Biomacromol. J., Vol. 3, No. 1, 1-4, December 2017.

Biomacromolecular Journal
www.bmmj.org

Single Nucleotide Polymorphisms and Association Studies: A Few Critical Points

S.A. Angaji*

Department of Cell and Molecular Sciences, Faculty of Biological Sciences, Kharazmi University, Tehran, Iran

(Received 9 June 2017, Accepted 16 October 2017)

ABSTRACT

Single nucleotide polymorphism (SNP, pronounced snip) represents a DNA sequence variant of a single base pair with the minor allele occurring in more than 1% of a given population. The broad field of applications for SNPs induced a pressing need for effective instruments for SNP detection. During the last decade a considerable number of methods for SNP discovery (search for new SNPs) and detection (recognition of already known SNPs) were developed. Studying the association between quantitative phenotype and SNPs has been a major challenge in genetics. To understand underlying mechanisms of complex phenotypes, it is often necessary to consider joint genetic effects across multiple SNPs. In this article, SNPs and their role in association studies were reviewed.

Keywords: Molecular markers, Candidate genes, Linkage

SNP AS THE THIRD GENERATION OF MOLECULAR MARKERS

Uncovering DNA sequence variations that correlate with phenotypic changes, *e.g.,* diseases, is the aim of sequence variation studies. Common types of sequence variations are SNPs. SNPs are the third-generation molecular markers. Their advantages, including high frequency across the whole genome, ease of detection, cost efficiency and co-dominance make SNP markers particularly popular (Angaji, 2011).

There are an approximately 56,000,000 SNPs in the human genome. Any two randomly chosen DNA molecular marker are likely to differ at one SNP site about every 1000 bp in noncoding DNA and at about one SNP site every 3000 bp in protein-coding DNA. Depending on their genomic locations, the phenotypic consequences of the SNPs differ. SNPs in coding regions of genes alter the amino acid sequence of the encoded proteins, thus affecting their structure and function, and consequently their physiological role. SNPs located in the regulatory regions of a gene may affect the binding of transcription factors, thereby influencing the expression level of the gene. SNPs located in non-coding regions of the genome have no known impact on the phenotype of an individual. These SNPs are useful as genetic markers in forensic identification, in tissue typing, for population genetic studies and evolutionary studies (Walker and Rapley, 2005).

The SNP detection technique can be divided into two areas: Scanning DNA sequences for previously unknown polymorphisms and screening (genotyping) individuals for known polymorphisms. Although the technologies capable of scanning DNA for new polymorphisms can be used in screening individuals for known polymorphism, there are many options for SNP genotyping (Angaji *et al*., 2017).

SNP scanning approaches are based on single-strand conformation polymorphism (SSCP) analysis, heteroduplex analysis and direct sequencing (Liu, 2007). SNP screening methods fall into one of four categories: allele-specific hybridization, primer extension, oligonucleotide ligation, and invasive cleavage (Cox Gad, 2007).

ASSOCIATION STUDIES

Genetic association studies test for a correlation between disease status and genetic variation to identify candidate genes or genome regions that contribute to a specific disease. A higher frequency of a SNP allele or genotype in a

*Corresponding authors. E-mail: angaji@khu.ac.ir

series of individuals affected with a disease can be interpreted as meaning that the tested variant increases the risk of a specific disease (although several other interpretations are also valid). SNPs are the most widely tested markers in association studies (Dudley and Karczewski, 2013).

There are two approaches for genetic dissections of complex and quantitative traits, *i.e.*, genome-wide scanning and candidate gene approach, each of which has specific advantages and disadvantages. Genome-wide scanning usually proceeds without any presuppositions regarding the importance of specific functional features of the investigated traits, and the principal disadvantages include both cost and being resource intensive (Zhu and Zhao, 2007).

