GeNeViSTA

# Population Genetic Variant Databases- Utility in Clinical Genetics

Ikrormi Rungsung, Ashwin Dalal

*Diagnostics Division, Centre for DNA Fingerprinting and Diagnostics, Hyderabad*

*Correspondence to:* Dr Ashwin Dalal *Email:* ashwindalal@gmail.com

## Abstract

Next generation sequencing based analysis has revolutionized the field of genetic diagnostics. However, these high throughput techniques reveal thousands of variants in individuals, many of which are non-disease-causing polymorphisms. Knowledge regarding the polymorphisms in each population is essential, so that these variants can be ignored in order to identify the disease-causing variant. This article focusses on various population databases which help us to know the frequencies of these polymorphisms in different populations throughout the world.

## Introduction

The genetic diversity between two unrelated human species is only 0.6% (Auton et al., 2015), which is far less diverse than the other mammalian species like apes. Studies have also reported that the genetic variation between two unrelated individuals is more diverse accounting for 87.6% of the variation, than the genetic variations between two different populations i.e., only 9.2% (Jorde 2003). The single nucleotide polymorphism (SNP) is the most abundant form of genetic variation which accounts for 90% and the other types are constituted by small insertions/deletions and large scale copy number variants (CNVs) (Collins et al., 1999; Gross et al., 2007). About 5% of the SNPs are non-synonymous with functional impact and are located in the coding regions. The SNPs from the non-coding regions also serve as important genetic markers throughout the human genome (Collins et al., 1999). The knowledge of the enormous genetic variation data has been pivotal in discovering new disease-causing genes, mapping the human genetic ancestry and evolution during the course of time. The genetic association studies of SNPs with disease phenotypic trait have been extremely successful. For instance, the association of HLA gene with several diseases such as type I diabetes, rheumatoid arthritis, coeliac disease, multiple sclerosis and ulcerative colitis, has been determined using the genome wide association strategies (Bell et al., 2002). In addition, Salla disease caused by mutation in the *SLC17A5* gene in an Old order Menonite child was identified, using the SNP arrays method (Strauss et al., 2005).

## Genetic databases in NGS analysis

With the emergence of massively parallel sequencing also known as Next generation sequencing technology (NGS) or deep sequencing and the increasing number of such population genetic variation databases, many more genes involved in rare diseases have been mapped (Fernandez-Marmiesse et al., 2017). The human disease association studies and clinical genetic testing are increasingly dependent on the deep sequencing-based evaluation. Population databases from across different geographical regions have become integral in the evaluation and interpretation of the genomic variants. The minor allele frequencies of variants have been shown to vary greatly in different populations. This has to be taken into consideration in the context of the disease prevalence in that particular population during variant filtering and prioritization. Some of the major population databases used in NGS-data analysis include 1000 Genomes project, Exome aggregation consortium (ExAC), Exome sequencing project (ESP), gnomAD, Complete genomics (CG69) and Great middle east (GME).

The present article describes some of these population genetic variation databases that are publicly available. The population genetic variation databases serve as sources of population allele frequency data for efficient filtering of the common variants from the rare disease-causing candidates

and therefore useful in mapping the disease phenotype. Most of the population genetic variation databases are generated by forming consortiums with many principal investigators across multiple laboratories from various parts of the world. Some of the population databases are restricted to a particular community whereas most of the databases are the collection of exomes and genomes data with participation from every continent.

- 1000 Genomes project: (https://www.internationalgenome.org/)

The 1000 genomes project consortium is a widely used catalogue of genetic variation across multiple populations. The project was initiated in January 2008, and carried out in three phases. The first pilot phase involved 1092 individuals from 14 populations spread across the five major populations; Europe, East Asia, South Asia, West Africa and the Americas. The pilot phase 1 reported 38 million SNPs, 1.4 million short insertions/deletions and more than 14000 large deletions (Altshuler et al., 2010; Altshuler et al., 2012). The phase 2 of the project focused on the technical development. The latest phase 3 of the project has compiled over 88 million variants in total, out of which 84.7 million are SNPs, 3.7 million are short insertions/deletions and 60,000 structural variants, from 2504 individuals spread across 26 populations (Auton et al., 2015).

- Exome Aggregation Consortium (ExAC): (http://exac.broadinstitute.org/)

The Exome Aggregation Consortium (ExAC) has the consolidated exome variants from 60,706 unrelated individuals. The ExAC is a combined effort of many principal investigators involved with various disease-specific and population genetic studies (Lek et al. 2016). The exome samples have 5 major clusters corresponding to European, African, South Asian, EastAsianand admixed American populations. The ExAC has released a total of 7,404,909 high-quality variants, including 317,381 insertions or deletions.

