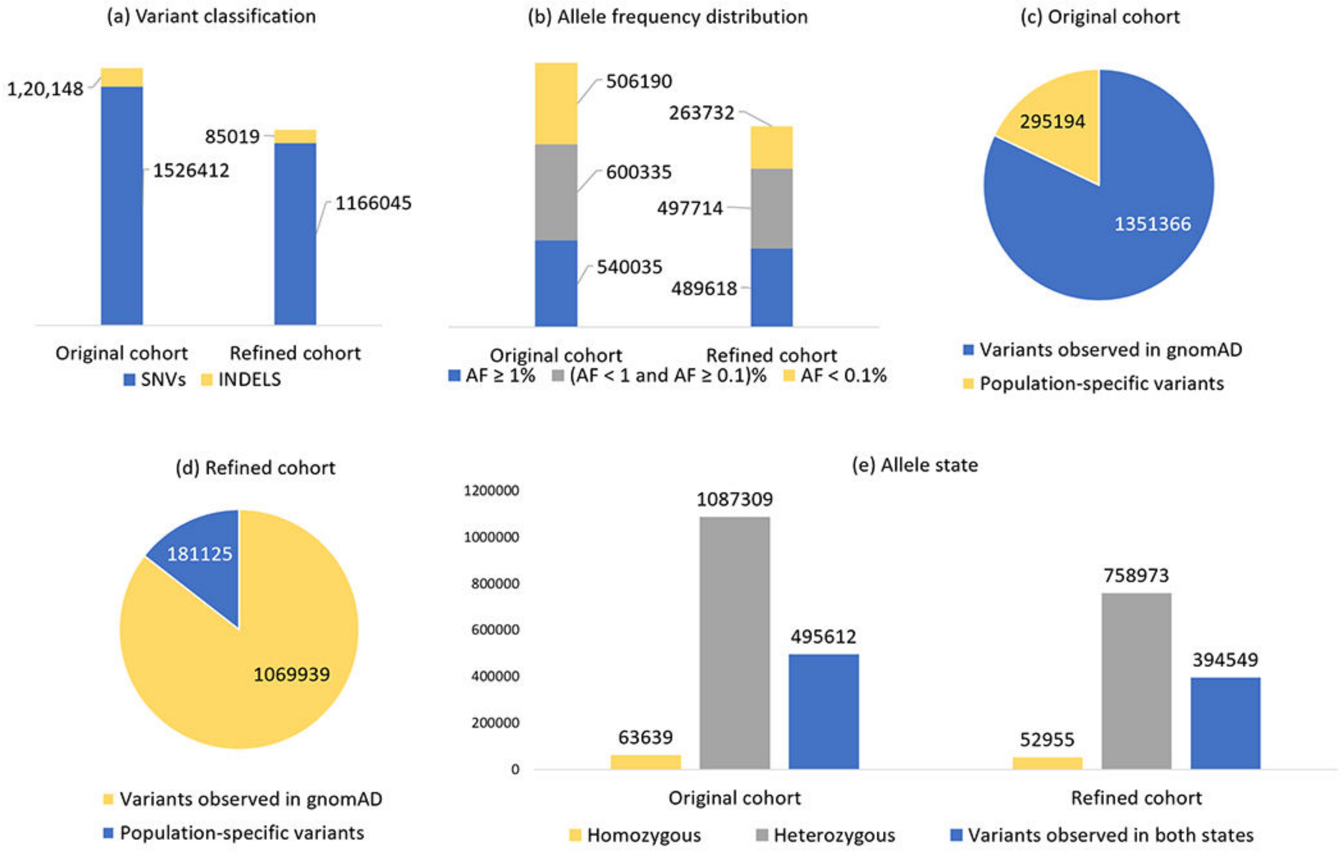

Figure 3:

Spectrum of genomic variants in the original (n=1455) and refined cohorts (n=836). **(a)** Distribution of SNVs and INDELs. **(b)** Distribution of common, rare and very rare variants in our cohort. 18% of the variants in the original cohort **(c)** and 14% in the refined cohort **(d)** were not observed in gnomAD. **(e)** We observed a significantly higher proportion of homozygous variants state in our population. AF: allele frequency, SNV: single nucleotide variants, INDEL: insertions and deletions.

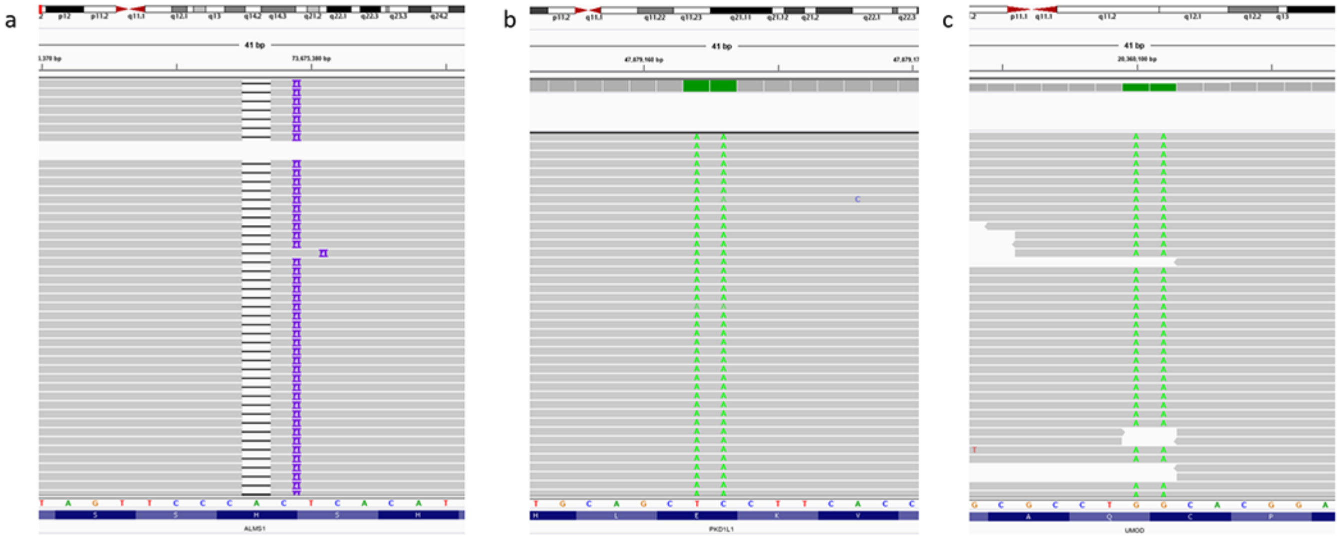
a
b
c
41 bp
73,675,380 bp
47,679,160 bp
20,360,100 bp
T A G T T C C C A C T C A C A T
S S H S H
ALMS1
T G C A G C T C C T T C A C C
H L E K V
PKD1L1
G C G C C T G G C A C G G A
A Q C P
UMOD

Figure 4:

(a) IGV snapshot of observed multi-nucleotide variant (MNV) in *AMLS1,* each of these variants were annotated separately as truncating variants but considering these as an MNV the observed consequence would be a non-frameshift variant, **(b)** The MNV in *PKD1L1* was annotated as 2 separate SNVs resulting in a truncating and a missense variant respectively, but the MNV is predicted to generate a missense variant. **(c)** The SNVs observed in the nearby codons of *UMOD* is annotated as a truncating and synonymous respectively, but considering this as an MNV still predicted to generate a truncating variant.

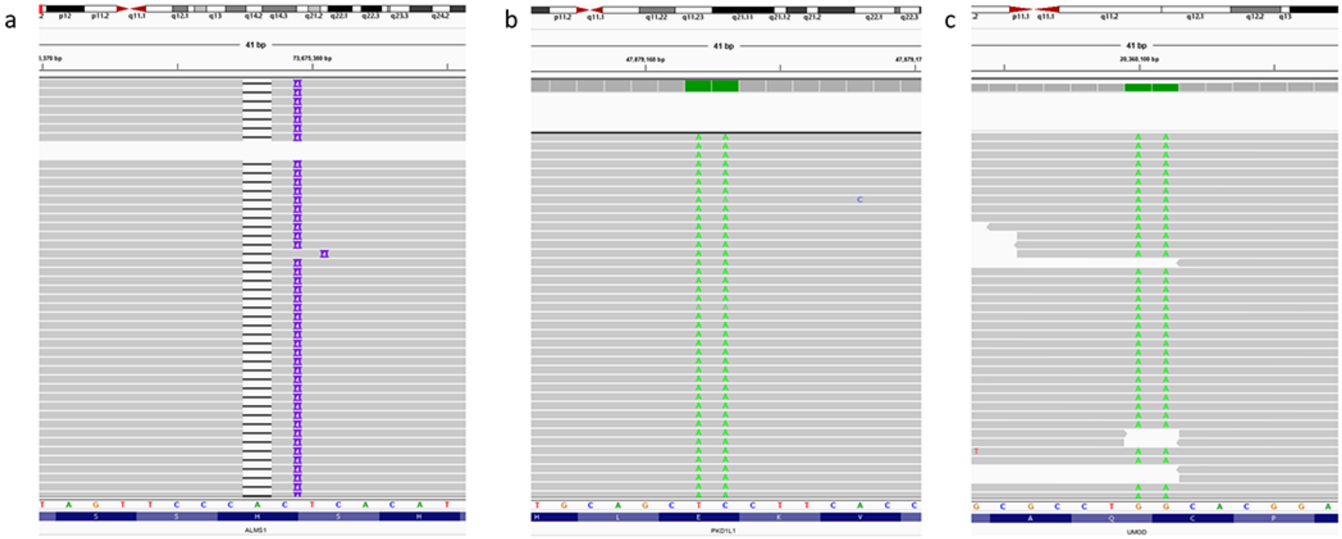

Figure 5:

Graphical demonstration of utility of refined cohort in variant prioritization for monogenic disorders. We observed significant reduction of **(a)** heterozygous variants and **(b)** homozygous variants per exome while applying the variant dataset from refined cohort alongside global population datasets.

Table 1.

A summary of global efforts to generate databases of genomic variants

S.No	Consortium/database	Cohort Size	Cohort description	Relatedness	Population/ population clusters	Age	Type of data	SNVs (million)	INDELS (million)	SVs (million)	Total variations (million)
1	The 1000 Genomes Project Consortium(A. Auton et al., 2015)	2504	All participants declared themselves to be healthy and self-reported gender and ethnicity	Unrelated and related samples	26 population	NA	WGS and targeted regions of 1000 genes	84.7M	3.6M	0.06M	88M
2	National, Heart, Lung and Blood Institute (NHLBI)-sponsored Exome Sequencing Project (ESP) (Fu et al., 2013)	6515	A multi-center study to deeply sequence the exomes of individuals segregating a variety of heart, lung, and blood disorders	Unrelated	European American and African Americans	NA	WES	1.14M	NA	NA	1.14M
3	Korean Personal Genomes Project (KPGP)(Zhang et al., 2014)	35	It is a participative research project established by Genome Research Foundation	NA	Korean population	NA	WGS	9.1M	NA	NA	9.1M
4	Greater Middle East Variome(Scott et al., 2016)	1111		Unrelated	1,794 self-reported nationals from GME regions participating in ongoing genetics studies. To minimize selection bias selected primarily healthy individuals from families and, wherever possible, removed from data sets the allele that brought the family to medical attention	NA	WES	NA	NA	NA	NA
5	Exome Aggregation Consortium (ExAC)(Lek et al., 2016)	60,706	Exomes are aggregated from various disease-specific and population genetic studies	Unrelated	7 population clusters	NA	WES	7.09M	0.31	NA	7.4M