There are two main types of genome-wide scanning: population-based case-control studies and family-based studies. Family-based association studies are often most efficiently aimed at finding rare variants underlying rare conditions or rare sub-phenotypes of a common condition. Their design is not the focus of this protocol. Population-based (defined here as non-family based) case-control studies have become the most popular design to find common polymorphisms thought to underlie complex traits (also termed 'common disease common variant' hypothesis). Genetic linkage and association between two loci are both related to recombination- in the former, recombination events are scored over a limited number of observed generations, whereas the latter relies on large numbers of unobserved recombination events in past generations. As generations go by after an initial disease mutation has occurred, recombination events with surrounding markers tend to occur closer and closer to the disease locus so that measurable association between disease and marker loci extends only over short distances of up to 100 kb, corresponding approximately to a recombination fraction (represented by θ) of 0.001, given 1 Mb ≈ 1 cM. Most differences between association and linkage analysis are due to this difference in the number of generations. Therefore, association analysis using common variants generally allows for finer mapping than linkage analysis (Ott *et al*., 2011; Ott *et al*., 2015). Several approaches may be used to identify candidate genes that deserve to be examined in further studies on association studies.

Positional Cloning: from Genetic Variation to Function

Positional cloning (also known as reverse genetics) is used to determine the location of a gene without the understanding of its function and isolating the gene starting from the knowledge of its physical location in the genome. This approach is classically used when there is little or no understanding of the function of a defective gene. The basic strategy of positional cloning is phenotype → genome location → cloned DNA → DNA sequence → protein sequence → protein. To search for relevant candidate genes involved in a specific trait, a successful approach may be to start with familial linkage analysis to find chromosomal locations containing one or more candidate genes that nonrandomly segregate with the trait.

Functional Cloning: from Protein Function to Genetic Variation

Functional cloning (also referred to as forward genetics), refers to identification of the gene causing a trait based on fundamental information about the basic biochemical defect, without reference to chromosomal map position (phenotype → function → gene → map).

Gene Expression Approach: from mRNA to Protein Function and Genetic Variation

Transcriptomics, or quantitative gene expression profiling is the large-scale study of the transcriptome, giving a global view of all transcripts simultaneously. It helps in identification of the complete set of transcripts in a particular biological sample and estimation of their abundances under specific physiological conditions or at developmental stages. Northern blotting, real-time PCR, serial analysis of gene expression (SAGE), and microarray were hitherto used for gene expression profiling. Application of next-generation sequencing technologies to cDNA sequencing is commonly called RNA-Seq. This method of transcriptome analysis is fast and simple because it does not require bacterial cloning of cDNAs. RNA-Seq technology, in conjunction with efficient bioinformatics tools, is now more widely used to support predicted gene models, extract differentially expressed genes, find novel

transcripts in *de novo* assemblies (Kimman, 2001).

REQUIREMENTS FOR ACCURATE ASSOCIATION STUDIES

Define Disease/Trait of Interest Accurately and Specifically

Non-specific definitions will increase heterogeneity of underlying causal factors and decrease power of the study. Secondly, replication of the study (a crucial part of the validation of the results found) will become impossible if the phenotype has not been adequately defined.

Measuring heritability. Exploring heritability of complex traits is a central focus of statistical genetics. Low heritability values mean a larger number of individuals will need to be recruited to have sufficient power of detection. However, high heritability values do not guarantee that genetic variants will be easier to detect.

Select controls derived from the same ethnic (sub) population as cases. A general guide to control selection for any case-control study is that controls should be selected from the same population in which cases arose, and should be representative of the population who would have become cases according to the case definition and recruitment strategies for the study. This has long been the golden rule in epidemiological study design, the reason being that it minimizes spurious findings ('false positives') due to information and selection biases, and confounding. In genetic association studies, bias due to environmental factors is not generally a problem; the most important type of bias - confounding - is related to the ethnic origin of cases and controls, and is often referred to as Population stratification. In this situation, a comparison of the frequency of the genetic variant between cases and controls will show a significant difference due to the underlying sampling scheme rather than to a real effect of the variant on disease risk (Zeggini and Morris, 2010; Zondervan and Cardon, 2007).

