Methods in
Molecular Biology 713

Springer Protocols

M. Dawn Teare
Editor

Genetic Epidemiology

Humana Press

METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For other titles published in this series, go to
www.springer.com/series/7651

Genetic Epidemiology

Edited by

M. Dawn Teare

University of Sheffield, Sheffield, UK

Humana Press

Editor
M. Dawn Teare
Health Services Research
School of Health and Related Research
University of Sheffield
Sheffield, UK
m.d.teare@sheffield.ac.uk

ISSN 1064-3745 e-ISSN 1940-6029
ISBN 978-1-60327-415-9 e-ISBN 978-1-60327-416-6
DOI 10.1007/978-1-60327-416-6
Springer New York Dordrecht London Heidelberg

While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Humana Press is part of Springer Science+Business Media (www.springer.com)

Preface

Genetic epidemiology is the study of the role of genes and environments on markers of health and disease risk in populations. It emerged as a mainstream discipline in the early 1980s, arising from firm foundations laid by mathematical population genetics, clinical genetics, and statistical epidemiology. Though genetic epidemiology attempts to identify the many components of risk attributable to genes, environments, and interactions between these two factors, the course of the research towards this goal can follow many diverse paths. In the last few years, the success of genome-wide association studies in their identification of hundreds of disease susceptibility loci has brought this specialist field to the forefront of biomedical research.

Advances in molecular genetics will soon offer affordable means to measure or observe study participant's genetic material at the sequence level as well as more detailed functional data, such as gene expression. It is evident that genetic epidemiology projects increasingly require long-term collaboration between bioinformaticians, geneticists, clinicians, statisticians, and epidemiologists. As with any field that is making rapid advances, technologies, and methodologies are both developed and superseded quickly. However, in spite of the rapid changes in techniques, much of the basic language, models, and principles have remained the same.

Interdisciplinary research requires good communication and understanding across the participating disciplines, and this book aims to provide a basic framework for this communication suited to newcomers to the field as well as experienced researchers and graduate level students. Statistical methods are applied in a wide range of disciplines, and this is one subject area that is well catered for by existing text books, particularly at the introductory level. This book assumes a basic level of competence with regard to statistical and probabilistic reasoning, so readers lacking confidence in this respect are guided towards more introductory texts [1–3].

Section 1 consists of three chapters covering the very basics of modern molecular genetics, the terminology and models frequently employed in genetic epidemiology, and an introduction to epidemiology. This section concisely presents most of the language and key concepts that are required to understand the more specific topics discussed in the subsequent sections.

Principles of genetic linkage analysis are outlined in Sect. 2. Section 3 contains five chapters that cover genetic association studies, including an overview chapter of genomic resources available through the Web.

Sections 4 and 5 contain some more specialist topics and three case studies where many of the concepts introduced in earlier chapters are illustrated with interpreted examples.

Those wanting more detail on how to apply statistical reasoning or how to use the necessary computational methods can move onto the more advanced range of textbooks which each have their own perspective. There are a good number of texts specifically for genetic epidemiology [4–10]. For many years, Jurg Ott's "The Analysis of Human Genetic Linkage" [11] was the key source for researchers in gene mapping, though more recent books also cover these methods [12–15]. The founder disciplines, population, and quantitative genetics, are well covered by many textbooks, and we refer to just a selection of these [16–20]. Statistical modellers and graduate researchers find the handbook of statistical genetics [21] and the encyclopaedia of biostatistics and genetic epidemiology comprehensive source material [22]. In addition to printed texts, there are extensive ranges of educational material available online. Resources supporting education in genetics are particularly well developed [23–25].

I thank sincerely all those who have helped to bring this book together, particularly the co-authors. The content and emphasis of this book has been strongly influenced by both colleagues on the academic staff and students of the MSc in Genetic Epidemiology that was lead by Professor Chris Cannings at the University of Sheffield.

M. Dawn Teare

References

1. Swinscow, T.D.V., Campbell, M.J. (eds) (2009) Statistics at Square One. BMJ Books, London.
2. Campbell, M.J. (ed) (2001) Statistics at Square Two. BMJ Books, London.
3. McKillip, S. (ed) (2005) Statistics Explained. Cambridge University Press, Cambridge.
4. Khoury, M.J., Beaty, T.H., Cohen, B.H. (eds) (1993) Fundamentals of Genetic Epidemiology. Oxford University Press, New York.
5. Thomas, D.C. (ed) (2004) Statistical Methods in Genetic Epidemiology. Oxford University Press, USA.
6. Ziegler, A., Koenig, I.R. (eds) (2010) A Statistical Approach to Genetic Epidemiology: Concepts and Applications. Wiley, Weinheim.
7. Sham, P. (ed) (1997) Statistics in Human Genetics. Wiley Blackwell, New York.
8. Palmer, L.J., Burton, P., Davey-Smith, G. (eds) (2010) An Introduction to Genetic Epidemiology. Policy Press, Bristol.
9. Kneale, B., Ferreira, M., Medland, S., Posthuma, D. (eds) (2007) Statistical Genetics: Gene Mapping Through Linkage and Association. Taylor and Francis, London.
10. Weiss, K.M. (ed) (1995) Genetic Variation and Human Disease. Cambridge University Press, Cambridge.
11. Ott, J. (ed) (1999) Analysis of Human Genetic Linkage. Johns Hopkins University Press, Baltimore.
12. Weir, B.S. (ed) (1997) Genetic Data Analysis. Sinauer Associates Inc., Sunderland, MA.
13. Siegmund, D., Yakir, B. (eds) (2006) The Statistics of Gene Mapping. Springer, New York.
14. Lang, K. (ed) (2003) Mathematical and Statistical Methods for Genetic Analysis. Springer, New York.
15. Camp, N.J., Cox, A. (eds) (2002) Quantitative Trait Loci: Methods and Protocols. Humana Press, New Jersey.
16. Falconer, D.S., Mackay, T.C. (eds) (1995) Introduction to Quantitative Genetics. Longman, New York.

17. Cannings, C., Thompson, E.A. (eds) (1981) Genealogical and Genetic Structure. Cambridge University Press, Cambridge.
18. Gillespie, J.H. (ed) (2004) Population Genetics: A Concise Guide. Johns Hopkins University Press, Baltimore.
19. Lynch, M., Walsh, B. (eds) (1998) Genetics and Analysis of Quantitative Traits. Sinauer Associates Inc., Sunderland, MA.
20. Gale, J.S. (ed) (1990) Theoretical Population Genetics. Springer, New York.
21. Balding, D.J., Bishop, M., Cannings, C. (eds) (2007) Handbook of Statistical Genetics. Wiley Blackwell, New York.
22. Elston, R.C., Olson, J.M., Palmer, L.J. (eds) (2002) Biostatistical Genetics and Genetic Epidemiology. Wiley, New York.
23. http://learn.genetics.utah.edu/content/begin/tour/
24. http://www.dnaftb.org/
25. http://www.eurogene.eu/

Contents

PART V CASE STUDIES

Contributors

Kristina Allen-Brady • *Genetic Epidemiology Division, Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA*

Jennifer H. Barrett • *Section of Epidemiology and Biostatistics, Leeds Institute for Molecular Medicine, University of Leeds, Leeds, UK*

Nicola J. Camp • *Genetic Epidemiology Division, Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA*

Françoise Clerget-Darpoux • *INSERM U535, Université Paris-Sud, Paris, France*

Andrew Collins • *Human Genetics Research Division, University of Southampton, Southampton, UK*

Karen Curtin • *Genetic Epidemiology Division, Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA*

Vanessa Didelez • *Department of Mathematics, University of Bristol, Bristol, UK*

Frank Dudbridge • *Department Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK*

Cother Hajat • *Public Health Programmes, Health Authority-Abu Dhabi, Dubai, United Arab Emirates*

Mark M. Iles • *Section of Epidemiology and Biostatistics, Leeds Institute for Molecular Medicine, University of Leeds, Leeds, UK*

Laura M. Johnson • *Cancer Research UK Health Behaviour Research Centre, Department of Epidemiology and Public Health, University College London, London, UK*

Jaakko Kaprio • *Department of Public Health and Institute of Molecular Medicine, University of Helsinki & National Institute for Health and Welfare, Helsinki, Finland*

Sha Meng • *Department of Health Sciences, University of Leicester, Leicester, UK*

Hervé Perdry • *INSERM U535, Université Paris-Sud, Paris, France*

Mauro F. Santibàñez Koref • *Institute of Human Genetics, University of Newcastle, Newcastle upon Tyne, UK*

Nuala A. Sheehan • *Department of Health Sciences, University of Leicester, Leicester, UK*

Karri Silventoinen • *Departments of Sociology and Public Health, University of Helsinki, Helsinki, Finland*

William J. Tapper • *Human Genetics Research Division, University of Southampton, Southampton, UK*

M. Dawn Teare • *Health Services Research, School of Health and Related Research, University of Sheffield, Sheffield, UK*

Martin D. Tobin • *Departments of Health Sciences and Genetics, University of Leicester, Leicester, UK*

Mohammed-Elfatih Twfieg • *Department of Probability and Statistics, University of Sheffield, Sheffield, UK*
Louise V. Wain • *Department of Health Sciences, University of Leicester, Leicester, UK*
Kevin Walters • *Department of Probability and Statistics, University of Sheffield, Sheffield, UK*

Part I

Introduction

Chapter 1

Molecular Genetics and Genetic Variation

Mohammed-Elfatih Twfieg and M. Dawn Teare

Abstract

This chapter contains brief notes on molecular genetics, focusing on those aspects most frequently encountered in genetic epidemiology. The main sections cover the organisation and physical structure of genetic material, the mechanisms involved in transmitting genetic material from one generation to the next, and forms of genetic variation.

Key words: DNA, RNA, Recombination, Mutation, Genetic polymorphism, Meiosis, Chromosomes

1. The Basic Building Blocks of Heritable Information

Genetics is defined as the study of heredity, that is, the study of units or characteristics that can be transmitted from parents to offspring. The term "gene" has been in use for over a century, and its definition has evolved as the field of molecular genetics has developed (1). When used in the context of molecular genetics "gene" commonly refers to a segment of a nucleic acid molecule, usually a deoxyribonucleic acid (DNA) that encodes a ribonucleic acid (RNA) molecule. These RNA molecules are then used to synthesise polypeptides (proteins).

In cellular organisms, the genetic information is stored in stable DNA molecules, though in viruses the heritable information can be transferred by RNA molecules. This genetic information is organised into chromosomes. A chromosome is defined as a long string of DNA which is capable of regulated replication and can be transmitted to its descendents through a reproductive cycle. The reproductive cycle is a process during which chromosomes are copied and passed from a parent organism to an offspring organism.

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_1, © Springer Science+Business Media, LLC 2011

DNA and RNA are long molecules composed of smaller molecular units called nucleotides. Each nucleotide is composed of a sugar residue (deoxyribose or ribose), a nitrogenous base, and a phosphate group. There are two classes of nitrogenous base; the purines, adenine(A), and guanine(G); and the pyrimidines, cytosine(C), and thymine(T). RNA differs from DNA in the sugar residue (ribose) and the nucleotide uracil (U) in place of thymine. The DNA double helix results from nucleotide base pair bonds causing two linear molecules to "pair" (see Fig. 1). These two paired molecules are said to be complementary, and this results because A specifically binds to T and C specifically binds to G.

