Indian J Med Res 145, June 2017, pp 715-717
DOI: 10.4103/ijmr.IJMR_660_17

Quick Response Code:

Commentary

Designing genetic association studies for complex traits in India

In human genome sequencing, single nucleotide variations (SNVs) have emerged as the most common polymorphisms. Except identical twins, all humans differ from one another in approximately 0.1 per cent of their genome and Single Nucleotide Polymorphism database (dbSNP, Build 132; *https://www.ncbi.nlm.nih.gov/projects/SNP/*) lists more than 37 million variants among the humans. During the last two decades, a large number of studies have been carried out to look for the association of genetic variations with human diseases, particularly complex polygenic conditions. In recent years, reliability and reproducibility of results in case-control genetic association studies increasingly show false positivity, especially for rare alleles in complex diseases. Factors such as effect size of susceptibility loci, frequency of disease alleles, frequency of marker alleles correlated with disease alleles and extent of linkage disequilibrium at or close to the region of the genome under investigation have been listed as potential problem areas[1].

There are several types of study designs (Table). Classical case-control association studies using candidate gene approach are hypothesis driven. Here, the candidate genes are selected based on risk factors or pathophysiology of the disease. In a typical study, modest number of cases and matched controls are recruited; the number depends on various factors such as population frequency of minor allele and disease prevalence. However, in most cases, major problem has been attributed to population stratification due to systematic differences in ancestry between cases and controls. In homogeneous stable populations such as White European or Chinese, there is little difficulty in selecting random non-related, age- and sex-matched controls. In populations where historical ethnic, religious and language barriers create community subdivisions, the organization of genetic association studies becomes somewhat complicated. In many populations of India, the divisions are further compounded by intra-community marriage and by marriage between close biological relatives. Due to these factors, failure to explicitly control for caste/*biraderi* (brotherhood) membership and the presence of consanguinity can seriously jeopardize and may totally invalidate the results of association/case-control studies and clinical trials[2].

India is a vast country with anthropological, linguistic, ethnic and religious diverse population. The successive waves of migration and numerous boundaries such as firm social customs and endogamy have shaped the population diversity in India. There are considerable SNP allele frequency differences between the groups in India. In systematic surveys of human variation in India focused on anthropometric traits, Reich *et al*[3] documented a high level of population substructure in India. According to their model, Indian population comprises two ancestral alleles: ancestral North Indian (ANI) which is genetically close to Middle Easterners, Central Asians and Europeans, and ancestral South Indian (ASI) as distinct from ANI and East Asians. Based on it, most of the mainland Indian population is an admixture of the two ancestral alleles and the admixture of ANI ranges from 39 to 71 per cent in most Indian groups. Thus, the extant Indian populations are admixtures of both ANI and ASI. In addition to the large number of indigenous populations, India has experienced immigration of several populations in the past, further adding to the complexities of Indian population structure[4,5].

When performing association studies in populations that have not been the focus of large-scale investigations of haplotype variation, researchers often rely on genomic databases in other populations for study design and analysis[6]. In case-control association study, minor allele frequency (MAF) of desired SNP is based on dataset of closely related population. To create such database on genetic variation in world populations, an

Table. Genetic association study designs

Type of study	Number of SNPs	Number of samples	Advantage	Limitations	Technique	Remarks
Candidate gene	Limited	Limited	Low cost and easy to carry out	Based on hypothesis, limited utility	RFLP/TaqMan probes, mass array, *etc*.	Possible in low cost and limited resource setting
GWA	Million or more	Thousand or more in primary and more for replication	Hypothesis independent	Mainly tagged (tSNPs) rather than casual SNPs, Expensive study	SNP microarray	Require multi-centric sample collection and large funds
Rare variants	Whole genome/ exome/ gene panels	In hundreds	Rare variants more likely to associate with complex phenotypes	Data interpretation requires extensive bioinformatics support	NGS	Same as GWAS

NGS, next-generation sequencing; GWAS, genome-wide association studies; tSNPs, tag single nucleotide polymorphisms; RFLP, restriction fragment length polymorphism

international HapMap project (*https://www.genome.gov/hapmap/*) was initiated in 2002 which served as a reservoir of information on common SNPs among the major world populations. As India has more than 1.25 billion populations, it is under-represented by genotypic dataset of small number of Gujarati Indians living in Houston (GIH). Subsequently, the HapMap project is largely replaced by 1000 genome which has expanded the number and ethnic communities. In addition to GIH, Indian Telugu in the UK dataset has also been added in the 1000 genome datasets (*http://www.1000genomes.org/*).

