INDEX-db: The Indian Exome Reference database (Phase-I)

Husayn Ahmed P[a,b], Vidhya V[c], Ravi P More[a], Mahendra S Rao[c], Biju Viswanath[d], Sanjeev Jain[s], Odity Mukherjee[c*], ADBS Consortium^

a National Centre for Biological Sciences, Bengaluru, India

b Institute of Bioinformatics and Applied Biotechnology, Bengaluru, India

c Institute for Stem Cell Biology and Regenerative Medicine, Bengaluru, India

d National Institute of Mental Health and Neuro Sciences, Bengaluru, India

* Corresponding author

Email: omukherjee@ncbs.res.in

Telephone: +91-80-26996180

Address : Institute of Stem Cells and Regenerative Medicine (InStem)

GKVK Post, Bellary Road, Bangalore 560065, India

^Membership of the ADBS (The Accelerator program for Discovery in Brain disorders using Stem cells) Consortium is provided in the Acknowledgment.

Abstract

Deep sequencing based genetic mapping has greatly enhanced the ability to catalog variants with plausible disease association. The bigger challenge now is to ascertain pathological significance to the array of identified variants to specific disease conditions. Differential selection pressure may impact frequency of genetic variations, and thus the detection of association with disease conditions, across populations. To understand the genotype to phenotype correlations, it thus becomes important to first understand the genetic variation spectrum of a population by creating a reference map. In this study, we report the development of phase I of a new database of coding variations, from the Indian population, with an aim to establish a centralized database of integrated information. This could be useful for researchers involved in studying disease mechanism at the clinical, genetic and cellular level.

Database URL: http://indexdb.ncbs.res.in

Keywords

Population-specific database; Genetic variations catalogue; Indian population; Whole exome sequencing

Introduction

Human population has increased significantly in numbers across all geographical regions in the recent past, resulting in population specific genetic architecture. Such

rapid population growth has significant impact on the occurrence and frequency of genetic variations, especially rare variants which may lie on conserved protein encoding sites, that may have a likely role in disease biology (Keinan and Clark, 2012). Next Generation Sequencing (NGS) strategies have greatly improved the ability to identify genetic variants, of varying frequencies. Recent studies to identify genetic variants associated with 'common non-communicable disease' suggest that these syndromes have high heritability, and that the risk arises from a polygenic contribution, caused by a combination of rare deleterious and common polymorphic modifier variants. NGS based evaluation of disease association thus becomes a useful way to identify disease genetic signature. A critical component of this analysis is the assignment of pathogenic relevance to the identified variants, done primarily by defining the frequency in affected individuals as compared to control, healthy samples. In this context, several genetic variation databases have been established incorporating different strategies and technological improvements (eg. haplotype mapping - HapMap project (The International HapMap Consortium, 2005); whole genome sequencing - 1000 Genomes project (The 1000 Genomes Project Consortium, 2015); whole exome sequencing - Exome Aggregation Consortium (Lek et al., 2016)). While the information gleaned from these databases improved our understanding of the complexities of the genetic architecture, it also reported that a significant proportion of genetic variations identified are population specific. We thus need a detailed evaluation in diverse populations, to better understand the genetic basis of epidemiology and semiology of human diseases by identifying modifier genetic

variations (Bamshad et al., 2011; Craddock and Owen, 2010; Higasa et al., 2016; Hindorff et al., 2011).

The Indian subcontinent is estimated to see an increase in the number of individuals needing care for adult onset common disorders due to improved health care and life expectancy. Identification of disease specific genetic signature is a critical first step in identifying – a) disease associated genetic variations, b) molecular sub-typing of complex human phenotypes and c) at risk individuals with improved efficiency. A comprehensive reference variation map, established from a clinically normal cohort that is representative of this population, will be of great benefit. There have been several reports of cataloging genetic variation from the Indian population which have suggested presence of distinct genome level sub-structuring, and its probable impact on disease Biology (Narang et al., 2010; Rustagi et al., 2017; The HUGO Pan-Asian SNP Consortium, 2011; The Indian Genome Variation Consortium, 2005; Upadhyay et al., 2016). However, there are a few limitations to these studies – a) these predominantly catalogue germline variants; b) are designed to capture high frequency common variations, which is sufficient for deciphering population structure, but lack information on rare mutations, CNVs and disallow haplotype analysis; and, importantly c) are not available as open access reference map.

In this study we report the development and completion of phase 1 of a new accessible database- the INDian EXome database (INDEX-db), that catalogues variations in exonic and regulatory regions from healthy control individuals, across

different geographical regions of southern India. To make the database a comprehensive resource for disease genetics studies, we have integrated WES derived SNV, CNV and phased LD information, along with expression data, on samples derived from a subset of individuals sequenced. We believe that such an integrated reference database may be valuable to understand the genomic architecture underlying susceptibility to disease, detect familial or geographical clustering of the population, and thus aid efforts to understand disease genetics.

Materials and methods

Samples information and ethical approval

Thirty one individuals tested to be asymptomatic for any adult onset common clinical illness were selected for the study at National Institute of Mental Health and Neuro Sciences, Bengaluru (Gender, age and other information of the individuals in Supplementary Table 1). The study was approved by the institutional ethics committee. Written informed consent was obtained from all participants prior to sampling. 10 ml of peripheral blood was collected under aseptic conditions and high molecular weight DNA isolated.

