예방의학회지 제42권 제6호(2009년 11월)
J Prev Med Public Health 2009;42(6):356-359
DOI: 10.3961/jpmph.2009.42.6.356

Special Issues

Discovering Gene-Environment Interactions in the Post-Genomic Era

Nirinjini Naidoo, Kee Seng Chia

Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, National University of Singapore, Singapore

In the more than 100 genome wide association studies (GWAS) conducted in the past 5 years, more than 250 genetic loci contributing to more than 40 common diseases and traits have been identified. Whilst many genes have been linked to a trait, both their individual and combined effects are small and unable to explain earlier estimates of heritability. Given the rapid changes in disease incidence that cannot be accounted for by changes in diagnostic practises, there is need to have well characterized exposure information in addition to genomic data for the study of gene-environment interactions. The case-control and cohort study designs are most suited for studying associations between risk factors and occurrence of an outcome. However, the case control study design is subject to several biases and hence the preferred choice of the prospective cohort study design in investigating gene-environment interactions. A major limitation of utilising the prospective cohort study design is the long duration of follow-up of participants to accumulate adequate outcome data. The GWAS paradigm is a timely reminder for traditional epidemiologists who often perform one- or few-at-a-time hypothesis-testing studies with the main hallmarks of GWAS being the agnostic approach and the massive dataset derived through large-scale international collaborations.

J Prev Med Public Health 2009;42(6):356-359

Key words : Genome wide association study, Gene environment interactions, Cohort study

INTRODUCTION

In the more than 100 genome wide association studies (GWAS) conducted in the past 5 years, more than 250 genetic loci contributing to more than 40 common diseases and traits have been identified [1]. Whilst many genes have been linked to a trait, both their individual and combined effects are small and unable to explain earlier estimates of heritability [2,3]. A major drive for GWAS to explore the genetic variance of common diseases was based on the hypothesis that "common diseases would be caused by common, low-penetrance variants when enough of them showed up in the same person," but this hypothesis is currently being challenged [2].

There may be genetic variants and mechanisms other than common low-penetrance single nucleotide polymorphisms like copy number variations [2,4,5] or epistasis that account for the missing hereditability. Rare variants either single-site or structural may have much larger effects than common variants but these cannot be elucidated through the current GWAS design [1].

Unfortunately, the current plethora of GWAS has diverted the role of non-genetic factors in the aetiology of common diseases. It is crucial to capitalize on the increased throughput and precision of measuring genetic factors to understand more precisely the role of gene-environment interactions in common diseases and traits.

ROLE OF NON-GENETIC FACTORS AND GENE-ENVIRONMENT INTERACTIONS

The role of environmental factors (broadly defined as non-genetic factors) in the aetiology of diseases have been well known in the pre-genomic era. Rapid changes in disease incidence within a few decades that cannot be accounted for by changes in diagnostic or notification practises suggests the prominent role of non-genetic factors. For example, the rapid rise in breast cancer incidence in Singapore compared to Sweden was attributed to the sharp decline in fertility in the 1970s [6]. Migrant studies have clearly shown that the migrants adopt the disease patterns of the host country within a short period of time [7,8]. Furthermore, reduction of environmental risk factors like smoking and hormone replacement therapy led to a corresponding decline in the related disease incidence [9-12].

There are also clear epidemiological evidence for interaction between genetic and environmental factors in disease causation [13,14]. The field of epigenetics has also introduced a new angle to the study of gene-environment interactions. Environmental factors have been shown to cause epigenetic changes like methylation and histone modifications resulting in heritable changes in gene function without a change in the DNA sequence [14]. However, not every gene-environment interaction results in an epigenetic

Received : 2 October 2009, Accepted : 2 November 2009
Corresponding author : Kee-Seng Chia (16 Medical Drive, Singapore 117597, Tel : (65)6874-4971, Fax : (65)6779-1489, E-mail : ephcks@nus.edu.sg)

modification or has easily accessible epigenetic markers. There is therefore a need to have well characterized exposure information in addition to genomic data for the study of gene-environment interactions.

THE COHORT STUDY DESIGN TO MEASURE GENE-ENVIRONMENT INTERACTIONS

The case-control and cohort study designs are most suited for studying the association between risk factors and the occurrence of an outcome. In a case-control study, cases and non-cases (controls) of the outcome of interest are selected from the same study-base. The exposure status is then determined retrospectively.