S.No	Consortium/database	Cohort Size	Cohort description	Relatedness	Population/ population clusters	Age	Type of data	SNVs (million)	INDELS (million)	SVs (million)	Total variations (million)
6	Simons Genome Diversity Project (SGDP)(Mallick et al., 2016)	300	Deep genome sequences of 300 individuals from 142 populations chosen to span much of human genetic, linguistic, and cultural variation	NA	142 population	NA	WGS	34.4M	2.1M	NA	36.5M
7	GnomAD exomes (v2.1.1) (Karczewski et al., 2020)	125,748	Exomes aggregated from various disease-specific and population genetic studies	Unrelated	6 global and 8 sub-continental ancestries	NA	WES	NA	NA	NA	14.9M
	GnomAD genomes (v2.1.1) (Karczewski et al., 2020)	15,708					WGS	NA	Na	NA	229.9M
8	Singapore Genome Project(Wu et al., 2019)	4810	The three major ethnicities in Singapore were sequenced	Unrelated	Singapore Chinese, Malays, and Indians		WGS	89M	9M	-	98M
9	GenomeAsia 100K Project(Wall et al., 2019)	1739	Includes publically available whole genome sequencing data as well as samples which are sequenced as a part of this project too	Unrelated	219 population groups and 64 countries across Asia	NA	WGS	63M	3M	-	66M
10	Genome of the Netherlands(Boomsma et al., 2014)	769 individuals from 250 families	A trio design where population is relatively healthy, although persons with severe obesity are also observed	Related and unrelated samples	Dutch	>=19Y	WGS	20.4M	1.1M	0.05M	21.5M
11	Japanese population reference panel (1KJPN) (Nagasaki et al., 2015)	1070	Healthy Japanese individuals	Unrelated	Japanese	NA	WGS	29.6	3.3M	0.05M	33.25
12	Australian Aboriginal Population(Tang et al., 2016)	72	Different subsets of these individuals are diagnosed with type 2 diabetes (T2D) and/or obesity (according to their BMI)	NA	Aboriginal Australians	NA	WES	0.32M	0.05M	NA	0.37M
13	Collaborative Spanish Variant Server(Dopazo et al., 2016)	267	Individuals of Spanish origin and phenotyped as healthy	Unrelated	Spanish	NA	WES	0.17M	NA	NA	0.17M
14	Human genetic variation database (exomes)(Higasa et al., 2016)	1208	Subjects have no clinical record associated with major diseases	Unrelated	Japanese	NA	WES	0.28M	NA	NA	0.69M

S.No	Consortium/database	Cohort Size	Cohort description	Relatedness	Population/ population clusters	Age	Type of data	SNVs (million)	INDELS (million)	SVs (million)	Total variations (million)
	Human genetic variation database (genotyping array) (Higasa et al., 2016)	3248					Genotyping array	NA	NA	NA	1.79M
15	Korean Variant Archive (KOVA)(Lee et al., 2017)	1,055	WES data from normal tissues from cancer patients and samples from healthy individuals with no apparent clinical history	Unrelated	Korean	NA	WES	NA	NA	NA	0.29M
16	Han Chinese genomes(Lan et al., 2017)	90	Healthy Chinese samples from the 1000GP	Unrelated	Chinese	NA	WGS	12.5M	2.1M	NA	0.026M
17	ABraOM(Naslavsky et al., 2017)	609	Elderly individuals with adult onset disorders	Unrelated	Brazilians	>=60Y	WES	NA	NA	NA	1.28M
18	SweGen(Ameur et al., 2017)	942	Individuals selected from Swedish biobanks	Unrelated	Swedish		WGS	29.2M	3.8M	NA	33M
19	Kuwaiti exome variants(John et al., 2018)	291	Healthy individuals	Unrelated	Kuwaiti	NA	WES	0.17M	.003M	NA	0.173M
20	Vietnamese human genetic variation database (genomes)(Le et al., 2019)	105	Self-declared healthy individuals	Unrelated	Vietnamese	NA	WGS	22.47M	2.34M	NA	24.81
	Vietnamese human genetic variation database (exomes) (Le et al., 2019)	200	Healthy parents whose children participated as cases in autism spectrum disorder study	Related and unrelated			WES				
21	Iranome(Fattahi et al., 2019)	800	Healthy individuals	Unrelated	Iranian	>30Y	WES	1.3M	0.2M	NA	1.5M
22	Italian genomic variation(Cocca et al., 2020)	926	Whole genome sequences from isolated populations localized in three different geographical areas of Italy	Unrelated	Italian	NA	WGS	24M	2M	NA	27M
23	Finnish isolates(Locke et al., 2019)	19292	Individuals with cardiometabolic disorders and related traits	Related as well unrelated	Finish	≥45	WES	1.31M	0.92M	NA	1.4M

NA: not available, WES: whole exome sequencing, WGS: whole genome sequencing, M: million, SNV: single nucleotide variant, INDEL: insertions and deletions, SV: structural variations

Table 2.

Representations of Indians in various datasets

S.No	Consortium/database	Total subjects	Subjects of Indian origin	Representation of Indian population	Type of data	Data availability (Restricted access/ open access)
1	1000 genomes project(Adam Auton et al., 2015)	2504	227	Healthy unrelated Gujarati Indians in Houston, Texas, USA and Indian Telugu in the UK	WGS	Open access
2	ExAC/gnomAD(Karczewski et al., 2020; Lek et al., 2016)	60,706/141,456	227	Inherited from 1000 genomes	WGS	Open access
3	SAGE(Hariprakash et al., 2018)	1213	334	Inherited from 1000 genomes, also integrated Singapore Sequencing Indian Project (SSIP) and Indian genomes from the study of Population Genetics of Andamanese	WGS	Restricted access
4	TMC-SNPdb(Upadhyay et al., 2016)	72	72	Non-cancerous samples derived from cancer patients	ES	Open access
5	INDEX-DB(Ahmed et al., 2019)	109	109	Individuals determined to be asymptomatic for adult-onset common clinical illnesses	ES	Restricted access
6	Singapore 10K Genome Project(Wu et al., 2019)	4810	1127	Healthy individuals	WGS	Restricted access
7	GenomeAsia 100K projects(Wall et al., 2019)	1739	598	This includes 38 publicly available whole-genome sequencing data as well as 560 samples which are sequenced as a part of this project too	WGS	Open access
8	ClinIndb(Narang et al., 2020)	2795	2795	This study catalogued the frequency profile of ~19K clinically relevant variants in multi-ethnic Indian population	Global screening array genotype data	Open access
9	IndiGenomes(Jain et al., 2020)	1029	1029	Self-declared healthy individuals	WGS	Open access
10	Current study	1455	1455	Individuals with suspected rare monogenic disorders and their parents and family members	ES	Open access

NA: not available, WGS: whole-genome sequencing, ES: exome sequencing

Table 3.

Contributing centers and technical details of exome sequencing

Center	DNA extraction		Sequencing platform (Illumina Inc. USA)	Read information	Capture kit	Target region (Mb)	Targeted sequencing coverage	Number of individuals	Total number of individuals
	Source	Method							
Kasturba Medical College, Manipal (KMC)	Whole blood/ Fetal tissue	Standard phenol-chloroform method/ QiAmp DNA Blood mini kit (250)/DNeasy Blood and tissue kit (250)-QIAGEN/QiAmp DNA FFPE tissue kit (50)	NextSeq500	PE 2x150	Nextera Rapid Capture Exome Kit	37	100x	327	**678**
			HiSeq2000	PE 2x150	Agilent SureSelect v6	60	100x	279	
			NovaSeq6000	PE 2x100	Agilent SureSelect CREv2	67	100x	55	
			HiSeq2000	PE 2x150	Kapa HTP along with Roche and Nimblegen (Customised kit)	8332 genes	80-100x	17	
Sanjay Gandhi Postgraduate Institute of Medical Sciences (SGPGIMS), Lucknow	Whole blood	Qiagen (Hilden, Germany)	NextSeq500	PE 2x150	Nextera Rapid Capture Exome Kit	36	100x	142	**252**
			HiSeq2000	PE 2x150	Agilent SureSelect v6	58	100x	87	
			HiSeq	PE 2x150	Agilent SureSelect v5	50	100x	5	
			HiSeq2000	PE 2x150	Kapa HTP along with Roche and Nimblegen (Customised kit)	8332 genes	80-100x	18	
Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad	Whole blood	Standard phenol-chloroform method/ DNeasy Blood and Tissue kit (Qiagen, Valencia, CA)	HiSeq2000	PE 2x150	Nextera Rapid Capture Exome Kit	37	100x	44	**76**
			HiSeq2000	PE 2x150	Agilent SureSelect v6	60	100x	32	
All India Institute of Medical Sciences (AIIMS), New Delhi	Whole blood	Qiasymphony Blood DNA midi kit	HiSeq2000	PE 2x150	Kapa HTP along with Roche and Nimblegen (Customised kit)	8332 genes	100x	46	**307**
			HiSeq	PE 2x150	Agilent SureSelect v5	50	80-100x	261	
Sir Gangaram Hospital (SGRH), New Delhi	Whole blood	Standard phenol-chloroform method	HiSeq 2500/3000/4000	PE 2x100	IDT Exome Research Panel	39	50x	7	**142**
				PE 2x100	Agilent SureSelect v5	50	80-100x	61	
				PE 2x150					
				PE 2x150	Agilent SureSelect v6	60	100x	28	
				PE 2x100	Agilent SureSelect CREv2	67		3	
				PE 2x100	Kapa HTP along with Roche and Nimblegen(Customised kit)	8332 genes	80-100x	43	
Total number of exomes									**1455**

PE: Paired end

Table 4.

Demographic profile of the cohort

Cohort characteristics	KMC	SGPGIMS	CDFD	AIIMS	SGRH	Total
Families with a monogenic disorder	522	215	65	293	108	**1203**
Affected individuals	567	217	66	255	102	**1207**
Unaffected family members	111	35	10	52	40	**248**
Individuals contributing their exomes to this study	678	252	76	307	142	**1455**
Male	355	140	36	179	61	**771**
Female	323	104	27	128	55	**637**
Gender not available	0	8	13	0	26	**47**
Referral pattern						
Neurocognitive disorders	221	76	4	73	35	**409**
Neuromuscular disorders	22	6	1	17	4	**50**
Skeletal disorders	159	52	11	34	11	**267**
Waardenburg syndrome	17	0	0	1	0	**18**
Albinism	11	1	1	0	0	**13**
Syndromes with multiple congenital anomalies	7	41	28	17	12	**105**
Others	77	39	16	123	16	**271**
Data not available	0	0	0	0	24	**24**
Age distribution						
Fetuses	40	4	5	0	26	**75**
0 – 1 year	65	47	2	27	9	**150**
1-5 years	185	65	9	110	26	**395**
5-18 years	201	78	13	94	27	**413**
18 years	187	49	5	75	49	**365**
Data not available	0	9	42	1	5	**57**
Exome sequencing approach						
Singletons	427	199	56	277	89	**1047**
Duo	36	0	7	0	0	**43**
Trio	40	13	0	#	16	**69**

Cohort characteristics	**KMC**	**SGPGIMS**	**CDFD**	**AIIMS**	**SGRH**	**Total**
Carrier-testing	6	0	0	16	3	**25**
Others	13	4	2	0	0	**19**
Families with a molecular diagnosis by exome sequencing	348	102	21	220	44	**735**

KMC: Kasturba Medical College, Manipal

SGPGIMS: Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow

CDFD: Centre for DNA Fingerprinting and Diagnostics, Hyderabad

AIIMS: All India Institute of Medical Sciences, New Delhi

SGRH: Sir Gangaram Hospital, New Delhi

Table 5.

Spectrum of genomic variants in the cohort

	Original cohort (N=1455)					Refined cohort (N=836)				
Variants	**Total**	**Common (AF 0.01)**	**Rare (AF <0.01)**	**Very rare (AF <0.001)**	**Population specific variants**	**Total**	**Common (AF 0.01)**	**Rare (AF <0.01)**	**Very rare (AF <0.001)**	**Population specific variants**
Raw variants	1844228	625023	1219206	556905	342608	1449306	615367	833940	290633	212323
'PASS' variants	1707888	591282	1116606	506190	301239	1348519	581542	766977	263732	188254
Variants with call rate of >8%	1646560 **(100%)**	540035 **(32.8%)**	1106525 **(67.2%)**	506190	295194 **(18%)**	1251064 **(100%)**	489618 **(39.1%)**	761446 **(60.9%)**	263732	181125 **(14.6%)**
SNVs	1526412 **(92.7%)**	482460	1043952	487692	268402	1166045 **(93.3%)**	442385	723660	256651	165899
INDELS	120148 **(7.3%)**	57575	62573	18498	26792	85019 **(6.7%)**	47233	37786	7081	15226
Transition/transversion (Ti:Tv) ratio	2.26					2.28				
Allele state										
Heterozygous	1087309 **(66.1%)**	77026 **(4.7%)**	1010283	493628	265196 **(16.1%)**	803560 **(64.3%)**	95275 **(7.6%)**	708285	258440	163142 **(13.1%)**
Heterozygous SNVs	1023740	66668	957072	476084	245264	758973	83632	675341	251843	151984
Heterozygous INDELs	163569	10358	53211	17544	19932	44587	11643	32944	6597	11158
Homozygous	63639 **(3.8%)**	26193 **(1.6%)**	37446	12522	20984 **(1.27%)**	52955 **(4.2%)**	24013 **(1.9%)**	28942	5274	13829 **(1.1%)**
Homozygous SNVs	56510	23289	33221	11585	18125	47574	21667	25907	4803	11947
Homozygous INDELS	7129	2904	4225	937	2943	5381	2346	3035	471	1882
Variants observed in homozygous and as well as in heterozygous state	495612 **(30.1%)**	436816 **(26.6%)**	58796	40	9091 **(0.06%)**	394549 **(31.6%)**	370330 **(30%)**	24219	18	4154 **(0.34%)**
SNVs observed in homozygous and as well as in heterozygous state	446162	392503	53659	23	5168	359498	337086	22412	5	1968
INDELs observed in homozygous and as well as in heterozygous state	49450	44313	5137	17	3923	35051	33244	1807	13	2186

Table 6.