Library preparation and exome sequencing

The genomic DNA was extracted from the blood and the Illumina Nextera Rapid Capture Expanded Exome kit was used for library preparation. Sequencing was carried out on Illumina Hiseq NGS platform. Quality check of the raw reads was

performed using FASTQC tool (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Only the paired-end raw reads with a score more than Q20 was filtered using Prinseq lite version 0.20.4 (Schmieder and Edwards, 2011) for further alignment to the reference genome. Reads were also checked for per base and per sequence quality scores, GC content, and sequence length distribution.

Alignment and mapping of reads

The raw reads were aligned to the Human reference genome hg19 (GrCh37) using BWA tool version 0.5.9 (Li and Durbin, 2009). PCR duplicates in the mapped reads were marked using Picard (http://broadinstitute.github.io/picard/). INDEL realignment was performed using GATK version 3.6 (Depristo et al., 2011). Conversion of the sequence alignment file (SAM to BAM), indexing and sorting were done by samtools version 1.5 (Li et al., 2009). The quality check for the alignment on the mapped reads was performed using Qualimap version 2.2.1 (Okonechnikov et al., 2015).

Detecting SNPs, indels and CNVs

SNPs and indels were called from the aligned files using Varscan2 version 2.3.9 (Koboldt et al., 2009; Koboldt et al., 2012) (with the criteria min coverage = 8, MAF >/= 0.25% and P </= 0.001). Depth of coverage was calculated using GATK version 3.8.0 (16) and this was used to detect copy number variations using XHMM (Fromer et al., 2012; Fromer and Purcell, 2014). XHMM employs principal component analysis to remove batch and target effects. Principal component analysis

was performed on the entire read-depth matrix (31 individuals by 336,037 targets) and a hidden Markov model was applied to the normalized data to detect CNVs.

Haplotype phasing

Haplotype pre-phasing was done for SNP genotypes from 31 individuals using SHAPEIT2 (v2.r837.GLIBCv2.12) (Delaneau et al., 2014; O'Connell et al., 2014). As a haplotype reference, we downloaded 1000Genome project Phase3 reference (http://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3/) and used only the SAS subgroup haplotype reference. The phased data was visualized and haplotype blocks were generated based on the Dprime values computed for every comparisons between markers (SNPs) which are present within a distance range of 500kb using Haploview version 4.2 (Barrett et al., 2005). Default parameters were used which includes markers having MAF>0.05, p-value cutoff of 0.001, with maximum Mendelian errors of 1, minimum genotype percentage of 75%, exclusion of individuals with >50% of missing genotypes, with 95% confidence bounds (Gabriel et al., 2002).

Development of INDEX-db

The SNPs and indels obtained from all the 31 individuals were merged using vcftools (Danecek et al., 2011) to create a merged SNPs and indels catalogue. This was annotated with ANNOVAR (reference assembly 65) (Wang et al., 2010). Copy number variations were pooled from all the individuals and used to create a reference copy number profile for the population. Pooling of data, functional analysis and other

downstream analysis were performed using in-house shell and python scripts. The entire workflow of developing INDEX-db is shown in Fig. 1. The graphical genome browser for the database was developed on JBrowse version 1.12.3 (Skinner et al., 2009).

Data availability

The raw sequence data has been deposited at the NCBI SRA database (SRA accession SRP135959). The entire database is hosted online at http://indexdb.ncbs.res.in and is freely accessible along with associated tools for querying and comparing user data to INDEX-db. The data is also available for download in standard formats at http://indexdb.ncbs.res.in/downloads.html. The SNPs are also deposited at the NCBI's dbSNP (https://www.ncbi.nlm.nih.gov/SNP/snp_viewTable.cgi?handle=OMUKHERJEE_ADBS).

Results

INDEX-db: Variant summary profile

A total of 397,336 single nucleotide variations were identified in this phase 1 of the INDEX-db with an average 96% of the reads mapping to the reference genome at a mean coverage of 54.6% with at least 20X depths (Fig. 2). There was no significant bias seen, in terms of sequencing and/or sample QC (Fig. 2). About ~23% of the total genetic mutations identified were in the coding region, of which nearly half (51.34%)

were a missense variation, followed by silent (43.36%), indel (1.8%), nonsense (0.81%) and splice sites (0.55%) (Fig. 3A). The ratio of non-synonymous (NS=49013) to synonymous variants (S=39876) was 1.23 (Fig. 2). The SNP profile observed in our study is comparable to exome sequencing reports published earlier (Lek et al., 2016; Rustagi et al., 2017; Upadhyay et al., 2016).

