

Development and application of Human Genome Epidemiology

Jingwen Xu^{1, a}

¹School of Clinical Medicine, Fudan University, Shanghai, 201803, China

^a13301050087@fudan.edu.cn

Abstract. Epidemiology is a science that studies distribution of diseases and health in population and its influencing factors, it also studies how to prevent and cure disease and promote health strategies and measures. Epidemiology has developed rapidly in recent years and it is an intercross subject with various other disciplines to form a series of branch disciplines such as Genetic epidemiology, molecular epidemiology, drug epidemiology and tumor epidemiology. With the implementation and completion of Human Genome Project (HGP), Human Genome Epidemiology (HuGE) has emerged at this historic moment. In this review, the development of Human Genome Epidemiology, research content, the construction and structure of relevant network, research standards, as well as the existing results and problems are briefly outlined.

1. Introduction

The progress of the Human Genome Project has led us into the era of genetic information. The completion of the human genome sequencing work is an important milestone in the history of medical and life sciences. Many scientists believe that this project can provide basic information for disease prediction and prevention and will have a profound impact on the 21st century medical and public health [1-3]. Almost every day, new reports about the findings on new genetic variants that may be relevant to disease pathogenesis emerge. These findings share various disease spectrum including major public health problems from adult chronic diseases like cancer and diabetes, to infectious diseases, autoimmune diseases and neonatal diseases. In order to apply the results of these genetic studies to disease prevention and public health promotion, the application of population-based epidemiological studies to quantify the effects of genetic variants on disease, death, and disability risks, identifying and quantifying the interaction of other risk factors with genetic variants is urgently needed. The results of such epidemiological studies help medical and public health practitioners to better conduct clinical care, behavioral care and psychosocial interventions.

2. Emergence and development of HuGE

Epidemiology has been defined in many ways and is often seen as the core of public health science. "Studying the distribution and determinants of health-related status and events in the population, and applying the relevant research to control health-related issues" is a widely used definition [4]. Epidemiologists not only investigate the outbreaks of disease in different populations, but also carry out research on the pathogenic factors of various diseases, identify high-risk groups in order to assess population health, and promote the cost-effectiveness of population health interventions.

The Human Genome Project is a joint initiative by many countries to sequencing all the human chromosomes and genes, completely unravelling the huge human genetic information [2]. So far, the



plan has been completed and has fruitfully obtained following results: (1) Has successfully drawn out the human chromosome map framework; (2) A large number of genes have been used for medical genetic testing; (3) Continuously emerging genetic publication of risk factors for adult chronic diseases such as cancer, diabetes and infectious, autoimmune diseases. But how are the research results of the Human Genome Project applied to the diagnosis and treatment and health promotion of human diseases as soon as possible? What are the combined effects of genetic methods that have been used in clinical testing and treatment? How large is the role of disease risk genes and their variants in human disease and health? These problems are urgently needed to be answered by population-based human genome epidemiological studies.

As Shpilberg et al. simply put it: "The sequencing of the human genome offers the greatest opportunity for epidemiology since John Snow discovered the Broad Street pump" [5]. Human genome epidemiology emphasizes the systematic application of epidemiological methods in human genome research, using quantitative indicators and statistical methods to describe the distribution of genes and their associated markers in the population and their association with disease. In 1998, Khoury and Dorman first introduced the term of HuGE [6] and defined it as: Human Genome Epidemiology is a discipline using the research methods combining epidemiology and genomic information to carry out population based-research, evaluate genomic information (gene or gene variants and its corresponding encoded product), which has great importance for population health. HuGE is the forefront of genetic epidemiology and molecular epidemiology.

Similar to the human genome epidemiology, in recent years we have seen the emergence of genetic epidemiology and molecular epidemiology, some argue that these concepts are synonyms, while others think they are different in evolution path, goals and priorities [7]. However, the importance of applying the discovery of human genome projects to population-based research is highly recognized.