The "ends" of the DNA are referred to as either the 3′ or 5′ end (pronounced "three primed" or "five primed"). This is due

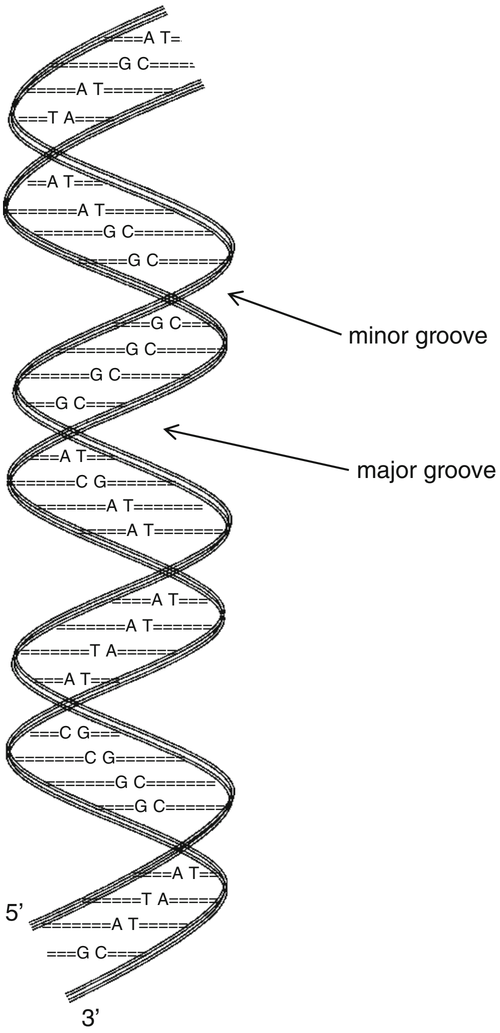


Fig. 1. The double helix.

to the conventional labelling of the carbon atoms constituting the sugar molecules. If a segment of DNA is described by writing down its sequence of nitrogenous bases (for example AGGTGTAAA), this is usually done for one of the paired strands and in the 5′ to 3′ direction, in the same orientation as genes are transcribed.

2. DNA Function from Transcription to Translation

The DNA alphabet consists of only four letters, A, C, G, and T. This language contains the basic instructions for building proteins. The protein building process has two steps, first the relevant section of DNA is transcribed into RNA, and then the RNA is translated into a sequence of amino acids which form a protein. The relation between specific base triplets (codons) of RNA and the associated amino acid is termed *the genetic code* (see Table 1). There are 20 amino acids and 64 possible codons. Several amino acids correspond to more than one codon, so there is redundancy, or we say the code is degenerate. Codons that correspond to the same amino acid are termed synonymous. The translation proceeds from the 5′ end to the 3′ end of the mRNA starting at an initiation codon (almost always AUG which specifies methionine). It proceeds one codon at a time until a termination or STOP codon is reached. The string of codons between the START and STOP codon is called the *open reading frame*. The physical and chemical properties of a protein molecule are heavily determined by its sequence of amino acids.

The linear sequence of DNA is first used as a template for the transcription of an RNA molecule. The RNA molecule then forms the basis for constructing a linear sequence of amino acids which constitute a polypeptide product, the translation step. However, before the translation step takes place the RNA transcripts are processed. The processing involves adding chemical markers to sites on the transcripts and removing “non-coding” regions of the RNA (splicing). The segments of the gene transcripts which are removed are termed introns, and these are flanked by exons (which are retained). The terms intron and exon can refer both to the RNA and DNA sequence. The processed RNA is called messenger RNA (mRNA).

The steps involved in producing mRNA are conducted in the cell nucleus, while the translation step involves ribosomes and occurs in the cytoplasm. In complex organisms, only a fraction of the DNA is *expressed* to give rise to an RNA product. Some transcribed RNA units are not destined to mRNA production and have a different cellular function. DNA sequence which is not translated directly into protein is termed non-coding

Table 1
The genetic code

	Second base			
First base	**U**	**C**	**A**	**w**
U	UUU,**F**	UCU,**S**	UAU,**Y**	UGU,**C**
	UUC,**F**	UCC,**S**	UAC,**Y**	UGC,**C**
	UUA,**L**	UCA,**S**	UAA,**X**	UGA,**X**
	UUG,**L**	UCG,**S**	UAG,**X**	UGG,**W**
C	CUU,**L**	CCU,**P**	CAU,**H**	CGU,**R**
	CUC,**L**	CCC,**P**	CAC,**H**	CGC,**R**
	CUA,**L**	CCA,**P**	CAA,**Q**	CGA,**R**
	CUG,**L**	CCG,**P**	CAG,**Q**	CGG,**R**
A	AUU,**I**	ACU,**T**	AAU,**N**	AGU,**S**
	AUC,**I**	ACC,**T**	AAC,**N**	AGC,**S**
	AUA,**I**	ACA,**T**	AAA,**K**	AGA,**R**
	AUG,**M**	ACG,**T**	AAG,**K**	AGG,**R**
G	GUU,**V**	GCU,**A**	GAU,**D**	GGU,**G**
	GUC,**V**	GCC,**A**	GAC,**D**	GGC,**G**
	GUA,**V**	GCA,**A**	GAA,**E**	GGA,**G**
	GUG,**V**	GCG,**A**	GAG,**E**	GGG,**G**

Codons and corresponding amino acids (indicated with single letter abbreviation). The codon codes for the amino acid indicated by the single letter in bold font. The single letter (and three letter) representation for the 20 amino acids is as follows: A (Ala) alanine, C (Cys) cysteine, D (Asp) aspartic acid, E (Glu) glutamic acid, F (Phe) phenylalanine, G (Gly) glycine, H (His) histidine, I (Ile) isoleucine, K (lys) lysine, L (Leu) leucine, M (Met) methionine, N (Asn) asparagines, P (Pro) proline, Q (Gln) glutamine, R (Arg) arginine, S (Ser) serine, T (Thr) threonine, V (Val) valine, W (Trp) tryptophan, Y (Tyr) tyrosine, and X labels a STOP codon

DNA. The ratio of coding to non-coding DNA varies among Eukaryotes. Only a small proportion (less than 5%) of human DNA is coding DNA. Until recently, non-coding DNA was termed “junk” DNA obviously implying it had no use. However, accumulating evidence strongly supports gene regulatory roles for non-coding DNA (2, 3).

In animal cells, DNA is found both in the nucleus and mitochondria. A very limited number of genes are contained within the mitochondrial DNA. The mitochondrial chromosome is circular in mammals and consists of a purine-rich *heavy* (H) strand and a pyrimidine-rich *light* (L) strand. The human mitochondrial genome codes for 37 genes. Cells contain different numbers of mitochondria and within each mitochondrion multiple copies of the chromosome occur. Unlike the nuclear DNA, the human mitochondrial genome is 93% coding sequence (4).

3. Ploidy

The full set of chromosomes for an organism is termed its genome. The number of chromosome sets per cell nucleus is termed ploidy. The human genome consists of 23 pairs of chromosomes and the mitochondrial chromosome. Most human cells, in common with the majority of mammals, are diploid (two copies of each chromosome). The adult human arises from a single diploid zygote, which results from the fusion of two gametes which are haploid cells (a single copy of each chromosome); one donated by each parent. Normal gametes contain 22 autosomes and 1 sex chromosome. Normal human zygotes consist of 22 pairs of autosomes and a pair of sex chromosomes.

There are two sex chromosomes labelled the X and Y. The Y chromosome is much shorter than the X and contains a short section which is homologous to a region on the X chromosome (the pseudoautosomal region). In this context, homologous means the regions exhibit highly specific similarity in DNA structure. When a pair of chromosomes is described as homologous, this generally means two distinct molecules which belong to the same chromosome group (for example, chromosome 8). Chromosome homology is very important during meiosis, and ensures that the equivalent or corresponding pairs of chromosomes align correctly. Females have two copies of X and males have 1 X and 1 Y. Hence, females can only transmit X chromosomes to their offspring, while males can transmit X or Y. The mitochondrial chromosome is passed to offspring through the maternal line only.

4. DNA Replication and Cell Division

Cell division is required for the growth and maintenance of an organism, and in reproduction. In order for a cell to undergo cell division, it must first make a full copy of its genome. In DNA replication, a double-stranded DNA (dsDNA) molecule is copied to produce two identical daughter molecules, referred to as a *pair of sister chromatids* (see Fig. 2). DNA replication is termed *semi-conservative* as each daughter helix consists of one parental strand and one new strand.

For a zygote (fertilised egg) to develop into a full adult diploid organism, mitotic cell division or *mitosis* is required. Among other things, mitosis requires that a single cell can make an exact copy or duplicate of its DNA so that the two resulting daughter cells contain the same full and identical complement of DNA that was present in the originator cell. So each daughter cell must receive exactly one copy of the DNA originally inherited from each parent.

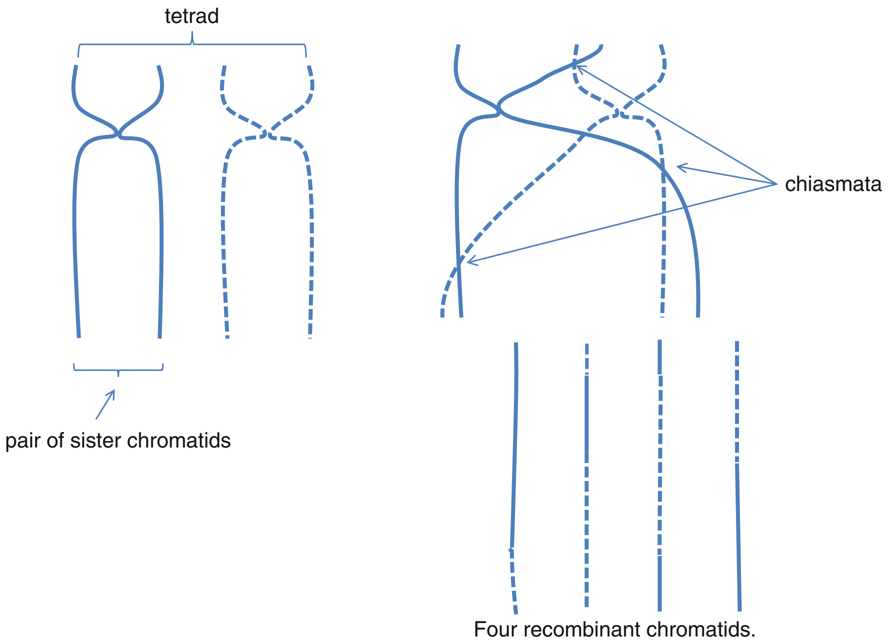


Fig. 2. A schematic of recombination and crossing over.

In mitosis, the two copies of the genome separate, the nucleus splits into two, and each daughter cell receives one nucleus.

Meiosis is also a form of cell division requiring DNA replication, but its primary function is to create haploid cells, containing exactly one copy of each chromosome. During meiosis, a single diploid cell is processed into four haploid cells. This requires a DNA replication step and cell division as in mitosis, but then a further cell division results giving rise to haploid cells. Unlike mitosis where the objective is to produce genetically identical daughter cells, meiosis results in genetically different haploid cells through two mechanisms, independent assortment of homologous chromosomes and recombination.

In multicellular organisms, meiosis occurs only in the germ cells. Cell division presents and opportunity for a copying error to be introduced to the DNA through the replication process so the germ cells are set aside early in the development of the organism. The primary function of the germ cells is to create haploid cells for the transmission of genetic material to the next generation. The non-germ cells are referred to collectively as somatic cells.

5. Recombination

There are three types of recombination that occur in normal cells; these are termed transposition, site-specific, and general. These three forms of recombination differ in their physical and molecular processes; however, here we only consider the products or outcomes of the processes. It is general recombination that is of the most interest in the study of heredity.

General recombination is the process by which parental DNA within a chromosome is shuffled during meiosis. Homologous pairs of dsDNA align and then exchange lengths of DNA by first cutting the strands in the same place and swapping the dsDNA at the breakpoint. The sites of the crossover or recombination are marked by chiasmata. General recombination occurs throughout the whole genome, though the frequency and chromosomal location of recombination events differs between males and females. Figure 2 shows an example of the product of recombination for one tetrad of sister chromatids. This random process occurs within meiosis for each pair of chromosomes and the longer the chromosome, the more likely recombinations occur. As can be seen in Fig. 2, recombination results in a shuffling of the genetic material so that for each chromosome potentially four distinct chromatids can be transmitted to the gamete or next generation. Aspects of recombination are further discussed in Chapters 4 and 5.

Site-specific recombination occurs as the name suggests at specific chromosomal sites and has the function of integrating DNA from non-homologous chromosomes, such as assembling genes in specific cells. Recombination by transposition inserts a fragment of DNA into a new chromosomal location (5).

6. Chromosome Structure and Nomenclature

The human autosomes are each referred to by the numbers 1–22. The chromosomes are numbered in decreasing size with the exception of chromosome 21 which is slightly smaller than chromosome 22. While all DNA is made up of sequences of nucleic acids, some parts of the chromosome contain highly repetitive sequences giving the DNA molecule-specific structural properties (6). All of the nuclear DNA chromosomes have a centromere and two telomeres (each end of the chromosome). Eukaryal telomeres are complex structures protecting the chromosome from degradation. The telomeres may also assist in sister-chromatid pairing during meiosis.