Although most researchers were aware of considerable differences between MAF, particularly in North and South Indian populations, very few studies had documented this fact. D'Cunha *et al*[7] in this issue reported the findings of genotyping 10 SNVs from genes associated with autoimmune disorders genotyped in 370 healthy individuals belonging to six different caste groups in southern India. From their data, they estimated genetic divergence and phylogenetic relationship within the various caste groups and other HapMap Phase 3 (*http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html*) populations [Yoruba people of Ibadan, Nigeria; Utah residents with Northern and Western European ancestry from the CEPH collection (CEU); Han Chinese from Beijing; Japanese from Tokyo and GIH, USA]. They did not find significant differences for MAF among different caste groups from Southern India included in their study. However, when all six caste groups were clumped together and compared with GIH population, the average Wrights's FST was 0.38 per cent. The MAF of Dravidian population for eight out of ten SNPs was significantly different from GIH in HapMap dataset[7]. For phylogenic relationship, the authors carried out a multidimensional scaling analysis between Dravidian and HapMap 3 populations. There was a clear separation between Dravidians and rest of the HapMap populations. The GIH population clearly aligned closer to CEU population. From their results, the authors suggested that Dravidian language-speaking populations of south India did not show specific patterns of genetic differentiation, but significant difference existed between the genetic architecture of non-tribal populations from the north and south of India. Therefore, in case-control association studies for complex diseases in Southern or Northern India, population controls should invariably be derived only from respective populations.

One of the persistent challenges of genetic association studies is the replication of genetic marker-disease associations across ethnic groups. It has been suggested to employ isolated population groups to conduct association studies of complex diseases to avoid various pitfalls, including those arising from population stratification[6]. In such a scenario, genotyping is carried out at tagSNP (tSNP) loci. However, tSNPs identified in one isolated population need to be used in another. Unless tSNPs are highly portable across populations, this strategy may result in loss of information in association studies. Sarkar Roy *et al*[8] examined the issue of tSNP portability by sampling individuals from 10 isolated ethnic groups from India. Their results showed that portability was low across the isolated Indian ethnic groups. By comparing their data with sequencing,

they recommended resequencing of a small number of individuals to discover SNPs and identify tSNPs, in which a disease association study is to be conducted.

In recent years, molecular genetic studies of complex diseases and candidate gene association studies involving small number of SNPs have largely been replaced by unbiased genome-wide association studies (GWAS) (Table). However, it requires larger number of cases and controls genotyped by tagger SNPs. Even in GWAS using tSNP microarrays, common identified alleles account for a fraction of heritability in complex phenotypes. Therefore, the current focus is shifting from common to rare variants which are believed to exert larger effect on phenotypes[9]. Genotyping of rare variants can be possible using whole genome or exome sequencing using next-generation sequencing (NGS) platforms. Although cost of DNA sequencing on NGS has significantly reduced but considering large sample size in association studies, it is still prohibitively expensive for developing countries. In future, as sequencing costs continue to decline, larger sequencing studies will yield clearer insights into the biological consequence of rare mutations and may reveal genes that are involved in the aetiology of complex multifactorial diseases.

Acknowledgment

The first two authors, (BM and RM) are Emeritus Medical Scientists of ICMR, and the third author (SK) is ICMR Senior research fellow.

Balraj Mittal[1,*], Rama Devi Mittal[2] & Surendra Kumar[3]

[1]Department of Biotechnology, Babasaheb Bhimrao Ambedkar University & Departments of [2]Urology & [3]Medical Genetics, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow 226 014, Uttar Pradesh, India

[*]*For correspondence:* balrajmittal@gmail.com

Received April 21, 2017

References

1. Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 2004; *5* : 89-100.
2. Bittles AH. Population stratification and genetic association studies in South Asia. *J Mol Genet Med* 2005; *1* : 43-8.
3. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature* 2009; *461* : 489-94.
4. Tamang R, Singh L, Thangaraj K. Complex genetic origin of Indian populations and its implications. *J Biosci* 2012; *37* : 911-9.
5. Tamang R, Thangaraj K. Genomic view on the peopling of India. *Investig Genet* 2012; *3* : 20.
6. Pemberton TJ, Jakobsson M, Conrad DF, Coop G, Wall JD, Pritchard JK, *et al.* Using population mixtures to optimize the utility of genomic databases: Linkage disequilibrium and association study design in India. *Ann Hum Genet* 2008; *72* (Pt 4) : 535-46.
7. D'Cunha A, Pandit L, Malli C. Genetic variations in the Dravidian population of South West coast of India - Implications in designing case - control studies. *Indian J Med Res* 2017; *145* : 753-7.
8. Sarkar Roy N, Farheen S, Roy N, Sengupta S, Majumder PP. Portability of tag SNPs across isolated population groups: An example from India. *Ann Hum Genet* 2008; *72* (Pt 1) : 82-9.
9. Kosmicki JA, Churchhouse CL, Rivas MA, Neale BM. Discovery of rare variants for complex phenotypes. *Hum Genet* 2016; *135* : 625-34.