PRINT: ISSN 0972-3757 ONLINE: 2456-6360
Int J Hum Genet, 9(2): 97-103 (2009)
DOI: 10.31901/24566330.2009/09.02.02

Genome-wide Association: "A Revolutionary Approach"

Vipin Gupta[1], K. N. Saraswathy[2], Rajesh Khadgawat* and M. P. Sachdeva[3]

Biochemical and Molecular Anthropology Laboratory, Department of Anthropology, University of Delhi, Delhi 110007, India
E-mail: [1]udaiig@gmail.com, [2]< knsaraswathy@yahoo.com>, [3]< mpsachdeva@rediffmail.com>
****Department of Endocrinology & Metabolism , All India Institute of Medical Sciences, Ansari Nagar, New Delhi, India***
Telephone: +91 11 26588641 Extn. 3237 (O), 4760 (W), Fax: +91 11 26589386
E-mail: rajeshkhadgawat@hotmail.com

KEYWORDS Genome-wide Association Studies. Reverse Genetics. Complex Disorders. Biology

ABSTRACT Genome-Wide Association studies (GWAS) have brought a revolutionary change or paradigm shift in detecting novel variants for complex disorders and shifting the burden of finding the biological relevance of these newly discovered variants on biochemists and physiologists, hence it is a movement from forward to reverse genetics. Here we discuss the role of such studies with GWAS designs from anthropological perspective for conducting futuristic India specific GWAS.

"Philosophy only demands creativity to solve complexity" –Udai G

INTRODUCTION

The evolution of gene mapping approaches for complex disorders began with the conventional ways of detecting disease genes through family based linkage scans, looking for portions of genome that moves along with the concerned trait in families, which may (but rarely) follow a particular pattern in their transmission, and then followed by variety of fine physical mapping techniques. This approach was highly successful for mapping disorders particularly with high penetration level i.e. Mendelian disorders, but its utility is still a matter of huge skepticism for identifying multiple low-penetrative variants involved in complex disorders. When Risch and Merikangas in 1999 argued that association studies should be more powerful than linkage studies, then geneticists started exploiting two alternative approaches to conduct association studies (alternative to Linkage studies) for complex disorders, first is the *Direct approach*, which involves testing each putative causal variant for correlation with the disease. At present, this approach is limited to sequencing the *functional parts of candidate genes* (selected on the basis of a previous functional or genetic hypothesis) for potential disease-associated candidate variants. Second is the *Indirect approach,* where a *set of sequence variants* [which are in linkage disequilibrium (LD)] in the genome could serve as genetic markers to detect association between a particular genomic region and the disease, whether or not the markers themselves had functional effects. This approach is based on the principle that unrelated individuals are more distantly related than subjects from large pedigrees, thus allowing sufficient recombinant events to have taken place. Unfortunately, these two approaches together were only partly successful in exploiting the allelic architecture of complex disorders, and were able to detect a few dozen genes. One fundamentally different approach (and less exploited) was Admixture Mapping (also known as "mapping by admixture linkage disequilibrium," or MALD) provides a pathway for localizing genes that cause disease in *admixed ethnic groups*. The strategy is to identify a genomic region with an *unusually high* contribution of ancestry from *one ancestral population*, usually the population with the highest incidence of disease (Collins-Schramm et al. 2002). Because of stringent and sensitive assumptions behind this design, and difficulty in finding subtle admixed populations makes this approach of rare utility.

The further extension of genetic association studies is the Genome-Wide Association (GWA) approach which is defined as an unbiased yet fairly comprehensive association study that surveys most of the genome for causal genetic variants even in the absence of convincing evidence regarding the function or location of the causal genes. That is, in contrast with

