Frequency spectrum of rare and clinically relevant markers in multi-ethnic Indian populations (ClinIndb): A resource for genomic medicine in India

Ankita Narang[1], Bharathram Uppilli[1], Vivek Anand[1], Salwa Naushin[1], Arti Yadav[1], Khushboo Singhal[1], Uzma Shamim[2], Pooja Sharma[1], Sana Zahra[1], Aradhana Mathur[1], Malika Seth[1], Shaista Parveen[1], Archana Vats[1], Sara Hillman[3], Padma Dolma[4], Binuja Varma[1], Vandana Jain[5], Trisutra Consortium[1], Bhavana Parasher[1], Shantanu Sengupta[1], Mitali Mukerji[6], and Mohammed Faruq[7]

[1]CSIR Institute of Genomics & Integrative Biology
[2]Institute of Genomics & Integrative Biology (IGIB)
[3]UCL Institute for Women's Health 25 Grafton Way, London UK
[4]Sonam Norboo Memorial Hospital, Leh Ladakh
[5]All India Institute of Medical Sciences, New Delhi
[6]Institute of genomics and integrative biology
[7]Institute of Genomics and Integrative Biology

April 28, 2020

Abstract

Purpose:There have been concerted efforts towards cataloging rare and deleterious variants in different world population using high throughput genotyping and sequencing based methods. The Indian populations are underrepresented or its information w.r.t. clinically relevant variants are sparse in public datasets. The aim of this study was to estimate the burden of monogenic disease causing variants in Indian populations. Towards this, we have assessed the frequency profile of monogenic phenotype associated ClinVar variants. Methods: The study utilized genotype dataset (global-screening-array, Illumina) from 2795 individuals (multiple in-house genomics cohorts) representing diverse ethnic and geographically distinct Indian populations. Results: Of the analyzed variants from GSA, ˜12% were found to be informative and were either not known earlier or underrepresented in public databases in terms of their frequencies. These variants were linked to disorders, viz. Inborn-errors of Metabolism, Monogenic-diabetes, hereditary cancers and various other hereditary conditions. We have also shown that our study cohort is genetically better representatives of Indian populations than its representation in1000 genome project (South-Asians). Conclusion: We have created a database, ClinIndb [(http://clinindb.igib.res.in) and (https://databases.lovd.nl/shared/variants?search_owned_by_=%3D%22Mohamed%20Faruq%22)], to help clinicians and researchers in diagnosis, counseling and development of appropriate genetic screening tools relevant to the Indian populations and Indians living abroad.

INTRODUCTION

Advancements over the last decade in genetic tools and high throughput detection methods has accelerated the pace of novel genes and variants associated with monogenic Mendelian diseases. Currently 7000 OMIM phenotypes with distinct genetic etiologies have been delineated (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005). These global efforts have significantly advanced our understanding of rare genetic disorders and monogenic diseases. Though there have been significant contributions of population genomics research of Indian populations very few studies have provided a comprehensive genome level understanding of

monogenic disorders in Indian populations. In the present era of precision medicine, there is an urgent need for nationwide genomics efforts to establish - a framework for genomic medicine guided healthcare delivery needs; provide extensive coverage of genomic biomarkers across populations that facilitate rapid diagnosis and affordable genomic healthcare solutions.

India comprises of 1.3 billion people from diverse ethnic, cultural and linguistic lineages and shared ancestries with many global populations. Further, the genetic diversity of the populations has also been shaped by socio-cultural factors such as endogamy and consanguinity, geographical clines, its vast history of migration events during intercontinental exchange of trade and art as also admixtures with local population (Basu et al., 2003; Basu, Sarkar-Roy, & Majumder, 2016; I. G. V. Consortium, 2008; Reich, Thangaraj, Patterson, Price, & Singh, 2009). This provides a unique gene-variant-pool and a reservoir for founder events in recent past, extensive nationwide genomic efforts have been undertaken to understand its genetic diversity. For instance, in IGVdb a consortium level efforts have provided a catalogue of single nucleotide polymorphisms of 900 genes that map to disease associated regions across 55 diverse Indian populations (I. G. V. Consortium, 2008; Narang et al., 2010). Genetic analysis revealed that ethnicity and language are major determinants than geography. These studies highlighted that Indian populations can be divided broadly into four genetic clusters (Austro-Asiatic (AA), Dravidian (DR), Indo-European (IE) and Tibeto-Burman (TB)) based on ethno-linguistic classification. DR and IE large are known to exhibit a large degree of admixture and there are multiple sub-clusters, however, isolated populations, specifically from DR and AA group are distinct and unique (I. G. V. Consortium, 2008). In addition, mitochondrial and Y-chromosome haplogroup based studies have also helped in characterization of gene pool of diverse Indian populations (Bamshad et al., 2001; Borkar, Ahmad, Khan, & Agrawal, 2011; Kivisild et al., 2003; Majumder et al., 1999; Thanseem et al., 2006). The utility of an India specific baseline variability has been demonstrated during pre-NGS days - in infectious diseases (For example, Malaria, HIV), pharmacogenomics studies, disease associations and identification of at-risk populations for various neurological, cutaneous and high altitude adaptation related disorders (Aggarwal et al., 2015; Aggarwal et al., 2010; Bhattacharjee et al., 2008; A Biswas et al., 2007; Arindam Biswas et al., 2010; Chaki et al., 2011; Giri et al., 2014; Grover et al., 2010; Gupta et al., 2007; P. Jha et al., 2012; Kanchan et al., 2015; Kumar et al., 2009; Sinha, Arya, Agarwal, & Habib, 2009; Sinha et al., 2008; Talwar et al., 2017).