The GWAS era has popularized the case-control study design [15]. It is well suited for GWAS as the risk factor is a genetic factor with germline mutations. This factor is not subjected to recall bias and is stable as it does not change with the onset of disease. However, it can still be potentially biased by subject selection, namely prevalence-incidence bias where rapid onset and fatal diseases (eg. coronary heart disease), mild or silent cases, and diseases with short episodes may be missed [16]. Respondent bias, where those participants with positive family histories are more likely to participate [17], may also occur. With the incorporation of environmental exposures, recall bias occurs when disease status influences the reporting of exposures, for example those with the disease may be asked many times about exposure to a potential cause, whereas those without disease may only be asked once [18]. This makes it difficult to establish a clear temporal relationship between the cause and the effect of the disease.

The appropriate selection of controls is challenging as the use of convenient controls (eg. hospital controls) has been shown previously to lead to erroneous conclusions leading to indentifying extraneous factors rather than risk factors [19,20].

Reducing bias is the principal reason for the choice of the prospective cohort study design in investigating gene-environment interactions [21]. Prevalence-incidence bias in case identification is minimised as all participants are followed in a systematic way with all cases having an equal likelihood of being detected [21]. Respondent bias and recall bias are avoided by collecting data from participants before the onset of disease and fatality [21].

Cohort studies can be used to answer multiple hypothesis questions defined at the beginning of the study, as well as yet-to-be formulated hypotheses [5]. This study design can be considered more scientficially rigorous as both biological samples and environmental exposures are collected before the onset of disease. This is considered to be important in establishing the link towards causality [22-24].

An additional advantage of this study design is that a case control study can be nested within the main cohort where only a small sample of non-diseased participants would be concurrently evaluated with the cases [21,25]. Cohort studies can also screen for pre-disease markers from samples collected prior to the onset of disease [21,26].

A major limitation of utilising the prospective cohort study design is the long duration of follow-up of participants to accumulate adequate outcome data [27,28]. This can be accommodated for by the use of pre-disease markers to define the outcomes or the use of continuous measurements as the outcome.

COHORT STUDIES IN THE PRE-GENOMICS ERA

There are a number of well-known cohort studies that have contributed significantly to the understanding of the etiology of common chronic diseases. The most notable is the British Doctors Study initiated by the late Sir Doll et al. [29]. This was a prospective cohort study conducted from 1951-2001 on 34,439 male doctors in the United Kingdom. The study aimed to compare the hazards of cigarette smoking in men who formed their habits at different periods, and the extent of the reduction in risk when cigarette smoking was stopped at different ages. The initial outcome measure was overall mortality by smoking habit [29] and this was subsequently expanded to incidence of various cancers and smoking-related diseases. Exposure data was through repeated questionnaires and outcomes were obtained through medical records and death certificates. There was no biological specimen collection.

The Framingham Heart study, which aimed at identifying cardiovascular risk factors, differed slightly from the British Doctors Study in that blood specimens were obtained and stored. This allowed for the identification of serum and plasma biomarkers and subsequently genomic analyses. Unfortunately, the GWAS that arose from the Framingham Heart Study was underpowered [30].

Other notable studies in the pre-genomic era would include the Physicians' Health Study [31,32], the Nurses' Health Study [33] and the European Prospective Investigation into Cancer and Nutrition (EPIC) [34]. There were also several cohorts in Asia like the Singapore Chinese Health Study [24], the Korean National Prospective Occupational Cohort Study [35] and the Korean National Health Service Prospective Cohort Study [36] and all these included blood specimens making it possible to explore gene-environment interactions.

COHORT STUDIES OF THE FUTURE

To maximize on the emerging technologies in genomics and other 'omics', Collins proposed a list of optimal characteristics of a prospective gene-environment cohort study [37]. These include:

1. A large number of participants, at least several hundred thousand, should be enrolled. This would ensure an adequate sample size for common disorders.
2. Minority groups should be intentionally over-sampled to permit meaningful inferences about these groups and for the study of health disparities.
3. A broad range of ages should be represented to provide information on disorders from infancy to old age, with over-sampling of age groups as needed.
4. A broad range of genetic backgrounds and environmental exposures should be included to provide enough variability to detect and compare associations and interactions.
5. Family-based recruitment, including multiple generations, should be used for at least part of the cohort to increase the power of genetic analyses.
6. A broad array of clinical and laboratory information, not limited to any single disease, should be collected at the beginning and at regular intervals thereafter.
7. Sophisticated dietary, lifestyle and environmental exposure assessments should be carried out, using both questionnaires and biological measures.
8. Biological specimens, including DNA, plasma and cells, should be collected and stored.
9. A highly sophisticated data-management system should be included.
10. Access to study data and biological materials should be free and open to allow research into many diseases by scientists in many sectors.
11. Investigations during the study should not be limited to hypotheses conceived at its inception.
12. Comprehensive community engagement should be a major feature in the design and implementation of the study.