Classification of variants based on genic regions

	Original cohort					**Refined cohort**				
	Total	**Common (AF ≥ 0.01)**	**Rare (AF <0.01)**	**Very rare (AF <0.001)**	**Population specific variants**	**Total**	**Common (AF ≥ 0.01)**	**Rare (AF <0.01)**	**Very rare (AF <0.001)**	**Population specific variants**
Intronic	776505 **(47.2%)**	333046 **(20.3%)**	443459	94798	139566 **(8.5%)**	592266 **(47.3%)**	293172 **(30%)**	299094	31653	85307 **(7%)**
Exonic	494628 **(30.1%)**	68626 **(4.2%)**	426002	319308	89977 **(5.5%)**	369040 **(30%)**	68937 **(5.6%)**	300103	194197	55300 (4.5%)
Intronic boundaries of exons (up to 20bp)	102757 **(6.3%)**	27214 **(1.7%)**	75543	45714	15393 **(0.1%)**	78685 **(6.2%)**	26683 **(2.3%)**	52002	20883	9123 **(0.8%)**
UTR3	62405	23546	38859	11202	11394	48150	21654	26496	4002	7008
UTR5	44027	12512	31515	12404	10094	33809	12095	21714	4608	6345
ncRNA_intronic	47466	22541	24925	4613	7826	36413	19655	16758	1420	4835
ncRNA_exonic	30871	11218	19653	8208	5547	24204	10698	13506	3593	3294
ncRNA_splicing	30871	11218	19653	8208	324	24204	810	865	113	227
Intergenic	48316	23384	24932	6053	7928	38083	20939	17144	2409	5195

Table 7.

Predicted functional consequences of the variants

	Original cohort					Refined cohort				
	Total	**Common** **(AF** $\geq$ 0.01)	**Rare** **(AF <0.01)**	**Very rare** **(AF <0.001)**	**Population specific variants**	**Total**	**Common** **(AF** $\geq$ 0.01)	**Rare** **(AF <0.01)**	**Very rare** **(AF <0.001)**	**Population specific variants**
Nonsynonymous SNV	278,376 **(16.9%)**	32,111 **(2%)**	246,104	187,354	52,741 **(3.2%)**	205413 **(16.6%)**	32318 **(2.6%)**	173095	114110	32315 **(2.5%)**
Synonymous SNV	184,799 **(11.2%)**	31,998 **(1.9%)**	152,709	112,469	28,240 **(1.7%)**	142615 **(11.5%)**	32131 **(2.6%)**	110484	70158	18751 **(1.5%)**
Stopgain	5,792 **(0.4%)**	395 **(0.02%)**	5,396	4,427	1,845 **(0.1%)**	3860 **(0.3%)**	371 **(0.03%)**	3489	2512	1016 **(0.08%)**
Stoploss	275 **(0.2%)**	44 **(0.002%)**	231	166	78 **(0.005%)**	183 **(0.01%)**	43 **(0.003%)**	140	79	32 **(0.003%)**
Frameshift deletion	6,008 **(0.4%)**	526 **(0.03%)**	5,475	4,270	2,613 **(0.2%)**	3442 **(0.3%)**	478 **(0.04)**	2964	2010	1162 **(0.09%)**
Frameshift insertion	3,449 **(0.2%)**	325 **(0.02%)**	3,118	2,464	1,718 **(0.1%)**	1782 **(0.1%)**	295 **(0.02%)**	1487	959	565 **(0.04%)**
Nonframeshift deletion	6,785 **(0.4%)**	1,121 **(0.07%)**	5,651	3,208	1,010 **(0.08%)**	4705 **(0.4%)**	1128 **(0.09%)**	3577	1629	501 **(0.04%)**
Nonframeshift insertion	3,146 **(0.2%)**	570 **(0.04%)**	2,570	1,720	808 **(0.06%)**	2096 **(0.2%)**	576 **(0.05%)**	1520	784	349 **(0.03%)**

Table 8.

Homozygous variants are common in Asians and Asian Indians

	gnomAD exome	gnomAD genome	GenomeAsia	Kuwaiti exomes	Original cohort	Refined cohort
Total variants	15,648,788	254,298,981	70,651,672	173,849	1,646,560	1,251,064
Homozygous	646,921 **(4.1%)**	25,928,447 **(10.19%)**	13,156,735 **(18.6%)**	35024 **(20%)**	559,251 **(34%)**	447,504 **(35.8%)**

Table 9:

Homozygous loss of function (HLoF) variants in genes without any phenotypic descriptions catalogued in OMIM

Chromosome	Coordinate	Reference allele	Altered allele	Gene	Predicted functional consequence	HGVS transcript	HGVS Predicted Protein	pLi score	Homozygotes	Allele frequency	Total number of alleles	Gene/Variant explains the phenotype of affected individual
15	84566628	C	T	*ADAMTSL3*	stopgain	NM_207517.2:c.1486C>T	NP_997400.2:p.(Arg496Ter)	9.55E-09	1	0.001205	1660	Deafness, autosomal recessive 1A; 220290; AR
6	46993710	G	A	*ADGRF1*	stopgain	NM_025048.3:c.157C>T	NP_079324.2:p.(Gln53Ter)	0	1	0.023	938	Wolcott-Rallison syndrome; 226980, AR
19	10206819	C	A	*ANGPTL6*	stopgain	NM_001321411.1:c.421G>T	NP_001308340.1:p.(Glu141Ter)	2.77E-08	1	0.003193	1566	Neurodegeneration with brain iron accumulation 4; 614298; AR
20	31598882	TC	T	*BPIFB2*	frameshift deletion	NM_025227.1:c.163del	NP_079503.1:p.(His55IlefsTer18)	7.69E-08	1	0.001706	1172	Bardet-Biedl syndrome 8; 615985; AR
20	31671213	AC	A	*BPIFB4*	frameshift deletion	NM_182519.2:c.218del	NP_872325.2:p.(Pro73GlnfsTer160)	3.47E-17	1	0.004274	1170	Metachromatic leukodystrophy; 250100; AR
16	80718540	C	A	*CDYL2*	stopgain	NM_152342.2:c.511G>T	NP_689555.2:p.(Gly171Ter)	0.62496	1	0.002793	358	Ciliary dyskinesia, primary, 5; 608647; AR
6	36343676	G	A	*ETV7*	stopgain	NM_001207037.1:c.61C>T	NP_001193966.1:p.(Arg21Ter)	2.01E-08	1	0.002564	1170	Unaffected family member
8	82439310	C	T	*FABP12*	stopgain	NM_001105281.2:c.293G>A	NP_001098751.1:p.(Trp98Ter)	0.002946	1	0.0008562	1168	Spondylocarpotarsal synostosis syndrome; 272460; AR
12	2970521	G	A	*FOXM1**	stopgain	NM_202002.2:c.1324C>T	NP_973731.1:p.(Arg442Ter)	0.80259	1	0.004225	1420	Unaffected family member
1	156713507	TC	T	*HDGF*	frameshift deletion	NM_001319188.1:c.556del	NP_001306117.1:p.(Glu186ArgfsTer107)	0.269699	1	0.002525	1188	Osteogenesis imperfecta, type VI; 613982; AR
4	57516913	G	C	*HOPX**	stopgain	NM_001145460.1:c.264C>G	NP_001138932.1:p.(Tyr88Ter)	0.146966	1	0.004008	998	Unaffected family member
16	67212203	CG	C	*KIAA0895L*	frameshift deletion	NM_001040715.1:c.1051del	NP_001035805.1:p.(Arg351ValfsTer62)	0.018433	1	0.0008503	1176	Eiken syndrome; 600002, AR
17	21188236	G	T	*MAP2K3*	stopgain	NM_145109.2:c.4G>T	NP_659731.1:p.(Glu2Ter)	0.000496	1	0.001205	1660	Rickets due to defect in vitamin

Chromosome	Coordinate	Reference allele	Altered allele	Gene	Predicted functional consequence	HGVS transcript	HGVS Predicted Protein	pLi score	Homozygotes	Allele frequency	Total number of alleles	Gene/Variant explains the phenotype of affected individual
												D 25-hydroxylation deficiency; 600081; AR
9	131475462	C	T	*PKN3*	stopgain	NM_013355.3:c.967C>T	NP_037487.2:p.(Arg323Ter)	1.47E-09	1	0.001203	1662	Hypophosphatemic rickets with hypercalciuria; 241530; AR
3	157289824	GA	G	*SLC66A1L*	frameshift deletion	NM_001099777.3:c.301del	NP_001093247.1:p.(Ile101PhefsTer60)	0	1	0.002564	1170	Short stature, brachydactyly, intellectual developmental disability, and seizures; 617157; AR
12	100706292	T	TA	*SCYL2*	frameshift insertion	NM_001330256.1:c.195dup	NP_001317185.1:p.(Leu66ThrfsTer6)	0.931829	1	0.002008	996	Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome; 609528; AR
12	51279076	C	T	*TMPRSS12*	stopgain	NM_182559.2:c.700C>T	NP_872365.1:p.(Arg234Ter)	0.003116	1	0.005682	352	Immunodeficiency 56; 615207; AR
4	69796349	TG	T	*UGT2A3*	frameshift deletion	NM_024743.3:c.1218del	NP_079019.3:p.(Gly408GlufsTer4)	5.00E-06	1	0.059	852	Orofaciodigital syndrome VI; 277170; AR
5	178454524	G	A	*ZNF879*	stopgain	NM_001353373.1:c.84G>A	NP_001340302.1:p.(Trp28Ter)	0.000705	1	0.0008562	1168	HMG-CoA lyase deficiency; 246450; AR

* The HLoF variants in FOXM1 and HOPX is observed in the exons that are barely expressed across the tissues and the pext score is near to zero

r	Inheritance pattern	Homozygotes	Allele frequency	Total number of alleles	Gene/Varaint explains the phenotype of affected individual	Genetic diagnosis/ observed phenotypes of the affected individual	MIM number (phenotype)	Inheritance pattern	Age (years) at the time of evaluation	Remarks
	AR	1	0.001202	1664	*COL2A1*: NM_001844.4:c.4135C>T: NP_001835.3:p.(Arg1379Cys)	Spondyloepiphyseal dysplasia	NA	NA	27	The identified frameshift SNV in *ALMS1* was classify as an MNV due to the presence of another INDEL nearby. These variants together predicted to lead to a nonframeshift variant (NM_015120.4(ALMS1_v001):c.1727_1728delinsCTAGT:p.(His576delinsProSer)) and rescue the truncating effect (figure 4 (a)). The individual with *ALMS1* variant is diagnosed with spondyloepiphyseal dysplasia and the identified disease-causing variant in *COL2A1* is a reported variant of unknown significance (Girisha et al., 2020).
	AR	1	0.001805	1662	*MTHFR*: NM_005957.4:c.202C>G: NP_005948.3:p.(Arg68Gly)	Homocystinuria due to *MTHFR* deficiency	236250	AR	2.5	The identified SNV lead to the stop gain variant was classify as an MNV due to the presence of another SNV in the same codon. These variants together predicted to rescue the truncating effect by generating a missense variant (NM_138295.3:c.5650_5651delinsTT, NP_612152.1:p.(Glu1884Leu)) (figure 4 (b)). The individual with *PKD1L1* variant is diagnosed with homocystinuria due to *MTHFR* deficiency and the identified variant is a reported pathogenic variant.
00;603860	NA;AD;AD	1	0.001205	1660	SERPINF1: NM_002615.5:c.248_249insA: NP_002606.3:p. (Ser84GlnfsTer28)	Osteogenesis imperfecta, type VI	613982	AR	1	The identified SNV lead to the stop gain variant was classify as an MNV due to the presence of another SNV observed in nearby codon. However, these variants together still predict this as truncating variant, NM_001008389.1:c.522_523delinsTT, NP_001008390.1:p.(Gln175Ter) (figure 4 (c)). In general, heterozygous missense variants in *UMOD* are known to cause the disease phenotype, a few reports are available in the literature reporting homozygous missense variants with more severe phenotype compared to heterozygotes (Edwards et al., 2017; Rezende-Lima et al., 2004). However, it is a progressive disorder and age of onset is usually young adulthood. We could not rule out the possibility of a blended phenotype of *UMOD* related kidney disease and osteogenesis imperfecta, as the child was aged one year and could not be assessed for renal phenotype
	AD,AR	1	0.001203	1662	Candidate gene (unpublished data)	Facial dysmorphism, ambiguous genitalia with short stature	NA	NA	1	*PRPH* have been associated with susceptibility to amyotrophic lateral sclerosis (ALS) and other than for juvenile ALS, rarely people will develop symptoms in early childhood Variable age of onset is also reported for *PRPH* disease phenotype.