Due to limited availability of high throughput platforms systematic efforts to understand the spectrum of Mendelian and monogenic variants have not carried out across the diverse Indian populations. With the advent of NGS, Indian other global research groups have put in additional efforts to provide variant information at the genome wide scale - SAGE (South Asian Genome and Exome) (Hariprakash et al., 2018), South Asian genomes from 1000 Genomes Project (G. P. Consortium, 2015), south Indians individuals (INDEX-db) (Ahmed P et al., 2019) and a few others. The Indian Genetic disease database v1.0 provides information on 1000 genetic disease in over 3500 Indian patients (#IGDD). Other noteworthy contributions have been made in the genetics of hemoglobinopathies (thalassemia and sickle cell anemia), Duchenne Muscular Dystrophy (DMD), cystic fibrosis (CF), spinocerebellar ataxias, Mitochondrial disorders, cardiomyopathies (Pradhan et al., 2010). There is now also representative knowledgebase of Indian genetic disorders that aggregate information from NGS and single sequencing based multiple case reports studies in Lysosomal storage disorders, skeletal dysplasias and disorders of primary immunodeficiencies, genodermatosis and other neurogenetic ailments (http://guardian.meragenome.com/). A recently published GenomeAsia 100k Project (GAsP) data provided a comprehensively covered genome level data of over 1700 individuals from different Asian countries, thus highlighting the need for adequate representation of Asian genome level information in public databases (GenomeAsia100K Consortium, 2019).

Multiple country wide efforts are ongoing from government funded basic and translational genomic research laboratories, genetics unit of tertiary hospitals and commercial enterprise to meet the needs of clinical genetics segment of healthcare system in India. Despite these there are a few unmet challenges for implementation of genomics medicine in Indian populations. Primarily, either **due to** lack of representation of different ethnic populations of India or low sample size in earlier studies conducted in Indian populations. Therefore, we have 1.) paucity of knowledge for mutations spectrum and their frequencies, 2.) lack of systematic characterization

of known pathogenic mutations linked to various monogenic disorders, 3.) scarcity of knowledge of genetic spectrum of 7000 OMIM phenotypes and other prevalent genetic disorders, 4.) characterization of novel mutations.

To address these issues primarily, our study provides a comprehensive catalogue of monogenic disease linked variants in diverse Indian populations (n=2795). Our study utilized a high throughput and affordable genomics tool that provides information of over 19,538 global clinical annotated variants using Global Screening Array (GSA) from Illumina. In brief, the content of our study is novel and unique as : i) it covers diverse multiethnic Indian cohorts with large sample size of 2795 healthy subjects, ii) provides frequency distribution of known pathogenic variants for Inborn errors of Metabolism, hematological disorders and other Mendelian disorders in Indian populations, (iii) representation of SAS pathogenic variants is higher in our study i when compared with other global repositories like 1000 Genome populations (G. P. Consortium, 2015), The Genome Aggregation Database (gnomAD) (K. Karczewski & Francioli, 2017) and The Exome Aggregation Consortium (ExAC) (K. J. Karczewski et al., 2016) and GenomeAsia100K (GenomeAsia100K Consortium (2019). We have created a unique database to catalogue and register the information of clinically relevant variants for Indian population. Further, we were able to demonstrate that our cohort is genetically much more diverse than representative South Asian populations in 1000 genome dataset to provide opportunities and gaps for future research.

MATERIAL AND METHODS

Study subjects/samples details

We have analyzed the frequencies of ClinVar reported pathogenic and likely pathogenic variants from genome-wide genotyping data (Global screening array, Illumina inc.) of 3132 Indian subjects who were part of four different in-house GWAS based cohort studies. The subjects included in these cohorts represent diverse ethnic and geographic background of India. These cohorts were – 1.) TRISUTRA Ayurgenomics cohorts (n= 958), 2.) CARDIOMED cohort (n= 1449), 3.) HAP (Hypoxia adaptation and pregnancy outcome) study cohort (n=438) and 4.) GOMED study cohort (n=287).

Since, this study aimed at curation and frequency analysis of Mendelian and mongenic phenotype associated variants in Indian populations from ongoing genomics based cohort studies (in-house) for various non-mendelian or polygenic or other related physiological conditions. The genotype datasets from following cohorts were analyzed: i) TRISUTRA Ayurgenomics cohorts included individuals from Indo-European (IE) and Dravidian (DR) linguistic lineages from IE-North (CBPACS), DR-South (KLE), IE –West (IPGTRA) and IE-East (JBR) for genetic study of various health and related other non-mendelian morbid conditions. There was near equal representation of both genders and the age group of the subjects were from 19-40 years (Prasher et al., 2017). ii) CARDIOMED cohort included 1449 subjects recruited for case control based GWAS study for coronary artery disease and it comprised897healthy individuals and 552patients of coronary artery disease (Table-S1 for cohort details). iii) The HAPS study included subjects from high altitude Tibeto -Burman lineage (TB) from Leh and Indo-European IE lineage from North Indian regions (AIIMS, Delhi). iv) GOMED study: to understand the utility of high throughput genotyping chip in clinical setting, 287 subjects referred from clinicians for genetic investigations of various hereditary disorders were included. Data about gender, age and disease status of samples included in Cardiomed and GOMED cohort is provided in **Table S1** .

Genotyping, Data processing and Quality control (QC)

Genotyping was performed in the Illumina iScan system using the high throughput Infinium Global Screening Array version 1.0 that contains 642,824 genome-wide probes. Experiments were performed as per the manufacturer's protocol. This chip has representation of clinically important markers from ClinVar, GWAS and pharmacogenomics information from PharmaGKB. (**Figure-1**)

To reduce genotype calling errors, we tried to follow best practices for illumina data processing. Raw data for 3,132 samples were loaded in the GenomeStudio version 2.0 for clustering and calling genotypes. Out

of 3,132, there were 2,976 samples that passed genotype call rate of >=0.95. Further, we selected 612,322 SNPs with GenTrain score >=0.7. SNPs with GenTrain >=0.7 are expected to be clustered correctly. Data filtered through GenomeStudio criteria was processed by zcall (Version3.4). Zcall is a variant caller that was suited for calling rare SNPs. We used z-score threshold of 8 and retrieved output in the PLINK format, preferable for QC and analysis (Goldstein et al., 2012).