Copy number variations (CNVs) contribute about one tenth of a percent of the total genetic variations of an individual, and it affects longer regions than both SNPs and/or short indels (The 1000 Genomes Project Consortium, 2015). The CNVs have a spectrum of phenotypic effects, from adaptive traits (Beckmann et al., 2007) to embryonic lethality (Hurles et al., 2008), and are implicated in many disorders including schizophrenia (Cook and Scherer, 2008), Down's syndrome (Korenberg et al., 1994), kidney diseases (Nagano et al., 2018), diabetes (Ascencio-Montiel et al., 2017; Prabhanjan et al., 2016), hypertension (Boon-Peng et al., 2016; Marques et al., 2014), cancer (Araujo et al., 2014; Liu et al., 2013), bipolar disorder (Grozeva et al., 2013) etc. While array comparative genomic hybridization (aCGH) is considered as the standard for molecular assessment of genome-wide copy number detection (Pinkel et al., 1998; Pinkel and Albertson, 2005), methods have now been developed to detect copy number variations from NGS based exome and genome sequencing data (Yoon et al., 2009). Using a hidden Markov model, we identified a total of 1,538 CNVs in the size range of 50 bp to 3 mb in the INDEX-db phase I analysis represented as a circos plot (Fig. 3B). The number, size range and distribution of the detected CNVs in INDEX-db is comparable to other published data (MacDonald et al., 2014).

The common pattern in which variants are inherited across a population have critical importance in studying genetic correlates to rare and complex human diseases (The International HapMap Consortium, 2005). As parental genotype information may not be available for all the samples, reference phased haplotypes imputed using population relevant reference is valuable for disease genetics investigations. In INDEX-db phase I, we identified a total of 3365 LD blocks spread across the autosomes with an average block length of ~3.6 kb. (Fig. 3C).

To identify if there exists any population specific mutation/recombination hotspots, we computed the mean density of the variants (SNPs and CNVs) and the resulting LD blocks across the chromosomes by calculating the number of variants and/or LD blocks per million base pairs for each chromosome (normalizing for chromosome size). Apart from chromosome 19 which showed a significantly higher number of SNPs and CNVs, we did not observe any significant clustering of variants in any other chromosome (Fig. 4A-B). The increased density of variants observed in chromosome 19 may be due to the highest gene density or the presence of many paralogues of immunoglobulins localized to this chromosome that may undergo repeated duplication and/or mutation as reported by earlier studies (Castresana, 2002; Grimwood et al., 2004). We also observed that chromosome 19 had the highest number of LD blocks compared to the other chromosome. This could be due to the increased number and polymorphic diversity of the variants localized to this region resulting in low levels of common haplotypes. (Fig. 4C and Supplementary Table 2).

Population genetics studies have shown that there is greater genetic drift in East Asian

populations, impacting the individual mutation burden load (Balick et al., 2015; Gao and Keinan, 2016; Simons and Sella, 2016). To ascertain the value of INDEX-db as a reference resource for disease genetics studies for the Indian population, we compared the INDEX-db phase I data with two publicly available databases. We used the ExAC as it is one of largest exome sequencing reference database with significant representation of South Asian population (although low representation from the pan-Indian population), and the AP-SAS, as it is a WGS based resource generated using samples from southern India. The variant parameter profile identified in INDEX-db are comparable to these other databases (Supplementary Table 3).

We found 12% (48732) of the variants identified were unique to INDEX-db phase I (Fig. 3A, Supplementary Table 3). Within the coding region, this translated to 8860 (~2.23%) variations, out of which 966 had a functional annotation of being 'deleterious' by two *in silico* algorithms (Fig. 4D) (Adzhubei et al., 2010; Ng and Henikoff, 2003). We found ~20% of coding variants identified in INDEX-db to be common between ExAC, and ~ 7% common to AP-SAS. The observation of low overlap between INDEX-db and AP-SAS could be attributed to the low coverage in the whole genome sequencing design of the AP-SAS study (~2X mean coverage). Differences between ExAC-SAS and INDEX-db could perhaps be attributed to population specific variation signature, especially since ExAC-SAS has a low representation from the Indian population. The mutational profile obtained in the phase I of this database is comparable to other databases, though currently limited by the number of individuals it represents.

Functional relevance of INDEX-db as a population specific reference database for disease association studies

Based on the protein perturbing impact, mutations can be classified as benign (tolerated), deleterious (loss of function) and neutral (no influence). Functional impact of genetic variants identified in INDEX-d, was analyzed using SIFT (Ng and Henikoff, 2003) and PolyPhen2 (Adzhubei et al., 2010) algorithms that predict the functional consequences based on conservation and protein structure. A total of 8345 non-synonymous variants (9% of the total protein-coding variants) spanning 5097 genes were predicted to be damaging by SIFT and Polyphen2 in INDEX-db (Fig. 4D). Of these, 11% (966) spanning 700 genes were novel to INDEX-db. The percentage of damaging and unique variants observed in our dataset is in the range as reported earlier, although the specific genes and/or pathways may differ (The 1000 Genomes Project Consortium, 2015). To evaluate the biological impact of these deleterious mutations, we performed enrichment analysis on the genes harbouring deleterious mutations, and found increased enrichment for Wnt signaling, Nicotinic acetylcholine receptor signaling, Integrin signalling, cytokine signaling and Cadherin signalling pathway (Fig. 2). These pathways are implicated in various disorders including severe mental illness, diabetes and cancer. (Supplementary Table 4). Assessing the frequency of these rare variants in the general, healthy population may be useful to understand the genetic contributions to risk for disease, and also the relation with particular clinical syndromes. The comprehensive profile of genetic variations cataloged in INDEX-db phase I is detailed in Fig. 2.

