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.

Another example is a study in which 213 SNPs previously associated with 14 other tumors were selected for a meta-analysis to validate their potential associations with endometrial cancer. The results showed that among these SNPs associated with other cancers, 14 were additionally associated with the risk of endometrial cancer, including 5 significantly associated with both breast cancer and endometrial cancer.[17] These results indicate that the occurrence and development of some types of cancer may share common pathological pathways. Therefore, joint analysis of GWAS data for multiple cancer types may help to elucidate the common molecular mechanism of different types of malignancies and establish therapies and/or prevention measures that are transferable between cancer types.[18]

Combined analysis of GWAS results with transcriptional or epigenetic data

In recent years, high-throughput sequencing techniques have been highly developed, and the newly emerging next-generation sequencing technology has been widely applied to generate big data at the genome, transcriptome, and epigenome levels. The explosive growth of such data has now made it possible to perform confluent analysis. Complex human diseases such as malignancies fit the model of interactions among multiple pathological processes at the genome, transcriptome, and epigenome levels. However, previous studies have often focused on only one of these levels at a time, which limits understanding of disease pathogenesis to a single aspect. Systemic analysis of data at multiple levels simultaneously, including the genome, transcriptome, and epigenome, would provide a more comprehensive understanding of certain diseases, and therefore help to inform strategies for early diagnosis, drug development, individualized treatment and prevention of the disease.

Combined analysis of GWAS data with whole-genome mRNA expression data

An expression quantitative trait locus (eQTL) is a basic statistic used in the combined analysis of GWAS data and whole-genome transcriptional results, and can be applied to identify the gene number in trait-associated areas of the genome.[19] PrediXcan is a commonly used statistical method that can integrate whole-genome mRNA expression data and genome-wide genotyping data to explore whether the genetic variants associated with a disease or trait have an impact on gene expression.[8] The association between a genetic variant and the disease or trait may be strengthened through eQTL analysis, because of the functional relevance. Schadt et al[20] developed five underlying inter-modulating models by further exploring the mode of action among SNPs, genes, and phenotypes in the process of analyzing the data of genome-wide variants and whole-genome mRNA expression profiles, and identified three novel susceptibility genes for obesity using a causal model based on a likelihood method.

Combined analysis of GWAS data with whole-genome epigenetic results

Methylation of quantitative trait loci (meQTLs) refers to genetic variations in the genome that may affect the DNA methylation status. A GWAS can identify SNPs that are associated with diseases or traits at the DNA level, whereas an epigenome-wide association study (EWAS) can identify epigenetic alterations that are associated with diseases or traits. By integrating these two types of data, the modulation caused by epigenetic change of genetic variants leading to gene expression alterations can be identified. This type of analysis may also strengthen the associations between the genetic variants and certain diseases or traits. For example, a modulating network model was developed, which has been used to analyze DNase I atlas and GWAS SNP data of 349 human cell lines and tissue samples with information of their loci. The results of this study showed that 93% of disease- or trait-associated SNPs are located in non-coding regions, especially those highly sensitive to DNase I.[21] The complexity of human diseases and the application of individualized medicine require more precise methods for data integration. Mining and integrative analyses of data at different levels have emerged as promising trends in the development of biomedical technology.

Analysis of genome-wide copy number variation in human diseases

Previous GWASs of human diseases have mainly focused on identification of SNPs. However, it is believed that other types of genome variations such as copy number variations (CNVs) also play important roles in the susceptibility to complex human diseases. In recent years, genome-wide analysis of the associations between CNVs and complex diseases or traits has attracted increased interest in the field of medical genetics.[22] The analysis of associations between CNVs

and certain diseases can be achieved using genome-wide SNP-array data.[23]

CNVs usually represent duplications, deletions, insertions, and other multi-site complex variations of genome segments between 1 kb and 3 Mb, constituting structural variations in the genome. Approximately 12% of the human genome contains CNVs, making CNVs a more common and potentially more important genetic polymorphism than SNPs, which occur in about 0.5% of the human genome. To date, the Database of Genomic Variants (DGV) has recorded a total of 66,741 CNVs, including 15,963 CNVs that may play important roles in the determination of individual phenotypes, especially with respect to disease susceptibility. Numerous studies have demonstrated that CNVs are highly associated with neurological, immunological, genetic, neoplastic, and many other complex human diseases. Thus, using the strategy of GWAS to identify CNVs that are associated with certain diseases is very important.[6] It is expected that many novel statistical models will be established for analyzing the association between CNVs and diseases.

Conflicts of interest

The authors declare that they have no conflicts of interest.

References

1. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308:385–389.
2. Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet*. 2011;43:339–344.
3. Hu Z, Wu C, Shi Y, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet*. 2011;43:792–796.
4. Bradfield JP, Taal HR, Timpson NJ, et al. A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat Genet*. 2012;44:526–531.
5. Wu C, Hu Z, He Z, et al. Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nat Genet*. 2011;43:679–684.
6. Wu C, Li D, Jia W, et al. Genome-wide association study identifies common variants in SLC39A6 associated with length of survival in esophageal squamous-cell carcinoma. *Nat Genet*. 2013;45:632–638.
7. Wu C, Miao X, Huang L, et al. Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nat Genet*. 2011;44:62–66.
8. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47:1091–1098.
9. Huang Q. Genetic study of complex diseases in the post-GWAS era. *J Genet Genomics*. 2015;42:87–98.
10. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet*. 2013;45:371–384, 384e 1-2.
11. Zheng J, Huang X, Tan W, et al. Pancreatic cancer risk variant in LINC00673 creates a miR-1231 binding site and interferes with PTPN11 degradation. *Nat Genet*. 2016;48:747–757.
12. Stadler ZK, Thom P, Robson ME, et al. Genome-wide association studies of cancer. *J Clin Oncol*. 2010;28:4255–4267.
13. Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*. 2008;40:955–962.
14. Kochi Y, Okada Y, Suzuki A, et al. A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. *Nat Genet*. 2010;42:515–519.
15. Jin Y, Birlea SA, Fain PR, et al. Common variants in FOXP1 are associated with generalized vitiligo. *Nat Genet*. 2010;42:576–578.
16. Quan C, Ren YQ, Xiang LH, et al. Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the MHC. *Nat Genet*. 2010;42:614–618.
17. Setiawan VW, Schumacher F, Prescott J, et al. Cross-cancer pleiotropic analysis of endometrial cancer: PAGE and E2C2 consortia. *Carcinogenesis*. 2014;35:2068–2073.
18. Scarbrough PM, Weber RP, Iversen ES, et al. A cross-cancer genetic association analysis of the DNA repair and DNA damage signaling pathways for lung, ovary, prostate, breast, and colorectal cancer. *Cancer Epidemiol Biomarkers Prev*. 2016;25:193–200.
19. Li Q, Seo JH, Stranger B, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*. 2013;152:633–641.
20. Schadt EE, Lamb J, Yang X, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. 2005;37:710–717.
21. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337:1190–1195.
22. Zhou XY, Zhang XG. Copy number variation based genetic association studies. *Chin Sci Bull*. 2011;56:370–382.
23. Zhang X, Du R, Li S, Zhang F, Jin L, Wang H. Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics*. 2014;15:50.

Edited by Pei-Fang Wei