Molecular epidemiology is applying the theory and technique of molecular biology, focusing on analysing the characteristics of pathogens that cause a particular disease at the genetic level, thus more accurately dealing with the source of infection, the route of transmission and the related epidemiological problems. On the other hand, genetic epidemiology is the combination of genetics and epidemiological methods, it focuses on research of genetic diseases, especially chronic diseases, genetic factors, environmental factors and their interaction. Human genome epidemiology, molecular epidemiology and genetic epidemiology are new epidemiological branches formed by epidemiology along with other disciplines. They are not absolutely separated, instead they are blended with each other, facing a common urgent task which is to focus on carrying out population-based research to apply the results of human genome project to medical and health care practice.

3. Topics of HuGE

Human Genome Epidemiology is to systematically apply epidemiology research methods in population-based research to explore the effects of human genetic variants on health and disease [8]. We see it as a integration between molecular epidemiology and genetic epidemiology. The topics include the following aspects, which cover the population-based epidemiological genetic variants research to the genetic testing evaluation, the structure diagram is shown in Figure 1.

- (1) Assess the prevalence of genetic variants in different populations.
- (2) Assess the association between genetic variants and disease risk in different populations (relative risk and absolute risk).
- (3) Assess the contribution of genetic variants to disease occurrence in different populations.
- (4) Assess the association between disease risk and gene-gene and gene-environment interactions in different populations.
- (5) Assess the validity of genetic testing in different populations. (positive predictive value and negative predictive value)
- (6) Evaluate the magnitude and determinants of genetic tests in different populations.

(7) Evaluate the effect of genetic tests on morbidity, disability, mortality and services and cost in different populations.

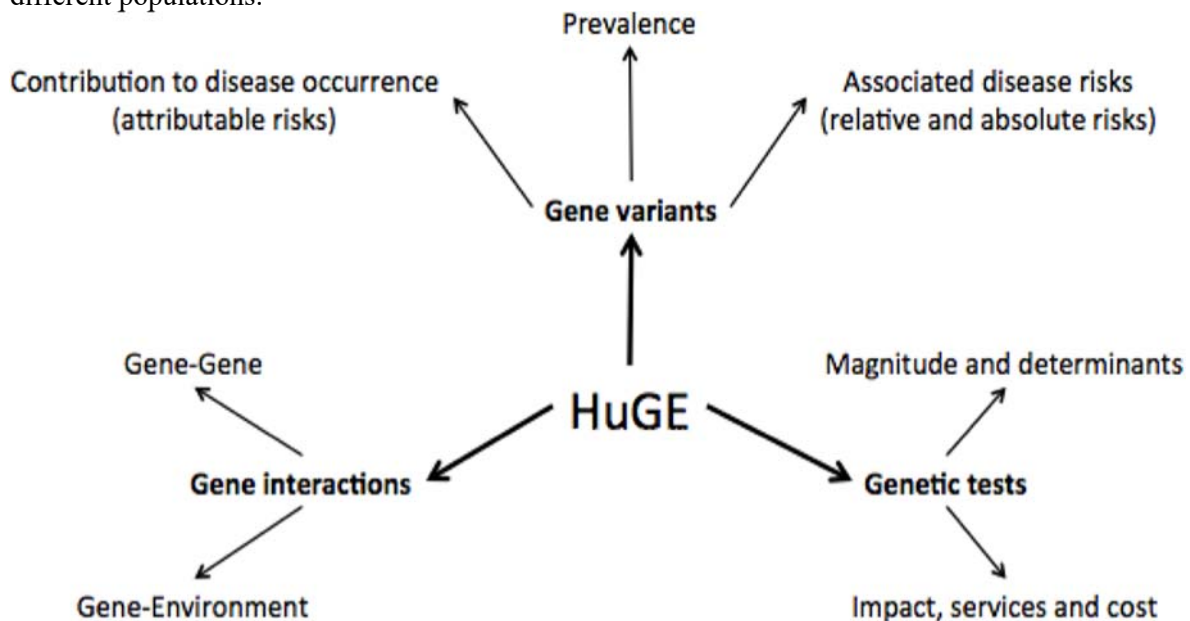


Figure1. Structure diagram of topics of HuGE

4. Research methods of HuGE

The research of human genome epidemiology requires to be conducted from both macro and micro aspects, including epidemiological studies and genetic information research methods.