Chromosomes can be studied under the microscope (cytogenetics) using staining techniques to identify regions on the chromosome. Each chromosome is partitioned into a p (short) arm and a q (long) arm by the location of the centromere. These two arms can be further subdivided by consideration of other landmark positions, for which a suite of banding techniques has been developed. The International System for Human Cytogenetic Nomenclature (ISCN) laid out the basic terminology for labelling banded chromosomes. Each region of each arm is partitioned into regions. For example, the p arm of chromosome 1 contains three distinct regions and each region may be further partitioned into subunits with the specific region labelled, for example, 1p31.2. The bands are labelled out from the centromere towards the telomeres. The terms, distal and proximal, describe the location relative to the centromere. For example, "distal 3p" means the portion of chromosome 3p nearest to the telomere, whereas "proximal 3p" refers to the portion closest to the centromere.

The physical manner in which the DNA is stored in the nucleus changes depending upon the state of the cell. When cells prepare to undergo cell division, the chromosomes link up in homologous pairs and appear packaged into very highly condensed units. The DNA double helix is capable of several levels of coiling, which require nucleosomes for the DNA to be wound around. The nucleosome consists of a core of eight histone proteins. Double-stranded DNA is wound around the nucleosomes, giving the DNA a "string of beads" appearance. The string of beads, which is around 10 nm in diameter is then itself coiled into a chromatin fibre of 30 nm in diameter.

7. Mutation

Though all adult cells descend from one single zygote cell, the many rounds of replication and cell division inevitably result in some changes (*mutations*) remaining undetected and hence unrepaired. The mutation may or may not have an impact on cellular function. This depends upon both the site and form of the mutation and the cell type in which it occurred. Of course, if the mutation occurs in the germ line, the mutation may be passed on to the next generation. The errors or mutations may be introduced during replication or be due to exposure to DNA damaging agents. An alteration in a single base of DNA may result in a codon specifying a different amino acid and hence result in a potential functional change in the resulting protein. Even with the extensive DNA repair machinery a small number of mutations do arise and they may become established. If a mutation has no functional consequences, it is termed neutral.

During DNA replication, nucleotide mispairing occurs approximately every 1,000 bp. However, the DNA repair mechanisms work within the replication process to ensure very low levels of mismatching (around 10^{-10}). Even this low rate of mutation leads to a significant number of accumulated mutations when considering all the cells in the adult organism.

Genetic mutation can take several forms, resulting in either a sequence state change or a change in the length of the sequence. A sequence change results where one or more base pairs have changed state compared to the original sequence. The other class of mutation is a change in length of the piece of DNA sequence, loss or gain of some material. The simplest form of a change in state is a single nucleotide change, for example, the base A is replaced by C at the equivalent sequence position. If instead the base A is lost (or duplicated) in the replicated DNA, this would be classed a deletion (or insertion).

Some sites within DNA are more prone to mutation, these sites include satellite DNA. The satellite DNA consists of highly repetitive sequences which are important for chromosome structure and function. The microsatellite class have been very frequently used as genetic markers due to their presence throughout the genome and polymorphic nature of the variants. Microsatellites typically consist of dinucleotide repeats (such as CA or TA). At the respective genomic site, there are then variable numbers of these repeats seen in individuals. When a repeat is copied with error, the descendent cell may receive more (inserted) or less (deleted) copies of this tandem repeat.

On a larger scale, a length of sequence may be cut from one genomic site and inserted into another (translocation). An inversion results when a length of sequence appears to cut and rotate but pasted at the same original site.

8. Genetic Polymorphism and Genetic Markers

The term mutation is frequently reserved for describing the direct result of a change, or the change itself. Hence, if a change in sequence is seen between, say, a normal cell and a cancer cell within an individual, it will be described as a mutation. Similarly, if an offspring has presented an allele not present in either parent, then a mutation is said to have occurred. Once a mutation has occurred and is potentially transmissible to offspring in it is termed a genetic variant. Variants with a population frequency higher than 1% are classed as polymorphisms; below this threshold they are termed rare variants. If a variant has no apparent functional consequences, it is therefore potentially useful as a neutral genetic marker. This means it can be used to effectively “mark” the state of the DNA at a specific location (i.e. where the variant resides in the genome) and thereby enables DNA molecules from different origins to be studied for similarities and differences. In order for a marker to be useful in this context, it needs to be highly polymorphic, that is to have many frequent alleles, resulting in a high probability that a randomly studied individual will be heterozygous at this marker. In the 1980s and early 1990s, microsatellite DNA markers were very popular in human genetic linkage studies as they were individually highly polymorphic and evenly spaced throughout the genome. However, their highly polymorphic nature was also associated with technical difficulties.

Over the last decade, there has been a general shift towards using the individually less informative single nucleotide polymorphism (SNP) markers. These variants generally have only two alleles, but due to their high density in the genome and the fact that SNP variants are technically easy to genotype using high throughput technology these are now routinely employed.

Structural genetic variation is used to describe variants where a large length of DNA sequence (typically greater than 1,000 bp) is duplicated or deleted (7). The extent of this variation in human genomes is still being characterised, and this form of genetic variation is considered in the later Chapter 13.

9. Methylation and Demethylation of DNA

Factors that are heritable (transmitted from a parent to offspring, or from cell to daughter cell) but are not seen at the DNA sequence level are termed epigenetic. The most well studied of these factors is DNA methylation (see Chapter 14). In mammals, methylation occurs preferentially at C residues. Cytosine bases that occur immediately upstream (i.e. 5′) of guanine residues are frequently methylated. The effect of this is to make the chromatin functionally inactive at these methylated sites. Methylation is an important mechanism for functional regulation. Studies of gene expression observe differential allelic expression, assumed to arise due to differential levels of allelic methylation (8). Imprinting relies on a form of methylation where the allelic effects on phenotype depend upon whether the allele was inherited from the mother or the father. The quantitative extent of methylation is currently quite difficult to measure with accuracy.

References

1. Gingeras, T.R. (2007) Origin of phenotypes: genes and transcripts. *Genome Res* **17**, 682–690.
2. Brent, M.R. (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet*, **9**, 62–73.
3. Mercer, T.R., Dinger, M.E., Mattick, J.S. (2009) Long non-coding RNAs: insights into functions. *Nat Rev Genet*, **10**, 155–159.
4. Strachan, T., Read, A.P. (eds.) (2004) Human Molecular Genetics, 3rd edn. Garland Science, New York.
5. Ringo, J. (ed.) (2004) Molecular events of recombination, in *Fundamental Genetics*. Cambridge University Press, Cambridge, pp 124–135.
6. Ringo, J. (ed.) (2004) Chromosomes in Eukarya, in *Fundamental Genetics*. Cambridge University Press, Cambridge, pp 34–42.
7. Feuk, L., Carson, A.R., Schere, S.W. (2006) Structural variation in the human genome. *Nat Rev Genet* **7**, 85–97.
8. Cheung, V.G., Spielman, R.S. (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression *Nat Rev Genet* **10**, 595–609.

Chapter 2

Terminology, Concepts, and Models in Genetic Epidemiology

M. Dawn Teare and Mauro F. Santibàñez Koref

Abstract

Genetic epidemiology brings together approaches and techniques developed in mathematical genetics and statistics, medical genetics, quantitative genetics, and epidemiology. In the 1980s, the focus was on the mapping and identification of genes where defects had large effects at the individual level. More recently, statistical and experimental advances have made possible to identify and characterise genes associated with small effects at the individual level. In this chapter, we provide a brief outline of the models, concepts, and terminology used in genetic epidemiology.

Key words: Population genetics, Mendelian segregation, Kinship, Identity by descent, Genetic components of variance

1. Introduction

Genetic epidemiology studies the influence of genes and environment on measures of health and disease susceptibility in populations. This discipline emerged relatively recently and brings together established methodologies arising from population genetics, quantitative genetics, medical genetics, and epidemiology. Much of the terminology currently used was conceived when little was known about the molecular mechanisms mediating inheritance (1). The term *gene* is now frequently used to refer to a functional segment of DNA, which is transcribed into RNA and may code for a protein. However, within the field of population genetics "gene" continues to be used in its original meaning, and refers to the basic unit of heredity. As with any speciality the terminology has become specialised, and this in itself can form a potential barrier to newcomers. The purpose of this chapter is to present the basic terminology and outline the basic models used in genetic epidemiology.

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_2, © Springer Science+Business Media, LLC 2011

2. Mendelian Genetics and Modes of Inheritance

Gregor Mendel was the first to propose a discrete model to explain the inheritance of genetic factors and their impact upon an organism's *phenotype* (2). By phenotype we mean an individual's measurable characteristics or *traits.* When Mendel reported the results of his experiments on pea plants in 1865, he focused his attention on qualitative phenotypes of his plants, such as pea seed coat shape (wrinkled vs. round) or flower colour (white vs. violet). He postulated a mechanism of inheritance in which each organism carries two factors that determine together the organisms' phenotype and that an adult organism can only transmit one of the two factors to each of its offspring.

With respect to the phenotype of *wrinkled seed coat* or *round seed coat*, he labelled the two possible factors r and w. For this phenotype, he proposed that the factor r was dominant to w (or conversely w was recessive to r). This means that the recessive phenotype of "wrinkled" is seen only in peas with two copies of the recessive factor w. Conversely, those pea plants with one or two copies of the dominant factor r would express the dominant phenotype of round seeds. We would now refer to these factors r and w as *alleles* of the gene determining the variation in seed coat shape. The physical location of this gene in the pea genome is referred to as the *locus* (plural *loci*). There are examples of alleles that are *codominant*, where in individuals carrying two different alleles the phenotypes characteristic for both alleles are displayed. For example, the ABO human blood group system has three classes of alleles, A, B, and O. The allele O is recessive with respect to A or B, but A and B are codominant, this gives rise to a four phenotype system (namely, blood groups A, B, AB, and O).

Mendel's extensive experiments on peas led him to propose two laws: the law of segregation and the law of independent assortment. The law of segregation stipulates that when an organism produces gametes the two copies of the gene separate so that each gamete randomly receives one allele. The law of independent assortment states that alleles at different loci are inherited independently from alleles at other loci, or that alleles of different genes segregate independently during gamete formation. We now know that genes are physically linked to other genes due to their location on chromosomes. Mendel happened to study traits arising from unlinked loci. The genes determining his seven phenotypes are each located on a different chromosome. Though this was true for the traits Mendel studied, the law of independent assortment is generally true when loci are not genetically linked. We now use the term *Mendelian segregation* to describe the pattern of allele transmission from one generation to the next, meaning that the probability of a parent transmitting one specific allele to one specific offspring is 50%.

The pair of alleles found at a single locus in a diploid organism is referred to as the *genotype*. At a multi-allelic locus, such as ABO described above, there are six possible genotypes (AA, AB, AO, BB, BO, OO). When the two alleles are of the same type (e.g. AA) the individual is said to be *homozygous* at that locus. When both alleles are not the same, the individual is described as *heterozygous* at that locus. Loci are also defined using descriptors of their position in the genome (see Chapter 4).

Mendel's model allows us to introduce the concept of a *penetrance function*. The penetrance function is used in both discrete and quantitative genetics, in the discrete setting it is the probability of having the trait or phenotype state of interest conditional on the genotype. For example, for Mendel's seed shape experiment we discussed earlier, there are three possible genotypes labelled ww, rw, and rr. If we define *round* as the normal or common state and *wrinkled* as the phenotype of interest, then the penetrance function is defined by the three conditional probabilities:

1. Prob(wrinkled/rr) = 0
2. Prob(wrinkled/rw) = 0
3. Prob(wrinkled/ww) = 1

Here, the *mode of inheritance* for the phenotype *wrinkled* is recessive.

3. Population Genetics

Sexual reproduction is a mechanism by which the genetic units are transmitted from one generation to the next. Mendel's model came from his experiments on peas. The diploid system that he discovered determines the distribution or patterns of phenotypes in a population. To illustrate these patterns, let us consider a locus that has two types of alleles which we designate as alleles of type A and of type B. The population relative frequency of the alleles of type A (termed *gene frequency* or *allele frequency*) is denoted as p. As there are only two types of alleles at this locus, and on a single chromosome the allelic state must be A or B, the frequency of alleles of type B is $(1 - p)$. It is important to remember that the population gene frequency refers to the population of chromosomes and not diploid organisms, whereas the term genotype refers to the type of the pair of alleles found at the locus in a single (diploid) organism. In large populations, when alleles are inherited independently, the expected frequencies of the genotypes in each generation are a simple function of the allele frequency, and they do not vary from one generation to the next. When such a state exists, the locus is said to be in Hardy–Weinberg Equilibrium

(HWE). Under these circumstances, the expected *genotype frequencies* can be derived from a binomial distribution, where the probability of success is p and the number of trials is two. The three genotypes AA, AB, and BB should be seen in the frequencies p^2, $2p(1-p)$, and $(1-p)^2$.

