Available online at www.jbr-pub.org.cn

Open Access at PubMed Central

The Journal of Biomedical Research, 2020 34(5): 323–327

JBR

Review Article

From human genome epidemiology to systems epidemiology: current progress and future perspective

Hongxia Ma[1,2], Hongbing Shen[1,2,✉]

[1]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu 211166, China;
[2]Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, Jiangsu 211166, China.

Abstract

The recent progress in human genome epidemiology (HuGE) is already having a profound impact on the practice of medicine and public health. First, the success of genome-wide association studies has greatly expanded the direction and content of epidemiological researches, including revealing new genetic mechanisms of complex diseases, identifying new targets for therapeutic interventions, and improving application in early screening of high-risk populations. At the same time, large-scale genomic studies make it possible to efficiently explore the gene-environment interactions, which will help better understand the biological pathways of complex diseases and identify individuals who may be more susceptible to diseases. Additionally, the emergence of systems epidemiology aims to integrate multi-omics together with epidemiological data to create a systems network that can comprehensively characterize the diverse range of factors contributing to disease development. These progress will help to apply HuGE findings into practice to improve the health of individuals and populations.

Keywords: human genome epidemiology, genome-wide association study, gene-environment interaction, systems epidemiology

Introduction

"Human genome epidemiology (HuGE)" was first proposed in 1998 by Khoury and Doman, and defined as an evolving field that uses epidemiologic methods and approaches in population-based studies to assess the impact of human genetic variations on health and diseases[1]. HuGE has been viewed as the intersection between molecular epidemiology and genetic epidemiology, which aims to translate human genetic research findings into meaningful actions to improve health and prevent disease[2]. In the past two decades, a large and rapidly increasing number of studies on HuGE have been carried out, along with the great progress in genomics technologies. These studies not only greatly promote people's understanding of the influence of genetic variations on disease occurrence, but also provide important theoretical basis for personalized healthcare and disease prevention. However, as a new and developing research field, it

✉Corresponding author: Hongbing Shen, Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Avenue, Nanjing, Jiangsu 211166, China. Tel/Fax: +86-25-86868439/+86-25-86508960, E-mail: hbshen@njmu.edu.cn.

Received 27 February 2020, Revised 21 April 2020, Accepted 29 April 2020, Epub 29 June 2020

CLC number: R181.3+3, Document code: A

The authors reported no conflict of interests.

https://doi.org/10.7555/JBR.34.20200027

also faces some difficulties and challenges. Here, we attempt to briefly describe the current progress, opportunities and challenges of HuGE, which will help medical and public health professionals integrate findings of genomics into practice to improve the health of individuals and populations.

