Published in final edited form as:

Surgery. 2010 April ; 147(4): 469–474. doi:10.1016/j.surg.2009.10.026.

Understanding Genetic Epidemiologic Association Studies Part 1: Fundamentals

Kaye M. Reid-Lombardo, MD^{*} and Gloria M. Petersen, PhD[†]

^{*}Division of Gastroenterologic and General Surgery, Department of Surgery

[†]Division of Epidemiology, Department of Health Sciences Research

Introduction

The scientific world has witnessed a fivefold increase in the number of published genomic studies of humans in the past two decades. Major advances in genetics, most notably, the completion of the Human Genome Project in 2001^{1, 2} and the International HapMap Project^{3–5} in 2003, in conjunction with the rapid advances of genotyping technology, has made it more feasible and less expensive for scientists to perform larger scale genetic epidemiologic association studies of common as well as rare diseases. The implications for patient care are not always readily apparent, but there is an underlying expectation that these types of studies will contribute to the development of individualized medicine, where genomic information is used during consultation to help tailor the treatment plan to the individual. More importantly, these studies may offer insight into the cause of disease under study and enhance our potential for risk stratification, which can ultimately help with local resource allocation.

Currently, we practice a “one-size-fits-all” approach to patients. With genomic information, we hope to offer in the future individualized screening recommendations and treatment plans that fall more in line with a “one-size-fits-one” approach. This “future” is currently being played out among several diseases, most notably breast cancer. We now understand and appreciate that in the two genes, breast cancer 1 (*BRCA1*)⁶ and breast cancer 2 (*BRCA2*)⁷ mutations are important risk factors for and predict high probability (about 70% lifetime risk) for early onset hereditary breast cancer. Indeed, women who undergo risk assessment and test positive for hereditary mutations in these genes may elect to have prophylactic bilateral mastectomies.^{8, 9} As for the medical oncologic treatment of breast cancer, there are compelling data based on genotyping information that suggests that patients with polymorphisms in the *CYP2D6* gene (an enzyme in the P450 pathway that helps to metabolize tamoxifen to its active form endoxifen) might alter the metabolism of the drug, causing low bioavailability of the drug. Women who have a specific *CYP2D6* polymorphism tend to have earlier disease recurrence and poorer survival.^{10, 11} Therefore, clinicians are now faced with the challenge of determining if the data available are robust enough to offer pre-treatment testing for this polymorphism, as it may help to determine if tamoxifen therapy, an accepted part of breast cancer care, will be a part of a patient’s treatment plan.

© 2009 Mosby, Inc. All rights reserved.

Address correspondence and reprint requests to: Kaye M. Reid-Lombardo, MD, Division of Gastroenterologic and General Surgery, Mayo Clinic, 200 First Street S.W., Rochester, MN 55905, Tel: 507-284-1529, Fax: 507-284-5196, reidlombardo.kaye@mayo.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Advances such as these require an understanding of the fundamentals of population-based genetic research in order to interpret the genetic data and apply the findings to patient care. This is the first of a two-part review that provides the fundamentals of genetic epidemiology association studies. Part 1 focuses on defining common terms and concepts used in these studies, and Part 2 focuses on how to critically analyze and interpret the findings of genetic epidemiology association studies.

What is Genetic Epidemiology?

Genetic epidemiology was originally defined as the study of the “etiology, distribution, and control of disease in groups of relatives and of inherited causes of disease in populations”.¹² This was later expanded to include a focus on “the role of genetic factors and their interaction with environmental factors in the occurrence of disease in human populations”. The goal of genetic epidemiology thus is to establish whether there is a genetic component to a disease observed in the population, taking into account environmental influences.

Why are these studies feasible?