The relation above (which can be extended to multi-allelic systems) holds under *random mating*, when alleles are inherited independently, and in the absence of *selection* or *mutation*. Random mating or *panmixia* means that sexual partners randomly select their partner, i.e. without reference to their genotypic state, their degree of relatedness or physical proximity. When sexual partners do exhibit some preference or selection, this is termed *non-random mating*. Mutation is the mechanism through which new alleles arise or one allele may change from one type to another.

HWE also implies that the gene frequency remains constant from one generation to the next. For this to be the case, all genotypes must be equivalent with respect to *viability* or *fitness* of the organism. Viability can be thought of as an individual's probability of survival or the fraction of the population surviving to reaching maturity. If organisms of one genotype have an advantage over another genotype (e.g. a better chance of survival to reproductive age), then that genotype group will be over represented in the parents of the next generation. The presence of HWE is often used to confirm that a locus is *neutral* (no variation in fitness associated with genotype variation), and hence may be useful as a genetic marker.

3.1. Effects of Population Size

The previous section requires a large population for these general properties to hold. However, all large populations must have gone through a small population phase at some time in their history. When the population is small, HWE may not hold even in the presence of random mating and equal viability. This is due to the random sampling of the gametes, and results in the population gene frequency varying from one generation to the next (*random genetic drift*). As the population size increases, the effect of genetic drift reduces and once the population becomes sufficiently large the gene frequency becomes effectively stabilised.

When a new population is established from a small number of founder individuals, the *founder effect* means the descendent population's genetic variation is limited by the genetic diversity of the *founder population*. When isolated populations remain at small numbers over several generations (through disaster or migration to new territory), this is described as a *bottleneck*, and a significant amount of genetic variation can be lost from the gene pool. The founder effect can be responsible for the different genetic profile in neutral markers seen between populations (3), and also is responsible for some of the different rates of genetic disease seen when comparing isolated populations (4).

3.2. Linkage Disequilibrium

When considering more than one genetic locus, alleles tend to be co-transmitted to the same gamete when they are located physically close on the same chromosome. Genetic linkage between loci is generally a consequence of their being located on the same chromosome. When the segregation of alleles is followed through the generations in a pedigree, the allelic states tend to be co-inherited as the loci are "linked" by both occurring on the same chromosome, forming a *haplotype* (see Chapters 4 and 5).

The term *linkage disequilibrium* (LD) is used in a slightly different context. Rather than relating to the probability of an exchange of information at meiosis, LD is observed at the population level. LD is a general term which exists when allelic association is seen between two loci. Sometimes, this is referred to as non-random association of alleles. LD can arise though several mechanisms: by chance in small populations, by new mutation and or selection, or by intermixture of previously isolated populations (5).

Such association arises when alleles at distinct loci are found together in gametic phase (the alleles originate from the same gamete) at frequencies different to those expected based on the allele frequencies alone. The existence of LD does not necessarily imply that the loci are "linked", i.e. are in close proximity on a chromosome, however, when two loci are in close physical proximity, LD implies that the population frequency of the two locus haplotypes are not as expected based on the allele frequencies. For example, consider two genetic loci with alleles labelled A and B at locus 1 and C and D at locus 2. At the population level, these alleles occur at the following frequencies A: 30%, B: 70%; C: 40%, D: 60%. While these loci may be "linked" and hence the probability of recombination between them at meiosis may be less than 0.5, this "linkage" is not seen at the population level. After many generations, you would expect the alleles to be randomly associated with haplotypes occurring at frequencies dictated by the product of their population allele frequencies AC: 12%, AD: 18%; BC: 28%, BD: 42%. If the population haplotype frequencies differ from these expected numbers, the loci are said to be in LD. The extent of LD is quantified by the disequilibrium parameter D (6). LD is discussed further in Chapter 6.

3.3. Kinship, Gene Identity by Descent and Inbreeding

Mendel's laws imply certain patterns of allele sharing between pairs of relatives. For example, consider the four alleles found in two siblings at one specific locus, we would expect the siblings to share alleles inherited from their shared or common ancestors, the probability that they would share 0, 1, or 2 alleles inherited in common is 0.25, 0.5, and 0.25, respectively. If we consider half siblings, they are equally likely to share exactly 0 or 1 allele through their common parent, but they cannot share both alleles. In this example, we are considering the probability that they share an allele inherited from a common ancestor, these alleles are said

Table 1
Kinship coefficients and IBD sharing probabilities for relative pairs

Relative pair	ψ	(k_2, k_1, k_0)
Full siblings	1/4	(1/4, 1/2,1/4)
Half siblings	1/8	(0, 1/2, 1/2)
Monozygous twins	1/2	(1, 0, 0)
Parent offspring	1/4	(0, 1, 0)
Cousins	1/16	(0, 1/4, 3/4)

to be *identical by descent* (IBD) (each allele is a descended copy from a common ancestor). Two alleles may be identical but have not been inherited from a recent common ancestor. In this case, the alleles are said to be *identical by state*. The relationship between a pair of individuals, labelled X_1 and X_2, can be summarised by the *coefficient of kinship* $\psi(X_1, X_2)$ (7). This coefficient is defined as the probability that an allele randomly sampled from X_1 and an allele randomly sampled from X_2 at the same locus are identical by descent. The *coefficient of inbreeding* $\alpha(X)$ for a single individual X, is defined as the probability that the pair of alleles that constitute the genotype of individual X at an arbitrary locus are IBD. The inbreeding coefficient for individual X is equal to the kinship coefficient for the parents of X. The *gene identity states* comprise the possible IBD sharing patterns for a pair of individuals. In the absence of inbreeding, pairs can share 2, 1, or 0 alleles IBD as argued above. The expected IBD sharing probabilities for each of these states are reported as a vector $\mathbf{k} = (k_2, k_1, k_0)$. Pairs of individuals with the same kinship coefficient do not necessarily have the same k vectors. Table 1 lists some kinship coefficients and IBD sharing probabilities. It is interesting to note, though obvious from Mendelian segregation, that although parents and offspring have the same expected kinship as full siblings, the parent offspring pairs always share exactly one allele IBD.

4. Quantitative Genetics

The terms phenotype and trait are often used interchangeably, however, trait is commonly used in the quantitative context, and phenotype in the qualitative context. A state of health such as diagnosis of diabetes (phenotype is "affected with diabetes") is often the result of consideration of a single quantitative trait, such as blood glucose levels; if the level of the trait is above a specified threshold, the individual is classed as affected. Most clinical conditions fall into the discrete or qualitative phenotype, though the diagnosis may reflect the presence of an extreme value for the underlying quantitative trait, such as the relationship between body mass index (BMI) and obesity.

The terminology introduced so far has focussed on the influence of genes on discrete or qualitative phenotypes. In the late nineteenth century, Francis Galton first used the term "regression" when describing the correlation he observed between traits measured in parents and offspring such as height (8). Galton's work laid the foundations for later researchers who made inferences about genetic models or trait inheritance by applying statistical methods to observations on pairs of relatives.

4.1. Components of Variance Models

While quantitative and Mendelian genetics use the same principles regarding the inheritance of genes, in the former the penetrance function (the relationship between genotype and phenotype) links a discrete with a continuous variable. A normally distributed quantitative trait can be summarily described by its mean and variance. Quantitative genetics models assume that genetic variation contributes to phenotype variation. Hence, the quantitative phenotype observed in an individual, *the phenotypic value*, can be thought of as made up of several components, one of which may be due to genes and another due to environment. This allows us to decompose the trait value seen in an individual into a linear expression.

$$\Upsilon_{\mathrm{g,c}} = \mu + G_{\mathrm{g}} + E_{\mathrm{c}}$$

where $\Upsilon_{\mathrm{g,c}}$ represents the phenotype value observed in a person with genotype g and environment c. The genetic (G_{g}) and environmental (E_{c}) contributions are generally represented as deviations from a population mean μ (3). By breaking down the phenotype into these components and using Mendelian segregation to derive the expected IBD sharing between pairs of relatives, we can model the "correlation" between pairs of relatives. The model can also be used to predict trait values but is more usually used to assess the evidence for a genetic component. In the simplest form, we might assume that the genetic contribution to the trait is due to a single locus with two alleles. A locus with two alleles, A and A′, has three associated genotypes AA, AA′, and A′A′. Each of these genotypes makes a specific contribution to the trait value. However, the contribution is rescaled so that the origin is at the value mid way between the two homozygote (AA and A′A′) values. If the alleles act in a simple *additive* fashion, the heterozygote value is exactly the mid-point between the two homozygote values. If there is an *interaction* between alleles at the same locus (*dominance*), then the value associated with the heterozygote, *d*, will deviate from this

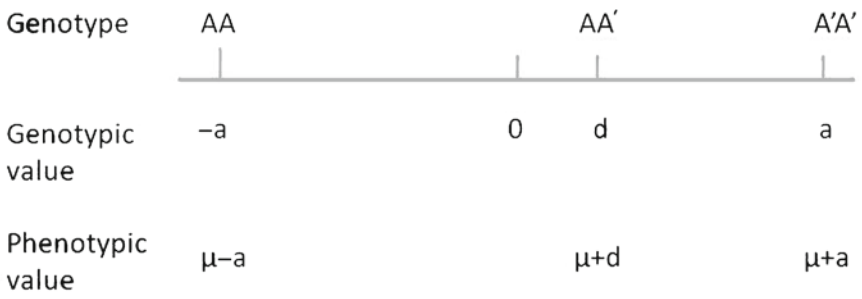


Fig. 1. Illustrating the relationship between the genotypes, the genotypic values and the phenotypic values. In this linear model, the impact of the genotype on the quantitative trait is described in terms of three parameters, μ, a, and d.

mid-point. According to our new scale illustrated on Fig. 1, the homozygote value ranges from $-a$ to $+a$. If $d=0$, we say there is no dominance, the alleles are codominant, or act additively. If $d=-a$, then A′ is recessive to A, if $d=a$, then A′ is dominant to A. If d is greater than $+a$ or less than $-a$, then we have *overdominance*. The degree of dominance is sometimes reported as d/a. This model can be extended to allow for multilocus genotypes, where each locus contributes additive and dominance effects.

We can extend this notation to multiple loci, say we have three loci with alleles A and A′, B and B′, and C and C′ adding a suffix to indicate the source of the genotype effects. The genotypic value associated with the *compound genotype* AA′, BB, and C′C′ would be $d_A - a_B + a_C$. This assumes no interaction between alleles at different loci. Interaction between genotypes at distinct loci is termed *epistasis*. When studying the correlation between pairs of relatives in a pedigree, it is important to remember that it is the allele that is transmitted from one generation to the next and not the genotype values directly. For this reason, “breeding values” are sometimes used when referring to the genotypic values of parents (3). The breeding value is the additive genotypic value, as the dominance effect arises only in the individual who receives the interacting alleles; it is not transmitted directly (though covariance due to dominance effects can be seen in some relative pairs).

This model describing the relationship between phenotype and genetic factors gives rise to a *variance components* framework. The total (or population) trait variance is made up of variance components attributable to the genetic component and the environmental component.

$$\sigma_p^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_c^2 + \sigma_e^2$$

The terms above for each component of population variance (σ_p^2) are defined as the *additive variance* (σ_A^2), *dominance variance*

(σ_D^2), *common environment variance* (σ_c^2) and random non attributed variance (σ_e^2). *Epistatic variance* (σ_I^2) or the variance attributable to interaction between loci, is included in the expression above for completeness but is very difficult to characterise in practice. The expression above assumes that there is no interaction between the environment and genotype. The *genetic variance* is the sum of all the genetic components ($\sigma_A^2 + \sigma_D^2 + \sigma_I^2$). While we have stated the model in terms of these components, these components cannot be identified by sampling from a population, unless relative pairs are studied. The expected sharing of alleles between pairs of relatives enables inferences to be made on the components of genetic variance. We can write down the expected covariance for pairs of relatives (see Table 2).