Potter [38] proposed the concept of 'The Last Cohort' - a cohort of a very large number of ethnically diverse individuals, who are well characterised genetically, whose exposures are diverse and well mapped, and whose illness pattern and mortality can be monitored over a long term using collected data, genomics and proteomics. While such proposals are extremely costly and would draw debate and criticisms [39], some countries like the UK had embarked on the ambitious UK Biobank Project which aimed at collecting data and samples from 500,000 subjects [40].

GWAS: A LESSON FOR TRADITIONAL EPIDEMIOLOGISTS

The GWAS paradigm is a timely reminder for traditional epidemiologists who often perform one-at-a-time or few-at-a-time hypothesis-testing studies. The main hallmarks of GWAS are the agnostic approach and the massive dataset derived through large-scale international collaborations [41]. The agnostic approach in studying non-genetic factors and gene-environment interactions may be debatable. The correlation structure in non-genetic factors are far more complex and the need for repeated sampling adds a further dimension to this high-dimensional data. However, the setting up of international consortia of cohort studies is potentially feasible. Examples of such efforts would include the Public Population Project in Genomics (P3G) and the Asia Cohort Consortium (ACC).

P3G was established in 2007 as a not-for-profit international consortium of leading public organisations involved in large-scale genetic epidemiological studies and biobanks. It was created for the development and management of a multidisciplinary collaboration and infrastructure for comparing and merging results and data sets from international population genomic studies eg. the Human Genome Project and the International HapMap Project [42].

The ACC was established in 2004, to understand the relationship between genetics, environmental exposures, and disease etiology through the formation of a cohort of at least 1 million healthy people internationally who will be followed over time to various disease outcomes. A current collaboration is a study focusing on the association of body mass index and total and all-cause mortality, and the role of a number of confounders, in Asian populations. Some relationships that can be explored include the association of exposure with disease, genome variability with disease and gene-environment interactions with molecularly defined disease [43].

REFERENCES

1. Goldstein DB. Common genetic variation and human traits. *N Engl J Med* 2009; 360(17): 1696-1698.
2. Maher B. Personal genomes: The case of the missing heritability. *Nature* 2008; 456(7218): 18-21.
3. Lango H, UK Type 2 Diabetes Genetics Consortium, Palmer CN, Morris AD, Zeggini E, Hattersley AT, et al. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* 2008; 57(11): 3129-3135.
4. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008; 455(7210): 237-241.
5. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; 455(7210): 232-236.
6. Chia KS, Reilly M, Tan CS, Lee J, Pawitan Y, Adami HO, et al. Profound changes in breast cancer incidence may reflect changes into a Westernized lifestyle: A comparative population-based study in Singapore and Sweden. *Int J Cancer* 2005; 113(2): 302-306.
7. Tominaga S. Cancer incidence in Japanese in Japan, Hawaii, and western United States. *Natl Cancer Inst Monogr* 1985; 69: 83-92.
8. McMichael AJ, Giles GG. Cancer in migrants to Australia: Extending the descriptive epidemiological data. *Cancer Res* 1988; 48(3): 751-756.
9. Ravdin PM, Cronin KA, Howlader N, Berg CD, Chlebowski RT, Feuer EJ, et al. The decrease in