PLINK (v1.9) was used (Chang et al., 2015) to filter out the variants and samples with 10% missing values, additional 1,373 SNPs were removed while there was no exclusion at sample level. We removed SNPs that significantly deviate from Hardy Weinberg equilibrium (HWE). 12,970 variants were excluded with p-value $<10^{-6.}$ We applied HWE filter only for common SNPs (–maf 0.05) as it is not appropriate to use this filter for rare SNPs. In addition, we also checked for relatedness among samples using –genome function. For this, autosomal SNPs with –maf 0.1 were LD pruned to estimate proportion of Identity-by-descent (IBD) between two individuals. We removed 181 individuals with PI_HAT >0.2, an estimate of IBD. Further, exclusion of samples was based on the rate of missing genotypes. Individual with high rate of missing genotypes was excluded among the two. After all QC, final set included 597,979 variants in 2795 individuals for which frequency was computed. Allele frequencies of the variants were computed using –freq option in Plink. Frequency was computed with respect to the alternate allele defined in the 1000 genomes (http://grch37.ensembl.org/Homo_sapiens/Info/Index). An in-house developed perl script was used to count the homozygous (ref/alt) and heterozygous genotypes.

Assessment of genetic structure/architecture of subjects

Principal component analysis (PCA) was performed using smartpca module in EIGENSOFT package to elucidate as well as to compare the genetic structure of our study cohorts with populations in the1000 genomes project (Price et al., 2006).. For genetic mapping of our study subjects w.r.t diverse Indian ethnic and linguistic groups, we used genome-wide reference dataset of 471 healthy individuals genotyped on OMINI array, Illumina Inc. (Data unpublished) as a representation of Indian genomic diversity. These samples were collected as part of Indian Genome Variation (IGV) Consortium study. We compared the representation of Indian genomic diversity with our study samples as well as with 1000 genomes data (n=2,504).

To perform PCA analysis, we used 161,484 markers common across three datasets i.e. samples in this study, 1000 genomes as well as reference IGV data. PLINK was used to merge the data for common markers. ggplot package in R was used to create customized PCA plots. F_{ST} values from smartpca results were also used to compare genetic differentiation among populations.

Mapping of clinically important / relevant variants in GSA with ClinVar database

We mapped the variants genotyped in our subjects with the variants in ClinVar database (ftp://ftp.ncbi.nlm.nih.gov/pub/clinv delimited/, downloaded on April 1, 2019 (Landrum et al., 2015) . Both the coordinates and dbSNP RSID names were used. In case of multi-allelic variants, we retained only those alleles with exact matches in ClinVar. After manual QC, we selected 19,538 variants (SNPs and Indels) with alternate allele frequency [?]0.05. As the focus of our analysis was only rare and clinically relevant variants we further narrowed our query to only pathogenic and likely pathogenic variants. To retrieve these variants, we used clinical significance value of 1 given in ClinVar database and then applied keyword filter of "Pathogenic or Likely pathogenic". Pathogenic or likely pathogenic variants are designated as pathogenic throughout this manuscript.

Variants with keywords "conflicting" and "no or uncertain interpretation" of pathogenicity and other such keywords as "uncertain significance, association, risk factor, affects" were selected and analysed using a combination of three tools to ascertain their effect. We used CADD scores, Polyphen_DIV and SIFT predictions from ANNOVAR (Wang, Li, & Hakonarson, 2010). A score of 3 has been assigned Variant of Uncertain Significance if all three tools predict pathogenicity with following criteria - deleterious in SIFT, Probably Damaging (D) in Polyphen , >=20 CADD and this we classified as (VUS-I). A score of 2.5 was assigned if the variant is deleterious in SIFT, Possibly Damaging (P) n Polyphen , >=20 CADD and was assigned as VUS-II.

Annotation of genes and variants associated rare and complex disorders: Inborn errors of metabolism (IEM), MODY, Cystic fibrosis, hereditary cancers and other hereditary conditions using different resources.

1. Genes associated with different IEM classes were retrieved from The Monarch Initiative database (https://monarchinitiative.org/) (Mungall et al., 2016). 419 unique genes for IEM related to four classes- carbohydrate, amino acid, thyroid and energy metabolism as well as subclasses defined under different every IEM class is provided in **Table S2** and**Figure S1** .
2. Maturity onset diabetes of the young (MODY) associated genes: This data is compiled from two sources. Source A – DiabetesGenes (https://www.diabetesgenes.org/tests-for-diabetes-subtypes/a-new-test-for-all-mody-genes/) houses 33 genes, implicated in MODY or its related form like MIDD (maternally inherited diabetes and deafness) or partial lipodystrophy and Source B: *Fidrous et al. 2018* compiled and classified genes into 14 MODY subtypes (Firdous et al., 2018). **Table S3** provides annotation of 35 genes associated with MODY.
3. Germline Variants in Hereditary cancers: List of 851 Genetic variants in 99 cancer predisposing genes that are associated with hereditary cancers is provided in the study by *Huang et al.* **Table S4**
4. Genetic Variants associated with Cystic Fibrosis **Table S5** : CFTR2 (https://www.cftr2.org/) database which reports pathogenic variants in cystic fibrosis transmembrane conductance regulator (CFTR) gene from 88,664 patients (Sosnay et al., 2013). Data was downloaded from - https://www.cftr2.org/sites/default/files/CFTR 11March2019%20%281%29.xlsx. We prioritized 28 pathogenic variants from cystic fibrosis transmembrane conductance regulator (CFTR) gene. This included classical Cystic Fibrosis (CF) causing Phenylalanine 508 (F508) deletion (rs113993960) which has ˜70% frequency in CFTR2 database. To investigate the haplotype origin of most common F508del mutation in CFTR gene, we performed haplotype analysis using genotype data on 4389 variants from 1000 genomes project. These genotype datasets were divided separately for the four major group of populations. We first selected those variants (209) that have frequency of [?]0.05 in European populations. Tagger was used to identify tag SNPs and we also included less frequent F508del variant with tag SNPs to identify the segregation of this variant on different haplotype backgrounds. The frequency of the inferred haplotypes was estimated using PHASE algorithm (Stephens, Smith, & Donnelly, 2001)**Table S6** .
5. Among other hereditary conditions, variants with high occurrence ([?]5) were analyzed for disorders viz. Neurological and other neuromuscular disorders, Cardiac disorders, Cornelia de Lange syndrome and other syndromic disorders.
6. We also shortlisted 30 variants relevant from pharmacogenomics perspective which are tagged with the keyword "drug response" in ClinVar (**Table S7**).