Discussion

We report the development of a new database, INDEX-db, which summarizes variations in coding and regulatory regions, identified from healthy control individuals. The first phase of the database consists of 31 individuals from southern India. We have integrated the exome sequencing results with expression data generated from a subgroup of the individuals constituting INDEX-db. The database is also layered with information regarding CNVs and phased LD mapping. The integrated database is available freely at http://indexdb.ncbs.res.in along with associated tools for querying and comparing user input data to INDEX-db.

The INDEX-db is in its first phase and thus in comparison to other public databases is limited in terms of the number of individuals sequenced to represent the population, but the variant profile we report in our pilot phase is comparable to population-based databases signifying its value in terms of giving population-specific information.

Genetic basis of complex disorders need to be better understood in India where the burden of these disorders is expected to increase significantly in the coming decades. In this context, we believe that an integrated reference database may be valuable to understand the genomic architecture underlying susceptibility to disease, familial or geographical clustering of the population and aid in disease genetics studies.

Acknowledgements

The authors are grateful to all the volunteers who participated in the study. We thank Drs. Lakshmi Narayanan Kota, Manasa Seshadri and Ravi Kumar Nadella for recruitment of the control individuals, their clinical assessments and initial sample processing. Ten individuals were recruited as part of a Center of excellence grant in collaboration with Geriatric clinic team of NIMHANS (Profs Mathew Varghese, Sivakumar PT and other clinical staff). The authors would like to thank the sequencing core facility at IGIB (Dr. Faruq Mohammed) and NCBS (Dr. Awadhesh Pandit) for sample processing and data generation. The authors would like to thank all investigators of ADBS consortia for providing valuable inputs to the study and the manuscript.

The ADBS consortium members: Biju Viswanath[#], Naren P. Rao[#], Janardhanan C. Narayanaswamy[#], Palanimuthu T Sivakumar[#], Arun Kandaswamy[#], Muralidharan Kesavan[#], Urvakhsh Meherwan Mehta[#], Ganesan Venkatasubramanian[#], John P. John[#], Odity Mukherjee[@], Meera Purushottam[#], Ramakrishnan Kannan[#], Bhupesh Mehta[#], Thennarasu Kandavel[#], Binukumar B.[#], Jitender Saini[#], Deepak Jayarajan[#], Shyamsundar A.[#], Sydney Moirangthem[#], Vijay Kumar G.[#], Jagadisha Thirthalli[#], Prabha S. Chandra[#], Bangalore N. Gangadhar[#], Pratima Murthy[#], Mitradas M. Panicker[*], Upinder S Bhalla[*], Sumantra Chattarji[*@],Vivek Benegal[#], Mathew Varghese[#], Janardhan YC Reddy[#], Padinjat Raghu[*], Mahendra Rao[@], Sanjeev Jain[#].

National Institute of Mental Health and Neuro Sciences (NIMHANS), Bengaluru,

India

* National Centre for Biological Sciences – Tata Institute of Fundamental Research (NCBS – TIFR), Bengaluru, India

@ Institute for Stem Cell Biology and Regenerative Medicine (InStem), Bengaluru, India

Conflicts of interest

There is no conflict of interest.

Funding Source

The study was supported by government funded research grant under the aegis of Department of Biotechnology (grant number BT/PR17316/MED/31/326/2015) and Pratiksha trust. Ten individuals were recruited as part of a Center of excellence grant from Department of Biotechnology (grant number BT/01/CEIB/11/VI/1). HAP is supported by a grant from the Department of Biotechnology (grant number BT/PR12422/MED/31/287/2014). The funding agencies had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods. https://doi.org/10.1038/nmeth0410-248

Araujo, A.N., Moraes, L., França, M.I.C., Hakonarson, H., Li, J., Pellegrino, R., Maciel, R.M.B., Cerutti, J.M., 2014. Genome-wide copy number analysis in a family with p.G533C RET mutation and medullary thyroid carcinoma identified regions potentially associated with a higher predisposition to lymph node metastasis. J. Clin. Endocrinol. Metab. 99, 1104–1112. https://doi.org/10.1210/jc.2013-2993

Ascencio-Montiel, I.D.J., Pinto, D., Parra, E.J., Valladares-Salgado, A., Cruz, M., Scherer, S.W., 2017. Characterization of large copy number variation in Mexican type 2 diabetes subjects. Sci. Rep. 7, 17105. https://doi.org/10.1038/s41598-017-17361-7

Balick, D.J., Do, R., Cassa, C.A., Reich, D., Sunyaev, S.R., 2015. Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck. PLoS Genet. 11, e1005436. https://doi.org/10.1371/journal.pgen.1005436

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., Shendure, J., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. 12, 745–755. https://doi.org/10.1038/nrg3031

Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263–265. https://doi.org/10.1093/bioinformatics/bth457

Beckmann, J.S., Estivill, X., Antonarakis, S.E., 2007. Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. Nat. Rev. Genet. 8, 639–646. https://doi.org/10.1038/nrg2149

Boon-Peng, H., Jusoh, J.A.M., Marshall, C.R., Majid, F., Danuri, N., Basir, F., Thiruvahindrapuram, B., Scherer, S.W., Yusoff, K., 2016. Rare copy number variants identified suggest the regulating pathways in hypertension-related left ventricular hypertrophy. PLoS One, 11, e0148755. https://doi.org/10.1371/journal.pone.0148755

Castresana, J., 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. Nucleic Acids Res. 30, 1751–6. https://doi.org/10.1093/nar/30.8.1751