The growth of genetic epidemiology has been fueled by two major developments. The first is the Human Genome Project, an international collaboration that began in 1990, and produced in 2001 a first comprehensive view of the deoxyribonucleic acid (DNA) sequence of the human genome, along with estimates that it comprises approximately 30,000 genes. The sequence data are now searchable via the Internet and offer an extremely accurate genomic sequence.² The second development was the International HapMap Project¹⁴ which had the goal of determining genotype, haplotype (consecutive alleles along a given human chromosome), and single nucleotide polymorphism (SNP) frequencies 3–5 from 30 trios (adult-and-both-parents) from Ibadan, Nigeria (YRI), 30 trios of United States Utah residents with ancestry from northern and western Europe (CEU), 44 unrelated individuals from Tokyo, Japan (JPT), and 45 unrelated Han Chinese individuals from Beijing, China (CHB). The SNP frequencies determined from this project are publicly available at the Single Nucleotide Polymorphism database (dbSNP) sponsored by the National Center for Biotechnology Information (NCBI),¹⁵ along with information on haplotypes. In the future, new initiatives including re-sequencing of genomes (i.e., 1000 genomes.org) will create a deep catalog of human genetic variation.

Understanding the attributes of DNA

The discovery of the double helix of DNA by Watson and Crick¹⁶ created the foundation that has helped to expand our knowledge to the point today of elucidating the entire sequence of the human genetic code. Scientists quickly deciphered the genetic code at the amino acid encoding level, and are now successfully identifying and characterizing the very genes that are associated with disease.

The basic unit of heredity is a gene, which is a contiguous region of DNA that is composed of combinations of four nitrogenous bases (adenine (A), guanine (G), thymine (T), and cytosine (C)) that by their very sequence along the DNA molecule encodes a protein.¹⁷

Deoxyribonucleic acid is linear molecule, configured as a double helix made up of two strands of a sugar-phosphate backbone and complement any base pairs A-G or C-T. (Figure 1 – DNA structure). Transcription and translation are the two primary processes which are required to assemble protein from a DNA template. Transcription is the process by which the DNA sequence is read by single stranded messenger ribonucleic acid (mRNA), which transports the code to ribosomes for building the protein; mRNA is then translated into protein via transfer ribonucleic acid (tRNA) which transports the appropriate amino acids to the ribosome for processing and protein assembly. A change in one or more bases (e.g., A to G or T to C) is known as a mutation which modifies the sequence, and can result in an alteration to the amino

acid sequence thus potentially creating a change, either as a substitution or frameshift in the genetic code, and consequently altering the protein structure and function which could lead to disease.

Genetic Variation

Genetic variation arises by mutations in DNA that can occur throughout the human genome and if not resulting in severe alterations, can persist over generations. If a given variant increases to >1% frequency in a population, it is called a *polymorphism*. Genetic variation can manifest in many ways; the more frequently studied variants/polymorphisms are insertion or deletions (indels), copy number variations (CNVs), and SNPs. Deletions, which are loss of nucleotides from the normal DNA sequence, and insertions, which are addition of nucleotide (s) to the DNA sequence, can involve a few base pairs or many base pairs.¹⁸ An example of one of the most common genetic diseases that results from a frameshift deletion is cystic fibrosis, which affects over 70% of those diagnosed with the disease.¹⁹

Copy number variation (CNV) is manifested as repeats of short sequences of DNA that can vary from individual to individual, and which are heritable. Copy number variants typically involve a DNA fragment that is one kilobase (kb) or larger, and occur at a frequency of about 12 per person.²⁰ It is estimated by Sebat et al. that more than half of copy number variants overlap DNA regions that encode proteins.²¹ Copy number variants can interact with each other or the environment to cause disease. In breast cancer, the overexpression of the *ERBB2* gene will lead to an aggressive type of cancer. Overexpression or underexpression of *ERBB2* is caused by CNV and other mutations. This gene is thus a good target in the treatment of breast cancer.