4.1 Epidemiological studies

All traditional epidemiological research methods including descriptive studies, analytical studies and experimental studies can be applied to human genome epidemiology.

4.1.1 Prevalence survey. Prevalence survey in HuGE is to deliver a cross-sectional study of disease or health status and associated genomic information in a given population for a specific period to analyze the association of disease or health with genomic information.

4.1.2 Case-control study. Case-control study for HuGE is mainly used to explore the association of genomic information with disease or health status. Nest case-control studies are particularly suitable for HuGE.

4.1.3 Cohort study. Cohort study can meet the requirement of being population-based, which has its advantage in HuGE research. However, cohort studies take long period of time and exhaust much more resources, thus there is limit use. Many studies have chosen combined cohort studies to compensate for the disadvantages of this approach.

4.2 Genetic information research methods

Genetic information research methods for HuGE include gene linkage, segregation analysis, the choose of genomic marker and technique to acquire gene information.

4.2.1 Linkage analysis. Linkage analysis is to detect whether a disease or certain phenotypic trait is transmitted from a generation to a next generation with a chromosome fragment in a family. This chromosome is determined by one or more genetic markers. The linkage analysis utilizes the current arrangement of human chromosomes. In the process of generation of gametes in meiosis, each pair of

homologous chromosomes pairing exchange DNA fragments, this exchange is also called recombination, which is the basis of linkage analysis. As for the linkage analysis, the first thing to do is to find candidate gene, which might be related to a disease according to pathophysiology knowledge and existing studies. Then use the candidate gene and its surrounding markers to do the linkage analysis.

4.2.2 Linkage disequilibrium analysis. The linkage disequilibrium analysis is performed to detect whether the frequency of certain gene in the diseased population is significantly higher than in the non-diseased population. Combining linkage disequilibrium analysis with linkage analysis can improve the efficiency of gene mapping.

4.2.3 Segregation analysis. Segregation analysis gives evidence about the relationship between a gene and possible relevant phenotype. Complex segregation analysis uses various kind of models, calculating the frequency at which the hypothesis test coincides with the observed value, in order to find out the most probable genetic pattern of a disease.

4.2.4 Genomic markers. The genomic information needed for HuGE is mainly from the detection of polymorphism markers. The earliest genetic polymorphism markers are called Restriction Fragment Length Polymorphism (RFLP). But RFLP is in small number in the genome distribution, the included polymorphism information is relatively low. The following developed marker is called STR, also known as microsatellites, whose information of chromosomal distribution is significantly better than RFLP, thus it became an extremely useful marker in genetic analysis. The third generation of polymorphic marker is SNP, which can detect a single base change. Currently SNP has become a new polymorphic marker for studying the genome diversity and identifying and locating disease-related gene. The use of SNP can provide HuGE study with more accurate information.

4.2.5 DNA and RNA analysis. The analysis of DNA and RNA includes extraction, isolation, amplification, cloning, identification, mutation analysis and sequencing analysis. The most widely used technique is PCR and mutation analysis. Using these technologies can not only analyze the target gene, but can also analyze the genome polymorphism and gene sequencing. In addition, recently developed DNA chips and DNA microarray technology has made multi-site screening in large samples simple and fast.

5. Network of HuGE

5.1 HuGE Net

Due to the nature of HuGE, the relevant research is not only limited to a single subject, but also requires joint cooperation to carry out large-scale scientific research to analyze and summarize the information from multiple population, thus gather information which cannot be obtained by single small-scale research [8,9].

The development of science and technology and the establishment of the Internet provides medical workers especially epidemiologists with access to the latest human genome project research progress, enabling the rapid translating of human genome project result into clinical application for human health improvement. At present, the CDC in collaboration with the American Journal of Epidemiology has launched a human genome epidemiology network which is undoubtedly playing an important role in promoting the development of the subject. HuGE Net represents the joint efforts of individuals and organizations dedicated to the development and dissemination of population-based human genome epidemiology.