Cook, E.H., Scherer, S.W., 2008. Copy-number variations associated with neuropsychiatric conditions. Nature, 455, 919–923. https://doi.org/10.1038/nature07458

Craddock, N., Owen, M.J., 2010. The Kraepelinian dichotomy - Going, going... but still not gone. Br. J. Psychiatry 196, 92–95. https://doi.org/10.1192/bjp.bp.109.073429

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Delaneau, O., Marchini, J., McVeanh, G. A., et al. (2014) Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel, Nat. Commun., 5, 3934. https://doi.org/10.1038/ncomms4934

Depristo, M.A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–501. https://doi.org/10.1038/ng.806

Fromer, M., Purcell, S.M., 2014. Using XHMM software to detect copy number variation in whole-exome sequencing data. Curr. Protoc. Hum. Genet., 81, 7.23.1-7.23.21. https://doi.org/10.1002/0471142905.hg0723s81

Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., McCarroll, S.A., O'Donovan, M.C., Owen, M.J., Kirov, G., Sullivan, P.F., Hultman, C.M., Sklar, P., Purcell, S.M., 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet. 91, 597–607. https://doi.org/10.1016/j.ajhg.2012.08.005

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D.,

2002. The structure of haplotype blocks in the human genome. Science. 296, 2225–2229. https://doi.org/10.1126/science.1069424

Gao, F., Keinan, A., 2016. Explosive genetic evidence for explosive human population growth. Curr. Opin. Genet. Dev. 41, 130–139. https://doi.org/10.1016/j.gde.2016.09.002

Grimwood, J., Gordon, L. a, Olsen, a, et al., 2004. The DNA sequence and biology of human chromosome 19. Nature 428, 529–535. https://doi.org/10.1038/nature02399

Grozeva, D., Kirov, G., Conrad, D.F., Barnes, C.P., Hurles, M., Owen, M.J., O'Donovan, M.C., Craddock, N., 2013. Reduced burden of very large and rare CNVs in bipolar affective disorder. Bipolar Disord. 15, 893–898. https://doi.org/10.1111/bdi.12125

Higasa, K., Miyake, N., Yoshimura, J., et al. , 2016. Human genetic variation database, a reference database of genetic variations in the Japanese population. J. Hum. Genet. 61, 547–553. https://doi.org/10.1038/jhg.2016.12

Hindorff, L.A., Gillanders, E.M., Manolio, T.A., 2011. Genetic architecture of cancer and other complex diseases: Lessons learned and future directions. Carcinogenesis 32, 945–954. https://doi.org/10.1093/carcin/bgr056

Hurles, M.E., Dermitzakis, E.T., Tyler-Smith, C., 2008. The functional impact of structural variation in humans. Trends Genet. 24, 238–245. https://doi.org/10.1016/j.tig.2008.03.001

Keinan, A., Clark, A.G., 2012. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants Alon Keinan and Andrew G. Clark. Science. 740, 740–744. https://doi.org/10.1126/science.1217283

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L., 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 2283–2285. https://doi.org/10.1093/bioinformatics/btp373

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., Mclellan, M.D., Lin, L., Miller, C. a, Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576. https://doi.org/10.1101/gr.129684.111

Korenberg, J.R., Chen, X.N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., Carpenter, N., Daumer, C., Dignan, P., Disteche, C., 1994. Down syndrome phenotypes: the consequences of chromosomal imbalance. Proc. Natl. Acad. Sci. U. S. A., 91, 4997–5001. https://doi.org/10.1073/pnas.91.23.11281a

Lek, M., Karczewski, K. J., Minikel, E. V., et al., 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291. https://doi.org/10.1038/nature19057

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, Y., Cope, L., Sun, W., Wang, Y., Prasad, N., Sangenario, L., Talbot, K., Somervell, H., Westra, W., Bishop, J., Califano, J., Zeiger, M., Umbricht, C., 2013. DNA copy number variations characterize benign and malignant thyroid tumors. J. Clin. Endocrinol. Metab. 98, E558-66. https://doi.org/10.1210/jc.2012-3113

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., Scherer, S.W., 2014. The Database of Genomic Variants: A curated collection of structural variation in the human genome. Nucleic Acids Res. 42, 986–992. https://doi.org/10.1093/nar/gkt958

Marques, F.Z., Prestes, P.R., Pinheiro, L.B., Scurrah, K., Emslie, K.R., Tomaszewski, M., Harrap, S.B., Charchar, F.J., 2014. Measurement of absolute copy number variation reveals association with essential hypertension. BMC Med. Genomics, 7. https://doi.org/10.1186/1755-8794-7-44

Nagano, C., Nozu, K., Morisada, N., et al., 2018. Detection of copy number variations by pair analysis using next-generation sequencing data in inherited kidney diseases. Clin. Exp. Nephrol. https://doi.org/10.1007/s10157-018-1534-x

Narang, A., Roy, R.D., Chaurasia, A., Mukhopadhyay, A., Mukerji, M., Dash, D., 2010. IGVBrowser-a genomic variation resource from diverse Indian populations. Database, baq022. https://doi.org/10.1093/database/baq022

Ng, P.C., Henikoff, S., 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814. https://doi.org/10.1093/nar/gkg509