r	Inheritance pattern	Homozygotes	Allele frequency	Total number of alleles	Gene/Varaint explains the phenotype of affected individual	Genetic diagnosis/ observed phenotypes of the affected individual	MIM number (phenotype)	Inheritance pattern	Age (years) at the time of evaluation	Remarks
	AR	1	0.013	1660	*CACNA1S*: NM_000069.2:c.4316G>T: NP_000060.2:p.(Cys1439Phe)	Arthrogryposis multiplex congenita, cleft palate and early death	NA	NA	16 months	Variable age of onset is reported for Xanthinuria, type II and the proband with variant in *MOCOS* was expired at very early age to rule out possibility of blended phenotype
53;610217	AR;AR;AR	1	0.0006849	1460	RAG2: NM_000536.3:c.1247G>T: NP_000527.2:p.(Trp416Leu)	Omenn syndrome	603554	AR	3 months	The individual with *PLA2G6* variant was diagnosed with Omenn syndrome and the identified variant in *RAG2* is a reported variant of unknown significance. The proband might have been too young to show features of *PLAG26* and the early death of the proband related neurodegeneration at four months of age Possibility of a blended phenotype couldn't rule out due to the death of the proband with bronchopneumonia

Validator

OMIM phenotype	Inheritance pattern	MIM number (phenotype)	Refined cohort: Allele counts| Homozygote| Hemizygote	Allele frequency	Total number of alleles	gnomAD: Allele counts| Homozygote| Hemizygote	Genetic diagnosis/observed phenotypes of the individuals and the reported variants observed in heterozygous state		Genetic diagnosis/observed phenotypes of the individual with the identified known disease-causing variants in homozygous state		Genetic diagnosis/ observed phenotypes of the individual with the identified known disease-causing variants in hemizygous state		Remarks
							Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype	Age (years) at the time of evaluation	
Keratoderma, palmoplantar, with deafness	AD	148350	18|0|0	0.011	1658	147|1|0	Sotos; 117550; AD	2	#	#	#	#	Clinical variability and variable age of onset known for Keratoderma, palmoplantar, with deafness. Also, this variant is observed in gnomAD database
							Osteogenesis imperfecta, type I; 166200; AD	11M					
							Bardet-Biedl syndrome 10; 615987; AR	7					
							Meckel syndrome 6; 612284; AR	Fetus					
							Spondylocarpotarsal synostosis syndrome; 272460; AR	3					
							Myasthenic syndrome, congenital, 5; 603034; AR	12					
							Mental retardation, stereotypic movements, epilepsy, and/or cerebral malformations; 613443; AD	9					
							congential hypothyroidism with developmental delay and regression; 253300; AR	5					
							Proband with jejunal atreria; 617598; AR	1					
							Waardenburg syndrome 1; 193500; AD	1					

OMIM phenotype	Inheritance pattern	MIM number (phenotype)	Refined cohort: Allele counts|Homozygote|Hemizygote	Allele frequency	Total number of alleles	gnomAD: Allele counts|Homozygote|Hemizygote	Genetic diagnosis/observed phenotypes of the individuals and the reported variants observed in heterozygous state		Genetic diagnosis/observed phenotypes of the individual with the identified known disease-causing variants in homozygous state		Genetic diagnosis/observed phenotypes of the individual with the identified known disease-causing variants in hemizygous state		Remarks
							Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype	Age (years) at the time of evaluation	
							Infantil neuroaxonal dystrophy-1, neurodegeneration with brain iron accumulation 2B; 610217; AR	10					
							Tuberous sclerosis; 613254; AD	2					
							Six unrelated unaffected family members	>20					
Keratoderma, palmoplantar, with deafness	AD	148350	1|0|0	0.0006766	1478	1737|10|0	Niemann-Pick disease, type C1; 257220; AR	4	#	#	#	#	Clinical variability and variable age of onset known for Keratoderma, palmoplantar, with deafness. Also, this variant is observed in gnomAD database. However now this variant has been reported in ClinVar with conflicting interpretation of pathogenicity
Tuberous sclerosis-2	AD	613254	1|0|0	0.0006024	1660	0|0|0	Lesch-Nyhan syndrome; 300322; XL	4	#	#	#	#	Clinical variability as well as non-penetrance is reported for Tuberous sclerosis-2

OMIM phenotype	Inheritance pattern	MIM number (phenotype)	Refined cohort: Allele counts|Homozygote|Hemizygote	Allele frequency	Total number of alleles	gnomAD: Allele counts|Homozygote|Hemizygote	Genetic diagnosis/observed phenotypes of the individuals and the reported variants observed in heterozygous state		Genetic diagnosis/observed phenotypes of the individual with the identified known disease-causing variants in homozygous state		Genetic diagnosis/ observed phenotypes of the individual with the identified known disease-causing variants in hemizygous state		Remarks
							Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype	Age (years) at the time of evaluation	
Hemolytic anemia, *G6PD* deficient (favism)	XLD	300908	9|3|0	0.005422	1660	471|6|256	Microcephalic osteodysplastic primordial dwarfism, type I; 210710; AR	2	Combined oxidative phosphorylation deficiency 3; 610505; AR	9	#	#	Variable severity as well as variable penetrance were reported for *G6PD* related anemia and they corresponding variant is observed in heterozygous, homozygous and hemizygous state in gnomAD
							Noonan syndrome-10, 616564, AD	3	Methylmalonic aciduria and homocystinuria, cblC type; 277400; AR	NA			
							Unaffected individual	27	Episodic ataxia/ myokymia syndrome; 160120; AD	30			
Hemolytic anemia, G6PD deficient (favism)	XLD	300908	5|2|0	0.003012	1660	31|0|20	Myasthenic syndrome, congenital, 11, associated with acetylcholine receptor deficiency; 616326; AR	47	Two unaffected individuals	>20	#	#	Variable severity as well as variable penetrance were reported for G6PD related anemia and they corresponding variant is observed in heterozygous and hemizygous state in gnomAD
{Breast-ovarian cancer, familial, 1}, Multifactorial	AD	604370	2|0|0	0.001205	1660	58|0|0	Rickets, vitamin D-resistant, type IIA; 277440; AR	4	#	#	#	#	Onset in adulthood and the corresponding variant is also

OMIM phenotype	Inheritance pattern	MIM number (phenotype)	Refined cohort: Allele counts| Homozygote| Hemizygote	Allele frequency	Total number of alleles	gnomAD: Allele counts| Homozygote| Hemizygote	Genetic diagnosis/observed phenotypes of the individuals and the reported variants observed in heterozygous state		Genetic diagnosis/observed phenotypes of the individual with the identified known disease-causing variants in homozygous state		Genetic diagnosis/ observed phenotypes of the individual with the identified known disease-causing variants in hemizygous state		Remarks
							Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype	Age (years) at the time of evaluation	
							Osteopetrosis, 259700, AR	8					observed in gnomAD
Amyloidosis, hereditary, transthyretin-related	AD	105210	1|0|0	0.0006075	1646	1|0|0	IFAP syndrome with or without BRESHECK syndrome; 308205; XL	5	#	#	#	#	Onset in adulthood and the corresponding variant is also observed in gnomAD
Factor XI deficiency	AD	612416	4|1|0	0.0002887	1662	26|0|0	Mucopolysaccharidosis type IIIA; 252900; AR	4	Muscular dystrophy, limb-girdle; 604286; AR	8	#	#	Usually factor XI deficiency lead to mild phenotypes and severity of this condition depends on other genetic and environment factors and the corresponding variant is also observed in gnomAD
							Robinow syndrome; 616894, AD	NA					
Fabry disease	XL	301500	0|0|1	0.001206	1658	10|0|6	#	#	#	#	Unaffected family member (male)	28	Significant phenotypic heterogeneity is known for Fabry disease and the same variant is observed in gnomAD
Pyruvate kinase deficiency	AR	266200	6|1|0	0.001443	1480	833|2|0	#	#	Eiken syndrome; 600002; AR	6	#	#	Clinical variability known for pyruvate

OMIM phenotype	Inheritance pattern	MIM number (phenotype)	Refined cohort: Allele counts| Homozygote| Hemizygote	Allele frequency	Total number of alleles	gnomAD: Allele counts| Homozygote| Hemizygote	Genetic diagnosis/observed phenotypes of the individuals and the reported variants observed in heterozygous state		Genetic diagnosis/observed phenotypes of the individual with the identified known disease-causing variants in homozygous state		Genetic diagnosis/ observed phenotypes of the individual with the identified known disease-causing variants in hemizygous state		Remarks
							Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype	Age (years) at the time of evaluation	
													kinase deficiency. Mild and asymptomatic individuals reported for this condition. Also, the corresponding variant is observed in gnomAD. However, now this variant has been reported in ClinVar with conflicting interpretation of pathogenicity
Laing distal myopathy	AD	160500	1|0|0	0.0006757	1480	34|0|0	Unaffected family member	35	#	#	#	#	This variant was reported as likely pathogenic in the ClinVar version which was used for annotation and now this has been reported in ClinVar with conflicting interpretation of pathogenicity and this has been

OMIM phenotype	Inheritance pattern	MIM number (phenotype)	Refined cohort: Allele counts| Homozygote| Hemizygote	Allele frequency	Total number of alleles	gnomAD: Allele counts| Homozygote| Hemizygote	Genetic diagnosis/observed phenotypes of the individuals and the reported variants observed in heterozygous state		Genetic diagnosis/observed phenotypes of the individual with the identified known disease-causing variants in homozygous state		Genetic diagnosis/ observed phenotypes of the individual with the identified known disease-causing variants in hemizygous state		Remarks
							Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype; MIM number (phenotype); Inheritance pattern	Age (years) at the time of evaluation	Phenotype	Age (years) at the time of evaluation	
													observed in gnomAD
Hypercholesterolemia, amilial, 1, 143890	AD	143890	1|0|0	0.0006024	1660	5|0|0	Epileptic encephalopathy, early infantile; 617132; AR	8	#	#	#	#	Onset in adulthood (only bi-allelic variants lead to coronary heart disease in childhood) and the corresponding variant is also observed in gnomAD. However, now this variant has been reported in ClinVar with conflicitng interpretation of pathogenicity
Duchenne muscular dystrophy	XLR	310200	2|0|2	0.0006017	1662	25|0|23	#	#	#	#	Waardenburg syndrome, type 4C; 613266; AD Osteogenesis imperfecta, type XV; 615220; AR	8 4.5	This variant was reported as likely pathogenic in the clinvar version which was used for annotation and now this has been reclassified as benign

Table 12:

Observed carrier status of ClinVar reported pathogenic/likely pathogenic variants in the cohort