The website is created to facilitate collaborative communication and information exchange. Its development and implementation is carried out under the guidance of the committee and the multidisciplinary working group worldwide. The website includes updated information on the

accumulation of human genetic epidemics. 1) Review of peer review information on human genetic epidemiology; 2) Original research articles on related topics; 3) Systematic review of technical assessment, large data analysis, seminar summary, academic conference; 4) Information about a specific gene on a timely update of the medical literature search; 5) Website links related to population-based specific genetic information; 6) Comments, editorial and summary of the views; 7) Announcement of HuGE related conferences, seminars, training opportunities; 8) Related sponsorship opportunities; 9) Communicating forum.

Since the establishment of HuGE Net, many medical and public health experts have expressed concern about the growing gap between gene discovery and disease prevention measures that genetic variants may apply to. The translation of genetic findings of disease prevention requires a lot of scientific information and requires high quality population-based research on gene-disease relationships and disease intervention assessments, which are exactly what HuGE is representing. The HuGE Net is designed to provide an open forum for sharing ideas on topic of statistical methods, experimental methods, policy and progress. Such a forum will create a core registry that contains basic information about the participating teams, research features and target groups. This broader knowledge base enhances the efficiency of research planning and allows the results to be tested by being replicated more quickly.

5.2 HuGE Navigator

Since the establishment of HuGE Net in 2001, it has always been able to obtain population-based human gene epidemiology information from PubMed and has already built a database. Some scholars use machine learning to reduce the time and effort required for manual retrieval and to increase the sensitivity and specificity of screening related literature. HuGE Navigator is a series of integrated applications, using PubMed as the core data source to establish knowledge base for the discovery of gene association, candidate genes and researcher network [10]. Genetic information can be displayed on the main gene databases (e.g. Entrez Gene, Swiss-Prot, OMIM and GeneCards), genetic variants and prevalence (e.g. dbSNP and HapMap Project), pathways (e.g. CGAP, KEGG and BioCarta) and others (e.g. Gene Ontology and Gene Clinics). HuGE Navigator builds on the principle of open source, standardization, interoperability, and extensibility, allowing new applications to be easily included. Currently, HuGE Navigator allows users to navigate through six applications (Table 1.) to navigate and retrieve the data. HuGE Navigator aims to be “the network of networks”, helping researchers understand the complex relationship between human disease related gene and environmental factors.

Table 1. HuGE Navigator database description

Database	Description
HuGE Literature Finder	HuGE Literature Finder is a search engine that can be used to discover human genome epidemiology published literature, including gene association studies. The query word may include disease terminology, environmental factors, gene or author name, institution. The results can be further refined by reusing the filter until the desired results are obtained. The result of the PubMed ID can be exported to the PubMed website for the next step of learning and downloading.
HuGE Investigator Browser	HuGE Investigator Browser is an investigator search engine for finding researchers or collaborators based on research interests (e.g. disease, risk factors, genes). The engine extracts the investigator's data by parsing the associated data provided by PubMed.
Gene Select Assist	Gene Select Assist is a search tool for finding candidate genes that may be relevant to a particular topic. Search terms may include disease or exposure. Based on the genetic association studies in the HuGE Navigator database, other PubMed abstracts, and animal model evidence in the NCBI Entrez Gene database to select and sort candidate genes.
HuGE Watch	HuGE Watch is a tool based on a database that dynamically tracks the progress of human genome epidemiology. It allows users to view the current published trends, according to genes, diseases, investigators, and the geographical distribution of the authors.
HuGEpedia	HuGEpedia is an online encyclopedia that summarizes research on gene-disease associations. It extracts data from meta-analyses and published genome-wide associations that will form the basis for a disease-specific synopsis written by domain experts. HuGEpedia can be searched by gene or disease.

HuGE Risk Translator	HuGE Risk Translator is a tool for assessing the effectiveness of genetic variation in predicting health by calculating population attributable risk, sensitivity, specificity, positive predictive value, negative predictive value.
----------------------	---

6. The standardization of HuGE research

Currently there has been increasing number of HuGE studies, thus the standardization is necessary for the research methods and content. In 2002, in order to promote the association of disease and gene polymorphism, Cooper et al. has put forward some methodological criteria, mainly focusing on the discovery of candidate genes [11]. In 2001, an expertise group from CDC and NIH suggested some advices to evaluate the data obtained from HuGE research, including: 1. Research on the association of gene polymorphism and gene-disease correlation; 2. Research on gene-environment and gene-gene interactions; 3. Research on genetic testing [12,13]. The criteria for the study of HuGE are also slowly being constructed, and we should carefully evaluate the feasibility of these criteria.