Database Implementation

The ClinIndb browser was designed using open source tools. The front end was developed using HTML5, PHP 5, Bootstrap.3.2.1, Javascript and Jquery. Highcharts was used to plot the graphs. Annotation, and frequency information of all variants are stored in mysql database. We used PHP to retrieve the information. Out of common 19,538 markers between ClinVar and GSA, 9853 markers belong to the pathogenic or likely pathogenic class.

RESULTS AND DISCUSSION

Genetic diversity analysis of the study cohorts

In this analysis, PCA used to compare and assess the genetic diversity of our study cohorts with respect to the 1000 genomes and IGV populations (**Figure-S2**). As expected, 1000 genomes European (EUR,) American(AMR) and African (AFR) super populations are distant while SAS is proximal to majority of the IGV large populations. Though TB group is closer to EAS super population than any other super population of 1000 genomes as well as IGV populations**(Figure-2).** OG-W-IP (an outgroup population of African descent), which was earlier demonstrated to be an admixed Indo-African population from western part of Indian is present in a cline between Indian and African populations (Narang et al., 2011; Shah et al., 2011). Further, we excluded the 1000 genomes AFR, AMR and EUR super populations as well as the

Indian outgroup population (OG-W-IP) to fine map genetic structure. We clearly observed that majority of the IE and DR large populations are proximal to the 1000 genomes SAS group (1kg_SAS). However, AA and DR isolated populations as well as TB genetic cluster are under-represented in the 1000 genomes. EAS group (1kg_EAS) in 1000 genomes is genetically distinct from populations in TB cluster (F_{ST}=0.01-0.02) **(Figure S3)** . Underrepresentation of Indian genomic diversity in 1000 genomes was earlier reported and also substantiated our findings (Sengupta, Choudhury, Basu, & Ramsay, 2016). Also, recently published GAsP project lacks representation from TB group and moreover, has comparatively less number of samples in SAS group (n=724) which might bias frequency estimations in SAS group.

Lastly, we compared the genetic diversity of our study cohorts with IGV populations as well as 1000 genomes SAS and EAS group. **Figure 3**shows we have representation of IE and DR large populations as well as from TB group (high altitude populations) in our cohorts. Representation of AA and DR isolated groups in our study samples is also lacking. However, F_{ST} analysis suggests that our study cohorts are more proximal to IGV populations than 1Kg_SAS. More specifically, AA and DR isolated groups as well as TB low altitude populations are genetically more closer to our study cohorts than 1kg_SAS

(Figure 3).

Spectrum of known pathogenic mutations in Clindb and other genomic databases

We have created a resource "Clindb" that houses frequency spectrum of known 9853 pathogenic variants (out of 19,538 mapped variants in ClinVar) in diverse Indian populations. Frequency distribution of 9853 pathogenic variants in Clindb was compared with SAS groups in 1000 genomes, gnomAD, ExAc and GAsP. **Figure 4a** – shows that Clindb has maximum unique variants (1128) with frequency of pathogenic variants in comparison to other databases. This number remains higher even if include variants with number of carriers >1. This necessitates the use of large and diverse cohorts from Indian populations in further genomic studies.

To evaluate the reliability of frequency estimates of Clindb, we compared the average frequency difference between Clindb and other databases under study and, also, compared the average frequency difference of of GAsP with other databases. Our analysis revealed that overall frequency difference between Clindb is lower than GAsP (**Figure 4b**).

We found 12 genes with carrier frequency [?] 1% (**Table S8**). MBL2, CBS and ZGRF1 are top genes with highest frequencies. Apart from cystathionine beta-synthase (CBS gene, category: Inborn errors of amino acid metabolism), there are few other genes with high carrier frequency that are related to different IEM classes. The distribution of variants in different IEM categories is discussed below.

Clinically relevant variants in Indian populations

We have analysed and compared pathogenic variants with [?] 5 carrier frequency in genes related to different phenotypic classes: Inborn errors of metabolism, Maturity onset diabetes of the young (MODY), Cystic fibrosis and hereditary cancers.

Inborn errors of metabolism: We have compiled genes related to thyroid, carbohydrate, energy and amino acid IEM classes. Table – provides list of pathogenic variants in different IEM classes. Majority of these disorders are treatable. Therefore, inclusion of variants implicated in IEM disorders in genetic screening might benefit patients by early treatment or dietary interventions. In amino acid metabolism, a pathogenic variant (rs5742905) in cystathionine beta-synthase (CBS) gene is associated with homocystinuria. This variant has highest carrier frequency of 4% (f=0.04) in our cohorts while other databases doesn't report any frequency for this variant. Frequency of another frequent variant (rs13078881) in BTD gene (f=0.03) is similar in our cohorts as well as SAS populations in other global databases. This variant is associated with Biotinidase deficiency whose partial deficiency is reported to be higher and is majorly asymptomatic in nature. In thyroid metabolism, a variant in SLC2A4 has frequency of 0.01% frequency in Clindb as well as SAS populations in other databases under study. This variant is associated with thyroid hypoplasia.

In carbohydrate metabolism, a variant (rs267606858) in GYG1 gene has frequency of 3% (frequency (f) =

0.03) in our cohorts while SAS populations in global resources reported has frequency of 0% Mutations in GYG1 are associated with glycogen storage disease XV disorders. Similarly, in energy metabolism, variant rs148639841 in ACAT1 gene has frequency of 0.04% (f=0.004) in Clindb while there is no representation from global populations.