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
11	5248155	C	G	*HBB*	NM_000518.5:c.92+5G>C	P	Beta-thalassemia	32	0.012	2860
13	20763650	C	T	*GJB2*	NM_004004.5:c.71G>A	P	Deafness, autosomal recessive	31	0.012	2888
9	135205481	G	A	*SETX*	NM_015046.6:c.1504C>T	LP	Ataxia-ocular apraxia 2	13	0.004492	2896
6	135726088	CT	C	*AHI1*	NM_001134830.1:c.2988del	P	Joubert syndrome-3	11	0.003817	2884
12	88512304	A	AT	*CEP290*	NM_025114.3:c.1666dup	P	Joubert syndrome	9	0.00311	2896
12	102159106	AT	A	*GNPTAB*	NM_024312.4:c.1613-25del	LP	Mucolipidosis II	9	0.005172	1742
11	5248159	C	A	*HBB*	NM_000518.5:c.92+1G>T	P/LP	Beta-thalassemia	9	0.003136	2872
18	44140185	CTCCTCTTCT	C	*LOXHD1*	NM_144612.6:c.2913_2921del	LP	Deafness, autosomal recessive	9	0.003119	2888
7	66459197	A	G	*SBDS*	NM_016038.2:c.258+2T>C	P	Shwachman-Bodian-Diamond syndrome	9	0.003129	2878
15	45393425	TGAAC	T	*DUOX2*	NM_014080.4:c.2895_2898del	P/LP	Thyroid dyshormonogenesis	8	0.00277	2890
7	107329499	T	C	*SLC26A4*	NM_000441.1:c.1003T>C	LP	Pendred syndrome	8	0.002762	2898
15	74635368	C	T	*CYP11A1*	NM_001099773.1:c.466G>A	LP	Adrenal insufficiency, congenital, with 46XY sex reversal, partial or complete	7	0.002424	2890
1	155261709	G	A	*PKLR*	NM_000298.6:c.1456C>T	P/LP	Pyruvate kinase deficiency	7	0.003692	2440
11	22283777	T	C	*ANO5*	NM_213599.2:c.1733T>C	P/LP	Muscular dystrophy, limb-girdle, type 2L	6	0.002095	2866
12	21721886	G	A	*GYS2*	NM_021957.3:c.736C>T	P/LP	Glycogen storage disease 0	6	0.002076	2892
7	107315505	T	A	*SLC26A4*	NM_000441.1:c.716T>A	P/LP	Pendred syndrome	6	0.002072	2898
9	133374932	G	A	*ASS1*	NM_000050.4:c.1168G>A	P/LP	Citrullinemia	5	0.002053	2438
11	108186796	G	A	*ATM*	NM_001330368.1:c.641-6998C>T	LP	Ataxia-telangiectasia	5	0.001729	2894
7	117188852	T	C	*CFTR*	NM_000492.3:c.1367T>C	LP	Cystic fibrosis	5	0.001725	2900

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
4	3494833	A	AGCCT	*DOK7*	NM_001164673.1:c.*345_*348dup	P	Congenital myasthenic syndrome	5	0.001724	2902
9	37429729	G	A	*GRHPR*	NM_012203.1:c.494G>A	P	Hyperoxaluria, primary, type II	5	0.001729	2894
12	110034320	G	A	*MVK*	NM_000431.2:c.1129G>A	P	Hyperimmunoglobulin D and periodic fever syndrome	5	0.001727	2898
13	51519581	G	A	*RNASEH2B*	NM_024570.3:c.529G>A	P/LP	Aicardi-Goutieres syndrome	5	0.002422	2892
15	48519315	C	T	*SLC12A1*	NM_000338.2:c.724+547C>T	LP	Bartter syndrome	5	0.001734	2886
5	131705912	G	T	*SLC22A5*	NM_001308122.1:c.248G>T	P/LP	Carnitine deficiency, systemic primary	5	0.002142	2336
6	32006858	C	G	*CYP21A2*	NM_001128590.3:c.203-13C>G	P	Adrenal hyperplasia, congenital, due to 21-hydroxylase deficiency	4	0.002782	2878
4	187206919	G	A	*F11*	NM_000128.3:c.1432G>A	P/LP	Factor XI deficiency	4	0.002072	2898
X	153762634	G	A	*G6PD*	NM_001042351.1:c.563C>T	P/LP,_other	Glucose-6-phosphate dehydrogenase deficiency	4	0.004161	2886
11	5248160	C	G	*HBB*	NM_000518.5:c.92G>C	P	Beta-thalassemia	4	0.001384	2892
3	120393748	CT	C	*HGD*	NM_000187.3:c.175del	P	Alkaptonuria	4	0.001381	2898
22	50523196	A	AG	*MLC1*	NM_139202.2:c.135dup	P	Megalencephalic leukoencephalopathy	4	0.001386	2888
16	8941651	C	T	*PMM2*	NM_000303.2:c.710C>T	P/LP	Congenital disorder of glycosylation, type Ia	4	0.001384	2892
15	43552349	C	A	*TGM5*	NM_201631.3:c.337G>T	P	Peeling skin syndrome, acral type	4	0.001385	2890
14	81528523	C	T	*TSHR*	NM_001018036.2:c.202C>T	LP	Hypothyroidism	4	0.001385	2890
11	88924382	C	T	*TYR*	NM_000372.4:c.832C>T	P	Albinism, oculocutaneous 1	4	0.002076	2892
7	117199644	ATCT	A	*CFTR*	NM_000492.3:c.1521_1523del	P	Cystic fibrosis	4	0.001725	2900
2	44050063	G	A	*ABCG5*	NM_001348912.1:c.*16-4462G>A	P/LP	Sitosterolemia	3	0.001036	2898
1	76215194	G	A	*ACADM*	NM_001286043.1:c.898G>A	P/LP	Medium chain acyl CoA dehydrogenase deficiency	3	0.001036	2898

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
12	76741493	C	CA	*BBS10*	NM_024685.3:c.271dup	P	Bardet-Biedl syndrome	3	0.001728	2896
4	15597800	C	G	*CC2D2A*	NM_001080522.2:c.4407C>G	LP	Joubert syndrome	3	0.001034	2904
2	233407739	CCT	C	*CHRNG*	NM_005199.4:c.753_754del	P	Pterygium syndrome	3	0.001034	2902
15	45403694	T	TC	*DUOX2*	NM_014080.4:c.602dup	P/LP	Thyroid dyshormonogenesis	3	0.001043	2878
13	20763490	C	T	*GJB2*	NM_004004.5:c.231G>A	P	Deafness, autosomal recessive	3	0.001039	2890
2	69168405	GTT	G	*LMOD3*	NM_001304418.1:c.1099_1100del	P	Nemaline myopathy	3	0.001066	2816
19	36326657	T	TG	*NPHS1*	NM_004646.3:c.3115dup	LP	Congenital nephrotic syndrome, Finnish type	3	0.00104	2888
12	103248932	C	T	*PAH*	NM_000277.1:c.688G>A	P/LP	Phenylketonuria	3	0.001037	2894
6	43581755	A	G	*POLH*	NM_006502.2:c.1603A>G	LP	Xeroderma pigmentosum	3	0.001036	2898
5	131705707	G	T	*SLC22A5*	NM_001308122.1:c.43G>T	P/LP	Carnitine deficiency, systemic primary	3	0.001231	2440
9	136218930	C	CTGCAGA	*SURF1*	NM_001280787.1:c.486_491dup	LP	Leigh syndrome, due to COX deficiency	3	0.001037	2896
21	43808545	G	T	*TMPRSS3*	NM_032404.2:c.32C>A	P/LP	Deafness, autosomal recessive	3	0.00104	2886
21	43808641	C	T	*TMPRSS3*	NM_032405.1:c.323-6G>A	P	Deafness, autosomal recessive	3	0.001251	2400
21	43809044	G	A	*TMPRSS3*	NM_001256317.1:c.316C>T	LP	Deafness, autosomal recessive	3	0.001237	2428
5	94848293	C	T	*TTC37*	NM_014639.3:c.2808G>A	P	Trichohepatoenteric syndrome	3	0.001036	2898
11	89017973	C	T	*TYR*	NM_000372.4:c.1217C>T	P/LP	Albinism, oculocutaneous 1	3	0.001038	2892
7	117267591	C	T	*CFTR*	NM_000492:c.3484C>T	P	Cystic fibrosis	3	0.001641	2440
7	117180312	GC	G	*CFTR*	NM_000492:c.1029delC	P	Cystic fibrosis	3	0.001035	2900
12	53715207	G	T	*AAAS*	NM_015665.5:c.43C>A	P	Achalasia-addisonianism-alacrimia syndrome	2	0.0006916	2894
1	94463488	G	A	*ABCA4*	NM_000350.2:c.6658C>T	P	Stargardt disease	2	0.0006897	2902
1	94473277	AC	A	*ABCA4*	NM_000350.2:c.5917del	P	Stargardt disease	2	0.0006897	2902

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
1	94512499	T	C	*ABCA4*	NM_000350.2:c.2894A>G	P	Stargardt disease	2	0.0006897	2902
1	100327080	T	G	*AGL*	NM_000028.2:c.104T>G	LP	Glycogen storage disease IIIa	2	0.0006906	2898
2	241808307	A	AC	*AGXT*	NM_000030.2:c.33dup	P	Hyperoxaluria, primary, type 1	2	0.0006906	2898
9	104189856	C	G	*ALDOB*	NM_000035.4:c.448G>C	P	Fructose intolerance	2	0.0006916	2894
2	73717233	C	CT	*ALMS1*	NM_015120.4:c.8155dup	LP	Alstrom syndrome	2	0.0006897	2902
13	52520472	G	A	*ATP7B*	NM_000053.3:c.3008C>T	LP	Wilson disease	2	0.000768	2606
11	62459856	AGTGAAGTGCGC	A	*BSCL2*	NM_001122955.3:c.844_854del	P	Berardinelli-Seip lipodystrophy	2	0.0006916	2894
15	42702128	G	T	*CAPN3*	NM_173088.1:c.515-1G>T	P	Muscular dystrophy, limb-girdle, type 2A	2	0.0006969	2872
10	73544798	C	T	*CDH23*	NM_022124.5:c.5653C>T	LP	Deafness, autosomal recessive	2	0.0006911	2896
10	73567342	G	A	*CDH23*	NM_001171933.1:c.1658G>A	LP	Usher syndrome, type 1D	2	0.0006911	2896
15	49030895	C	T	*CEP152*	NM_014985.3:c.4516G>A	LP	Seckel syndrome	2	0.0006983	2866
12	88449443	A	AT	*CEP290*	NM_025114.3:c.6869dup	P	Joubert syndrome	2	0.0007052	2838
12	88471040	C	A	*CEP290*	NM_025114.3:c.5668G>T	P	Joubert syndrome	2	0.002764	2896
7	117304834	G	T	*CFTR*	NM_000492.3:c.4056G>T	LP	Cystic fibrosis	2	0.0006901	2900
17	4805974	T	TC	*CHRNE*	NM_000080.4:c.130dup	P	Congenital myasthenic syndrome	2	0.0006916	2894
15	68506515	C	G	*CLN6*	NM_017882.2:c.297+113G>C	LP	Ceroid lipofuscinosis, neuronal, 6	2	0.001326	1508
3	15497518	AG	A	*COLQ*	NM_005677.3:c.1082del	P	Congenital myasthenic syndrome	2	0.0006897	2902
1	53668099	C	T	*CPT2*	NM_001330589.1:c.338C>T	P	Carnitine palmitoyltransferase 2 deficiency	2	0.0006901	2900
2	71901372	C	T	*DYSF*	NM_001130985.1:c.5767C>T	P	Miyoshi muscular dystrophy 1	2	0.0006901	2900
19	44015618	C	T	*ETHE1*	NM_001320868.1:c.107G>A	LP	Ethylmalonic encephalopathy	2	0.000693	2888
15	80472572	G	A	*FAH*	NM_000137.2:c.1062+5G>A	P	Tyrosinemia, type I	2	0.0006974	2870