While Mendelian segregation dictates how the alleles are shared among relatives, the degree of shared environment is more open to discussion. In Table 2, you can see that only full sibs are assumed to share a common sibling environment. This type of shared environment is commonly assumed, but other models can be proposed depending upon the characteristic of interest (9).

We have presented a framework where a trait value is made up of contributions from many sources. The genetic component may arise from additive effects of alleles, an interaction between alleles at the same locus and interaction between genotypes at different loci. Similarly, the influence of the environment can be dissected in more detail. Particularly, if we want to know how much of the correlation in relatives is due to shared environment.

Table 2
Expected co-variances between relative pairs

Relative pair	Expected covariance
Full siblings	$\frac{1}{2}\sigma_A{}^2 + \frac{1}{4}\sigma_D{}^2 + \sigma_c{}^2$
Half siblings	$\frac{1}{4}\sigma_A{}^2$
Monozygous twins	$\sigma_A{}^2 + \sigma_D{}^2 + \sigma_c{}^2$
Parent offspring	$\frac{1}{2}\sigma_A{}^2$
Cousins	$\frac{1}{8}\sigma_A{}^2$

4.2. Heritability

The magnitude of the genetic contribution is frequently summarised as *heritability*. Heritability is defined as the proportion of the trait variance that is attributable to genetic variation. It is therefore the ratio of the genetic variance compared to the total variance. *Heritability in the broad sense* (H^2) includes additive, dominant, and epistatic effects. Heritability in the narrow sense (h^2) restricts attention to additive effects only. Given these definitions heritability must always lie between 0 and 1. Values close to 1 suggest a strong genetic component with most trait variation due to genetic variation. Conversely, values close to zero suggest that genetic variation only weakly contributes to trait variation. Caution needs to be used when comparing heritability estimates from different populations as the heritability is defined relative to the population phenotypic variation and hence is population specific.

4.3. Twin Studies

In human genetics, twin studies are commonly used to establish and identify the strength of a genetic component. This study design uses the variance component framework and is frequently used to estimate heritability. Monozygous (MZ) twins are genetically identical, whereas dizygous (DZ) twins can be thought of as age-matched siblings. So the classic twin study design (contrasting covariances between MZ and DZ twins) offers a means to estimate the components of genetic variance. If the MZ and DZ correlation are similar, then this would be evidence that any genetic component is weak. However, if the MZ correlation is greater than the DZ correlation, this is evidence for a genetic component. As can be seen from the table the co-variance is a function of three parameters, but there are only two equations to link the observations and the model. Hence, only two of the three components of variance can be estimated and investigators can only report if the evidence for a shared environmental component is stronger than the evidence for a dominance component (10). If the basic twin design can be extended to include observations on other relatives, such as additional siblings or parents, more specific components of variance can be modelled and potentially estimated (see Chapter 11).

One extension to the twin study is to compare co-variances between twins reared together and those reared apart (*adoption* studies). This design allows the estimation of both the dominance and shared environment component. A common criticism of the twin study design is the validity of the key assumption that the shared environment is equivalent for MZ and DZ twins. This may not be valid for the analysis of behavioural traits as MZ twins, and DZ twins can have socially very different experiences (10).

4.4. Major Genes and Polygenes

When a single gene has a strong influence on a trait, i.e. a large a, then this gene is called a *major gene* and the allele-specific effects of the gene can be identified, through the model outlined above. If, however, many genes are involved, it becomes difficult to isolate the allele-specific effects, and it is more common to then assume that the several (unlinked) genes involved all have a small but equivalent effect. As we allow more loci to contribute to the variation, the individual allelic effect must reduce. If we then further assume that all the alleles at these unlinked loci have equivalent and only additive effects, then the distribution of the compound genotypic values will approach the normal distribution. This leads to the *polygenic model*, the joint effect of an infinitely large number of loci results in *polygenic values* distributed about a mean of zero and variance σ^2_{PG}.

The *mixed model* (11) allows for both a major gene and a polygenic effect, assuming no interaction between these two components. Hence, the variation in a phenotype can be attributable to a major gene effect, a polygenic effect and environmental effects. It is important to note here that the source of the environmental sharing is not directly measured but is often assumed due to familial factors.

5. Familial Aggregation, Segregation Analysis, and Qualitative Traits

The framework described in Section 4 relates to the variation in quantitative phenotypes, which lend themselves naturally to a variance component model. However, the same approach can be used to make inferences about binary traits with one extension to the framework. Instead of assuming that the model predicts phenotype, we allow the model to predict an underlying latent variable *liability*. The link between the model and our binary phenotype is established by defining a *threshold*. If an individual's liability value exceeds a threshold, the individual becomes affected with the disease. This extension enables the calculation or estimation of risk of disease or penetrance function. In some variable age at onset models a log-normal distribution of risk is assumed rather than the liability threshold (12).

Approaches to identify the genetic component for binary phenotypes frequently take a different form than for quantitative traits. If a major gene is suspected, the genotype-specific penetrance estimates will be reported along with an estimate of disease allele frequency. These models can be fitted without the need for a variance component framework. Often families have been selected due to the presence of at least one relative with the disease or phenotype of interest. This individual is designated as the *proband*. Proband-based sampling is common when a disease is rare, and it is expensive to study and record information on families who have no cases of disease occurring. The manner in which such families are identified and the members of the family studied is called the *ascertainment scheme*. This biased sampling scheme, resulting in an oversampling of affected individuals, needs to be taken account of in any subsequent analysis so that statistical inferences are not biased. Further constraints can be applied in segregation analysis to ensure that the model predicts incidence or prevalence rates consistent with population data.

Simple Mendelian traits or simple genetic models imply that one genotype determines one phenotype, such as the dominant and recessive examples above for Mendel's peas. A deviation from this simple one to one correspondence is termed "complex". A disorder is called a *single gene disorder* when it only arises when mutations occur in a specific gene. However, if the probability of being affected with the disease conditional on the risk genotype is less than 1, then the term *incomplete penetrance* is used. Cystic Fibrosis (CF) is an example of a single recessive gene disorder with variable severity of phenotype and showing extensive *allelic heterogeneity*. Over 1,000 distinct mutations (alleles) in the CFTR gene have been described, and the clinical phenotype varies from severe when detected soon after birth, to mild and clinically undetectable until well into adulthood. The term *locus heterogeneity* is used when several genes can each independently give rise to the same phenotype. In qualitative phenotype analysis, the term *sporadic case* or *phenocopy* is used to indicate an affected individual whose phenotype has arisen due to an environmental cause and not the genetic predisposition. When the model permits phenocopies and incomplete penetrance, both phenotypes (e.g. affected and unaffected) are possible for all genotypes, hence all penetrance probabilities are greater than 0.

Rather like the study of correlations between pairs of relatives, observed familial aggregation of binary phenotypes is often reported as the *familial relative risk* or *familial recurrence ratios* (FRR) (13). These are simply defined as the risk of disease in relatives of a case compared to the risk in the general population. The FRR can be reported for all relatives within kinship groups, such as first degree relatives, or by the specific form of the relationship, for example, sibling. Though a genetic model gives rise to predictable patterns of FRR, the FRR merely summarises the pattern of risk and does not necessarily imply a genetic cause to a correlation in risk. The FRR are often referred to as the "lambda" risks (Greek letter λ), with a subscript indicating which relative of the case is considered. Commonly considered relative types are the sibling (λ_s), parent (λ_p), and offspring (λ_0).

6. Prospects for Phenotype Studies

The use of segregation analysis and the variance components approaches rely only on measuring the phenotypes in relatives. Large extended pedigrees or observations on many different types of relative pair enables the exploration of more complex models than those outlined above. However, when only phenotype data is available, these models lack the power to distinguish between common genes with low penetrances and the polygenic components. Advances in both molecular genetics and statistical computing are now making it feasible to identify and characterise locus-specific effects, by incorporating measured genotypes into the analysis. It is the identification and characterisation of the environmental components that present the next major challenge to the field.

References

1. Gingeras, T.R. (2007) Origin of phenotypes: genes and transcripts. *Genome Res.* 17, pp 682–690.
2. Mendel, G. (2008) Experiments in Plant Hybridisation, in *Ending the Mendel-Fisher Controversy.* (Franklin, A., Edwards, A.W.F., Fairbanks, D.J., Hartl, D.L., Seidenfeld, T. eds.), University of Pittsburgh Press, Pittsburgh, PA. pp 79–116.
3. Falconer, D.S., Mackay, T.F.C. (eds.) (1996) *Introduction to Quantitative Genetics*, Fourth Edition. Longman, Essex, UK.
4. Conrad, D.F., Pritchard, J.K. (2007) Population genetics and disease, in *Genes and Common Diseases.* (Wright, A., Hastie, N. eds.), Cambridge University Press, Cambridge. pp 44–58.
5. Cordell, H.J., Clayton, D.G. (2005) Genetic association studies. *Lancet* 366, pp 1121–1131.
6. Weir, B.S. (ed.) (1996) *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.
7. Cannings, C., Thompson, E.A. (eds.) (1981) *Genealogical and Genetic Structure.* Cambridge University Press, Cambridge.
8. Galton, F. (1886) Regression towards mediocrity in hereditary stature. *J Anthr Inst.* 15, pp 246–63.
9. Lynch, M., Walsh, B. (eds.) (1998) *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.
10. Sham, P. (ed.) (1998) *Statistics in Human Genetics.* Wiley, New York.
11. Morton, N.E., MacLean, C.J. (1974) Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. *Am J Hum Genet.* 26, pp 489–503.
12. Pharoah, P.D.P., Antoniou, A.C., Easton, D.F., Ponder, B.A.J. (2008) Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med.* 358, pp 2796–2803.
13. Risch, N. (1990) Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet.* 46, pp 222–228.

Chapter 3

An Introduction to Epidemiology

Cother Hajat

Abstract

Epidemiology as defined by Last is “the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to the prevention and control of health problems”. Traditional epidemiological studies include quantitative and qualitative study designs. Quantitative study designs include observational and interventional methodology. Observational methods describe associations that are already present at population (descriptive) or individual (analytical) level. Although they form the mainstay of epidemiological studies, observational methods are prone to bias and confounding. These can be dealt with by various means involving both the study design and statistical analysis. Interventional methods involve changing variables in one or more groups of people and comparing outcomes between those with the changed and unchanged variable. Interventional studies can more readily account for bias (such as through randomisation) and confounding (such as through controlling) as is seen in randomised, controlled trials. Qualitative studies employ non-numeric methods to obtain “richer” information on how people perceive or experience situations. Much of epidemiology and epidemiological methods have been stable for many years. There are, however, emerging issues in epidemiology, including those of causal inference, counterfactuals and Mendelian randomisation, among others. There are also several modern and emerging uses of traditional epidemiological techniques in the fields of infectious disease, environmental, molecular and genetic epidemiology.

Key words: Observational studies, Interventional studies, Bias, Confounding, Emerging epidemiology

1. Traditional Epidemiology

Epidemiology was described by Last in 1995 as “the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to the prevention and control of health problems” (1). Epidemiological studies are usually focused on a geographical unit, such as a section of the population and are one of the major tools employed in improving public health. The scope of epidemiology is large and includes both interventional and observational studies.

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_3, © Springer Science+Business Media, LLC 2011

1.1. Interventional Study Designs

Interventional studies involve changing a variable in one or more groups of people and then comparing outcomes between those with the changed and unchanged variable. The commonest form of interventional study used in epidemiology is that of the randomised controlled trial although other forms of trial are also commonly found. Randomised controlled trials are not covered in detail here.

1.2. Observational Study Designs

The mainstay of epidemiology is the observational study design, which may be descriptive or analytical.

1.2.1. Descriptive Studies

Descriptive studies describe patterns of disease occurrence within a population and can include study designs, such as ecological (also termed correlational) studies, cross-sectional surveys, case-reports and case-series. As the data used are often routinely collected, these are the least time and resource intensive study designs. Although commonly used, it is difficult to convincingly demonstrate causation between an exposure and outcome using descriptive study designs. Also, as the findings are at population level, they do not necessarily reflect what is true at the level of the individual and may lead to the phenomenon of "ecological fallacy", a form of bias that is the major limitation of descriptive studies. It is also usually not possible to control for other types of bias or confounding in descriptive studies. For all of these reasons, descriptive studies are often used as the first step in generating a hypothesis for further testing using analytical or interventional study designs.