Okonechnikov, K., Conesa, A., García-Alcalde, F., 2015. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32, 292–294. https://doi.org/10.1093/bioinformatics/btv566

O'Connell, J., Gurdasani, D., Delaneau, O., et al., 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet., 10, e1004234. https://doi.org/10.1371/journal.pgen.1004234

Pinkel, D., Segraves, R., Sudar, D., et al., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. 20, 207–211. https://doi.org/10.1038/2524

Pinkel, D., Albertson, D.G., 2005. Array comparative genomic hybridization and its applications in cancer. Nat. Genet., 37, S11–S17. https://doi.org/10.1038/ng1569

Prabhanjan, M., Suresh, R. V., Murthy, M.N., Ramachandra, N.B., 2016. Type 2 diabetes mellitus disease risk genes identified by genome wide copy number variation scan in normal populations. Diabetes Res. Clin. Pract. 113, 160–170. https://doi.org/10.1016/j.diabres.2015.12.015

Rustagi, N., Zhou, A., Watkins, W.S., Gedvilaite, E., Wang, S., Ramesh, N., Muzny, D., Gibbs, R.A., Jorde, L.B., Yu, F., Xing, J., 2017. Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. BMC Genomics 18, 369. https://doi.org/10.1186/s12864-017-3767-6

Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864. https://doi.org/10.1093/bioinformatics/btr026

Simons, Y.B., Sella, G., 2016. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Curr. Opin. Genet. Dev. 41, 150–158. https://doi.org/10.1016/j.gde.2016.09.006

Skinner, M.E., Uzilov, A. V., Stein, L.D., Mungall, C.J., Holmes, I.H., 2009. JBrowse: A next-generation genome browser. Genome Res. 19, 1630–1638. https://doi.org/10.1101/gr.094607.109

The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393

The HUGO Pan-Asian SNP Consortium, 2011. Mapping Human Genetic Diversity in Asia. Science (80-.). 1541, 1541–1546. https://doi.org/10.1126/science.1177074

The Indian Genome Variation Consortium, 2005. The Indian Genome Variation database (IGVdb): A project overview. Hum. Genet. 118. https://doi.org/10.1007/s00439-005-0009-9

The International HapMap Consortium, 2005. A haplotype map of the human genome. Nature 437, 1299–1320. https://doi.org/10.1038/nature04226

Upadhyay, P., Gardi, N., Desai, S., Sahoo, B., Singh, A., Togar, T., Iyer, P., Prasad, R., Chandrani, P., Gupta, S., Dutt, A., 2016. TMC-SNPdb: an Indian germline variant database derived from whole exome sequences. Database, baw103. https://doi.org/10.1093/database/baw104

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164. https://doi.org/10.1093/nar/gkq603

Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J., 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19, 1586–1592. https://doi.org/10.1101/gr.092981.109

Supplementary data

Supplementary Table 1: Sample IDs, age and gender of the individuals of INDEX-db

Supplementary Table 2: Mean density of the variants (SNPs and CNVs) and the resulting LD blocks across the chromosomes

Supplementary Table 3: Comparison of INDEX-db with other public databases

Supplementary Table 4: Deleterious variants in genes (known to be implicated in a spectrum of diseases) identified in the samples and comparison with public databases