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
X	153764383	G	C	*G6PD*	NM_001042351.1:c.131C>G	P	Glucose-6-phosphate dehydrogenase deficiency	2	0.002778	2882
3	33138500	T	TA	*GLB1*	NM_000404.2:c.75+2dup	P	Gangliosidosis GM1	2	0.0006969	2872
12	102179919	TG	T	*GNPTAB*	NM_024312.4:c.441del	LP	Mucolipidosis II	2	0.000692	2892
11	5247992	CAAAG	C	*HBB*	NM_000518.5:c.126_129del	P	Beta-thalassemia	2	0.000692	2892
11	5248301	T	G	*HBB*	NM_000518.5:c.−50A>C	P	Beta-thalassemia	2	0.0008292	2414
3	120369690	G	A	*HGD*	NM_000187.3:c.365C>T	LP	Alkaptonuria	2	0.0006906	2898
X	148568514	G	A	*IDS*	NM_000202.6:c.1122C>T	P	Mucopolysaccharidosis II	2	0.0006935	2886
18	21487603	C	T	*LAMA3*	NM_001127717.1:c.6640C>T	P	Epidermolysis bullosa, junctional	2	0.0006925	2890
17	56283862	G	GTGCC	*MKS1*	NM_001330397.1:c.1274-125_1274-122dup	P/LP	Meckel syndrome	2	0.0006916	2894
1	45974001	C	T	*MMACHC*	NM_015506.2:c.394C>T	P	Methylmalonic aciduria and homocystinuria, cblC type	2	0.001381	2898
6	49419383	TGC	T	*MUT*	NM_000255.3:c.1126_1127del	P/LP	Methylmalonic aciduria, mut(0) type	2	0.0006906	2898
1	45797228	C	T	*MUTYH*	NM_001293190.1:c.1148G>A	P/LP	MUTYH-associated polyposis	2	0.0006906	2898
1	45798130	G	A	*MUTYH*	NM_001293190.1:c.682C>T	P/LP	MUTYH-associated polyposis	2	0.0006906	2898
22	42457056	C	T	*NAGA*	NM_001362848.1:c.973G>A	P	N-acetylgalactosaminidase alpha deficiency	2	0.0006935	2886
2	152353454	C	T	*NEB*	NM_001271208.1:c.24498+1G>A	LP	Nemaline myopathy	2	0.000692	2892
18	21118573	C	G	*NPC1*	NM_000271.4:c.2974G>C	P/LP	Niemann-Pick disease type C1	2	0.0006925	2890
2	220432785	G	GT	*OBSL1*	NM_001173431.1:c.1273dup	P	3-M syndrome	2	0.0006906	2898
16	8905010	G	A	*PMM2*	NM_000303.2:c.422G>A	P	Congenital disorder of glycosylation, type Ia	2	0.000692	2892
5	131726524	C	T	*SLC22A5*	NM_001308122.1:c.1267C>T	P/LP	Carnitine deficiency, systemic primary	2	0.0006906	2898
7	107334918	T	G	*SLC26A4*	NM_000441.1:c.1334T>G	P	Pendred syndrome	2	0.0006906	2898

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
7	107334921	A	G	*SLC26A4*	NM_000441.1:c.1337A>G	LP	Pendred syndrome	2	0.0006906	2898
9	140128361	AGAACAGCACA GCC CCGGCGGACAG GCTG CCCTGTGAGGCC CGG CCCACCCCAAGC CCCC TACACCCCCCAC ACTC CCCCTCACCGGC CCC TACATGGAGAG	A	*SLC34A3*	NM_080877.2:c.925+20_926-48del	P	Hypophosphatemic rickets with hypercalciuria	2	0.0006916	2894
15	43893593	C	T	*STRC*	NM_153700.2:c.4701+1G>A	P	Deafness, autosomal recessive	2	0.0008673	2308
3	48508110	T	TG	*TREX1*	NM_033629.5:c.58dup	P	Aicardi-Goutieres syndrome 1	2	0.0006892	2904
8	100865697	T	TA	*VPS13B*	NM_017890.4:c.10156dup	P/LP	Cohen syndrome	2	0.0006916	2894
3	14200140	G	A	*XPC*	NM_001354726.1:c.664C>T	P	Xeroderma pigmentosum	2	0.0006906	2898
7	117199517	G	A	*CFTR*	NM_000492.3:c.1393-1G>A	P	Cystic fibrosis	2	0.0006949	2880
13	20763685	AC	A	*GJB2*	NM_004004.5:c.35delG	P	Deafness, autosomal recessive	1	0.0004119	2430
9	34648167	A	G	*GALT*	NM_000155.3:c.563A>G	P	Galactosaemia	1	0.0006916	2894
9	136220806	A	C	*SURF1*	NM_001280787.1:c.−4-11T>G	P	Leigh syndrome, due to COX deficiency	1	0.001396	2868
1	94466627	C	T	*ABCA4*	NM_000350.2:c.6317G>A	LP	Stargardt disease	1	0.0004095	2444
1	94471025	C	T	*ABCA4*	NM_000350.2:c.6119G>A	LP	Stargardt disease	1	0.0003448	2902
1	94509018	C	T	*ABCA4*	NM_000350.2:c.3064G>A	LP	Stargardt disease	1	0.0003448	2902
1	94526230	C	T	*ABCA4*	NM_000350.2:c.2023G>A	LP	Stargardt disease	1	0.0003448	2902
1	94528819	G	A	*ABCA4*	NM_000350.2:c.1609C>T	LP	Stargardt disease	1	0.0003448	2902
1	94577093	G	C	*ABCA4*	NM_000350.2:c.203C>G	LP	Stargardt disease	1	0.0003482	2874
16	16251519	CCTT	C	*ABCC6*	NM_001351800.1:c.3538_3540del	P	Pseudoxanthoma elasticum	1	0.0003484	2872
16	16256943	C	T	*ABCC6*	NM_001351800.1:c.3071G>A	P	Pseudoxanthoma elasticum	1	0.000346	2892

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
16	16272711	C	T	*ABCC6*	NM_001351800.1:c.2017G>A	P	Pseudoxanthoma elasticum	1	0.000346	2892
16	16295902	G	A	*ABCC6*	NM_001351800.1:c.790C>T	P	Pseudoxanthoma elasticum	1	0.000346	2892
X	152991242	A	G	*ABCD1*	NM_000033.3:c.521A>G	P	Adrenoleukodystrophy	1	0.0003465	2888
3	128625054	C	T	*ACAD9*	NM_014049.4:c.1240C>T	LP	ACAD9 deficiency	1	0.0003453	2898
1	76190504	T	C	*ACADM*	NM_001286043.1:c.30+2T>C	LP	Medium chain acyl CoA dehydrogenase deficiency	1	0.0003479	2876
2	241814542	C	T	*AGXT*	NM_000030.2:c.697C>T	LP	Hyperoxaluria, primary, type 1	1	0.0003453	2898
17	19566799	C	T	*ALDH3A2*	NM_000382.2:c.1094C>T	LP	Sjogren-Larsson syndrome	1	0.0003463	2890
9	104184173	G	A	*ALDOB*	NM_000035.4:c.1013C>T	P/LP	Fructose intolerance	1	0.0003458	2894
2	73828531	CTACT	C	*ALMS1*	NM_015120.4:c.12086_12089del	LP	Alstrom syndrome	1	0.0003448	2902
2	202589115	G	A	*ALS2*	NM_020919.3:c.3415C>T	P/LP	Amyotrophic lateral sclerosis	1	0.0003451	2900
3	49459565	G	A	*AMT*	NM_001164711.1:c.90+229C>T	LP	Hyperglycinaemia, non-ketotic	1	0.0003451	2900
3	49459869	T	TA	*AMT*	NM_001164711.1:c.14dup	P	Hyperglycinaemia, non-ketotic	1	0.0003451	2900
11	22284590	G	A	*ANO5*	NM_213599.2:c.1898+1G>A	P	Muscular dystrophy, limb-girdle, type 2L	1	0.0004108	2436
4	80905984	CA	C	*ANTXR2*	NM_001145794.1:c.1074del	P	Hyaline fibromatosis syndrome	1	0.0003453	2898
22	51064581	C	T	*ARSA*	NM_001085428.2:c.721+1G>A	LP	Metachromatic leukodystrophy	1	0.0003489	2868
5	78181606	G	A	*ARSB*	NM_198709.2:c.943C>T	LP	Mucopolysaccharidosis type VI (Maroteaux-Lamy)	1	0.0003453	2898
7	65554101	A	G	*ASL*	NM_001024944.1:c.857A>G	P	Argininosuccinic aciduria	1	0.0003451	2900
11	108200944	C	A	*ATM*	NM_001330368.1:c.641-21146G>T	P	Ataxia-telangiectasia	1	0.0003837	2608
12	124203239	C	T	*ATP6V0A2*	NM_012463.3:c.187C>T	P	Cutis laxa, autosomal recessive, type IIA	1	0.000346	2892

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
2	71190325	C	T	*ATP6V1B1*	NM_001692.3:c.943C>T	LP	Renal tubular acidosis & hearing loss	1	0.0003453	2898
13	52511620	G	A	*ATP7B*	NM_000053.3:c.3895C>T	P/LP	Wilson disease	1	0.000346	2892
13	52515217	C	T	*ATP7B*	NM_000053.3:c.3556G>A	P/LP	Wilson disease	1	0.000346	2892
13	52516633	C	T	*ATP7B*	NM_000053.3:c.3301G>A	LP	Wilson disease	1	0.000346	2892
13	52518306	C	T	*ATP7B*	NM_000053.3:c.3182G>A	P	Wilson disease	1	0.000346	2892
4	123663048	A	C	*BBS12*	NM_152618.2:c.1A>C	LP	Bardet-Biedl syndrome	1	0.0003448	2902
16	56536294	G	A	*BBS2*	NM_031885.3:c.1015C>T	LP	Bardet-Biedl syndrome	1	0.0003458	2894
16	56543916	G	A	*BBS2*	NM_031885.3:c.565C>T	P	Bardet-Biedl syndrome	1	0.0003458	2894
19	41928539	C	T	*BCKDHA*	NM_001164783.1:c.856C>T	P	Maple syrup urine disease	1	0.0003463	2890
6	80982916	C	T	*BCKDHB*	NM_000056.3:c.1016C>T	P	Maple syrup urine disease	1	0.0003451	2900
6	81053406	CT	C	*BCKDHB*	NM_000056.3:c.1065del	P	Maple syrup urine disease	1	0.0003451	2900
15	91292792	AAC	A	*BLM*	NM_000057.3:c.298_299del	P/LP	Bloom syndrome	1	0.0003463	2890
3	15686795	G	C	*BTD*	NM_000060.4:c.1432G>C	P	Biotinidase deficiency	1	0.0003451	2900
15	42695077	G	A	*CAPN3*	NM_173088.1:c.86G>A	P/LP	Muscular dystrophy, limb-girdle, type 2A	1	0.000346	2892
15	42703156	G	C	*CAPN3*	NM_173088.1:c.802G>C	P/LP	Muscular dystrophy, limb-girdle, type 2A	1	0.000346	2892
1	116247829	G	A	*CASQ2*	NM_001232.3:c.923C>T	P/LP	Ventricular tachycardia, catecholaminergic polymorphic	1	0.0003453	2898
1	116260490	A	G	*CASQ2*	NM_001232.3:c.809T>C	LP	Ventricular tachycardia, catecholaminergic polymorphic	1	0.0003453	2898
4	15575830	C	T	*CC2D2A*	NM_001080522.2:c.3652C>T	P	Joubert syndrome	1	0.0003446	2904
4	15587793	G	A	*CC2D2A*	NM_001080522.2:c.3989G>A	P	Joubert syndrome	1	0.0003446	2904
12	88452645	C	T	*CEP290*	NM_025114.3:c.6798G>A	P	Joubert syndrome	1	0.0003453	2898
12	88487680	A	AT	*CEP290*	NM_025114.3:c.3175dup	P	Joubert syndrome	1	0.0003489	2868
12	88519134	G	A	*CEP290*	NM_025114.3:c.1078C>T	P	Joubert syndrome	1	0.000347	2884