1.2.2. Analytical Studies

Analytical studies analyse the relationship between health status and other variables, typically risk factors for disease causation, at the individual level within a section of the population. It is easier to make causal inferences from analytical, compared with descriptive, studies as the study can be controlled by varying degrees for bias and confounding. The most frequently used designs are those of case–control and cohort studies.

2. Case–Control Studies

Case–control studies are a type of longitudinal study in which subjects are selected on the basis of their disease status relating to the disease of interest. The two groups (cases with disease and controls without disease) are compared for exposure to the characteristic of interest.

2.1. Case Selection

Cases should ideally be representative of cases in the source population. If all cases are not to be included, then a random sample of cases should be chosen, for example, those attending a disease clinic. However, the latter is likely to introduce some form of bias in the method of recruitment of cases if they are no longer representative of all cases in the source population. For example, clinic attendees for asthma may have a more severe form of the disease than those who do not attend asthma clinics, thereby introducing a bias in terms of severity of disease.

2.2. Control Selection

Controls should ideally also be from the same (source) populations as for cases and be as similar as possible in all other ways, other than the disease status, to the cases. The study is said to be population based if controls are sampled directly from the source population and both cases and controls are representative of those with and without disease in the source population. Commonly used sources for selection of controls include:

- hospital or clinic attendees
- friends/spouses
- randomly derived individuals from a GP register
- random household sampling

It is usual for there to be one control per case. However, if the control group chosen are not sufficiently comparable to the cases then up to four groups of controls, often from different sources or with different attributes, may be chosen. For example, if the cases are selected from inpatients with coronary heart disease and controls from inpatients with lung cancer, there may be something inherent in the selection method (both are inpatients) that link them with the exposure of interest. In this case, other types of inpatients, or attendees at a hospital clinic, may provide a broader range of controls that are more comparable. A ratio of up to four to six controls per case may also be used in order to increase the power of the study. Controls may or may not be matched to cases but the results should confirm that the groups are similar in aspects that are pertinent to the research question, such as age, sex, and other key factors.

Case–control studies are often termed “retrospective” because the investigator looks back in time after the disease state is known to establish past exposures. However, case–control studies can actually be either prospective or retrospective. In a prospective case–control study, the data collection continues prospectively, whereas a retrospective case–control study deals only with exposures and outcomes that have already taken place.

2.3. Analyses in Case–Control Studies

The analyses in case–control studies compare the frequencies of exposure to a particular “risk factor” (or risk factors) between those who have the disease outcome (cases) and those who do not (controls). The association between exposure and outcome is estimated using an odds ratio (OR), with relevant confidence intervals which estimates the odds of exposure among cases compared with the odds of exposure amongst controls. The estimates obtained can only, therefore, provide a picture of relative risk and not of absolute risk from that risk factor in the population. However, if the disease is rare, the OR approximates to the risk ratio of the disease in the population.

Advantages of the case–control design in epidemiology studies include:

1. A relatively simple method of looking at disease causation
2. They are good for studying rare outcomes
3. They are less time and resource intensive relative to cohort studies
4. They are better than cohort studies for investigating conditions with long latent periods following exposure
5. Multiple exposures can be investigated quite readily

Limitations of the case–control design in epidemiology studies include:

1. They are not optimal for the study of rare exposures.
2. They can only provide an estimate of relative, and not absolute, risk in the population.
3. They cannot provide information on the temporal relationship between exposure and outcome. As a result, any association may be prone to reverse causation. For example, a patient with a terminal condition such as lung cancer may be less likely to stop smoking than someone without this condition. This would increase the apparent association between smoking and lung cancer without the possibility of detecting this reverse causation with a case–control design.
4. They are particularly prone to some forms of bias, especially selection and observation bias.
5. There is less scope to control for confounding factors compared with cohort studies.

2.4. Bias in Case–Control Studies

2.4.1. Selection Bias

This is particularly problematic if both the exposure and disease outcome have already occurred (retrospective case–control design) leading to bias in the inclusion of either cases and/or controls depending on the exposure of interest. For example, if a low response rate is achieved for either cases or controls, there may be something inherently different between those who participate and those who either refuse or for other reasons do not participate in the study resulting in bias in the association under investigation.

2.4.2. Observation Bias

Types of observation bias particularly seen in case–control studies include recall bias and misclassification.

2.4.3. Recall Bias

Recall bias refers to a difference in reporting of the same exposure by cases, which have the disease outcome, and the controls that do not. Those with the disease are more likely than controls to recall past exposures, which they may have already considered as contributing to disease causation, and are more likely to confer greater significance on them. For example, a childhood vaccination, if followed by the onset of illness, is more likely to be attributed by the parent as causative than if no illness followed. Later, recall of the vaccination is likely to be higher in the parent of the child who developed the illness compared with those who did not.

2.4.4. Misclassification

Misclassification refers to the incorrect classification of either exposure or disease status. If the misclassification affects cases and controls similarly, then it is termed non-differential or random misclassification and the consequence on any association would be regression to the null. However, if it affects either cases or controls to a greater or lesser extent than the other, this leads to differential or non-random misclassification and the consequence is either an increase or decrease in apparent size of the true association between exposure and disease outcome.

2.5. Variations on the Case–Control Study Design

Various other types of case–control study designs exist, the commonest ones of which include:

2.5.1. Nested Case–Control Design

A nested case–control design is one where the cases and controls are chosen from a cohort in which information on exposures is already available. Additional information for the subset of the cohort selected for the case–control study is often collected. This type of design makes efficient use of cohort data.

2.5.2. Cumulative (Epidemic) Case–Control Design

In situations where the study is conducted after all of the potential cases have occurred, for example, after an epidemic of a short-lived condition following a single specific exposure, the remainder of the population who did not contract the disease would for the source for selection of controls.

2.5.3. Case-Only Design

The control is replaced by a (known or assumed) prior distribution of exposure in the source population, such as with the distribution of genotypes in genetic studies (2). Gene environment interaction may also be studied with the case-only design (3). Further details of genetic epidemiology study designs are provided in Chapter 3.2.

2.6. Summary

Case–control studies are particularly useful for the investigation of rare diseases and are less time and resource intensive than cohort studies but are particularly prone to bias and cannot determine temporal relationships between exposures and outcome. They produce estimates of relative rather than absolute risk in the form of ORs.

3. Cohort Studies

Cohort study is another type of longitudinal study in which individuals are defined on the basis of their exposure status, i.e. the presence or absence of exposure to the characteristic of interest, and are followed up over time to monitor disease outcome. The two groups (cases with exposure and controls without exposure) are compared for occurrence of the disease outcome.

3.1. Selection and Definition of Exposure

The selection of the exposure and its source population depends on many factors relating to the research question including:

- the frequency of exposure
- the feasibility of recording exposure status
- the feasibility of recording disease outcomes after follow-up
- the type of comparison to be made, for example, incidence rate ratios or standardised mortality rate.

The duration of follow-up is often measured in person-time years. Each person in the cohort contributes one person-year for each year (or other time such as day, week, or month) of observation before the person leaves the study due to the development of the disease outcome or loss to follow-up.

The population from which the exposed cohort is obtained also varies. They may be selected from an open population, also termed a dynamic population (4) in which the person-time years of exposure are derived from changing individuals rather than a fixed cohort of people. The alternative is of a closed population or fixed cohort in which the cohort is fixed from the outset and followed up for a certain duration, with little migration of individuals from the study. This type of selection of exposed individuals is particular prone to loss to follow-up (4).

3.2. Definition of Exposure

In its simplest form, an exposure would be recorded as a binary variable, such as ever-smoked versus never-smoked. More often, the research question warrants further, varying degrees of information, for example the duration of exposure, such as number of years smoked, or quantity of exposure, such as number of cigarettes smoked per day, or a combination such as for cigarette pack years. If quantity of the exposure is the most informative measure, a continuous measurement of amount, rather than binary or categorical measurements, would provide the greatest amount of information for study analyses. If the duration of exposure is the main measure, consideration should be given to the confounding effects of age on the duration of any exposure. A combination of quantity and exposure may provide a neat summary measure of the exposure. However, if either the quantity or duration is of greater importance, the result will be of "diluting" its effect by use of this composite measure.

3.3. Selection of the Unexposed Group

As with case–control studies the unexposed group should be as similar as possible to the exposed group in all but exposure status. Sources for selection of the unexposed group can include:

1. General population cohorts
   Also called internal comparison groups, a single cohort is split into exposure categories. An internal cohort allows greater control of the exposure measure to be compared, e.g. quantity or duration, rather than just a binary measure of exposure. It also enables greater accounting for potential bias and confounding.
2. Special Exposure Cohorts
   These are employed when discrete and identifiable sections of the population are known to have been exposed to a particular factor, such as an occupational hazard. There may not be an obvious unexposed group for comparison. The controls may then be obtained from pre-collected data, such as from population rates known as external comparison groups. External comparison groups have less random variation compared with internal controls. However, there are many limitations with their use. They are only available for a restricted range of outcomes, in particular mortality. External comparison groups for morbidity are uncommon but do exist in certain forms, such as from cancer and chronic disease registries. There is usually less data available for the comparison group than if collected as part of an internal comparison group which may restrict the research question. For these reasons, it is difficult to ensure that the group are comparable with the exposed cohort in other aspects, although age and sex can usually be accounted for using standardisation. Other confounding factors may include ethnicity and socio-economic status. There may be bias in the selection of either the study or reference population. One example of this is the "healthy worker" effect – a form of selection bias whereby occupational cohorts are fitter than the general population resulting in lower rates or risk of illness. Not only is it more difficult to control these, but also there may be a greater degree of unknown bias and confounding.

They are often termed prospective studies as they start by identifying exposure and follow-up subjects to determine whether the disease outcome has occurred. However, they may truly be prospective or retrospective depending on whether the disease outcome has already occurred in the cohort (retrospective or historical cohort study design) or whether they are prospectively followed up to determine outcomes in the future (prospective cohort study design). They may also be ambi-directional whereby some (shorter term) outcomes have already taken place, but the cohort is also followed up for another outcome with a longer latency period.

3.4. Analyses

The analysis in a cohort study is essential to determine the difference in the occurrence of the disease outcome between the exposed and unexposed groups. The association between the exposure and outcome(s) can be estimated using a form of relative risk/rate or time to event if using an internal comparison group. The two groups should also be compared to ensure that they are sufficiently comparable in all other relevant aspects. Some studies may choose to look at when an event (disease outcome) happens rather than just the total number of events that occur. With this method, if a person leaves the study due to developing the event under investigation, death or loss to follow-up for other reasons, then they would no longer contribute to the person-time years of follow-up recorded. This type of analysis is also termed survival analysis, event history analysis, duration analysis, or hazard modelling.

If comparing with an external reference group, a standardised mortality or morbidity ratio is determined. Age and sex are the most common factors to be standardised.

Advantages of the cohort study design include:

1. In general, cohort studies provide a better indication of causation than case–control studies and are more direct measures of the risk of developing disease following a particular exposure.
2. The cohort can be used to investigate several disease outcomes.
3. Cohort studies are efficient at following rare exposures which would be difficult to investigate using the case–control design.
4. They afford greater control over confounding factors and are less prone to certain types of bias, such as selection and recall bias, compared with case–control studies.
5. A temporal relationship between the timing of the exposure and the onset of the disease outcome can usually be established in cohort, but not case–control, studies.

Limitations of the cohort study design include:

1. Cohort studies can be very time and resource intensive. In order to obtain sufficient person-time years of follow-up, the study would need to either include a large source population or have a long follow-up period. The resource implications for both of these often determine whether a cohort study is feasible. There are several means by which these can be reduced such as:

   (a) use of existing mechanisms or tools for recording of disease outcomes, such as cancer and chronic disease registers

   (b) use of a retrospective or historical cohort design

   (c) use of an external comparison group

2. Cohort studies are not optimal for investigating rare disease outcomes or for diseases where multiple exposures may be causally related.
3. Cohort studies, while less prone to selection bias and recall bias, are prone to loss to follow-up if the design is prospective (survivor bias).