1. MODY: Maturity-onset diabetes of the young (MODY), one of the diseases inherited as an autosomal dominant trait. This form of diabetes is underrepresented either because of misdiagnosis or sometimes it remains undetected (Nair, Chapla, Arulappan, & Thomas, 2013). It has high prevalence rate in Asian Indians, is genetically heterogeneous and has many subtypes (Table 4). We screened 14 genes related to 14 MODY subtypes. We found three variants in GCK gene (Maturity-onset diabetes of the young, type 2) with our defined frequency criteria. These variants having frequency ranging from 0.0009-0.04 have frequency in our study with absence in global databases. Majority of the cases are known to be have either GCK or HNF1A mutation in European populations. However, a recent study highlighted that they observed GCK mutations in <1% of Indian patients and HNF1A is commonly mutated among patients (˜7%) (Mohan et al., 2018). However, GCK mutations are more represented in our cohorts than reported earlier. It is possible that the earlier studies have been conducted on fewer samples. Equally possible is that the so-called pathogenic variants are non-consequential and does not necessarily need pharmacological intervention and therefore, remains underrepresented in patients.
2. Cystic Fibrosis: We have prioritized 6 pathogenic variants from cystic fibrosis transmembrane conductance regulator (CFTR) gene. This included classical Cystic Fibrosis (CF) causing Phenylalanine 508 (F508) deletion (rs113993960, f=0.0016) which has ˜70% frequency in CFTR2 database. It causes misfolding of CFTR and is known to be the most common cause of autosomal recessive Cystic Fibrosis (CF). Our analysis suggests that other variants in CFTR are as frequent as this classical deletion in our cohorts (f=0.001-0.002). Representation of these variants is either less or nearly absent in global databases expect rs193922500 (f=0.001-0.03 in SAS groups in global databases). The CFTR variants are also linked to other phenotypes. For example, pancreatic insufficiency, male infertility and sinopulmonary disease. Heterogeneity in CFTR phenotypes depends upon the type of mutations in CFTR gene (Noone & Knowles, 2001). Mutations that doesn't lead to complete loss of function has less severe phenotypes as compared to classical cystic fibrosis, where there is a complete loss of function of CFTR gene.
3. Hereditary Cancers: On mapping 853 pathogenic genetic variants (in 99 cancer predisposing genes) from 33 cancer subtypes in 10,389 cases with our dataset, we retrieved set of 18variants (13 genes) in our study. Number of carriers in these variants' ranges from n=1 to n=5, which is less than our described threshold. Number of carriers for variants in GJB2 (Deafness related) and CHEK2 (Familial cancer of breast) genes are 5 and 4 respectively. On mapping 99 cancer predisposing genes, we retrieved 26other pathogenic variants with high carrier frequency in ALK, Neuroblastoma (f=0.01), MAP2K2, multiple tumor types (f=0.007) and BRAC2, Breast-ovarian cancer, familial 2 (f=0.005) and GJB2, deafness related (f=0.005).
4. Variants in gene related to other system disorders: the observed data of the monogenic disease related genes responsible for other critical organs, brain, heart etc (**Table-2**). Thus, it highlights the gap in the knowledge and their true occurrence of the genetically confirmed cases of various rare childhood and adult onset Mendelian disorders in Indian population. For instance, 133 heterozygous occurrences of 11 pathogenic variants in genes related to cardiac arrhythmias (Channelopathy genes, *KCNH2* , *KCNQ1* and*SCN5A*), cardiomypathies (*MYBPC3* and *TNNT2*) were observed. Seven variants with 145 heterozygous count in various muscular dystrophy genes; mainly in *CAPN3* and *SELENON*were observed. In addition multiple heterozygous calls were observed in the genes linked to various syndromic, neurological, renal and hemoglobinpathies disorders. Thus this frequency data of multiple rare pathogenic variants necessitates the relevant identification of patients carrying such defects in clinics and also allow appropriate mutation centric approach for rapid diagnosis.

Validation

Validation of genotying results were done by doing GSA-genotyping of selected smaples in replicates (n=13)

and also by concordance analysis with the whole exome sequencing data (n=48), out of which two samples has been eliminated due to poor genotyping quality score HWE significance test. The concordance analysis between GSA and Exome data, 1274 a common set of variants have shown, 99.69% genotyping concordance

Table S10.

ClinIndb database

We have developed this database to facilitate clinicians and researchers using various search options such as gene, location, rsid, batch gene, batch rsid to explore the database. Links to UCSC genome browser (hg19), ExAC, 1000 Genome and dbSNP has also been provided. There are 1974 unique variants of VUS category which have frequency in our cohorts. This list is provided in supplementary as well as in database**Table S11.** Variants of uncertain significance were also evaluated and cataloged in the database for users.

Clinical Utility

GSA based genetic investigations of patients samples which were referred for various clinical diagnosis of rare genetic diseases under GOMED cohort, yielded us a very low positivity rate for known clinical markers on GSA chip. In total for only 9 of 287 (3%) patients' samples we could arrive at the diagnosis (**Table-S12**). This could be due to low representation of clinical variants for the referred clinical diagnosis and or low abundance of Indian specific mutations on GSA chip. For clinical genetic application, GSA could have further application for common pathogenic variants identified from our overall cohort (**Table-2**)

DISCUSSION

In this study, we cataloged and estimated the mutational burden of known pathogenic or clinically relevant variants in different Indian populations. Our study showcases its applicability by evaluating the mutational load in rare and complex disorders in Indian populations. As an elaborated example, we have analysed frequency spectrum of mutations in cystic fibrosis and different classes of Inborn errors of metabolism in Indian context. In addition, our study provides, a comprehensive knowledge of underrepresented prevalent and common genetic variants in hereditary cancer associated genes, monogenic diabetes and Neuro and Neuromuscular disorders.

This study suggests that pathogenic variants with high carrier frequency are important candidates for prioritization in genetic testing. On the other hand, absence of frequency for ˜88% of known pathogenic variants in Indian populations demonstrate the need of cataloging and including population specific rare pathogenic variants in public databases as well as in commercial assays or genotype chips. Importantly, this study is an a priori guide for conducting genetic screening studies which will benefit clinicians and researchers for decision making as well as may aid in reducing genetic screening costs using informative / uninformative estimates from our catalog.

This study also highlighted the fact that reliable estimates of carrier frequency are required to estimate the real mutation load which cannot be accurately estimated from literature based cataloging as well as only patient data.