Genome-wide association studies

In the recent 10 to 15 years, perhaps the most important development of HuGE has been the emergence of genome-wide association studies (GWAS)[3], which tests hundreds of thousands to millions of genetic variants across the genomes to identify genotype-phenotype associations[4]. Since the first GWAS for age-related macular degeneration was published in 2005[5], more than 50 000 genetic loci related to complex diseases or traits have been identified[6]. These findings have advanced our current knowledge in genetic architecture of complex trait or disease (identification of novel susceptibility genes and biological mechanisms) and promoted the practice in clinical care (discovery of disease progress biomarkers and new targets for therapeutic interventions) and personalized medicine (risk prediction and optimization of therapies)[4]. For example, recent studies have shown that polygenic risk score (PRS) can quantify the cumulative effect of genetic variants discovered by GWAS to identify high-risk individuals, thereby improving disease screening or clinical outcomes through early detection, prevention or treatment[7]. In 2019, our research team first built a PRS based on 19 lung cancer susceptibility loci in Chinese populations and evaluated the utility and effectiveness of the generated PRS in predicting subpopulations at high risk of lung cancer in a large-scale prospective cohort[8]. The results showed that GWAS-derived PRS can be effectively used in discriminating subpopulations at high risk of lung cancer, who might benefit from lung cancer screening for precision prevention in Chinese populations[8]. Similar findings have been reported in cardiovascular diseases, diabetes, inflammatory bowel disease, other types of cancers, *etc.*[9]. In addition, the discovery of genetic variations based on GWAS has been used to identify several novel candidate drugs that are now being applied in clinics or evaluated in clinical trials[5]. A recent study also reported that patient-specific human leukocyte antigen class I genotype could influence the efficacy of immune checkpoint inhibitors targeting cytotoxic T-lymphocyte-associated protein 4 and programmed cell death protein 1/programmed cell death 1 ligand 1 in cancer patients, suggesting an important role of genetic variations in immunotherapy[10]. Thus, the Clinical Pharmacogenetics Implementation Consortium has developed a rigorous approach to evaluate the clinical value and interpretation of genetic variants associated with drug response, providing clinical decision supports for physicians. However, despite the great success in identifying disease loci, GWAS still has some limitations, such as limited sample size, "missing" heritability and interpretation of GWAS associations[4]. Substantial efforts have been made to overcome these difficulties. First, larger sample size has been the most effective way of increasing the power of GWAS to detect the small or moderately sized effects. Studies have indicated that many common variants with relatively weak effects on human disease were missed by GWAS due to a lack of statistical power[11]. Now the global multi-center data integration and collaboration for GWAS become the trend and the research subjects can reach more than 100 000 people to identify all causal genetic variants and measure how much trait variation they explain. Secondly, rare variants may make a major contribution to missing heritability, which are not captured by common SNPs on current genotyping arrays[12]. Thus, the research scope has expanded to low-frequency or rare genetic variations by using exome or whole-genome sequencing studies[13]. These sequencing findings would also improve the application value of PRS. However, very large samples will be needed for rare variants in such studies unless the effect size of the variant is particularly large. Additionally, it is a great challenge to identify causal genes or SNPs although GWAS have identified thousands of variants associated with common diseases and complex traits. Remarkably, with the promotion of large-scale resource-based projects such as ENCODE, TCGA, and GTEx, the biological significance of genetic loci in the development of human diseases is gradually uncovered[14].

Gene-environment interactions

It has been widely accepted that both genetic variations and environmental exposures affect disease risk, and individuals with different genotypes may respond differently to environmental exposures and generate an array of phenotypic landscape[15]. Such gene-environment (G×E) interactions may be responsible for a large fraction of the unexplained variances in heritability and disease risk[14]. In the meantime, the study of G×E interaction is important

for better understanding the biological pathways, estimating population-attributable risk(s), and identifying individuals who may be more susceptible to diseases[16]. In the past, G×E interactions have been investigated for a wide range of candidate genes and exposures for many complex traits. Limited by factors such as small sample size, representativeness of genes or loci and improper correction of multiple comparisons, these studies only provide little evidence[17]. At present, progress of the large-scale genomic studies makes it possible to explore the G×E interactions through more dense panels of genetic variants and larger sample sizes, and even detect interactions with small effect sizes, rare frequencies, and higher order interactions[18]. In 2012, the National Institutes of Health had launched the Genetic Associations and Mechanisms in Oncology (GAME-ON) Initiative for five common malignant tumors, including breast, prostate, ovarian, lung, and colorectal cancers. It aimed to "rapidly move forward promising leads from initial cancer GWAS by ….unraveling the function of genetic variants and how environmental factors may influence the genetic effect …"[19]. Since then, studies focusing on G×E interactions have begun to yield interesting findings. For example, our research team first identified several novel loci that were significantly associated with lung cancer risk in the Chinese Han population, including 13q12.12, 22q12.2, 1p36.32, and 5q31.1[20–21]. Among those, a well-replicated GWAS risk SNP rs753955 within 13q12.12 is situated in the gene desert region, about 150 kb away from the nearest upstream gene, *TNFRSF19*. When compared to non-smokers with wild genotype of this variant, the lung cancer risk increased to 3.8 times for those smokers with variant genotypes, suggesting a potential gene-smoking interaction[20]. Functional assays indicated that this susceptible locus in 13q12.12 could decrease the expression of *TNFRSF19* and promote the malignant transformation of lung epithelial cells caused by NNK[22]. These results revealed the potential biological mechanism underlying the interaction between susceptible genes and tobacco carcinogens. It should be noted that bias can be induced in case-control studies of genotype effects if the underlying population is genetically stratified or admixed. Investigators have extended family-based studies to G×E interaction studies, which can provide strategies for testing genetic effects that are robust to undetected/unaccounted for population substructure[23]. The Framingham Heart Study is a well-known example of a family-based study. However, challenges still exist in G×E interaction studies: the complexity of measuring environment exposures, limited range of genetic and/or environmental variation, limitation of different statistical methods for interaction analysis, and lack of data on the biological significance of most genetic variants[18]. More importantly, the risk of complex disease is a consequence of multiple genes in multiple biologic pathways interacting with each other and with cumulative environmental factors over a lifetime. It will require new paradigms for interdisciplinary collaborative research with very large-scale studies, as well as new analysis tools to help scientists reveal the complex multi-gene-environment interactions involved in human diseases.