3.5. Variations of the Cohort Study Design

3.5.1. Historical Cohort Studies

Historical cohort studies use past exposure data and are termed retrospective if all of the exposure and outcome data are retrospectively obtained. Otherwise, the study can be extended to further follow-up for secondary outcomes with longer latency periods following exposures. The advantage is that it reduces the cost of the study, which is often a limiting factor in cohort study feasibility and the time taken for results to be obtained. However, it is uncommon to find sufficiently detailed information from past records. One example of a retrospective cohort study might use records of drug therapy from general practice to establish an association with a particular disease outcome. In this case, both the exposure status (drug therapy) and the disease outcome should already be recorded with reasonable accuracy and completeness in the records.

3.5.2. Nested Case–Control Study Design

Nested case–control studies use data already collected as part of a cohort study and are described earlier in the chapter (see page 2).

3.6. Summary

Cohort studies are well suited for investigating the effects of rare exposures on disease outcomes and usually provide stronger evidence for a causal effect compared with case–control studies. It is easier to control for bias and confounding in cohort studies and the temporal relationship between exposure and outcome is evident. However, they can be time and resource intensive and are not suited to the study of rare outcomes.

4. Qualitative Study Designs

While the above study designs are all quantitative, it is also worth noting the role of qualitative study designs in epidemiology. Qualitative studies "employ non-numeric information to explore individual or group characteristics producing findings not arrived at by statistical procedures or other quantitative means" (1).

Qualitative research is quite frequently used to supplement quantitative health services research data with "richer" information that relates to how people perceive or experience situations. It is also particularly good for investigating complex or sensitive issues where subjects might be less forthcoming with information in response to a structured questionnaire. In these and other situations, qualitative research can provide a "deeper" understanding of individuals' perspectives.

Qualitative data can be very varied but often takes the form of interviews (one-one interviews or focus groups), direct observation (such as in field research), or use of written documents.

There are four main approaches used.

1. Ethnography
   This method stems from anthropology and is concerned with studying the entire culture surrounding the subject of interest. An example is of participant observation whereby the researcher immerses themselves in the culture too.
2. Phenomenology
   Uses a philosophical approach and is interested in the subjective experience and interpretation of the world.
3. Field Research
4. Grounded Theory
   The researcher uses participant observation to form extensive field notes which are then coded and analysed.

The researcher uses generative questions to guide the direction of the inquiry until a core theoretical concept is identified.

5. Emerging Issues in Epidemiology

5.1. Causal Effect and Counterfactuals

The determination of causal effect is the main goal of epidemiological studies. The Bradford Hill criteria of causation (5) identified nine components that gave weight to an observed association being causal. However, an emerging theme is that of the counterfactual approach that states that the above causal criteria are an over-simplification of the approaches necessary to establish causation. The term counterfactual refers to the situation where an event A can only occur if a counter event B does not. The "counterfactual theory" insists that there should be a well-defined answer, in this situation, to the question of what would have happened if A had not been performed. The opposing "predictive theory" considers only what can be predicted before the choice between A and B is made.

In terms of causal inference, causal effects are defined as a contrast of the values of counterfactual outcomes but only one of those values is observed (6). Some epidemiologists argue that it is not possible to directly prove causation because at least one of the disease frequency parameters must be counterfactual and, therefore, unobservable. They postulate that under the counterfactual approach, causal contrasts are the only meaningful effect measures for aetiological studies (7). Thus, causal rate ratios and rate differences should be used rather than measures that are not causal contrasts, such as correlation coefficients, percentage of variance explained, *p*-values, chi-squared statistics, and standard regression coefficients (7).

However, other epidemiologists argue that the original Bradford Hill Criteria for proving causation already incorporate the supposition of counterfactual outcomes and that no further accounting is required for this (8).

5.2. Mendelian Randomisation

Mendelian Randomisation relies on the use of intermediate phenotypes (genetic variants that can influence an individual's response to an environmental factor). This method assesses causality in associations observed between these intermediate phenotypes and disease and, thereby, whether interventions to modify the intermediate phenotype could be expected to influence risk of disease. There are many advantages of using these intermediate phenotypes over conventional disease end-points, such as dealing with the problems of confounding, reverse causation, selection bias, and regression dilution bias. However, in order to undertake such a study, a clear association between the genotype and disease end-point would need to be established. There is also often insufficient understanding of the function of such genetic variants in the disease process. This emerging epidemiological methodology is discussed further in Chapter 6.4.

6. Emerging Uses of Epidemiology

6.1. Infectious Disease

Among the earliest epidemiological observations was John Snow's hypothesis on the causation of cholera by contaminated drinking water (9). However, newer methods of epidemiology are now routinely employed in the surveillance, prevention, and control of outbreaks of infectious disease.

6.2. Environmental Epidemiology

Most diseases are caused or influenced by environmental factors. Environmental epidemiology produces a scientific basis for studying and interpreting the relationships between the environment and population health.

6.3. Molecular Epidemiology

Molecular epidemiology employs the addition of molecular tools to traditional epidemiological approaches and aims to use molecular markers to establish associations between exposures and disease.

6.4. Genetic Epidemiology

The rapidly emerging field of genetic epidemiology aims to use the systematic methods of epidemiology described above to investigate the influence of human genetic variation on health and disease and increasingly on the relationship between environmental factors and disease. Genetic epidemiological study designs have been largely based on traditional study designs, such as case–control and cohort studies. Case–control studies are naturally suited to the study of genetic risk because they can be used for uncommon disease outcomes, such as those seen with single gene disorders. They also allow the simultaneous investigation of multiple genetic risk factors alongside environmental risk factors and their gene-environmental interaction.

Previously, large cohort studies that have been used to undertake genetic analyses have been designed with the primary aim of investigating environmental risk factors, such as the Framingham Study (10) and Atherosclerosis Risk in Communities Study (ARIC) (11). Increasingly, large longitudinal studies are now being designed and undertaken with the genetic basis of common diseases as their primary aim, such as the UK biobank Study (12). Emerging study designs are also being adapted to account for characteristics specific to genetic, rather than environmental, factors such as family structure. Family-based study designs form the most commonly used study design for genetic epidemiological studies. This is further discussed in Chapter 3.2.

References

1. Last JM. A Dictionary of Epidemiology, 3rd Ed. New York: Oxford University Press, 1995.
2. Self SG, Longton G, Kopecky JK, Ling KY. On estimating LHA/disease association with application to a study of aplastic anemia. Biometrics 1991;47:53–61.
3. Khoury MJ and Flanders WD. Nontraditional epidemiologic approaches in the analyses if gene-environment interactions: Case-control studies with no controls! Am J Epidemiol 1996;144:207–213.
4. Rothman JR, Greenland S, Lash TL. Modern Epidemiology, 3rd Ed. Philadelphia: Lippincott Williams & Wilkins, 2008.
5. Hill AB. The environment and disease: Association or causation? Proc R Soc Med 1965;58:295–300.

6. Hernán MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health 2004;58(4):265–271.
7. Maldonado G and Greenland S. Estimating causal effects. Int J Epidemiol 2002;31: 422–429.
8. Höfler M. The Bradford Hill considerations on causality: A counterfactual perspective. Emerg Themes Epidimol 2005;2:11.
9. Snow J. On the Mode of Communication of Cholera. London: Churchill, 1855. (Reprinted in: Snow on cholera: A reprint of two papers. New York, Hafner Publishing Company, 1965).
10. Dawber TR, Kannel WB. An epidemiologic study of heart disease: The Framingham Study. Nutr Rev 1958;16(1):1–4.
11. ARIC Investigators. The decline of ischaemic heart disease mortality in the ARIC Study communities. Int J Epidemiol 1989;18: 588–598.
12. Burton PR, Hansell A. UK Biobank: The expected distribution of incident and prevalent cases of chronic disease and the statistical power of nested case-control studies. Technical Report for UK Biobank, 2005. http://www.ukbiobank.ac.uk/.

Part II

Genetic Linkage Mapping

Chapter 4

Genetic Distance and Markers Used in Linkage Mapping

Kristina Allen-Brady and Nicola J. Camp

Abstract

In this chapter, we focus on maps and markers used for a linkage analysis. More detail regarding the actual linkage analysis methodology follows in Chapter 11. There are two major types of maps: genetic maps and physical maps. Genetic maps indicate the expected number of meiotic crossover events between two loci on a chromosome and are measured in centiMorgans (cM). Physical maps use molecular techniques, such as DNA sequencing, to determine the location of markers on a chromosome and are measured in base pairs (bp). Linkage analysis relies on genetic maps, and hence they are discussed in this chapter. In addition to a discussion of genetic maps and various map functions, we also discuss the selection of markers for a linkage analysis, including the more traditional microsatellite markers and the newer single nucleotide polymorphism (SNP) markers.

Key words: Recombination, Genetic map, Map functions, Microsatellite markers, SNP markers

1. Introduction

1.1. Recombination

Meiosis is the primary mechanism by which haploid gametes (eggs or sperm) are formed from parental germ cells. Meiosis contains a number of stages, including prophase I during which chromosome pairs line up and sometimes exchange material, otherwise known as "crossover" or "chiasma" (plural chiasmata). As a consequence of crossover events, new combinations of alleles are generated along the newly formed haploid chromosomes. The process of crossing over or "recombining" and forming new patterns of alleles is known as recombination. Through recombination, the haploid gametes can receive a mixture of genetic contributions from both chromosomes in the parental pair. Essentially, an infinite number of genetically different gametes can be produced depending on where on the chromosome recombination occurs. Recombination is more likely to occur

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_4, © Springer Science+Business Media, LLC 2011

between loci that are far apart on a chromosome than between loci that are close together. Hence, loci that are close together on the chromosome are often inherited together from parent to offspring and are said to be linked. Loci on different chromosomes segregate independently, and there is no linkage between them. Recombination events do not occur uniformly across a chromosome. Crossover rates vary by chromosomal region and chromosome. In general, recombination is more likely to occur near the tail of a chromosome (i.e., the telomere) and less likely near the middle of the chromosome (i.e., the centromere). Furthermore, the likelihood of a new crossover in the region of an existing crossover is much smaller than expected, a phenomenon known as chiasma interference.

2. Genetic Maps

The genetic distance between two loci located on the same chromosome is defined as the number of crossover events between those two loci averaged from all meioses (parent to offspring transmissions) across multiple families. In simple terms, all the meioses in the family data are inspected for recombinant events and the positions are noted. For example, a study may find that recombination occurs between two loci 10% of the time. The recombination rate between those two loci then is 10%.

Genetic distance between two loci is measured in units of Morgans (M), in honor of the American geneticist Thomas Hunt Morgan who hypothesized the process of crossover in 1910. One Morgan unit between two markers indicates an expected rate of one recombination between the loci per meiosis. A more commonly used unit for recombination, however, is the centiMorgan (cM). Two markers are 1 centiMorgan (cM) apart if a crossover event occurs during meiosis only 1 time in 100 (0.01), in other words a rate of 1%. Morgan's student, Alfred Sturtevant, recognized that the percentage of recombinants across multiple families could be used as quantitative index of the linear distance between two genes. He realized that the greater the distance between two linked genes, the greater the chance that recombination would occur between them, and importantly that genes are arranged in some linear order. Sturtevant created the first genetic map in 1913 (1).

Figure 1 illustrates a simple example of a genetic map. In this map, it is estimated that the recombination rate between loci X and loci Y is 5%, or 5 cM, and the recombination rate between loci Y and loci Z is 2%, or 2 cM. If the physical order of the loci is known to be X–Y–Z, then a simple additive rule can be used and it can be inferred that loci X and Z are 7 cM apart.

