Available online at www.sciencedirect.com

ScienceDirect

Chronic Diseases and Translational Medicine 2 (2016) 231–234

www.keaipublishing.com/en/journals/cdtm/

www.cdatm.org

Perspective

Data analysis in the post-genome-wide association study era

Qiao-Ling Wang [a], Wen-Le Tan [a], Yan-Jie Zhao, Ming-Ming Shao, Jia-Hui Chu, Xu-Dong Huang, Jun Li, Ying-Ying Luo, Lin-Na Peng, Qiong-Hua Cui, Ting Feng, Jie Yang, Ya-Ling Han*

Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

Received 31 July 2016

Available online 21 December 2016

Abstract

Since the first report of a genome-wide association study (GWAS) on human age-related macular degeneration, GWAS has successfully been used to discover genetic variants for a variety of complex human diseases and/or traits, and thousands of associated loci have been identified. However, the underlying mechanisms for these loci remain largely unknown. To make these GWAS findings more useful, it is necessary to perform in-depth data mining. The data analysis in the post-GWAS era will include the following aspects: fine-mapping of susceptibility regions to identify susceptibility genes for elucidating the biological mechanism of action; joint analysis of susceptibility genes in different diseases; integration of GWAS, transcriptome, and epigenetic data to analyze expression and methylation quantitative trait loci at the whole-genome level, and find single-nucleotide polymorphisms that influence gene expression and DNA methylation; genome-wide association analysis of disease-related DNA copy number variations. Applying these strategies and methods will serve to strengthen GWAS data to enhance the utility and significance of GWAS in improving understanding of the genetics of complex diseases or traits and translate these findings for clinical applications.

Keywords: Genome-wide association study; Data mining; Integrative data analysis; Polymorphism; Copy number variation

Introduction

During the last decade, genome-wide association study (GWAS) has been widely employed in case-control settings to identify the genetic variants [mostly single-nucleotide polymorphisms (SNPs)] associated with complex human diseases or traits. Since the first GWAS on human age-related macular degeneration was reported in 2005,[1] numerous GWASs on other diseases have been documented, including coronary heart disease,[2,3] diabetes,[4] and several forms

* Corresponding author.
E-mail address: hyl1288@aliyun.com (Y.-L. Han).

[a] The first two authors contributed equally to this study.

Peer review under responsibility of Chinese Medical Association.

Production and Hosting by Elsevier on behalf of KeAi

http://dx.doi.org/10.1016/j.cdtm.2016.11.009

of cancer such as esophageal cancer,[5,6] lung cancer,[3] and pancreatic cancer,[7] resulting in the establishment of a massive genotyping database. According to the US National Human Genome Research Institute, to date, there have been at least 1751 GWAS papers published, which have collectively identified 11,912 SNPs associated with various diseases.[6] Although many loci have been identified for many diseases, the underlying mechanisms of action of these loci in disease development and progression are largely unknown, which limits the clinical applications of GWAS results. In recent years, several strategies to make the GWAS findings more useful have been proposed.[7,8] In this minireview, we summarize and discuss the strategies available for the deep analysis of GWAS data to obtain further insight into the function and underlying mechanisms of associated loci and their biological actions.

Fine-mapping of susceptibility regions and their functional characterization

The major challenges in the post-GWAS era are to determine the function of identified susceptibility variants, characterize the biological action of the susceptibility genes, and clarify the regulatory mechanism if the variants are located within non-coding elements such as the gene promoter region, untranslated region, enhancer region, or regions that generate non-coding RNAs. Characterization of the biological mechanism underlying associations between genetic variants and diseases can provide a better understanding of disease pathogenesis, and therefore lead to better clinical care of patients.

Numerous studies have shown that genetic variants such as SNPs that are associated with diseases may not always be located in coding regions that produce proteins. In fact, the majority of disease-associated variants are located in non-coding regions, including the introns of genes. Although such variants in non-coding regions would not cause an amino acid change in the protein, they can nevertheless affect regulation of gene expression.[6] However, the regulatory functions of SNPs can be complex, involving effects on RNA splicing, transcription factor binding, DNA methylation, and microRNA (miRNA) recruitment.[9] For example, it has been found that SNPs in the telomere-related gene are associated with an increased risk of developing various types of human cancers and influence the length of the telomere, with direct associations between specific variants and telomere length. For example, the rs10069690 SNP in the telomere gene is significantly associated with the risks of breast, prostate, and invasive ovarian cancer linked to the *BRCA1* mutation. Functional analysis of Bojesen et al[10] showed that the risk genotype regulates alternative splicing, resulting in a truncated telomerase reverse transcriptase transcript that may affect telomerase activity. Thus, this study illustrated the functional relevance of a non-coding variant associated with multiple types of cancer, providing a potential strategy for targeted therapy. Another good example of such combination analysis of GWAS with functional characterization is a study conducted by Zheng et al,[11] who identified an SNP (rs11655237G>A) located within a "gene" producing long intergenic non-coding RNA (lincRNA), LINC00673, whose variant genotype is associated with pancreatic cancer risk. This lincRNA could reinforce the interaction of PTPN11 with PRPF19, an E3 ligase, in turn inducing PTPN11 degradation through ubiquitination, which causes diminished Src/extracelluar signal-regulated kinase (ERK) oncogenic signaling and enhanced activation of the signal transducer and activator of transcription 1 (STAT1)-dependent antitumor response. The G to A change at rs11655237 in the *LINC00673* exon creates a target site for miRNA-1231 binding, which diminishes the effect of LINC00673 in an allele-specific manner, and thus confers susceptibility to tumorigenesis. These findings shed new light on the important role of LINC00673 in maintaining cell homeostasis, and demonstrate that functional germline variation might confer susceptibility to pancreatic cancer.

Joint analysis of susceptibility variants associated with multiple diseases

Many disease-associated susceptibility loci or regions identified by GWAS are disease-specific; however, some susceptibility regions or loci shared by multiple diseases have also been found. For example, the locus at chromosomal region 8q24 was first identified as a susceptibility region for prostate cancer, but was subsequently associated with susceptibility to other types of cancer, including colorectal, breast, and bladder cancer.[12] Similarly, the locus at 6q27 has been reported to be associated with susceptibility to Crohn's disease[13] and rheumatoid arthritis,[14] as well as vitiligo[15,16] and other related diseases. This suggests that different types of diseases may share a common genetic susceptibility mechanism. Therefore, it is interesting and important to analyze the GWAS data obtained for multiple diseases jointly, which would improve the efficiency in finding common susceptibility loci for common diseases and reveal the underlying mechanisms for some diseases that may share the same genetics.