Systems epidemiology

Despite some successes at identifying genetic and environmental risk factors for complex diseases, they still represent only the tip of the iceberg and much of the etiology remains unexplained. This may be due in part to the limitation of many studies on a single or small set of risk factors or data types. With the availability of high throughput -omics technologies, researchers gradually realize that a more comprehensive and systematic analysis with multiple dimensions, integration of genomics, transcriptomic, metabolomic, and other omics data, is needed to better understand their contributions to diseases at multiple levels as well as their interactions. Therefore, "systems epidemiology" emerged as a new research discipline that integrates multi-omics together with epidemiological data to create a systems network that can be used to better characterize the diverse range of factors influencing disease development[24–25]. Remarkably, systems epidemiology involves not only the measurement of biomarkers, such as genomic, transcriptomic, proteomic, and metabolomic profiles, but also a variety of environmental interaction components including smoking, behavior, socio-demographic factors, and group levels that may affect health and disease[26]. Some researchers have proposed a globolomic study design for systems epidemiology, which will collect biological samples at the beginning of the follow-up and the time at diagnosis based on large-scale prospective cohort studies to detect DNA, RNA and other biomarkers, and then combine with the outcomes of the cohort study to evaluate complex interactions between multiple exposures and their dynamics encompassing human diseases[24]. However, the most compelling challenge will be to integrate multi-level data. Recently, some statistical methods have been developed to integrate different types of

data, such as TCGA analysis platform which has comprehensively used the multi-dimensional genomic data (including variations, copy numbers, epigenetic data, gene expression, and miRNA sequencing data) of more than 30 kinds of tumors to mine gene networks related to cancers[27–28]. Although such analysis does not provide a complete understanding of the human system, the findings make us more deeply aware of the pathogenesis of disease.

As described above, new knowledge and technology have given genome epidemiology the possibility of a new research discipline, which could form an important scientific foundation for using genetic information to improve health and prevent disease. However, the development in genome epidemiology still requires continued technological advances in high-throughput methods, enhanced bioinformatics and analytical tools, coordinated efforts that span multiple disciplines of laboratory sciences, medicine and public health. In addition, well-designed prospective cohort studies with large sample size, long-term follow-up, availability of archived biological samples, and detailed measures of exposures, are also necessary for the successful application of genome epidemiology into medicine and public health. Finally, it is difficult to appropriately evaluate the utility of genetic information based solely on measures such as relative risks of genetic loci and PRS. Other factors such as environmental and lifestyle factors, also contribute to the risk prediction models. Moreover, clinicians and scientists need to engage with the public and get across the fact that, for many complex diseases, metrics like PRS are probabilistic at the population level. Although a person's genetic makeup cannot be altered, some lifestyle and environmental modifications may reduce disease risk in people with a genetic predisposition.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 81521004 and No. 81922061).

References

[1] Khoury MJ, Dorman JS. The human genome epidemiology network[J]. *Am J Epidemiol,* 1998, 148(1): 1–3.

[2] Khoury MJ. Human genome epidemiology: translating advances in human genetics into population-based data for medicine and public health[J]. *Genet Med,* 1999, 1(3): 71–73.