candidate gene approach, no prior information regarding gene function is required in Genome-Wide Association Studies (GWAS), not only that no assumption is necessary regarding type of variant involved, and hence variation of allele frequencies among populations also is not a prerequisite. The strength of GWAS is that it has sufficient power to detect common alleles (minor allele frequency (MAF)>5%) and also has the ability to detect rare (MAF<1%) alleles (Thomas et al. 2005). The decade long research efforts after the completion of Human Genome Project pave the way for primarily this indirect agnostic approach i.e. GWAS. These efforts incorporate the developments of various logistic and statistical tools for the genotyping of millions of SNPs per individual and analyzing this huge amount of raw data respectively. For instance, because it is impossible to scrutinize and monitor every single variable for its quality in an individual, quality control needs to be highly automated. Therefore the statistical software of genotype calling algorithms (BRLMM and CHIAMO) for assigning discrete genotype call to each genotype cluster and getting reliable signal intensity plots (without error in measurement) for the evaluation of quality of genotype calls were developed (McCarthy et al. 2008, Ziegler et al. 2008). Then the frequency of missing genotype of each SNP per individual is another important aspect of quality control that helps in the exclusion of SNPs if they have missingness beyond the chosen criteria (for example in Wellcome Trust Case Control Consortium study [WTCCC] it was >3%). Various multipoint imputation methods were also developed to predict (or impute) the missing SNPs statistically and thus gaining added value for the analysis by combining information with the help of HapMap reference data (Marchini et al. 2007). The availability of IInd generation HapMap data of 3.1 million SNPs also helped in the selection of informative tag SNPs or the evaluation of genomic coverage in terms of linkage disequlibrium and power of commercially available genotyping chips (Affymetrix, Illumina etc.) for uncovering genome-wide associations (Barret and Cardon 2006; International HapMAp Consortium 2007). In the light of the problem of multiple hypotheses testing and high rate of ingenuine associations, some researchers also suggest that the use of Bayesian statistical approach would be more reliable than Frequentist in detecting spurious associations because this approach avoids a single threshold for genome-wide significance but depends on its ability to assign plausible probabilities to each alternative hypothesis (McCarthy et al. 2008).

An alternative complementary to GWAS is deep resequencing (which is a costly affair) of candidate genes which are often selected on the basis of probable biological function. The importance of deep resequencing can be seen in that none of genes identified by this strategy would have been found by GWAS because of low frequency of the causative disease alleles. This implies that both approaches (GWAS and deep resequencing) are needed for comprehensive finding of causal genes for complex traits (Petretto et al. 2007).

Thomas 2006, questioned that "Does this development marks the end of pathway driven research?" and suggested that "it is possible to marry the hypothesis driven and exploratory approaches in a way that will make better use of this novel but expensive technology". Now at least for sometime it has been hoped that the emergence of GWAS would provide efficient new tools for exploiting the genetic basis of many of these common causes of human morbidity and mortality.

FORWARD TO REVERSE GENETICS: A PARADIGM SHIFT

Earlier the trend was that we selected the defect at amino acid level (*candidates*) detected by biochemists and then tried to find the genetic root of these protein defects. This is known as *Forward* genetics. Due to the advent of GWAS, there is a paradigm shift i.e. now we have the knowledge of mutations in geneic regions because of inbuilt principle of design itself, but we do not have any information regarding their biochemical or physiological effect. Therefore to look forward for the biochemical or physiological effects of these known mutations on genome is known as *Reverse* genetics. Hence, instead of going from phenotype to sequence as in forward genetics, reverse genetics works in the opposite direction – a gene sequence is known, but its exact function is uncertain. To make this trend from forward to reverse genetics more explicit let us take the example of Type 2 Diabetes Mellitus (T2DM).

To date the well replicated markers for T2DM

are limited to only three variants i.e. PPARG (encoding the peroxisomal proliferative activated receptor gamma; P12 A); KCNJ11 (the inwardly rectifying Kir.6 component of the pancreatic beta cell K-ATP channel; E23K); and TCF7L2 (transcription factor 7-like 2). PPARG was first isolated in 1990 as a member of nuclear hormone receptor superfamily of ligand activated transcription factors and its role in T2DM was suggested. Beamer et al. 1997, reported the chromosomal location of PPARG gene on chromosome 3, band 3p25, and its role was confirmed by reporting the presence of missense mutation i.e. Pro12Ala in 1997. These developments explicitly suggest how functionally known candidate gene forwardly direct the genetic studies. Similarly, for KCNJ11, a candidate gene association was discovered after the discovery of its functional manifestation i.e. responsible for the ATP-sensitive K+ channel is critical because the channel links glucose metabolism to electrical activity of the cells (Nielsen et al. 2003).