- Genome Aggregation database (gnomAD): (https://gnomad.broadinstitute.org/)

The Genome Aggregation database (gnomAD) consists of human sequencing studies of 125,748 exomes and 15,708 genomes, generating over 270 million variants (Karczewski et al. 2019). In addition, the gnomAD project identified 443,769 high confidence predicted loss-of-function (pLoF) variants, which have been validated using animal models and engineered human cells. Besides, the gnomAD has also reported 498,257 unique structural variants (SVs) including 5,729 multi-breakpoint complex SVs and other types of SVs in the general population (Collins et al., 2019).

- CG69: (https://www.completegenomics.com/public-data/69-genomes/)

The cg69 database contains the genetic data of 69 non-diseased samples along with two matched tumour and normal sample pairs, sequenced by the complete genomics (Drmanac et al., 2002). The genetic variations include SNPs, small insertions/deletions, substitutions and complex small variants, CNVs and structural variants (SVs).

- Exome Sequencing Project (ESP): (https://evs.gs.washington.edu/EVS/)

The US National Institutes of Health (NIH) Heart, Lung and Blood Institute (NHLBI)-sponsored Exome Sequencing Project (ESP) has catalogued 1,146,401 autosomal single nucleotide variants (SNVs) in 15,336 protein coding genes from 6515 individuals of European Americans and African Americans. The other objectives of the ESP project were to estimate the age of mutation segregating in the contemporary human populations. The study concluded that 73.2% of SNVs are 5000 years old and are present both in the European Americans and African Americans, whereas the SNVs which are more than 50,000 years old are observed more frequently in the African American samples suggesting a genetic drift in European Americans as they move out of the Africa continent (Fu et al., 2013).

- Great Middle East (GME): (http://igm.ucsd.edu/gme/)

The Great Middle East (GME) genetic variation database is an whole-exome variant database from 1111 unrelated individuals of the Great Middle East countries; Persian Gulf region, North Africa, and Central Asia. This database is essentially useful for finding recessive variants among the GME populations, as consanguinous marriage is common in those countries (Scott et al., 2016).

- Catalogue of Somatic Mutations in Cancer (COSMIC): (https://www.sanger.ac.uk/science/tools/cosmic)

The Catalogue of Somatic Mutations in Cancer (COSMIC) contains 40,67,689 observed coding mutations, 9,175,462 gene expression variants, 1,271,436 CNVs and 1,87,429 structural mutations from 1,235,846 tumour samples (Forbes et al., 2017). Each mutation in the COSMIC data is tagged with SNPs status which infers if the mutation is present as polymorphism in the 1000 genomes