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
7	117120150	T	C	*CFTR*	NM_000492.3:c.2T>C	LP	Cystic fibrosis	1	0.0003451	2900
2	233406133	CCT	C	*CHRNG*	NM_005199.4:c.401_402del	P	Pterygium syndrome	1	0.0003448	2902
15	68504182	C	CG	*CLN6*	NM_017882.2:c.316dup	P	Ceroid lipofuscinosis, neuronal, 6	1	0.0004112	2434
3	15495353	G	A	*COLQ*	NM_005677.3:c.1281C>T	LP	Congenital myasthenic syndrome	1	0.0003448	2902
11	68579904	C	T	*CPT1A*	NM_001876.3:c.281+1G>A	P	Carnitine palmitoyltransferase I deficiency	1	0.0003484	2872
8	143958607	G	A	*CYP11B1*	NM_000497.3:c.427C>T	LP	Adrenal hyperplasia, congenital, due to 11-beta-hydroxylase deficiency	1	0.0003458	2894
2	219678911	G	A	*CYP27A1*	NM_000784.3:c.1184+1G>A	P	Cerebrotendinous xanthomatosis	1	0.0003475	2880
X	32429987	G	A	*DMD*	NM_004009.3:c.4103C>T	LP	Becker muscular dystrophy	1	0.001733	2888
16	84209776	CT	C	*DNAAF1*	NM_178452.4:c.1937del	P	Primary ciliary dyskinesia	1	0.0003497	2862
7	21675609	C	T	*DNAH11*	NM_001277115.1:c.4621C>T	P	Primary ciliary dyskinesia	1	0.0003453	2898
5	13753598	C	T	*DNAH5*	NM_001369.2:c.10616G>A	P	Primary ciliary dyskinesia	1	0.0003453	2898
15	45391576	C	A	*DUOX2*	NM_014080.4:c.3515+5G>T	LP	Thyroid dyshormonogenesis	1	0.0003497	2862
19	45855493	G	A	*ERCC2*	NM_000400.3:c.2164C>T	P	Xeroderma pigmentosum	1	0.0003465	2888
19	44015606	C	T	*ETHE1*	NM_001320868.1:c.119G>A	LP	Ethylmalonic encephalopathy	1	0.0003465	2888
4	187205296	C	T	*F11*	NM_000128.3:c.1186C>T	LP	Factor XI deficiency	1	0.0003453	2898
X	154185236	T	C	*F8*	NM_000132.3:c.1748A>G	LP	Hemophilia A	1	0.0003467	2886
4	155533035	G	C	*FGG*	NM_000509.5:c.323C>G	P	Afibrinogenaemia	1	0.0003453	2898
1	241667423	G	A	*FH*	NM_000143.3:c.1027C>T	P	Fumarase deficiency	1	0.0003451	2900
9	108366537	C	A	*FKTN*	NM_001351500.1:c.15C>A	P/LP	Congenital muscular dystrophy-dystroglycanopathy	1	0.001037	2896

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
							with brain and eye anomalies			
X	153761205	C	T	*G6PD*	NM_001042351.1:c.1003G>A	P	Glucose-6-phosphate dehydrogenase deficiency	1	0.0003467	2886
X	153762710	C	T	*G6PD*	NM_001042351.1:c.487G>A	P	Glucose-6-phosphate dehydrogenase deficiency	1	0.0003467	2886
17	78086764	C	T	*GAA*	NM_000152.3:c.1978C>T	LP	Glycogen storage disease II	1	0.0003463	2890
17	78087149	C	T	*GAA*	NM_000152.3:c.2173C>T	P/LP	Glycogen storage disease II	1	0.0003465	2888
14	88401093	C	T	*GALC*	NM_001201401.1:c.1972G>A	LP	Krabbe disease	1	0.000346	2892
14	88412026	A	G	*GALC*	NM_001201401.1:c.1472T>C	LP	Krabbe disease	1	0.000346	2892
9	34647525	A	G	*GALT*	NM_000155.3:c.289A>G	LP	Galactosaemia	1	0.0003458	2894
9	34648376	C	T	*GALT*	NM_000155.3:c.610C>T	P/LP	Galactosaemia	1	0.0003458	2894
9	34648763	G	A	*GALT*	NM_000155.3:c.692G>A	P	Galactosaemia	1	0.0003458	2894
1	155207932	A	T	*GBA*	NM_001171811.1:c.493T>A	P	Gaucher disease 1	1	0.0003451	2900
1	155208006	T	C	*GBA*	NM_001171811.1:c.419A>G	P/LP	Gaucher disease 1	1	0.0003451	2900
1	155210420	C	T	*GBA*	NM_001171811.1:c.−146-552G>A	P	Gaucher disease 1	1	0.0003475	2880
19	13007058	G	A	*GCDH*	NM_013976.2:c.675G>A	LP	Glutaricaciduria, type I	1	0.0003463	2890
19	13007153	G	A	*GCDH*	NM_013976.2:c.770G>A	P	Glutaricaciduria, type I	1	0.0003463	2890
19	13008600	TG	T	*GCDH*	NM_013976.2:c.1173del	P/LP	Glutaricaciduria, type I	1	0.0003463	2890
X	100652999	C	T	*GLA*	NM_001199973.1:c.408+2554C>T	P	Fabry disease	1	0.001734	2886
9	6550845	G	A	*GLDC*	NM_000170.2:c.2527C>T	P/LP	Glycine encephalopathy	1	0.0004105	2438
9	36218221	G	A	*GNE*	NM_001190383.2:c.1670C>T	P	Inclusion body myopathy	1	0.0003455	2896
12	102147247	TGA	T	*GNPTAB*	NM_024312.4:c.3503_3504del	P	Mucolipidosis II	1	0.000346	2892
12	21715851	C	A	*GYS2*	NM_021957.3:c.1062+1G>T	LP	Glycogen storage disease 0	1	0.000346	2892
12	21727209	G	A	*GYS2*	NM_021957.3:c.547C>T	P	Glycogen storage disease 0	1	0.000346	2892

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
2	26416536	CAT	C	*HADHA*	NM_000182.4:c.1793_1794del	LP	Mitochondrial trifunctional protein deficiency	1	0.0004102	2440
11	5246970	A	C	*HBB*	NM_000518.5:c.316-14T>G	P	Beta-thalassemia	1	0.0003511	2850
11	5248200	TG	T	*HBB*	NM_000518.5:c.51del	P	Beta-thalassemia	1	0.0003494	2864
11	5248224	A	AC	*HBB*	NM_000518.5:c.27dup	P	Beta-thalassemia	1	0.0004112	2434
15	72636480	G	A	*HEXA*	NM_001318825.1:c.1561C>T	P/LP	Tay-Sachs disease	1	0.000346	2892
15	72638612	T	A	*HEXA*	NM_001318825.1:c.1418A>T	P	Tay-Sachs disease	1	0.000346	2892
15	72638920	G	GGATA	*HEXA*	NM_001318825.1:c.1307_1310dup	P	Tay-Sachs disease	1	0.000346	2892
5	74009409	C	T	*HEXB*	NM_000521.3:c.850C>T	P	Sandhoff disease, infantile, juvenile, and adult forms	1	0.001035	2900
8	43014188	G	A	*HGSNAT*	NM_001363228.1:c.493+1G>A	P	Mucopolysaccharidosis IIIC	1	0.0005513	1814
X	148585007	C	T	*IDS*	NM_000202.5:c.253G>A	P	Mucopolysaccharidosis II	1	0.001239	2424
4	996890	T	C	*IDUA*	NM_001363576.1:c.1073T>C	P	Mucopolysaccharidosis Ih	1	0.0003458	2894
11	68673577	C	T	*IGHMBP2*	NM_002180.2:c.127C>T	P	Spinal muscular atrophy with respiratory distress	1	0.0003458	2894
11	68682370	G	T	*IGHMBP2*	NM_002180.2:c.791G>T	LP	Spinal muscular atrophy with respiratory distress	1	0.0004108	2436
11	68685249	C	T	*IGHMBP2*	NM_002180.2:c.958C>T	LP	Spinal muscular atrophy with respiratory distress	1	0.0003458	2894
11	68701976	G	A	*IGHMBP2*	NM_002180.2:c.1582G>A	LP	Spinal muscular atrophy with respiratory distress	1	0.0003458	2894
19	7267654	GTAGT	G	*INSR*	NM_000208.2:c.350_353del	P	Leprechaunism	1	0.0003465	2888
17	73738661	A	G	*ITGB4*	NM_001005731.1:c.2783-2A>G	LP	Epidermolysis bullosa, junctional, with pyloric atresia	1	0.0003484	2872
6	129371234	G	A	*LAMA2*	NM_000426.3:c.283+1G>A	P/LP	Muscular dystrophy, congenital merosin-deficient	1	0.0003475	2880

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
6	129475649	GCATGCAATTGT	G	*LAMA2*	NM_000426.3:c.1032_1042del	P	Muscular dystrophy, congenital merosin-deficient	1	0.0003475	2880
6	129637234	C	T	*LAMA2*	NM_000426.3:c.3976C>T	P	Muscular dystrophy, congenital merosin-deficient	1	0.0003451	2900
1	225599113	G	A	*LBR*	NM_194442.2:c.1114C>T	LP	Pelger-Huet anomaly	1	0.0003453	2898
10	90982268	C	T	*LIPA*	NM_001288979.1:c.546G>A	P/LP	Wolman syndrome	1	0.000346	2892
1	156105059	C	T	*LMNA*	NM_005572.3:c.892C>T	P	Charcot-Marie-Tooth disease	1	0.0003451	2900
5	68728915	C	T	*MARVELD 2*	NM_001038603.2:c.1498C>T	P/LP	Deafness, autosomal recessive	1	0.0003831	2612
20	10394044	G	C	*MKKS*	NM_170784.2:c.119C>G	LP	Bardet-Biedl syndrome	1	0.000346	2892
1	45973954	T	C	*MMACHC*	NM_015506.2:c.347T>C	LP	Methylmalonic aciduria and homocystinuria, cblC type	1	0.0003453	2898
6	49416553	G	A	*MUT*	NM_000255.3:c.1420C>T	P	Methylmalonic aciduria, mut(0) type	1	0.0003453	2898
6	49419405	C	T	*MUT*	NM_000255.3:c.1106G>A	P	Methylmalonic aciduria, mut(0) type	1	0.0003453	2898
6	49419406	G	A	*MUT*	NM_000255.3:c.1105C>T	P	Methylmalonic aciduria, mut(0) type	1	0.0003453	2898
6	49427089	G	A	*MUT*	NM_000255.3:c.91C>T	P	Methylmalonic aciduria, mut(0) type	1	0.0003453	2898
1	45796890	TTCC	T	*MUTYH*	NM_001293190.1:c.1398_1400del	P	MUTYH-associated polyposis	1	0.0003453	2898
1	45796892	C	A	*MUTYH*	NM_001293190.1:c.1399G>T	P	MUTYH-associated polyposis	1	0.0003453	2898
1	45797201	G	A	*MUTYH*	NM_001293190.1:c.1175C>T	P	MUTYH-associated polyposis	1	0.0003453	2898
12	110029080	T	C	*MVK*	NM_000431.2:c.803T>C	P	Hyperimmunoglobulin D and periodic fever syndrome	1	0.0003453	2898
11	76867967	G	A	*MYO7A*	NM_001127179.2:c.652G>A	LP	Usher syndrome	1	0.0003455	2896
2	152350725	CAG	C	*NEB*	NM_001271208.1:c.24632_24633del	P	Nemaline myopathy	1	0.0003451	2900
2	152354850	T	TTTTC	*NEB*	NM_001271208.1:c.24232_24235dup	LP	Nemaline myopathy	1	0.0003451	2900