Recently published four GWAS explains how these robust designs can uncover novel variants and shift the burden of finding the biological (biochemical or physiological) relevance on clinical biochemists or physiologists. Sladek et al. (2007) conducted a two-stage, genome-wide association study to identify T2D susceptibility loci. In first stage, they tested 392,935 SNPs in 1363 French case control cohort. They initially studied diabetic subjects with BMI < 30 kg m^{-2} with at least one affected first degree relative, and in order to decrease phenotypic heterogeneity and to enrich variants determining insulin resistance on beta cell dysfunction through mechanism other than obesity. Because of unequal male/female ratio in their case controls, they additionally genotyped 12666 sex chromosome SNPs, separately for each gender. They also conducted population stratification analysis (using component analysis) to find out any spurious association and they found spurious association for only one SNP. None of the other previously identified T2D genes (like PPARG, KCNJ11) were found to be significant, because of limited power of stage 1 to detect their modest effect. From their stage 1 results they prioritized 57 SNPs showing significantly associated in stage 2 analysis in case/control: 2617/2894 (cases may or may not have a family history or to be lean like stage 1). In total 8 SNPs representing five unique loci (TCF7L2; SLC30A8; HHEX; LOC387761; EXT2) showed significant association after Bonferroni correction was applied for the 57 SNPs. The population attributable risk (PAR) of four novel loci was 70%. It is worth noting that for three of the four novel loci, the risk allele was the major one. The findings also suggested that allelic heterogeneity did not seem to be large.

The GWAS conducted by Scott et al. (2007) and Zeggini et al. (2007) along with largest association study i.e. WTCCC (2007) independently confirmed the association of these novel gene regions with T2DM. Their findings offer new avenues for exploring the pathophysiology of this complex disorder.

Except TCF7L2, the seven variants SLC30A8, HHEX, FTO, CDKAL1, CDKN2A, CDKN2B, and IGF2BP2 are novel and the product of robust GWAS designs. Although these GWAS mentioned some *broad or distantly overlapping* functions of these novel variants, but it seems to be very much similar to "*Kite Flying*", that is saying hypothetically without discovering firm and absolute pathway involved. Hence to discover the biological importance of these promising novel variants, researchers may look for animal models to exploit the clinical relevance of these upcoming variants. Although with straight forward genome-wide resequencing of exonic regions to screen non-synonymous SNPs (nsSNPs) as done by WTCCC and The Australo-Anglo American Spondylitis Consortium 2007 could also facilitate in detecting the causal nsSNPs directly.

But to maximize the biological information gathered from GWAS we need to develop novel experimental designs so that gene expressions data could be used to identify causal mechanisms in different cell lines. These novel causal pathways provide new opportunities for clinical advancement of genetic benefit to all suffering from concerned disease and also in the field of personalized medicine (McCarthy et al. 2008).

STRATEGIC USE OF ANTHROPOLOGY IN GWAS RESEARCH DESIGNS

Genome-wide association studies despite of their due advantages, still suffer from the conventional limitations of typical case/control association designs. The first problem is population stratification which is the "mixing of

chromosomes from two different populations; typically, haplotype frequencies differ in these populations" (Gordon and Finch 2005). In simple words it is equivalent to selecting cases and control from two different populations, having different ancestry. Meaning thereby that there must be substantial variation across ethnicities in the frequency of the variant genotype concerned. But this stratification is partly hidden, and many a times researchers are not able to discern such existing stratification, and make mistakes while recruiting subjects. The immediate consequence of this bias is that it increases the chances of false positives due to inflated type I statistical error, and does not become smaller with increase in sample size; on the contrary it inflates more in much larger GWAS (Thomas et al. 2005). It could lead to three relatively distinct problems: confounding; cryptic relatedness, resulted in overdispersion of the test statistic; and selection bias (Clyton and McKeigue 2003). Hence this problem could be one of the major cause of lack of reproducibility of association studies on other populations. Second problem is of non homogeneity of the sample, cases as well as controls must share various environmental aspects like lifestyle, socioeconomic status, or greater similarity in the presence of one or more risk-conferring alleles. In epidemiological designs it is important to understand that not the existence of some bias that is intrinsically troublesome but its potential magnitude is relevant. Therefore, it is urgent to understand that "ethnicity per se does not explain the risk, it is only a marker for individuals at a similar risk" (Wacholder et al. 2002).