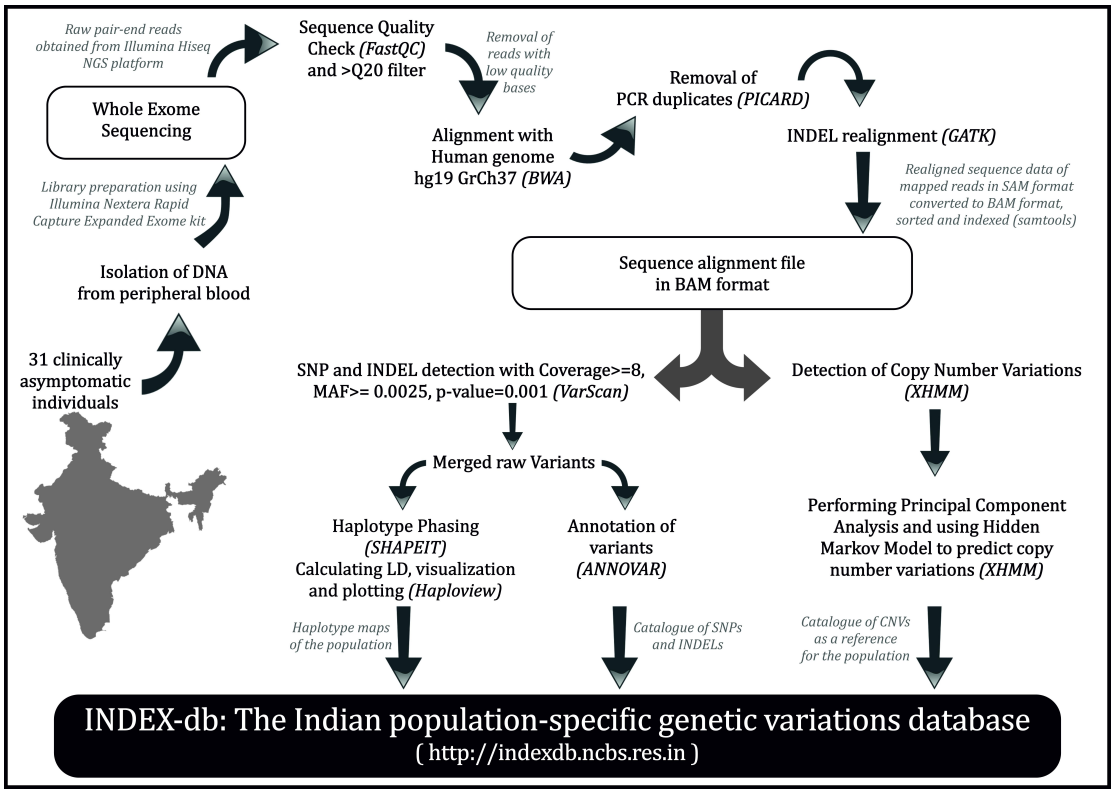

Fig. 1: The workflow of the development of phase 1 of INDEX-db. The steps involved in the development of INDEX-db. The tools used in every step are mentioned in the brackets.

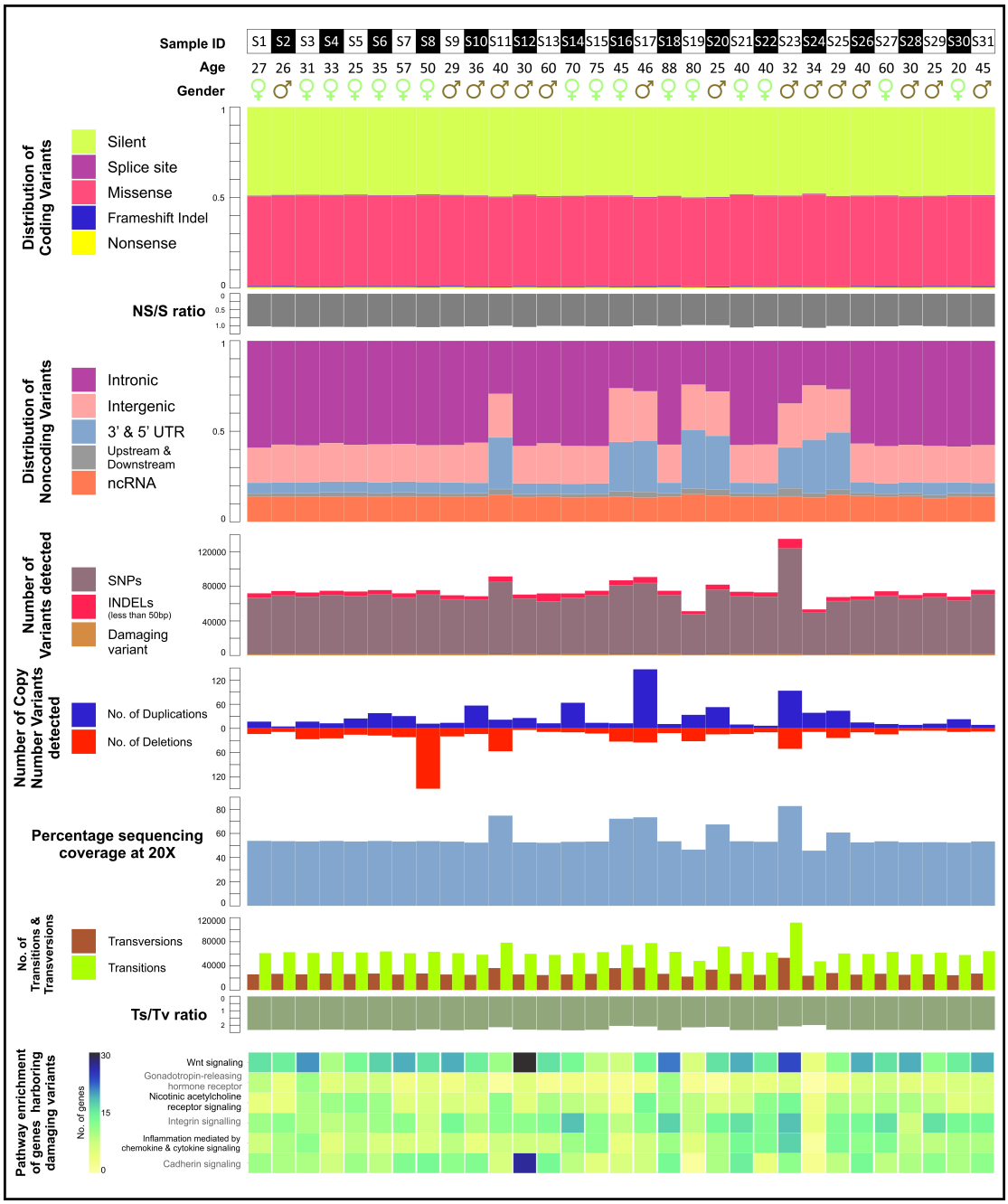

Fig. 2: Variant summary profile. The distribution of coding and non-coding variants, the nonsynonymous to synonymous and the transitions to transversions ratios, and the percentage coverage of sequencing at 20X of 31 individuals catalogued in INDEX-db. The number of SNPs and CNVs detected in every individual with the pathway enrichment of genes harbouring damaging SNPs predicted by SIFT and PolyPhen2. Abbreviations: NS-Non-synonymous; S-Synonymous; UTR-Untranslated Region; ncRNA-Non-coding RNA; Ts-Transitions; Tv-Transversions.