In addition to pathogenic variants, we also evaluated the effect of those variants whose clinical significance is uncertain in ClinVar. We used consensus of three predictive tools to ascertain their effect. Out of 5,984 such variants, there are 3,751 variants that are predicted to be detrimental. These variants are not part of the main analysis but a list is provided in the database for the users. In addition, we also cataloged variants that are important from pharmacogenetics perspective

(Table S8).

To benefit clinicians and researchers with this knowledge, we developed a compendium named ClinIndb which houses frequency data of clinically relevant variants in diverse Indian populations. Cohorts included in this study are closer representatives of IGVC populations than any world population even SAS group in 1000 genomes and GAsP. Therefore, frequency estimates from this study are anticipated to be more reliable.

There were earlier efforts to create such catalogs however they have their own limitations. Indian genetic disease database (http://www.igdd.iicb.res.in/home.htm) houses data of 6647 mutations from 52 diseases in 5760 individuals (Pradhan et al., 2010). This data was collated from literature published during 1993-2010 as well as personal communications. These individual studies might suffer from biases. Importantly, data is collated from patients, therefore carrier frequency estimates cannot be computed. Frequency comparison with other global populations is absent. Moreover, there has been exponential growth of clinically relevant variants after the advent of next generation sequencing and in India, this growth is quite evident after 2010 and this database is not updated yet.

The data content of our study has direct implications for evolving rapid genetic diagnostics and in determining clinically actionable variants in patients with suspected genetic ailments.

CONFLICT OF INTERST

Authors declare that there is no conflict of interest

Data Sharing: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions

ACKNOWLEGEMNTS

Authors acknowledge the funding from grants - MLP1601, MLP1802, BSC0122,

HAPS Cohort – We acknowledge the contribution of Mitali Mukerji and Bhavana Prasher in Sample collection and coordination from Leh and AIIMS. For Sample QC, quantitation, genotyping and analysis, Mitali Mukerji, Aniket Bhattacharyya and Kalpana Kunj, Sangeeta Khanna are duly acknowledged.

TRISUTRA Ayurgenomics Cohort – Contribution of individuals (mentioned below) is duly acknowledged for their respective work.

1. Bhavana Prasher and Mitali Mukerji for coordination of the study
2. TRISUTRA Ayurgenomics consortium for resource development from field studies
3. Binuja Varma, Bhavana Prasher, Mitali Mukerji for field establishment, sample identification and collection
4. Binuja Varma and Arti Yadav for sample collection, coordination, processing, DNA repository, quantitation
5. Archana Vats, Khushboo Singhal and Sangeeta Khanna for GSA array genotyping and analysis
6. Mitali Mukerji for supervising the genomic study and Bhavana Prasher for supervising the field study

Indian Genome Variation Data – Contribution of individuals (mentioned below) is duly acknowledged for their respective work.

1. Indian Genome Variation Consortium and Mitali Mukerji for resource development and management
2. Mitali Mukerji, Binuja Varma, Roshni Thomas, Anubhuti Triparthi and Ankita Narang - OMNI data genotyping and analysis

We acknowledge IARI for their genotyping facility. CSIR supported MLP901 for all the financial assistance as a project grant and I would give the grant no for Leh once I hear from Sara.

REFERENCES

Aggarwal, S., Gheware, A., Agrawal, A., Ghosh, S., Prasher, B., & Mukerji, M. (2015). Combined genetic effects of EGLN1 and VWF modulate thrombotic outcome in hypoxia revealed by Ayurgenomics approach. *Journal of translational medicine, 13* (1), 184.

Aggarwal, S., Negi, S., Jha, P., Singh, P. K., Stobdan, T., Pasha, M. Q., . . . Mukerji, M. (2010). EGLN1 involvement in high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda.*Proceedings of the National Academy of Sciences, 107* (44), 18961-18966.

Ahmed P, H., More, R. P., Viswanath, B., Jain, S., Rao, M. S., Mukherjee, O., & Consortium, A. (2019). INDEX-db: The Indian Exome Reference Database (Phase I). *Journal of Computational Biology, 26* (3), 225-234.

Alves, M. M., Sribudiani, Y., Brouwer, R. W., Amiel, J., Antiñolo, G., Borrego, S., . . . Garcia-Barcelo, M.-M. (2013). Contribution of rare and common variants determine complex diseases—Hirschsprung disease as a model. *Developmental biology, 382* (1), 320-329.

Azad, A. K., Sadee, W., & Schlesinger, L. S. (2012). Innate immune gene polymorphisms in tuberculosis. *Infection and immunity, 80* (10), 3343-3359.

Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B., . . . Rasanayagam, A. (2001). Genetic evidence on the origins of Indian caste populations. *Genome research, 11* (6), 994-1004.

Bancone, G., Chu, C. S., Somsakchaicharoen, R., Chowwiwat, N., Parker, D. M., Charunwatthana, P., . . . Nosten, F. H. (2014). Characterization of G6PD genotypes and phenotypes on the northwestern Thailand-Myanmar border. *PloS one, 9* (12), e116063.

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., . . . Bhattacharyya, N. P. (2003). Ethnic India: a genomic view, with special reference to peopling and structure. *Genome research, 13* (10), 2277-2290.

Basu, A., Sarkar-Roy, N., & Majumder, P. P. (2016). Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proceedings of the National Academy of Sciences, 113* (6), 1594-1599.

Bhattacharjee, A., Banerjee, D., Mookherjee, S., Acharya, M., Banerjee, A., Ray, A., . . . Consortium, I. G. V. (2008). Leu432Val polymorphism in CYP1B1 as a susceptible factor towards predisposition to primary open-angle glaucoma. *Molecular vision, 14* , 841.

Biswas, A., Maulik, M., Das, S., Consortium, I. G. V., Ray, K., & Ray, J. (2007). Parkin polymorphisms: risk for Parkinson's disease in Indian population. *Clinical genetics, 72* (5), 484-486.

Biswas, A., Sadhukhan, T., Majumder, S., Misra, A. K., Das, S. K., Ray, K., . . . Consortium, I. G. V. (2010). Evaluation of PINK1 variants in Indian Parkinson's disease patients. *Parkinsonism & related disorders, 16* (3), 167-171.