A *single nucleotide polymorphism* (SNP) is a change of a nucleotide at a single base pair (Figure 2), and is the most common type of polymorphism. For example, in one person, without disease, a portion of DNA sequence could read GCTACTGCC while in a patient with disease, the sequence could be GCTACTACC. Single nucleotide polymorphisms will have only two alternate forms, called *alleles*, which could thus contain different information at a specific locus within a gene. The frequency of the least frequent allele of a SNP in a population is termed the *minor allele frequency* (MAF) and can vary among populations. The *wild-type* allele is considered the “normal” allele in a specific population.

Single nucleotide polymorphisms occur throughout the genome in the coding and noncoding portions of the DNA. If a change in a SNP does not lead to a change in the polypeptide sequence, then that SNP is referred to as a *synonymous* SNP (sSNP). Single nucleotide polymorphisms that lead to a change in the protein or protein structure and thus have more potential to represent a disease causing mechanism, are called *non-synonymous* SNPs (nsSNPs). Synonymous SNPs are thus more common than nsSNPs.

The natural process of linkage disequilibrium (nonrandom association of two or more alleles at two or more physically proximate loci) and recombination during meiosis allows for an efficient analysis of resulting patterns of SNPs in the genome. Regions of DNA that are associated or remain during recombination (and thus “travel together” in transmitting the genomic material from parent to offspring), efficiently permit investigation of only one representative SNP that can serve as a “tag” or “marker” for two or more nearby SNPs, by decreasing the number of total SNPs that need to be directly tested. Tag SNPs are routinely utilized for a cursory overview of the genome when searching for “hot spots” of interest, thus identifying regions of DNA that might be associated with the disease under study. Once identified, the region of interest can be scrutinized with additional genomic techniques at a more granular level in the hope of determining what is the *causal* gene.

Types of Studies

Currently, there are two general approaches available for rapid evaluation of the genome to determine if there is a genetic cause of a disease. Both typically involve a case-control design to determine whether a SNP is found in higher frequency among the cases when compared to controls. The first approach is the *genome-wide association study* (GWAS). Tag SNPs from across the entire genome are selected to give a first step high level assessment of possible interesting regions. Genome-wide association studies are typically hypothesis-generating studies, in that the entire genome is “scanned” to identify disease susceptibility genes in patients with the disease or trait of interest (cases) compared to subjects without the disease (controls). Because of this, the sample size required is typically quite large, with 1,000 or more pairs of cases and controls. This makes this method costly, since hundreds or thousands of SNPs are evaluated in large samples. Genome-wide association studies also typically require a second phase, or replication step, in which an independent sample of cases and controls is tested for a smaller number of interesting SNPs that were identified through statistical analyses of the initial phase data set. An illustrative example is a genome-wide association study by Tomlinson et al., who examined 550,000 tag SNPs in 930 familial colorectal tumor cases and 960 controls. They identified a susceptibility variant for colorectal cancer at chromosome 8q24.21.²² Statistical analysis is applied to identify significant differences between cases and controls, enabling focus on SNPs and haplotypes in potential genes of interest.

The ultimate goal of GWAS is to identify novel genomic targets for disease intervention. The primary disadvantage at present is cost, but this is becoming less of a barrier as the technologies for high throughput processing improve. Another consideration is the computational support that is required. Genome-wide association studies generate massive amounts of data, and a sophisticated electronic and bioinformatics infrastructure is needed to support data storage and analysis.

The second most common approach is the *candidate gene association study*. Unlike the GWAS, this method is hypothesis driven where there is *a priori* selection of genes for examination. Typically, the investigator focuses on genes in a pathway of interest thought to be involved in the disease of interest. The major advantage of this study is that fewer genes are examined and sample sizes can be smaller than required for a GWAS, so the cost is lower. The major limitation in the candidate gene association study is that the number of genes analyzed is pre-established and limited, therefore novel pathway or critical genes can be missed.