7. Existing achievement and problems of HuGE

There are thousands of research published every year on the topic of genetic variants and phenotypical relationship. Transparent research methods and results allows readers to better evaluate the credibility of HuGE research achievement. HuGE has had a considerable number of case studies. Khoury et al. has listed many detailed examples in their review, including the use of pesticides and oral contraceptives, chronic diseases (colon cancer, Alzheimer's disease, cardiovascular disease, breast cancer, iron overload), occupational disease (beryllium poisoning), birth defects (fragile X chromosome syndrome, congenital deafness) and infectious disease (HIV), etc. [14] However, in the early stages of development, most of the studies were limited to smaller populations and only tested one or a few more genetic variants within one gene. The methods used in articles for evaluating selective bias, reducing false classification and population stratification are various. In recent years, researchers have gradually formed a unified standard for evaluating statistical power, using copy samples and all other details. Guideline for research reports and online supplemental materials can be used to enhance the transparency and clearance of the literature.

8. Conclusions

HuGE plays an important role in the study of causes of disease and disease distribution in the population. However, HuGE research requires advanced molecular biology techniques and large scale population study. It requires people to study the interaction between gene-gene and gene-environment. The establishment of network like HuGE Net and HuGE Navigator offers valuable help to the development of HuGE. The research still faces many problems and challenges which requires multiple countries and regions to work together. In the process of translating the great achievements of the HGP into medical practice, it needs the joint efforts of many disciplines in the field of medicine and public health, among which epidemiology is the one that plays a crucial role. We firmly believe that in the future HuGE will develop into an influential discipline.

Reference

- [1] JC Venter, MD Adams, EW Myers, et al. The sequence of the human genome [J]. Clinical Chemistry, 2001, 61 (9): 1207-8.
- [2] GS Consortium. Initial sequencing and analysis of the human genome [J]. Nature, 2001, 409 (6822): 860-921.
- [3] FS Collins, AE Guttmacher. Genetics moves into the medical mainstream [J]. Jama the Journal of the American Medical Association, 2001, 286 (18): 2322-4.
- [4] JM Last. A Dictionary of Epidemiology, third edition [M]. New York: Oxford University Press, 1995.
- [5] O Shpilberg, JS Dorman, RE Ferrell, et al. The next stage: molecular epidemiology [J]. Clinical Epidemiology, 1997, 50: 633-638.

- [6] MJ Khoury, JS Dorman. The human genome epidemiology network [J]. American Journal of Epidemiology, 1998, 148 (1): 1-3.
- [7] AE Guttmacher, FS Collins. Genomic medicine--a primer [J]. New England Journal of Medicine, 2009, 347 (19): 1512-20.
- [8] JP Ioannidis, M Gwinn, J Little, et al. A road map for efficient and reliable human genome epidemiology [J]. Nature Genetics, 2006, 38 (1): 3-5.
- [9] AJ Frodsham, JP Higgins. Online genetic databases informing human genome epidemiology [J]. BMC Medical Research Methodology, 2007, 7 (1): 1-6.
- [10] W Yu, M Gwinn, M Clyne, et al. A navigator for human genome epidemiology [J]. Nature Genetics, 2008, 40 (2): 124.
- [11] BS Jr, J Concato, AR Feinstein. Clinical epidemiological quality in molecular genetic research: the need for methodological standards [J]. Jama, 1999, 281 (20): 1919.
- [12] J Little, L Bradley, MS Bray, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations [J]. American Journal of Epidemiology, 2002, 156 (4): 300-310.
- [13] W Burke, D Atkins, M Gwinn, et al. Genetic test evaluation: information needs of clinicians, policy makers, and the public [J]. American Journal of Epidemiology, 2002, 156 (4): 311.
- [14] MJ Khoury, J Little, W Burke. Human genome epidemiology : a scientific foundation for using genetic information to improve health and prevent disease [M]. Oxford University Press, 2004.