Borkar, M., Ahmad, F., Khan, F., & Agrawal, S. (2011). Paleolithic spread of Y-chromosomal lineage of tribes in eastern and northeastern India. *Annals of human biology, 38* (6), 736-746.

Cao, Y., Wang, X., Cao, Z., Wu, C., Wu, D., & Cheng, X. (2018). Genetic polymorphisms of MBL2 and tuberculosis susceptibility: a meta-analysis of 22 case-control studies. *Archives of medical science: AMS, 14* (6), 1212.

Chaki, M., Sengupta, M., Mondal, M., Bhattacharya, A., Mallick, S., Bhadra, R., & Ray, K. (2011). Molecular and functional studies of tyrosinase variants among Indian oculocutaneous albinism type 1 patients. *The Journal of investigative dermatology, 131* (1), 260-262.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4* (1), 7.

Consortium, G. P. (2015). A global reference for human genetic variation. *Nature, 526* (7571), 68.

Consortium, I. G. V. (2008). Genetic landscape of the people of India: a canvas for disease gene exploration. *Journal of genetics, 87* (1), 3-20.

Daya, M., Van der Merwe, L., Van Helden, P. D., Möller, M., & Hoal, E. G. (2015). Investigating the role of gene-gene interactions in TB susceptibility. *PloS one, 10* (4), e0123970.

Firdous, P., Nissar, K., Ali, S., Ganai, B. A., Shabir, U., Hassan, T., & Masoodi, S. R. (2018). Genetic testing of maturity-onset diabetes of the young current status and future perspectives. *Frontiers in endocrinology, 9* .

GenomeAsia100K Consortium (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia.Nature. Dec;576(7785):106-111.

Giri, A. K., Khan, N. M., Grover, S., Kaur, I., Basu, A., Tandon, N., . . . Brahmachari, S. K. (2014). Genetic epidemiology of pharmacogenetic variations in CYP2C9, CYP4F2 and VKORC1 genes associated with warfarin dosage in the Indian population.*Pharmacogenomics, 15* (10), 1337-1354.

Grillet, N., Schwander, M., Hildebrand, M. S., Sczaniecka, A., Kolatkar, A., Velasco, J., . . . Kimberling, W. J. (2009). Mutations in LOXHD1, an evolutionarily conserved stereociliary protein, disrupt hair cell function in mice and cause progressive hearing loss in humans. *The American Journal of Human Genetics, 85* (3), 328-337.

Grover, S., Gourie-Devi, M., Baghel, R., Sharma, S., Bala, K., Gupta, M., . . . Kaur, K. (2010). Genetic profile of patients with epilepsy on first-line antiepileptic drugs and potential directions for personalized treatment. *Pharmacogenomics, 11* (7), 927-941.

Gupta, A., Maulik, M., Nasipuri, P., Chattopadhyay, I., Das, S. K., Gangopadhyay, P. K., & Ray, K. (2007). Molecular diagnosis of Wilson disease using prevalent mutations and informative single-nucleotide polymorphism markers. *Clinical chemistry, 53* (9), 1601-1608.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research, 33* (suppl_1), D514-D517.

Hariprakash, J. M., Vellarikkal, S. K., Verma, A., Ranawat, A. S., Jayarajan, R., Ravi, R., . . . Kashyap, A. K. (2018). SAGE: a comprehensive resource of genetic variants integrating South Asian whole genomes and exomes. *Database, 2018* .

Huang, K.-l., Mashl, R. J., Wu, Y., Ritter, D. I., Wang, J., Oh, C., . . . Oak, N. (2018). Pathogenic germline variants in 10,389 adult cancers. *Cell, 173* (2), 355-370. e314.

Jha, A. N., Sundaravadivel, P., Singh, V. K., Pati, S. S., Patra, P. K., Kremsner, P. G., . . . Thangaraj, K. (2014). MBL2 variations and malaria susceptibility in Indian populations. *Infection and immunity, 82* (1), 52-61.

Jha, P., Sinha, S., Kanchan, K., Qidwai, T., Narang, A., Singh, P. K., . . . Sharma, S. K. (2012). Deletion of the APOBEC3B gene strongly impacts susceptibility to falciparum malaria. *Infection, Genetics and Evolution, 12* (1), 142-148.

Kanchan, K., Pati, S., Mohanty, S., Mishra, S., Sharma, S., Awasthi, S., . . . Consortium, I. G. V. (2015). Polymorphisms in host genes encoding NOSII, C-reactive protein, and adhesion molecules thrombospondin and E-selectin are risk factors for Plasmodium falciparum malaria in India. *European Journal of Clinical Microbiology & Infectious Diseases, 34* (10), 2029-2039.

Kapoor, S., & Kabra, M. (2010). Newborn screening in India: Current perspectives. *Indian pediatrics, 47* (3), 219-224.

Karczewski, K., & Francioli, L. (2017). The Genome Aggregation Database (gnomAD). *MacArthur Lab* .

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., . . . Cummings, B. B. (2016). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research, 45* (D1), D840-D845.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., . . . Stepanov, V. (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *The American*

Journal of Human Genetics, 72 (2), 313-332.

Kumar, J., Garg, G., Kumar, A., Sundaramoorthy, E., Sanapala, K. R., Ghosh, S., . . . Sengupta, S. (2009). Single nucleotide polymorphisms in homocysteine metabolism pathway genes: association of CHDH A119C and MTHFR C677T with hyperhomocysteinemia. *Circulation: Cardiovascular Genetics, 2* (6), 599-606.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., . . . Hoover, J. (2015). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research, 44* (D1), D862-D868.

Lim, R. M., Silver, A. J., Silver, M. J., Borroto, C., Spurrier, B., Petrossian, T. C., . . . Silver, L. M. (2016). Targeted mutation screening panels expose systematic population bias in detection of cystic fibrosis risk. *Genetics in medicine, 18* (2), 174.