There are various ways to protect association designs from population stratification, for example usage of family based case control design, but it suffers from the limitation of loss of some power because of overmatching (Gauderman et al. 1999) along with extra genotyping cost; second is the method of genomic control (highly used and proposed by Devlin et al. 1999) which uses unlinked markers to control stratification; third is method of structured association (Pritchard et al. 2000), and fourth is multivariate approaches (generally used in genome-wide scans) like principal component analysis which is used to infer continuous axes of genetic variation (eigenvectors) that reduce the data to a small number of dimensions, whilst describing as much of the variability between individuals as possible (Price et al. 2006). Alternative strategy is to detect highly informative markers and use them in the testing for population stratification in case-control genetic association studies (Pritchard and Rosenberg 1999). Excluding family based method which reduces stratification but does not eliminate, rest of these methods are used after the sample collection has been completed. In contrast to this we suggest that it will be better to use designs which recruit subjects without stratification bias and that too before completion of field work because "Prevention is always better than cure".

The recent examination of GWAS again underlined the conventional challenges of association studies i.e. non-replication and inconsistency especially in the framework of cumulative meta-analysis. The ideal situation of the existence and detection of gene-phenotype associations through GWA, along with successful replications would still remains to be realized practically (Ioannidis John PA 2007). Although literature is filled with various epidemiological research designs to reduce the limitations of association studies like cohort studies, nested case control designs, case cohort designs, and family based designs, but the basic principle behind all these different strategies is to select a homogenous population group. However, Hardy 2002, argued that there is no evidence that population stratification is the real problem for the lack of reproducibility of association studies; and suggested that currently we need simpler research designs with less tests because statistical sophistication is blinding us to terrible problem. Therefore, we recommend strategic usage of anthropology in research designs concerned with GWAS. For instance, in India there is a huge existence of number of anthropologically well defined communities which follow clan exogamy and caste endogamy (i.e. socio-cultural groups who marry among themselves) which are different in terms of their cultural pattern of living. And there exists only negligible deviation in terms of marrying beyond the clan boundaries of a particular community. The recent Indian genome variation consortium, 2008 revealed a high degree of genetic differentiation among Indian ethnic groups and suggested that pooling of endogamous populations without regard to endogamous nature of subdivided Indian population will result in false inferences in association studies. They also

criticized heavily the reference of people of India as 'Indian' in many population genetic studies.

Thus, these well defined cultural groups can be a wonderful resource for exploiting GWAS with more power and better clarity as in these endogamous groups the genetic basis of disease susceptibility is likely to be less complex than in heterogeneous populations, so there would be more chances of getting a well defined or limited allelic spectrum of the concerned disease. Although improbable, but if the evidence of non-replication still persist, then this time non-replication would be expected to be informative about the cause of heterogeneity of the concerned population. McCarthy et al. (2008) suggest that "we should be wary of heterogeneity as a rationale for failure to replicate as over-eagerness to deploy such an explanation would mean that no report of association could be refuted". Therefore we should look for informative heterogeneity (identification of its source) and this could only be possible when we assure the reasonably high power of our studies to detect genuine associations. Furthermore in the context of meta-analyses which has become a standard practice for the publication of genome-wide association studies, the explanation of between-study heterogeneity is an utmost important goal for meta-analysis because in the presence of between-study heterogeneity in genetic effects, the precision derived from combining data may be spurious (Ioannidis et al. 2007).