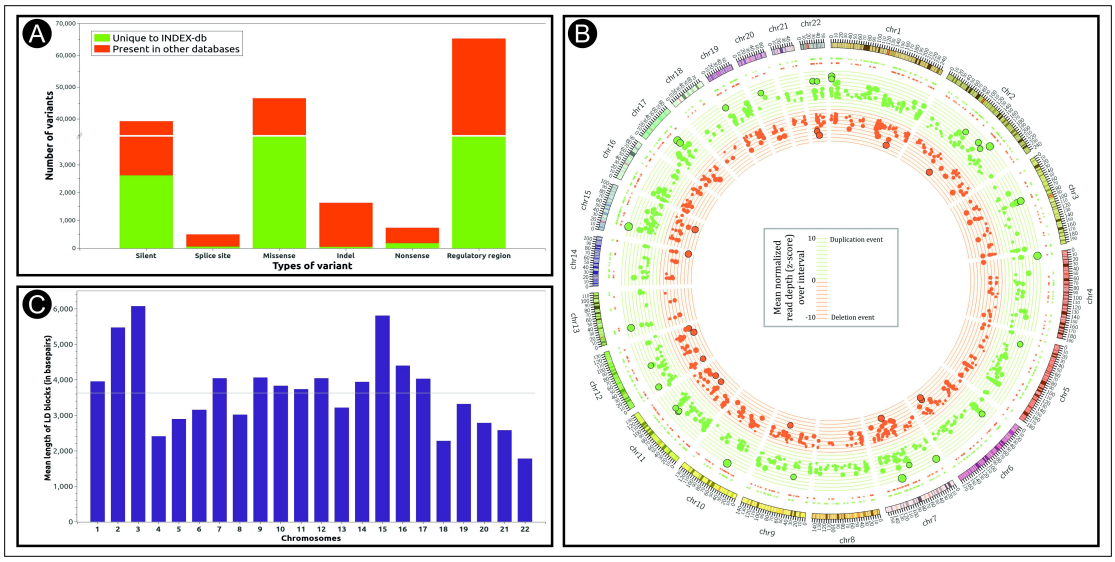

Fig. 3: INDEX-db genetic catalogue (A) Comparison of INDEX-db with other public databases. (B) The circos plot showing the copy number variation events catalogued in INDEX-db. The duplication and deletion events has been coloured green and red respectively. (C) Mean length of linkage disequilibrium blocks identified in autosomes.

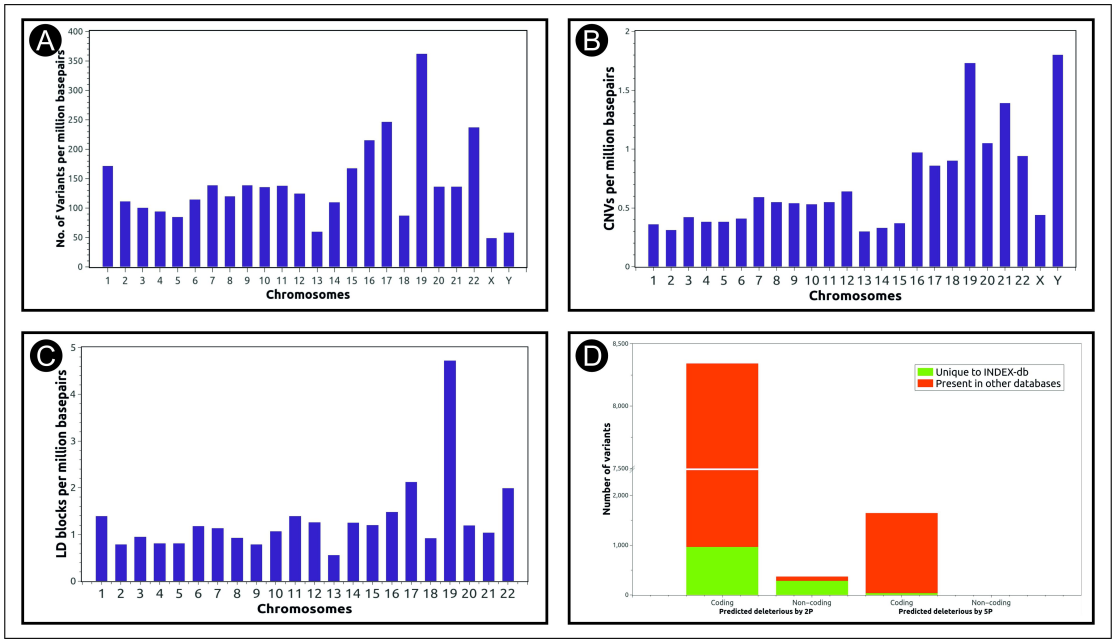

Fig. 4: The distribution of genetic variations across chromosomes. The number of (A) SNPs, (B) CNVs and (C) LD blocks catalogued in INDEX-db per million basepairs in each chromosome. (D) In silico prediction of deleteriousness on coding and non-coding variants by 2 predicting algorithms (SIFT and PolyPhen2) and 5 algorithms (SIFT, PolyPhen2, Mutation Taster, Mutation Accessor, LRT predictor).