GenomeIndia: Cataloguing the genetic variations in Indians

GenomeIndia consortium

Aim:

The human genome, comprising 3.2 billion DNA base pairs, exhibits around 0.1% variation among any two individuals, influencing disease predispositions, drug responses, and population evolution. The GenomeIndia project aims to create a comprehensive catalog of genetic variations in the Indian population. This initiative helps to create a reference haplotype structure, aiding future research by facilitating missing genetic variation imputation, designing cost-effective genome-wide arrays, and creating a biobank for DNA and plasma collection for Indian population.

Methods:

In the GenomeIndia project, we have conducted genome-wide microarray-based genotyping for more than 9000 samples and taken forward 7303 samples for whole genome sequencing after rigorous QC of array data while also considering relatedness among individuals of the same population group, using the Illumina NovaSeq 6000 at four centers: CCMB, CBR, IGIB, and NIBMG. Following stringent quality control measures, we have selected 5750 samples for joint genotyping for population level variant identification. Among these, we have, 2587, 1055, 572, 1536 samples from CBR, CCMB, IGIB, and NIBMG, respectively. These 5750 individuals represent 69 distinct ethnicities prevalent within diverse population groups across India.

Results:

Our current focus is to identify genetic variants, carrying out detailed genomic annotation, and assessing the deleterious impact and investigating the medical relevance of identified genetic variations, with in-depth inferences from specific population subgroups. Our initial analysis identified 135.48 million variations, including 117.8 million biallelic SNVs, 10.6 million INDEL biallelic, and 6.97 million multiallelic genetic variants. After applying stringent filtering criteria, we retained 50.98 million SNVs and 4.36 million INDEL. Most of these variants (65% of SNVs and 64% of INDELs) fall into the ultra-rare category, with a minor allele frequency of less than 0.1% in the overall population. We found >300,000 missense variants and splice-region-impacted SNVs, some of them causing frameshift alterations, potentially leading to significant phenotypic changes. For medically relevant findings, for example, *LDLR* gene linked to familial hypercholesterolemia, our dataset contains at least 10 unique missense variants present only in Indian population as assessed against gnomAD.. A genome-wide reference imputation panel has been constructed with the variant call dataset, showing improved imputation accuracy and allelic concordance for Indian population genotypes compared to that of TOPMed and Haplotype Reference Consortium panels.

Conclusion:

In the future, our goal is to complete 10,000 India samples in the final phase of the GenomeIndia project. The outcomes from this project will serve as a valuable national resource for the country. They will collectively facilitate future large-scale human genetic studies for researchers across India. The database of harmonized genetic variants will empower worldwide variant interpretation efforts and will likely serve as the foundational resource for next-generation basic and clinical research in India.

Krithika Subramanian (Lab of Prof. Bratati Kahali) on behalf of the GenomeIndia consortium