Genotyping Platforms

The design of genomic studies includes the genotyping platform (technology used to assay the genotypes of the study participants), selection of which is influenced by the number of SNPs to be examined, the available budget, resources, and sample size, whether for GWAS or candidate gene study. High throughput genotyping platforms evaluate many SNPs, from 384 to as high as 1 million SNPs. Platforms can be customized to allow for individualized selection of SNPs or can be fixed panels focusing on a disease type (such as cancer) or biologic processes (such as inflammation). Various technologies also exist that allow an investigator to evaluate a small number of SNPs at a time.

Selection of SNPs

Candidate gene association studies or GWAS that use a commercially available customizable platform allow for individualized selections of SNPs. There are many variables that must be considered in selecting SNPs for a custom panel. The most important is cost. Budgets may require that an investigator limit the amount of SNPs for study. In addition, the less common,

or minor allele frequency (MAF) is usually determined prior to the selection of SNPs; a minimal MAF threshold, for example 1%, is selected for *nonsynonymous* SNPs.

A given gene can contain dozens of SNPs. In order to ensure full coverage of potential variation of a specific gene during the SNP selection process, the regions of DNA that are upstream and downstream of the known chromosomal location of the gene are also eligible for analysis. Often, 1000 base pairs upstream and down stream of the gene is included. However, not all possible SNPs would be included in an analysis. *Synonymous* SNPs are given preference over nsSNPs, and SNPs that are close together (≤ 60 base pairs apart), are selectively excluded, as it is difficult to genotype both on some platforms. Large genes will require the selection of multiple tag SNPs to ensure adequate coverage of the entire gene. These tag SNPs are selected on the basis of ease of and success rate of genotyping for that SNP (*design score*), and then genotyped using a commercially available genotyping platform. During the SNP selection process, SNPs of a particular gene are selected by interrogating available databases such as HapMap³ and Seattle SNPs.²³

The potential for utilizing genomics for personalized medicine is great, and genetic epidemiologic analysis of large amounts of genomic data is one avenue to achieve this. As we learn more about formal analysis of SNP-SNP interactions and SNP-environment interactions, the hope is to be able to predict associations more accurately. The technologic costs continue to decrease; it is reported that by the end of 2009 the cost of performing a full genomic analysis will be approximately \$3,000 per person. We can therefore anticipate an increase of these types of studies in the scientific literature. Reviews such as this will help to prepare clinicians to ensure that new genomic information can be integrated into the clinical practice as seamlessly as possible. In Part II, the focus will be on critically analyzing the methodology of genetic association studies, allowing for improved interpretation.

Acknowledgments

This publication was made possible by Grant Number 1 UL1 RR024150 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and the NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. Information on NCRR is available at <http://www.ncrr.nih.gov/>. Information on Reengineering the Clinical Research Enterprise can be obtained from <http://nihroadmap.nih.gov>.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921. [see comment][erratum appears in *Nature* 2001 Aug 2;412(6846):565 Note: Szustakowki, J [corrected to Szustakowski, J]]. [PubMed: 11237011]
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* 2001;291:1304–1351. [see comment][erratum appears in *Science* 2001 Jun 5;292(5523):1838]. [PubMed: 11181995]
3. The International HapMap Consortium. The International HapMap Project. *Nature* 2003;426:789–796. [PubMed: 14685227]
4. The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437:1299–1320. [PubMed: 16255080]
5. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–861. [PubMed: 17943122]
6. Futreal PA, Liu Q, Shattuck-Eidens D, Cochran C, Harshman K, Tavtigian S, et al. BRCA 1 Mutations in Primary Breast and Ovarian Carcinomas. *Science* 1994;266:120–122. [PubMed: 7939630]
7. Gayther SA, Mangion J, Russell P, Seal S, Barfoot R, Ponder BA, et al. Variation of risks of breast and ovarian cancer associated with different germline mutations of the BRCA2 gene. *Nat Genet* 1997;15:103–105. [PubMed: 8988179]