[3] Shen HB, Jin GF. Human genome epidemiology, progress and future[J]. *J Biomed Res,* 2013, 27(3): 167–169.

[4] Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies[J]. *Nat Rev Genet,* 2019, 20(8): 467–484.

[5] Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS discovery: biology, function, and translation[J]. *Am J Hum Genet,* 2017, 101(1): 5–22.

[6] Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019[J]. *Nucleic Acids Res,* 2019, 47(D1): D1005–D1012.

[7] Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores[J]. *Nat Rev Genet,* 2018, 19(9): 581–590.

[8] Dai JC, Lv J, Zhu M, et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations[J]. *Lancet Respir Med,* 2019, 7(10): 881–891.

[9] Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations[J]. *Nat Genet,* 2018, 50(9): 1219–1224.

[10] Chowell D, Morris LGT, Grigg CM, et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy[J]. *Science,* 2018, 359(6375): 582–587.

[11] Young AI. Solving the missing heritability problem[J]. *PLoS Genet,* 2019, 15(6): e1008222.

[12] Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies[J]. *Proc Natl Acad Sci USA,* 2014, 111(4): E455–E464.

[13] Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies[J]. *Mol Cell,* 2015, 58(4): 586–597.

[14] Wang QG, Armenia J, Zhang C, et al. Unifying cancer and normal RNA sequencing data from different sources[J]. *Sci Data,* 2018, 5(1): 180061.

[15] Favé MJ, Lamaze FC, Soave D, et al. Gene-by-environment interactions in urban populations modulate risk phenotypes[J]. *Nat Commun,* 2018, 9(1): 827.

[16] Idaghdour Y, Awadalla P. Exploiting gene expression variation to capture gene-environment interactions for disease[J]. *Front Genet,* 2013, 3: 228.

[17] Aschard H, Lutz S, Maus B, et al. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies[J]. *Hum Genet,* 2012, 131(10): 1591–1613.

[18] McAllister K, Mechanic LE, Amos C, et al. Current challenges and new opportunities for gene-environment interaction studies of complex diseases[J]. *Am J Epidemiol,* 2017, 186(7): 753–761.

[19] Hutter CM, Mechanic LE, Chatterjee N, et al. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report[J]. *Genet Epidemiol,* 2013, 37(7): 643–657.

[20] Dong J, Hu ZB, Wu C, et al. Association analyses identify multiple new lung cancer susceptibility loci and their

interactions with smoking in the Chinese population[J]. *Nat Genet,* 2012, 44(8): 895–899.

[21] Hu ZB, Wu C, Shi YY, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese[J]. *Nat Genet,* 2011, 43(8): 792–796.

[22] Shao LP, Zuo XL, Yang Y, et al. The inherited variations of a p53-responsive enhancer in 13q12.12 confer lung cancer risk by attenuating TNFRSF19 expression[J]. *Genome Biol,* 2019, 20(1): 103.

[23] Shi M, Umbach DM, Weinberg CR. Family-based gene-by-environment interaction studies: revelations and remedies[J]. *Epidemiology,* 2011, 22(3): 400–407.

[24] Lund E, Dumeaux V. Systems epidemiology in cancer[J]. *Cancer Epidemiol Biomarkers Prev,* 2008, 17(11): 2954–2957.

[25] Jacobs L, Thijs L, Jin Y, et al. Heart 'omics' in AGEing (HOMAGE): design, research objectives and characteristics of the common database[J]. *J Biomed Res,* 2014, 28(5): 349–359.

[26] Haring R, Wallaschofski H. Diving through the "-omics": the case for deep phenotyping and systems epidemiology[J]. *OMICS,* 2012, 16(5): 231–234.

[27] Thorsson V, Gibbs DL, Brown SD, et al. The immune landscape of cancer[J]. *Immunity,* 2018, 48(4): 812–830.

[28] Berger AC, Korkut A, Kanchi RS, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers[J]. *Cancer Cell,* 2018, 33(4): 690–705.