Lack of power is always a pertinent question in disease association studies (especially for gene-gene and gene-environment interactions), which can be increased by using flexible matching strategies i.e. the degree of matching in case-control studies with frequency matching for the environmental exposure (Stürmer and Brenner 2002). Research based on endogamous groups also enhances the power of the study significantly in terms of finding rare variants involved (if any) in addition to the common ones, and thus highlighting the population attributable risk more credibly. Comings 2003, argues that the real problem in replication of association studies related with complex disorders is a combination of the two, first is each gene accounts for only small percent of the variance and second is genetic heterogeneity i.e. the greater heterogeneity powerfully interacts with the low variance attributed to each gene to produce a situation in which variation from study to study is the expected outcome of association studies. Our recommended design, which is based on endogamous groups, almost eliminates the genetic heterogeneity to a finer extent. But the only downside of using endogamous groups is that we cannot generalize the outcome of these population-wise compact designs to general heterogeneous populations or to other endogamous groups. Another important aspect on genome wide association study is to detect gene-environment effects (G x E effects) but, because the mapping and target populations may differ substantially in their environments, these effects confound the replication efforts on heterogeneous populations. In case of GWAS where we spend huge money (few years back it was prohibitive to even think of such on amount), we should first think of increasing the breadth of our sample size i.e. by incorporating various endogamous groups rather than restricting to only one heterogeneous group. In other words, we should redesign over our fieldwork strategies instead of spending huge amount of public money without appropriate hard work during and before fieldwork, and then stupidly correcting these errors with the help of available statistical tools to justify our results. Hence, by modifying our designs slightly, we would easily be able to limit the drawback of non-generalization of results.

The proof for common disease common variant hypothesis (which assumes much of the genetic variation of a common disease is due to few common variants) was one of the reasons for the onset and rising of GWAS, but unfortunately there is no great deal of evidence in support of this hypothesis. But if we go for anthropologically recognized populations the chances of detecting rare variants are high because of good sharing of phenotype in culturally homogenous population with restricted mating pattern.

Finally, this study design also helps in better detection of highly transferable tag SNPs for populations of Indian subcontinent so that optimum utilization of linkage disequilibrium (LD) based association studies could be achieved. This would also enhance the power of multipoint imputation analysis (statistical way of predicting missing SNPs) because of the availability of local pattern of LD.

In conclusion we would like to suggest a unique design for India specific GWAS: to first

incorporate different endogamous groups (at least 5 to 10 - previously well studied), and then after properly planned multistage testing, which considerably reduces the complexities (phenotypic level, genetic level, and model complexities) of genetic-epidemiologic datasets (Dube et al. 2007), we should focus on another set of endogamous groups in the second stage. This will further help in systematic meta-analysis of genome-wide association data on different endogamous populations. It is based on the principle that because of different and sometimes stringent socio-cultural obligations makes these groups reproductively isolated from other groups, and therefore different in their evolutionary features of genetic predisposition.

REFERENCES

Beamer BA 1997. Chromosomal locakization and partial genomic structure of the human peroxisome proliferator activated receptor-γ (hPPAR γ**)** gene. *Biochemical Biophysical Research Communications*, 233: 756-759.

Barret JC and Cardon LR 2006. Evaluating coverage of genome-wide association studies. *Nature Genetics*, 38(6): 659-662.

Clayton DG, McKeigue PM 2001. Epidemiological methods for studying genes and environmental factors in complex disorders. *Lancet*, 358: 1357-1360.

Collins-Schramm, Heather E, Phillips Carolyn M, Operario DJ et al. 2002. Ethnic-Difference Markers for Use in Mapping by Admixture Linkage Disequilibrium. *American Journal of Human Genetics*, 70: 737–750.

Comings David E 2003. The Real Problem in Association Studies. *Am Med Genet, (Neuropsychiatric Genetics)*, 116B: 102.

Devlin B and Roeder K 1999. Genomic control for association studies. *Biometrics*, 55: 997–1004.

Dube MP, Schmidt S, Hauser E, Barhdadi A, Wang X 2007. Multisatge designs in the Genomic Era: Providing Balance in Complex Disease Studies. *Genetic Epidemiology*, 31: S118-S123.

Frayling Timothy M 2007. Genome–wide association studies provide new insights into type 2 diabetes aetiology. *Nature Review Genetics*, 8: 657-662.

Gauderman W James, Witte John S, Thomas Duncan C 1999. Family based Association Studies. *Journal of the National Cancer Institute Monographs*, 26: 31-37.