8. Blanchard DK, Hartmann LC. Prophylactic surgery for women at high risk for breast cancer. *Clin Breast Cancer* 2000;1:127–134. discussion 35. [PubMed: 11899651]
9. Ray JA, Loescher LJ, Brewer M. Risk-reduction surgery decisions in high-risk women seen for genetic counseling. *J Genet Couns* 2005;14:473–484. [PubMed: 16388328]
10. Goetz MP, Knox SK, Suman VJ, Rae JM, Safgren SL, Ames MM, et al. The impact of cytochrome P450 2D6 metabolism in women receiving adjuvant tamoxifen. *Breast Cancer Res Treat* 2007;101:113–121. [PubMed: 17115111]
11. Goetz MP, Rae JM, Suman VJ, Safgren SL, Ames MM, Vissche DW, et al. Pharmacogenomic determinants of outcome with tamoxifen therapy: findings from the randomized North Central Cancer Treatment Group adjuvant breast cancer trial 89-30-52. *Breast Cancer Res Treat* 2004;88:S35.
12. Morton, N. Outline of genetic epidemiology. Basel (Switzerland): S. Karger AG; 1982.
13. Khowry, M.; Beaty, T.; Cohen, B. Fundamentals of Genetic Epidemiology. New York (NY): Oxford U. Press; 1993.
14. Thorisson GA, Smith AV, Krishnan L, Stein LD. The International HapMap Project Web site. *Genome Res* 2005;15:1592–1593. [PubMed: 16251469]
15. Single Nucleotide Polymorphism (dbSNP) [Internet]. Bethesda (MD): National Center for Biotechnology Information (NCBI); 2008 [cited 2009]. Available from: <http://www.ncbi.nlm.nih.gov/>.
16. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 1953;171:737–738. [PubMed: 13054692]
17. Lewis, B. *Genes VIII*. Saddle River (NJ): Prentice Hall; 2003.
18. Nussbaum, R.; McInnes, R.; Willard, H. *Thompson & Thompson Genetics in Medicine*. 6th. Philadelphia (PA): WB Saunders Co; 2007.
19. Harris A. Cystic fibrosis gene. *Br Med Bull* 1992;48:738–753. [PubMed: 1281033]
20. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 2006;7:85–97. [PubMed: 16418744]
21. Sebat. Large-scale copy number polymorphism in the human genome. *Science* 2004;305:528–528. [PubMed: 15247437]
22. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 2007;39:984–988. [PubMed: 17618284]
23. SeattleSNPs [Internet]. Seattle (WA): The University of Washington; 2000 [cited February 19, 2009]. Available from: <http://pga.gs.washington.edu/>.
24. Attia J, Ioannidis JP, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association: A: Background concepts. *JAMA* 2009;301:74–81. [PubMed: 19126812]
25. Attia J, Ioannidis JP, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association: C: What are the results and will they help me in caring for my patients? *JAMA* 2009;301:304–308. [PubMed: 19155457]
26. Attia J, Ioannidis JP, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association: B: Are the results of the study valid? *JAMA* 2009;301:191–197. [PubMed: 19141767]

Figure 1. Components and Structure of DNA

Reproduced by permission: JAMA January 7, 2009, 301(1):74–81. Copyright © 2009 American Medical Association. All rights reserved.²⁴

The building blocks of DNA are the nucleotides, adenine (A), thymine (T), guanine (G) and cytosine (C). Base pairs are linked to a sugar phosphate backbone in the DNA strand. The nucleotides form base pairs via hydrogen bonds from two strands of DNA to create the DNA helix. The DNA helix is comprised of two single strands of DNA both with a 5' and 3' end that bind in opposite directions. The pairing of base pairs are always as follows: adenine binds with thymine, and guanine binds with cytosine.^{24–26}

Figure 2. Single Nucleotide polymorphism (SNP)

Single Nucleotide polymorphism (SNP), in this example A-T is replaced by G-C, thus creating two potential forms or alleles at this site.