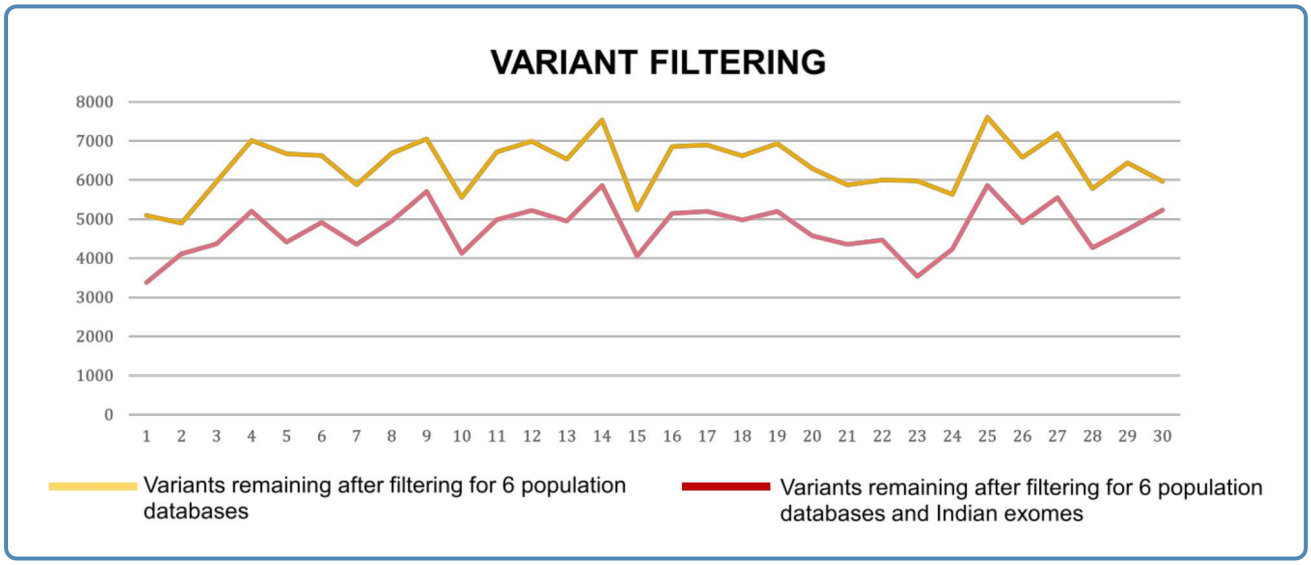


**Figure 1** Number of variants remaining in each exome after filtering for variants with MAF<0.01.

database or in a panel of control samples used in the International cancer genomic consortium (ICGC) and a pathogenicity value, enabling the identification of disease driver mutations. The COSMIC data also has information about the genetic variants conferring drug resistance to various cancer drugs.

## Absence of Indian population database and its consequences

The Indian population is not comprehensively represented in many population genetic variation databases. This is a limitation while analyzing NGS data in Indian patients since a large number of variants remain even after filtering for variants with minor allele frequency of 1 in 100 or 1 in 1000. This leads to increased time and effort needed to identify disease causing variants in diagnostics as well as novel disease-causing gene in research. We analysed 30 exomes sequenced using Agilent SureSelect V5 exome capture kit with an average of ~5,09,427 variants per sample. After filtering against the 1000 Genomes, ExAC, ESP, gnomAD, cg69, GME databases with MAF of 0.01, a total of 2,31,858 variants remained (about 5000 to 7000 variants in each sample). Further filtering with the in-house database consisting of exomes of ~800 Indian individuals revealed about 30% reduction in number of variants (Fig 1) (Unpublished data).

This shows that a significant proportion of the genetic variations are population specific. Thus, it is very important to have a detailed population allele frequency database for Indian population which will help in better and efficient diagnosis of genetic disorders using NGS technologies.

## GenomeIndia Project

The Department of Biotechnology, Government of India is planning to launch a GenomeIndia project for sequencing of 10,000 Indian genomes with partnership of about 22 institutes. (https://www.thehindu.com/sci-tech/science/biotechnology-department-will-scan-20000-indian-genomes/article28815520.ece). The data generated in this project is likely to revolutionize the field of genetic diagnostics and research in India.

In summary, the 1000 Genomes, GME, ExAC, gnomAD, CG69, ESP, COSMIC are the population genetic variation databases that contain the genetic variations data such as SNPs, indels, SVs and CNVs, generated from thousands of exomes and genomes of normal and disease individuals across multiple populations. These databases primarily can be used for filtering the polymorphisms of allele frequency in populations, which is a fundamental process for the detection of disease-causing variant(s)/gene(s) in rare Mendelian disorders and other genetic disorders.

## References


1. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010; 467: 1061-1073.
2. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491: 56-65.
3. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015; 526: 68-74.
4. Bell JI. Single nucleotide polymorphisms and disease gene mapping. Arthritis Res Ther 2002; 4: 273-278.
5. Collins FS, et al. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 1998; 8: 1229-1231.
6. Collins RL, et al. An open resource of structural variation for medical and population genetics. BioRxiv 2019; 578674.
7. Drmanac R, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 2010; 327: 78-81.
8. Fernandez-Marmiesse A, et al. NGS technologies as a turning point in rare disease research, diagnosis and treatment. Curr Med Chem 2018; 25: 404-432.
9. Forbes SA, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 2016; 45: D777-783.
10. Fu W, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 2013; 493: 216-220.
11. Gross L. A new human genome sequence paves the way for individualized genomics. PLoS Biol 2007; e266.
12. Jorde LB. Genetic variation and human evolution. American Society of Human Genetics. February 2003; 7: 2019: 28-33.
13. Karczewski KJ, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 2019: 531210.
14. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016; 536: 285.
15. Scott EM, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet 2016; 48: 1071–1079.
16. Strauss KA, et al. Genome-wide SNP arrays as a diagnostic tool: Clinical description, genetic mapping, and molecular characterization of Salla disease in an Old Order Mennonite population. Am J Med Genet A 2005; 138: 262-267.