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
6	31828365	C	T	*NEU1*	NM_000434.3:c.649G>A	P/LP	Sialidosis	1	0.0004098	2442
5	156895736	C	A	*NIPAL4*	NM_001172292.1:c.470C>A	P	Ichthyosis, autosomal recessive	1	0.001231	2440
2	26700593	C	A	*OTOF*	NM_194248.2:c.2239G>T	P	Deafness, autosomal recessive	1	0.0003453	2898
12	103237484	G	A	*PAH*	NM_000277.1:c.1139C>T	P/LP	Phenylketonuria	1	0.0003458	2894
12	103245479	C	A	*PAH*	NM_000277.1:c.898G>T	P	Phenylketonuria	1	0.0003458	2894
12	103246653	C	T	*PAH*	NM_000277.1:c.782G>A	P/LP	Phenylketonuria	1	0.0003458	2894
12	103260393	T	C	*PAH*	NM_000277.1:c.490A>G	P/LP	Phenylketonuria	1	0.0003458	2894
12	103288698	T	C	*PAH*	NM_000277.1:c.169-2A>G	P	Phenylketonuria	1	0.0003482	2874
20	3899342	G	A	*PANK2*	NM_001324191.1:c.688G>A	P	Neurodegeneration with brain iron accumulation 1	1	0.0003458	2894
13	100962159	C	T	*PCCA*	NM_001352605.1:c.1426C>T	P/LP	Propionicacidemia	1	0.0003463	2890
6	51497362	C	T	*PKHD1*	NM_138694.3:c.11665+1G>A	LP	Polycystic kidney and hepatic disease	1	0.0003477	2878
22	38508566	G	A	*PLA2G6*	NM_001349869.1:c.1525C>T	P	Infantile neuroaxonal dystrophy 1	1	0.0003465	2888
16	8900255	C	T	*PMM2*	NM_000303.2:c.338C>T	P	Congenital disorder of glycosylation, type Ia	1	0.000346	2892
16	8904955	C	T	*PMM2*	NM_000303.2:c.367C>T	P	Congenital disorder of glycosylation, type Ia	1	0.000346	2892
19	50364799	AGGGGTCAGGG GAGGAGG	A	*PNKP*	NM_007254.3:c.1386+49_1387-33del	P	Microcephaly - seizures - developmental delay	1	0.0003472	2882
1	46657979	C	G	*POMGNT1*	NM_001243766.1:c.1413+1G>C	P	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies)	1	0.0003475	2880
2	128186061	G	A	*PROC*	NM_000312.3:c.925G>A	P	Thrombophilia due to protein C deficiency	1	0.0003453	2898
10	73587809	CCTT	C	*PSAP*	NM_002778.3:c.679_681del	LP	Metachromatic leukodystrophy	1	0.0003458	2894
11	112101362	C	T	*PTS*	NM_000317.2:c.200C>T	P	Hyperphenylalaninemia , BH4-deficient, A	1	0.0003458	2894

Chromosome	Co-ordinate	Reference allele	Altered allele	Gene	Variant (transcript level HGVS nomenclature)	Clinical significance (ClinVar)	Disease	Carriers	Allele frequency	Total number of alleles
11	112104157	C	T	*PTS*	NM_000317.2:c.317C>T	LP	Hyperphenylalaninemia, BH4-deficient, A	1	0.0004108	2436
14	51378873	C	T	*PYGL*	NM_002863.4:c.1768+1G>A	P	Glycogen storage disease VI	1	0.0003487	2870
11	36596040	C	T	*RAG1*	NM_000448.2:c.1186C>T	P	Omenn syndrome	1	0.000346	2892
19	12921137	C	T	*RNASEH2A*	NM_006397.2:c.556C>T	LP	Aicardi-Goutieres syndrome	1	0.0003463	2890
11	65487856	G	A	*RNASEH2C*	NM_032193.3:c.205C>T	LP	Aicardi-Goutieres syndrome	1	0.00242	2894
6	116938246	C	T	*RSPH4A*	NM_001161664.1:c.460C>T	P	Ciliary dyskinesia, primary	1	0.0003453	2898
17	48244791	C	T	*SGCA*	NM_001135697.2:c.100C>T	P/LP	Muscular dystrophy, limb-girdle, type 2D	1	0.000346	2892
17	48244792	G	A	*SGCA*	NM_001135697.2:c.101G>A	P/LP	Muscular dystrophy, limb-girdle, type 2D	1	0.000346	2892
5	149359991	C	T	*SLC26A2*	NM_000112.3:c.835C>T	P	Achondrogenesis 1B	1	0.0003453	2898
7	107334849	T	C	*SLC26A4*	NM_000441.1:c.1265T>C	LP	Pendred syndrome	1	0.0003453	2898
7	107341576	AAG	A	*SLC26A4*	NM_000441.1:c.1741_1742del	LP	Pendred syndrome	1	0.0003453	2898
7	107344785	G	T	*SLC26A4*	NM_000441.1:c.2044G>T	P	Pendred syndrome	1	0.0003453	2898
19	33353427	C	T	*SLC7A9*	NM_001126335.1:c.544G>A	P/LP	Cystinuria	1	0.0003465	2888
9	136219371	C	T	*SURF1*	NM_001280787.1:c.354G>A	P	Leigh syndrome, due to COX deficiency	1	0.0003455	2896
11	67811770	C	T	*TCIRG1*	NM_006053.3:c.331C>T	LP	Osteopetrosis, infantile malignant	1	0.000346	2892
15	43552684	C	T	*TGM5*	NM_201631.3:c.104G>A	P/LP	Peeling skin syndrome, acral type	1	0.0003463	2890
8	94777801	CAG	C	*TMEM67*	NM_153704.5:c.579_580del	P/LP	Joubert syndrome|Meckel syndrome	1	0.0003453	2898
21	43795896	C	T	*TMPRSS3*	NM_032404.2:c.895G>A	LP	Deafness, autosomal recessive	1	0.0003467	2886
21	43808633	G	A	*TMPRSS3*	NM_001256317.1:c.325C>T	P/LP	Deafness, autosomal recessive	1	0.0004122	2428
11	6635918	C	T	*TPP1*	NM_000391.3:c.1552-1G>A	LP	Neuronal ceroid lipofuscinosis	1	0.0003482	2874

Chromosome	**Co-ordinate**	**Reference allele**	**Altered allele**	**Gene**	**Variant** **(transcript level HGVS nomenclature)**	**Clinical significance** **(ClinVar)**	**Disease**	**Carriers**	**Allele frequency**	**Total number of alleles**
11	6638858	G	A	*TPP1*	NM_000391.3:c.379C>T	P	Neuronal ceroid lipofuscinosis	1	0.0004108	2436
22	46749726	G	A	*TRMU*	NM_001282784.1:c.415G>A	P/LP	Liver failure, transient infantile	1	0.0003467	2886
11	88924475	A	AC	*TYR*	NM_000372.4:c.929dup	P	Albinism, oculocutaneous 1	1	0.000346	2892
11	88961018	C	T	*TYR*	NM_000372.4:c.1064C>T	P	Albinism, oculocutaneous 1	1	0.000346	2892
11	89017960	C	T	*TYR*	NM_000372.4:c.1204C>T	P	Albinism, oculocutaneous 1	1	0.000346	2892
1	215848385	G	A	*USH2A*	NM_206933.2:c.12868C>T	P	Usher syndrome 2	1	0.0003448	2902
1	215956215	G	A	*USH2A*	NM_206933.2:c.10450C>T	P/LP	Usher syndrome 2	1	0.0003448	2902
1	215972392	G	A	*USH2A*	NM_206933.2:c.9815C>T	LP	Usher syndrome 2	1	0.0003448	2902
1	216052272	CT	C	*USH2A*	NM_206933.2:c.8391del	LP	Usher syndrome 2	1	0.0003448	2902
1	216592035	C	T	*USH2A*	NM_007123.5:c.486-14G>A	P/LP	Usher syndrome 2	1	0.0003482	2874
8	100880515	A	G	*VPS13B*	NM_017890.4:c.11291-2A>G	LP	Cohen syndrome	1	0.0003482	2874
19	36574073	G	A	*WDR62*	NM_001083961.1:c.1480G>A	LP	Microcephaly 2, primary, autosomal recessive, with or without cortical malformations	1	0.0003497	2862
9	100451874	C	A	*XPA*	NM_000380.3:c.331G>T	P	Xeroderma pigmentosum	1	0.001037	2896

Table 13.

List of monogenic disorders with at least 10 observed carriers in the cohort

Gene	Disease	MIM number	Number of carriers
HBB	Beta-thalassemia	613985	44
GJB2	Deafness	220290	35
SLC26A4	Pendred syndrome	274600	21
CFTR	Cystic fibrosis	219700	20
CEP290	Joubert syndrome	610188	16
SETX	Ataxia-ocular apraxia 2	602433	13
ABCA4	Stargardt disease	248200	12
DUOX2	Thyroid dyshormonogenesis	607200	12
GNPTAB	Mucolipidosis II	252600	12
AHI1	Joubert syndrome-3	608629	11
TMPRSS3	Deafness, autosomal recessive	601072	11
SLC22A5	Carnitine deficiency, systemic primary	212140	10
TYR	Albinism, oculocutaneous 1	606952	10

Table 14:

Efficiency of dataset from refined cohort for variant prioritization

Variant prioritization for monogenic disorders		**Variable**	**Average number of variants in a test set of 50 exomes**	**Percentage of variants getting prioritized after each filtering criteria (%)**	**Filtering efficiency of different strategies**		**Shapiro-Wilk (p-value)**
					Formula	**Value (%)**	
Number of variants called		-	110759.9	-	-	-	-
Exonic or splicing variants (+/− 20bp from the exonic boundaries)		-	35092.42	-	-	-	-
Filtering for heterozygous variants	Heterozygous variants	a	22500.02	100.00	-	-	-
	Rare variants with <1% frequency or absent in gnomAD	b	2030.92	9.03	[(a-b)/a]*100	91.0	-
	Variants that are not observed in homozygous state in gnomAD	c	3466.24	15.41	[(a-c)/a]*100	84.6	-
	Apply filters (b) and (c)	d	1178.24	**5.24**	[(a-d)/a]*100	**94.8**	0.4328
	Rare variants with <1% frequency or absent in GenomeAsia	e	8000.74	35.56	[(a-e)/a]*100	64.4	-
	Variants that are not observed in homozygous state in GenomeAsia	f	7931.4	35.25	[(a-f)/a]*100	64.7	-
	Apply filters (e) and (f)	g	7830.18	34.80	[(a-g)/a]*100	65.2	-
	Rare variants with <1% frequency or absent in refined cohort	h	1380.5	6.14	[(a-h)/a]*100	**93.9**	-
	Variants that are not observed in homozygous state in refined cohort	i	4023.64	17.88	[(a-i)/a]*100	82.1	-
	Apply filters (h) and (i)	j	1290.84	**5.74**	[(a-j)/a]*100	**94.3**	0.01752
	Apply filters (b), (c),(e) and(f)	k	1152.7	**5.12**	[(a-k)/a]*100	**94.9**	0.4134
	Apply filters (b), (c),(e), (f), (h) and (i)	l	774.22	**3.44**	[(a-l)/a)*100	**96.6**	0.1133
Filtering for presumable de-novo variants	Apply filters (h) and (i) and variants not observed in heterozygous state in the refined cohort	m	553.08	**2.5**	[(a-m)/a]*100	**97.5**	-
Filtering for homozygous variants	Homozygous variants	n	12361.54	100.00	-	-	-
	Rare variants with <1% frequency or absent in gnomAD	o	79.18	0.64	[(n-o)/n)*100	99.6	-
	Variants that are not observed in homozygous state in gnomAD	p	199.46	1.61	[(n-p)/n)*100	99.1	-
	Apply filters (o) and (p)	q	40.32	**0.33**	[(n-q)/n)*100	**99.8**	2.046e-05

Variant prioritization for monogenic disorders		**Variable**	**Average number of variants in a test set of 50 exomes**	**Percentage of variants getting prioritized after each filtering criteria (%)**	**Filtering efficiency of different strategies**		**Shapiro-Wilk (p-value)**
					Formula	**Value (%)**	
	Rare variants with <1% frequency or absent in GenomeAsia	r	1857.08	15.02	[(n-r)/n)*100	91.7	-
	Variants that are not observed in homozygous state in GenomeAsia	s	1854.7	15.00	[(n-s)/n)*100	91.8	-
	Apply filters (r) and (s)	t	1851.48	14.98	[(n-t)/n)*100	91.8	0.5727
	Rare variants with <1% frequency or absent in refined cohort	u	60.5	0.49	[(n-u)/n)*100	**99.7**	-
	Variants that are not observed in homozygous state in refined cohort	v	51.5	**0.42**	[(n-v)/n)*100	99.8	-
	Apply filters (u) and (v)	w	33.86	0.27	[(n-w)/n)*100	**99.8**	7.964e-07
	Apply filters (o), (p), (r) and (s)	x	39.5	**0.32**	[(n-x)/n)*100	**99.8**	2.933e-05
	Apply filters (o), (p), (r), (s), (u) and (v)	y	19.52	**0.16**	[(n-y)/n)*100	**99.9**	8.074e-06

Table 15:

Wilcoxon signed rank test with continuity correction

	Pairwise filters	**p-value**
Prioritization of heterozygous variants	d vs. j	1.067e-08
	d vs. k	7.693e-10
	k vs. l	7.775e-10
Prioritization of presumable de-novo variants	i vs m	5.296e-10
Prioritization of homozygous variants	q vs. t	0.0009909
	q vs. w	1.181e-05 *
	x vs. y	1.098e-09

* 26 ties are observed in the data so the exact p-values can’t be calculated