breast-cancer incidence in 2003 in the United States. *N Engl J Med* 2007; 356(16): 1670-1674.
10. Peto R, Lopez AD, Boreham J, Thun M, Heath C Jr. Mortality from tobacco in developed countries: Indirect estimation from national vital statistics. *Lancet* 1992; 339(8804): 1268-1278.
11. Wingo PA, Ries LA, Giovino GA, Miller DS, Rosenberg HM, Shopland DR, et al. Annual report to the nation on the status of cancer, 1973-1996, with a special section on lung cancer and tobacco smoking. *J Natl Cancer Inst* 1999; 91(8): 675-690.
12. Colditz GA, Stampfer MJ, Willett WC, Hennekens CH, Rosner B, Speizer FE. Prospective study of estrogen replacement therapy and risk of breast cancer in postmenopausal women. *JAMA* 1990; 264(20): 2648-2653.
13. Ordovas JM, Corella D, Demissie S, Cupples LA, Couture P, Coltell O, et al. Dietary fat intake determines the effect of a common polymorphism in the hepatic lipase gene promoter on high-density lipoprotein metabolism: Evidence of a strong dose effect in this gene-nutrient interaction in the Framingham Study. *Circulation* 2002; 106(18): 2315-2321.
14. Liu L, Li Y, Tollefsbol TO. Gene-environment interactions and epigenetic basis of human diseases. *Curr Issues Mol Biol* 2008; 10(1-2): 25-36.
15. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; 447(7145): 661-678.
16. Schlesselman JJ. *Casecontrol Studies: Design, Conduct and Analysis*. New York: Oxford University Press; 1982.
17. Neyman J. Statistics: Servant of all sciences. *Science* 1955; 122(3166): 401-406.
18. Austin H, Hill HA, Flanders WD, Greenberg RS. Limitations in the application of case-control methodology. *Epidemiol Rev* 1994; 16(1): 65-76.
19. Grimes DA, Schulz KF. Compared to what? Finding controls for case-control studies. *Lancet* 2005; 365(9468): 1429-1433.
20. West DW, Schuman KL, Lyon JL, Robison LM, Allred R. Differences in risk estimations from a hospital and a population-based case-control study. *Int J Epidemiol* 1984; 13(2): 235-239.
21. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet* 2006; 7(10): 812-820.
22. Delco F, Sonnenburg A. Birth-cohort phenomenon in the time trends of martality from ulcerative colitis. *Am J Epidemiol* 1999; 150(4): 359-366.
23. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci USA* 2008; 105(44): 17046-17049.
24. Stern DA, Morgan WJ, Halonen M, Wright AL, Martinez FD. Wheezing and bronchial hyper-responsiveness in early childhood as predictors of newly diagnosed asthma in early adulthood: A longitudinal birth-cohort study. *Lancet* 2008; 372(9643): 1058-1064.
25. Manolio TA. Cohort studies and the genetics of complex disease. *Nat Genet* 2009; 41(1): 5-6.
26. Ioannidis JP. Genetic and molecular epidemiology. *J Epidemiol Community Health* 2007; 61(9): 757-758.
27. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; 358(9290): 1356-1360.
28. Seng CK, Wong CS, Lim WY, Loy EN, Wee S. *The Future Development and Funding of Population-Based Research in Singapore*. Singapore: NUS-GIS Centre for Molecular Epidemiology; 2007.
29. Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* 2004; 328(7455): 1519.
30. Kathiresan S, Manning A, Demissie S, D' Agostino R, Surti A, Guiducci C, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet* 2007; 8(Suppl 1): S17.
31. Physicians' Health Study I [Internet]. Boston: Physicians' Health Study; 2009 [cited 2009 Oct 6]. Available from: URL:http://phs.bwh.harvard.edu/phs1.htm.
32. Physicians' Health Study II [Internet]. Boston: Physicians' Health Study; 2009 [cited 2009 Oct 6]. Available from: URL:http://phs.bwh.harvard.edu/phs2.htm.
33. Nurses' Health Study (original cohort) [Internet]. Boston: Nurses' Health Study; c2008 [cited 2009 Oct 7]. Available from: URL:http://www.channing.harvard.edu/nhs/index.php/history.
34. The European Prospective Investigation into Cancer and Nutrition (EPIC). International Agency for Research on Cancer. Available from://epic.iarc.fr/.
35. Song YM, Sung J, Lawlor DA, Davey Smith G, Shin Y, Ebrahim S. Blood pressure, haemorrhagic stroke, and ischaemic stroke: The Korean national prospective occupational cohort study. *BMJ* 2004; 328(7435): 324-325.
36. Song YM, Ferrer RL, Cho SI, Sung J, Ebrahim S, Davey Smith G. Socioeconomic status and cardiovascular disease among men: The Korean national health service prospective cohort study. *Am J Public Health* 2006;96(1): 152-159.
37. Collins FS. The case for a US prospective cohort study of genes and environment. *Nature* 2004; 429(6990): 475-477.
38. Potter JD. Toward the last cohort. *Cancer Epidemiol Biomarkers Prev* 2004; 13(6): 895-897.
39. Willett WC, Blot WJ, Colditz GA, Folsom AR, Henderson BE, Stampfer MJ. Merging and emerging cohorts: Not worth the wait. *Nature* 2007; 445(7125): 257-258.
40. UK Biobank Limited. UK Biobank: What is it? [Internet]. Cheshire: UK Biobank Limited; [cited 2009 Oct 8]. Available from; URL: http://www.ukbiobank.ac.uk/about/what.php.
41. Seminara D, Khoury MJ, O' Brien TR, Manolio T, Gwinn ML, Little J, et al. The emergence of networks in human genome epidemiology: Challenges and opportunities. *Epidemiology* 2007; 18(1): 1-8.
42. Knoppers BM, Fortier I, Legault D, Burton P. Population genomics: The Public Population Project in Genomics (P3G): A proof of concept? *Eur J Hum Genet* 2008; 16(6): 664-665.
43. The Asia Cohort Consortium. About the Asia Cohort Consortium [Internet]. Seattle: The Asia Cohort Consortium; [cited 2009 Oct 6]. Available from: URL:http://www.asiacohort.org/Pages/ContactUs.aspx.