GWAS as the Detective to Find Genetic Contribution in Diseases

Simanti Bhattacharya
University of Kalyani, India

Amit Das
University of Kalyani, India

INTRODUCTION

In the era of unprecedented advancement in medical and technological sciences, Garrod A.E., a physician to the Hospital of Sick Children, in the year of 1902, reported a case of alkaptonuria that he described as "not the manifestation of disease but is rather of the nature of an alternative course of metabolism…". That was the first report where the possibility of underlying "molecular evidence" behind human disease came in the lime light. Human diseases and their genetic contribution share a complex and intricate relationship, yet to be explored fully. For few cases, phenotypes (diseased condition) could directly be associated with gene, experimentally, whereas a large number of genetic associations behind disease state remained hidden.

That called for a situation where mapping of gene knowledge between afflicted and un-afflicted individuals can be mapped to find out the difference at each point (each nucleotide position or *allele* or *variant*) and the number of occurrences of the mismatched alleles in the diseased individuals (*allelic frequency*) with an assumption that if any allele has higher frequency to appear in the diseased individual then that is associated with the diseased trait. That can be translated as 'scan through entire gene' for gene to disease relation mapping or rather, as genetic language, G*enome* W*ide* A*ssociation* S*tudy (GWA or GWAS).*

These GWAS data not only provide us with the information on the disease association with the gene level knowledge but also enable a deeper understanding of the entire scenario generating a landscape of gene with its minute changes that can be extrapolated to genes coding (impact on protein production) or non-coding regions (impact on protein production regulation), the transcription factor binding sites (regulating transcription), epigenetic modification probabilities (regulation in genetic coding), pathways involved (visualizing the upstream or downstream possible effects) extending to heritability of the diseases. These all impose final impact on the phenotype which is nothing but the diseased state to us. Thus, insight generated with GWAS leads us to understand the actual reason or mechanism behind disease onset that, in turn, guides scientists to find novel druggable targets for more efficient medications or some times, to look for personalized medications (Bush & Moore, 2012).

With the completion of Human Genome Project (human DNA sequence) in 2003 (International HapMap Consortium, 2003) and the International Hapmap project (haplotype map of the human genome) in 2005, scientists are well-equipped with resources to correlate genetic contribution to disease onset. Success of the GWAS reflects in identifying the genetic factor contributing to Parkinson's disease, Crohn's disease, type 2 diabetes and obesity to name a few. These GWAS data can also be accessed through various repositories. However, its smaller variant size, unavailability of replicated reports, smaller population size under study stand as limiting factors to uncover a larger

DOI: 10.4018/978-1-5225-2255-3.ch041

portion of genetic information to understand properly. With the rapid advancement in research, these limitations will be overcome to generate a better understanding of the entire scenario of disease with GWAS concept.

BACKGROUND

Life material nucleus i.e. DNA is composed of four basic entities and they are: the purines: Adenine (A), Guanine (G) and the pyrimidines: Cytosine and Thymine (T) arranged in a specific pattern that carries information of life. DNA is double helical element where these purines and pyrimidines pair up with each other (A with T and C with G). The complex of different DNA stabilizing and regulatory molecules (protein, RNA etc) along with DNA is collectively known as Genome. Since not only the presence of these basic entities but also their relative position is crucial to maintain the information for life and its sustainability, small variation of these elements (mutation or polymorphism or SNPs) at any of the otherwise conserved position on DNA has been found to bring change in the system. In Pre-GWAS (Ertekin-Taner, 2010) era researchers managed to characterize the genetic association behind disease that followed a Mendelian pattern of inheritance that largely depended on generation wise co-segregation of causal variants with marker alleles that simply followed Hardy-Weinberg Equilibrium. Cause to genetic variation was thought to be chromosomal cross over at chaismata during meiosis. To explain, suppose, a family or rather a pedigree of a population was studies for a disease that was found to be occurring due to certain enzyme mutation. This, at that time, was considered to be the candidate gene and was followed by characterization, identification of mutated allele as well as developing therapeutics against it. But there could be number of possible reason behind genetic variation within or among the population, random mutations being the ultimate source of genetic variations. Along with that not only inheritance but also other effects including environment, living style, age etc contribute significantly to genetic changes. And these also can contribute to disease state if the changed allele gets a higher frequency is diseased population than in healthy population. But during GWAS era scientists actually received the scanned report of entire genome that help understanding the complex relation of DNA level mutation and disease phenotypes. As for human cases, each disease may require looking at hundreds or thousands of positions to identify SNPs and associated genes that may contribute to risk of developing a certain disease. Increasing evidences show that GWAS represents a really powerful technique to identify these marker alleles at relatively higher speed and precision. GWAS is based on linkage disequilibrium (LD) principle where loci that are physically close together show stronger LD than the loci that are far apart (Visscher, Brown, McCarthy, & Yang, 2012). The strength of LD for an effective population size also decides the number of genetic marker required to specify the haplotype. GWAS also hypothesizes the 'allelic frequency' of a rare variant will be in low LD with nearby common variant. As time advanced, GWAS started exploring different type of variations and linkage to disease for example rare variant identification through evolutionary model, copy number variation (CNV) that include deletion and/or duplication of DNA segments of diverse size and frequency.

WORKFLOW FOR GWAS METHODS

For any experiment to be successful, a well-structured experimental design is essential. GWAS is not exceptional in this case. It systematically follows basic steps (Figure 1), given below:

- **Sample Collection:** A large cohort collection (>1000) for case and control is essential for initial set up. Chip based microarray technology that assays 1 million or more SNPs from population under study