Majumder, P. P., Roy, B., Banerjee, S., Chakraborty, M., Dey, B., Mukherjee, N., . . . Sil, S. K. (1999). Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications. *European Journal of Human Genetics, 7* (4), 435.

Mohan, V., Radha, V., Nguyen, T. T., Stawiski, E. W., Pahuja, K. B., Goldstein, L. D., . . . Bhangale, T. (2018). Comprehensive genomic analysis identifies pathogenic variants in maturity-onset diabetes of the young (MODY) patients in South India.*BMC medical genetics, 19* (1), 22.

Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., . . . Engelstad, M. (2016). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species.*Nucleic acids research, 45* (D1), D712-D722.

Nair, V. V., Chapla, A., Arulappan, N., & Thomas, N. (2013). Molecular diagnosis of maturity onset diabetes of the young in India. *Indian journal of endocrinology and metabolism, 17* (3), 430.

Narang, A., Jha, P., Rawat, V., Mukhopadhayay, A., Dash, D., Basu, A., . . . Consortium, I. G. V. (2011). Recent admixture in an Indian population of African ancestry.*The American Journal of Human Genetics, 89* (1), 111-120.

Narang, A., Roy, R. D., Chaurasia, A., Mukhopadhyay, A., Mukerji, M., & Dash, D. (2010). IGVBrowser–a genomic variation resource from diverse Indian populations.*Database, 2010* .

Noone, P. G., & Knowles, M. R. (2001). 'CFTR-opathies': disease phenotypes associated with cystic fibrosis transmembrane regulator gene mutations. *Respiratory research, 2* (6), 328.

Peter, B., Wijsman, E. M., Nato Jr, A. Q., Matsushita, M. M., Chapman, K. L., Stanaway, I. B., . . . Raskind, W. H. (2016). Genetic candidate variants in two multigenerational families with childhood apraxia of speech. *PloS one, 11* (4), e0153864.

Pradhan, S., Sengupta, M., Dutta, A., Bhattacharyya, K., Bag, S. K., Dutta, C., & Ray, K. (2010). Indian genetic disease database. *Nucleic acids research, 39* (suppl_1), D933-D938.

Prasher, B., Varma, B., Kumar, A., Khuntia, B. K., Pandey, R., Narang, A., . . . Kukreti, R. (2017). Ayurgenomics for stratified medicine: TRISUTRA consortium initiative across ethnically and geographically diverse Indian populations.*Journal of ethnopharmacology, 197* , 274-293.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics, 38* (8), 904.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature, 461* (7263), 489.

Sarker, S. K., Islam, M. T., Eckhoff, G., Hossain, M. A., Qadri, S. K., Muraduzzaman, A., . . . Tahura, S. (2016). Molecular analysis of Glucose-6-phosphate dehydrogenase gene mutations in Bangladeshi individuals. *PloS one, 11* (11), e0166977.

Sengupta, D., Choudhury, A., Basu, A., & Ramsay, M. (2016). Population stratification and underrepresentation of Indian subcontinent genetic diversity in the 1000 genomes project dataset. *Genome biology and evolution, 8* (11), 3460-3470.

Shah, A. M., Tamang, R., Moorjani, P., Rani, D. S., Govindaraj, P., Kulkarni, G., . . . Reddy, A. G. (2011). Indian siddis: African descendants with Indian admixture. *The American Journal of Human Genetics, 89* (1), 154-161.

Sinha, S., Arya, V., Agarwal, S., & Habib, S. (2009). Genetic differentiation of populations residing in areas of high malaria endemicity in India. *Journal of genetics, 88* (1), 77-80.

Sinha, S., Qidwai, T., Kanchan, K., Anand, P., Jha, G. N., Pati, S. S., . . . Sharma, S. K. (2008). Variations in host genes encoding adhesion molecules and susceptibility to falciparum malaria in India. *Malaria journal, 7* (1), 250.

Sosnay, P. R., Siklosi, K. R., Van Goor, F., Kaniecki, K., Yu, H., Sharma, N., . . . Zielenski, J. (2013). Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nature genetics, 45* (10), 1160.

Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics, 68* (4), 978-989.

Talwar, P., Kanojia, N., Mahendru, S., Baghel, R., Grover, S., Arora, G., . . . Singh, M. (2017). Genetic contribution of CYP1A1 variant on treatment outcome in epilepsy patients: a functional and interethnic perspective. *The pharmacogenomics journal, 17* (3), 242.

Thanseem, I., Thangaraj, K., Chaubey, G., Singh, V. K., Bhaskar, L. V., Reddy, B. M., . . . Singh, L. (2006). Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC genetics, 7* (1), 42.

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.

Goldstein, J. I., Crenshaw, A., Carey, J., Grant, G. B., Maguire, J., Fromer, M., ... Neale, B. M. (2012). zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinformatics (Oxford, England), 28(19), 2543–2545. doi:10.1093/bioinformatics/bts479

Figure legends:

Figure-1 : Work flow the study and details of Genotyping, data processing and quality controls and screen shot ClinIndb databse.

Figure 2: Diversity and relatedness of Indian populations with respect to 1000 genomes SAS and EAS super populations. Majority of the IE and DR large populations are covered by SAS group. EAS group is distinct from TB populations of India. Isolated AA and DR populations not covered by any of the 1000 genomes populations. Keys outside the main figure represent different populations: Circles - IGV populations, plus sign - SAS and EAS super populations from 1000 genomes.

Figure 3: Sufficient coverage of IE and DR large populations as well as TB cluster (by highlanders) by our cohorts in this study with the exception of isolated groups. Keys outside the main figure represent different populations: Filled circles - Cohorts in this study, Unfilled circles - IGV populations, plus sign -1000 genomes SAS and EAS super populations.

Figure-4: Spectrum of Pathogenic Variants in Clindb and other genomic databases (SAS Populations) A) Number of variants containing either "pathogenic" or "likely pathogenic" term were compared across different SAS populations in global databases. gnomAD_SAS is representative of both gnomAD and ExAc SAS groups. B) Absolute average frequency differences of Clindb and GAsP (SAS) were compared with 1000G, ExAc and gnoMAD SAS groups