Gordon D, Finch Stephan J 2005. Factors affecting statistical power in the detection of genetic association. *Journal of Clinical Investigation*, 115(6): 1408-1417.

Hardy John 2002. The Real Problem in Association Studies. *American Journal of Human Genetics*, 114: 253

Hirshhorn Joel N, Daly Mark J 2005. Genome-Wide Association Studies for Common diseases and Complex traits. *Nature Reviews*, 6: 95-108.

Ioannidis John PA 2007. Non-replication and inconsistency in the genome-wide association studies. *Human Heridity*, 64: 203-213.

Ioannidis JP, Patsopoulos NA, Evagelou E 2007. Heterogeneity in Meta-analyses of genomewide association investigations. *Plos one*, Sept 5; 2(9) e841: 1-7.

INDIAN GENOME VARIATION CONSORTIUM 2008. Genetic landscape of the people of India: A canvas for disease gene exploration. *Journal of Genetics*, 87(1): 3-20.

Kruglyak Leonid 2008. The road to genome-wide association studies. *Nature Reviews Genetics*, 9(4): 314-8.

Marchini J, Howie B, Myres S, McVean G, Donnelly P 2007. A new multipoint method for genome-wide association by imputation of genotypes. *Nature Genetics*, 39(7): 906-913.

McCarthy MI, Abecasis GR, Cordon LR, Goldstein DB, Little J et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Review Genetics*, 9: 356-369.

Nielsen EM, Hansen L, Carstensen B, Echwald SM, Drivsholm T et al. 2003. The E23K Variant of Kir6.2 Associates with Impaired Post-OGTT Serum Insulin Response and Increased Risk of Type 2 Diabetes. *Diabetes*, 52: 573-577.

Pritchard JK, Stephans M, Rosenberg NA, Donnelly P 2000. Association mapping in structured populations. *American Journal of Human Genetics*, 67**:** 170-181.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA et al. 2006. Principal Component Analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8): 904-909.

Petretto E, Liu ET, Aitman TJ 2007. A gene harvest reavealing the archeology and complexity of human disease. *Nature Genetics*, 39(11): 1299-1301.

Reisch N and Merkingas K 1996. The future of genetic studies of complex human diseases. *Science*, 273: 1616-1617.

Rosenberg, Li, Ward, Pritchard 2003. Informativeness of Genetic Markers for Inference of Ancestry. *American Journal of Human Genetics*, 73: 1402–1422

Stürmer T, Brenner H 2002. Flexible Matching Strategies to Increase Power and Efficiency to Detect and Estimate Gene-Environment Interactions in Case-Control Studies. *American Journal of Epidemiology*, 155: 593–602

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li et al. 2007. A Genome-Wide Association Study of Type 2 Diabetes in Finns Dtetects Multiple Susceptibility Variants. *Science*, 316: 1341-1345

Sladek R, Rocheleau G, Rung J, Dina C, Shen L et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445: 881-885.

Thomas DC, Haile RW, Duggan D 2005. Recent Developments in Genomewide Association Scans: A Workshop Summary and Review. *American Journal of Human Genetics*, 77: 337-345**.**

Thomas Duncan C 2006. Are we ready for Genome-Wide Association Studies? *Cancer Epidemiology Biomarkers and Prevention*, 15(4): 595-598.

The Wellcome Trust Case Control Consortium (WTCCC) 2007. Genome-Wide Association Study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature*, 447: 661-678.

The International HapMap Consortium 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(18): 851-862.

Wacholder S, Rothman N, Caporaso N 2002. Counterpoint: Bias from Population Stratification Is Not a Major Threat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer. *Cancer Epidemiology, Biomarkers & Prevention*, 11: 513-520.

Wellcome Trust Case Control Consortium & The Australo-Anglo American Spondylitis Consortium 2007. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genetics*, 39(11): 1329-1337.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS et al. 2007. Replication of Genome-wide Association studies in UK samples Reveals Risk Loci for Type 2 Diabetes. *Science*, 316: 1336-1339.

Zeigler A, Konig IR, Thompson JR 2008. Biostatistical aspects of Genome-wide association studies. *Biometrical Journal*, 50(1): 8-28.