REASONS FOR PRESENCE OF ASSOCIATION

If we find an association between a particular allele and the disease we are studying, it may be for one of four reasons: 1) It could be a false positive due to chance (Type I error); 2) It could be a false positive due to confounding because of population stratification; 3) It may be that the allele in "linkage disequilibrium" with the true trait allele, meaning that the allele we found more frequently in cases than in controls, is located physically close enough to the true disease allele that the two alleles tend to be inherited together and co-occur in affected individuals; 4) There really is a true causal association of the allele we studied and the disease (Wassertheil-Smoller and Smoller, 2015).

ASSOCIATION AND CAUSATION

An observed statistical association between a molecular marker and a disease does not necessarily lead us to infer a causal relationship. Conversely, the absence of an association does not necessarily imply the absence of a causal relationship. The judgment as to whether an observed statistical association represents a cause-effect relationship between exposure and disease requires inferences far beyond the data from a single study and involves consideration of criteria that include the magnitude of the association, the consistency of findings from other studies and biologic credibility (Hennekens and Buring, 1987).

The Bradford-Hill criteria are widely used in epidemiology as providing a framework against which to assess whether an observed association is likely to be causal (Rothman, 2002). The features defined by Bradford-Hill were:

1) Biological plausibility of association and its consistency with existing knowledge about biology and disease etiology are evaluated. Is the candidate gene likely to be involved in the phenotype? Are the SNPs likely to have functional effects on the protein?

2) The strength of the association between the risk factor and the disease is examined. When considering multiple SNPs in a candidate gene, the ones with strongest association are most likely to be causally related.

3) The dose-response relationship of the association is considered .For example, individuals with two copies of a variant might be at a greater risk of disease than individuals with one copy of the variant.

4) The consistency of the association across past and future studies, and across different populations, is an

important consideration. Consistent replication in different populations is strong evidence of causality. Lack of replication does not necessarily imply lack of causality, but might point to the need for more studies in certain populations or more detailed study of the function of a particular gene (Lucas *et al*., 2005; Tabor *et al*., 2002).

REFERENCES

[1] S.A. Angaji, Genomics at a Glance: Molecular Markers & Genome Mapping. Lambert Academic Publishing, 2011.

[2] S.A. Angaji, M. Ahmadzadeh, S.E. Yazdi Rouholamini, S. Mohammadi, Single Nucleotide Polymorphisms (Applications & Techniques). Asar-e-Nafis, 2017.

[3] J.T. Dudley, K.J. Karczewski, Exploring Personal Genomics. Oxford University Press, 2013.

[4] C.H. Hennekens, J.E. Buring, Epidemiology in Medicine, Lippincott Williams & Wilkins, 1987.

[5] T.G. Kimman, Genetics of Infectious Disease Susceptibility. Springer Science & Business Media, 2001.

[6] Z. Liu, Aquaculture Genome Technologies, John Wiley & Sons, 2007.

[7] S. Cox Gad, Handbook of Pharmaceutical Biotechnology. Wiley-Interscience, 2007.

[8] R.M. Lucas, J. Mcmichael Anthony, Bull World Health Organ 83 (2005) 792.

[9] J. Ott, Y. Kamatani, M. Lathrop, Nat. Rev. Genet. 12 (2011) 465.

[10] J. Ott, J. Wang, S.M. Lea, Nat. Rev. Genet. 16 (2015) 275.

[11] K.J. Rothman, Epidemiology: An Introduction, Oxford University Press, USA, 2002.

[12] H.K. Tabor, N.J. Risch, R.M. Myers, Nat. Rev. 3 (2002) 1.

[13] J.M. Walker, R. Rapley, Medical Biomethods Handbook. Humana Press Inc., 2005.

[14] S. Wassertheil-Smoller, J. Smoller, Biostatistics and Epidemiology: A Primer for Health and Biomedical Professionals. Springer, 2015.

[15] E. Zeggini, A. Morris, Analysis of Complex Disease Association Studies. Elsevier, 2010.

[16] M. Zhu, S. Zhao, Int. J. Boil. Sci. 3 (2007) 420.

[17] K.T. Zondervan, L.R. Cardon, Nat Protoc. 2 (2007) 2492.