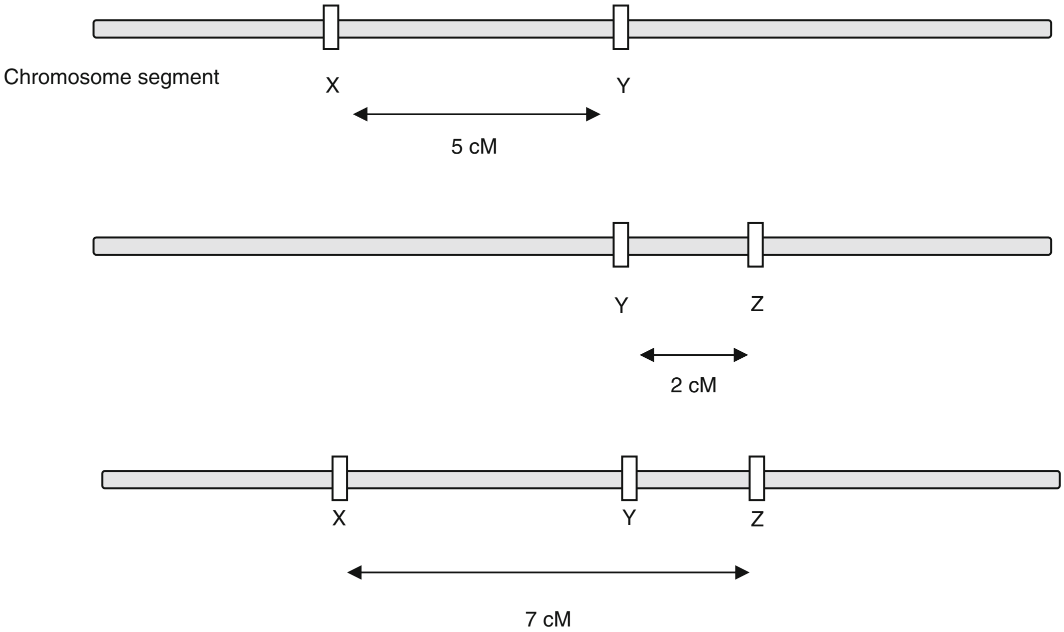


Fig. 1. A chromosomal segment is shown with two loci X and Y being separated by 5 cM. Loci Y and Z are determined to be 2 cM apart. If the physical location of X, Y, and Z are known, it can be determined that loci X and loci Z are 7 cM apart.

There are several issues that need to be considered with genetic maps. First, genetic distance is estimated and relies on the ability to recognize crossover events in family data. Uninformative genetic markers and genotyping errors can create problems. Crossover events are best observed if at least one of the parents is doubly heterozygous for both loci of interest. If the parents are homozygous for the loci of interest, a recombinant in the offspring cannot be recognized. This is illustrated in Fig. 2 where the father has a genotype of CT at one locus and GG at another locus. If the resulting offspring inherits C and G alleles at the two loci from this parent, it cannot be determined if a recombination occurred between the two loci since the alleles at the second locus are identical. Second, crossover rates vary by the sex of the parent. Females form approximately 1.5 times the number of crossover events as compared to males, therefore it is best to estimate sex-specific genetic maps. However, sex-averaged genetic maps are most often used for convenience and because historically, some of the most widely used linkage analysis packages did not support sex-specific maps. Lastly, it should be noted that crossover events do not necessarily always lead to an observable recombination. If two crossovers (a double crossover) occur between two loci, then this does not result in a new combination of alleles at the two loci. Only an odd number of crossover events result in an observable recombinant event. Hence, relying on observable recombinant events can underestimate map distance. For loci that are close together, it is highly probable that only

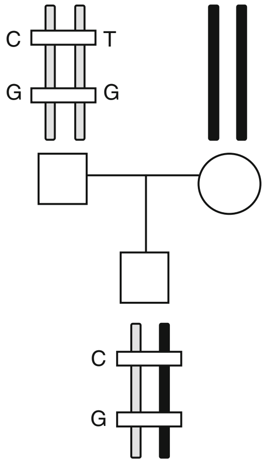


Fig. 2. This figure represents a simple pedigree of two parents and their offspring. Males are represented by *squares* and females are represented by *circle*. The father has alleles C/T at the first marker and alleles G/G at the second marker. When the father transmits a chromosomal segment to his son, it cannot be determined if a recombination between the two markers occurred or not because the alleles at the second marker are homozygous. The situation where parents are heterozygous at both markers is the most informative for a linkage analysis to determine where a recombinant occurs.

one crossover occurs between them. For markers that are farther apart, the probability of more than one recombination events increases.

For markers that are closer together, a related measure to the recombination rate is the recombination fraction. The recombination fraction, θ, indicates the probability that a recombination occurs between two markers. Although a probability, its maximum value is 0.5 indicating a 50:50 chance of recombination, or that two loci sort independently and are unlinked. A recombination fraction less than 0.5 indicates that two loci are not sorting independently and there is linkage between them. The minimum value for θ is zero. For example, if recombination occurs between two loci with probability 0.1, that is, 10% of the time, then $\theta = 0.1$.

Mathematically, the expected number of recombination events between two loci, the recombination rate (measured in Morgans), can be written as:

$$E(\text{recomb events}) = 1 \times P(1 \text{ recomb event}) +$$

$$2 \times P(2 \text{ recomb events}) + 3 \times P(3 \text{ recomb events}) + \cdots$$

where E= expected and P= probability.

For small distances where the probability of multiple recombination events is extremely unlikely, only the first term in the equation above is nonzero, and the recombination rate becomes equivalent to the recombination fraction, θ.

$$E(\text{recomb events}) = P(1 \text{ recomb event}).$$

3. Map Functions

The advantage of using map distance is that it is an additive function, as illustrated in Fig. 1. Recombination fractions are not additive; they have an upper bound of 0.5. Hence, map distances are preferred for mapping chromosomes and are used for linkage analysis. However, recombination fractions are more convenient to estimate because they are determined over small distances where the simplifying assumption of no multiple recombination events can be made. Map functions are used to define the relationship between genetic distance and recombination fraction. The simplest genetic mapping function is that the genetic distance (*d*) between two loci is equal to the recombination fraction (θ).

$$d = \theta$$

This simple equation holds true if the two loci are close together, for example, when $\theta < 0.10$.

When multiple recombination events occur over a longer interval, the simple formula listed above no longer applies. Multiple different map functions have been suggested. In this chapter, we mention only the two most common: the Haldane map function and the Kosambi map function. For more details behind these functions, we refer interested readers to Ott (2). The Haldane map function derived in 1919 assumes that recombination events are rare and follow a Poisson distribution (3). There is no correction made for a previous recombination occurring in the same region.

Haldane map function:

$$\theta = \frac{1}{2}\left(1 - e^{-2d}\right)$$

and solving for the distance,

$$d = -\frac{1}{2}\ln(1 - 2\theta).$$

While the Haldane mapping function assumes that a previous recombination does not increase or decrease a subsequent recombination, the Kosambi map function published in 1944 (4), adjusts the map distance based on a level of interference from a nearby recombination. In essence, the Kosambi map function adjusts the proportion of double crossovers because it modifies the map distance based on interference. The adjustment for interference has been found to more adequately reflect the actual location of markers observed in humans and other mammals and has been found to produce more realistic map distance values. For this reason, the Kosambi map function has been widely used in genetic maps.

Kosambi map function:

$$\theta = \frac{1}{2}(e^{2d} - 1)/(e^{4d} + 1)$$

and solving for distance,

$$d = \frac{1}{4}\ln[(1 + 2\theta)/(1 - 2\theta)]$$

3.1. Genetic Maps Versus Physical Maps

As has been described above, genetic maps are built on counting the number of recombinants that can be inferred in a large number of families genotyped across a large number of genetic markers. Physical maps are based on sequence data to determine the order and spacing of markers and genes. While an increase in genetic map distance translates directly into an increase in physical map distance, the actual distance between two loci may not correlate well between the genetic map and the physical map. A rough rule of thumb is that 1 cM corresponds to 1 megabase (Mb), or 1,000,000 bp, of DNA. However, this estimate is very rough and has been shown to vary considerably depending on the chromosomal region examined.

4. Genetic Map Resources

Much work has been done to construct genetic maps that can be used for a linkage analysis. The first reference genetic map was proposed and implemented by Jean Dausset in Paris. He proposed that a human genetic map be based on a reference group of families and that the results are made available to all genetic researchers (5). The set of families selected were from France and Utah and became known as CEPH families (Centre d'Etudes du Polymorphisme Humain). The 40 original families selected each consisted of an average of 8.3 siblings, their parents and their four grandparents. Thus, each family contributed approximately 16 meioses (from each parent to all offspring) with the grandparents used to establish the haploid chromosomes in the parental generation. As more genetic markers were discovered, these were genotyped and mapped; however, many were not genotyped in all 40 families, but rather in only 8 or 9 families. Genome-wide genetic maps predate physical maps, and therefore genetic maps were also used to establish marker order. Early genetic maps were typically built to map a locus order that required the smallest number of recombination events between markers. The first published human genome-wide genetic map was based on biallelic (i.e., two alleles) polymorphisms that have rather low informativeness (6). Genetic maps based on more informative multiallelic short tandem-repeat (microsatellite markers) are the classical linkage genetic maps. More recently, integrated genetic maps of microsatellite and single nucleotide polymorphism (SNP) markers have been developed (e.g., MAP-O-MAT (7)). Here, we briefly mention some of the genetic maps that are currently available.

4.1. Marshfield Genetic Map

The Marshfield genetic map was built using eight CEPH families and based on genotype data from polymorphic microsatellite markers (8). Within these eight families there are 188 meioses. When markers are close together and there are no recombination events separating them, the markers are listed in an arbitrary order. Hence, in the Marshfield map, the marker order is not well determined when marker distance is small, although physical maps can be used to order the markers appropriately. The total number of microsatellite markers incorporated in the Marshfield map to date is 8,325.

4.2. Généthon Genetic Map

The Généthon genetic map was prepared from 1990 to 1996 under the direction of Jean Wissenbach (9–11). The Généthon map was also built using the microsatellite marker $(AC)_n$ repeats and the same eight families as used for the Marshfield map. The group originally built a map using 814 microsatellite markers in 1992 and progressed up to 5,264 microsatellite markers by 1996.

4.3. deCODE Genetic Map

Just as Marshfield and Généthon maps are based on individuals with European ancestry, the deCODE genetic map also uses individuals with European ancestry, Icelandic ancestry specifically (12). The deCODE genetic map is considered the most accurate to date as it is based on a larger sample size compared to other genetic maps. The deCODE map utilizes genotype data for 870 individuals in 146 Icelandic two-generation families (1,257 meiotic events). The map is based on 5,136 microsatellite markers.

4.4. Other Maps

In addition to genetic linkage maps based on estimating recombination in family data, other types of genetic maps have also been constructed. We mention only one, the radiation hybrid map. Radiation hybrid maps are maps that are created based on irradiating chromosomes with X-rays, such that they break into several fragments. The breaks are somewhat analogous to crossovers in genetic linkage maps and are measured in Rays (R), or centiRays (cR), where 1 cR is equivalent to a 1% probability that a chromosome break has occurred between two markers after irradiation. The chromosomal breaks are recovered in rodent cells and the rodent–human cells are cloned. The farther apart two markers are, the more likely it is that a break occurs between them. In radiation maps, the loci need not be polymorphic as do markers for genetic maps.

4.5. Integrated Genetic Maps

Much work has been done to integrate information between the various types of maps to produce a map that contains the most informative information from each type of map. These types of maps are called integrated maps. As different metrics are used to create different maps, for example, cM distances from genetic maps, bp positions from physical maps, and cR distances from radiation hybrid maps, it is difficult to combine map information. The first level of integration typically focuses on ensuring that a locus is on the same chromosome for all maps. The next level of integration focuses on ensuring that locus order on a correct chromosome is the same across all different types of maps. Finally, integration of interlocus distance across maps is considered. Reconciling differences is not trivial. Sequencing errors or missing sequence can be problematic for physical maps. For genetic maps, the number of meioses studied, the informativeness of the DNA markers used, and genotyping errors impact its reliability. Map misspecification can have serious negative consequences on gene-mapping studies.

4.5.1. MAP-O-MAT

A combined genetic linkage and physical map is available from the Human Genetics Institute at Rutgers University, USA. The resource is an integrated genetic linkage map that incorporates sequence-based positional information. Currently, it includes 28,121 polymorphic markers (including both miscrosatellites and SNPs) with physical positions corroborated by recombination-based data (7). Radiation hybrid data is not incorporated. The map data is called MAP-O-MAT and is available at http://compgen.rutgers.edu/mapomat/.

5. Marker Selection for Linkage Analysis

Linkage analysis relies on the ability to model the positions of recombination events in families. Hence, it is imperative that the marker maps used in a linkage analysis are designed with this in mind. As was discussed above and illustrated in Fig. 2, recombinants are best identified if all markers are heterozygous in the parents. A marker has an increased likelihood of being heterozygous if it has a substantial number of alleles and that each allele has a high enough frequency in the population being studied. Markers meeting these criteria are considered to have high heterozygosity and are often referred to as being highly polymorphic. Microsatellite markers are the traditional marker of choice for a linkage analysis because they are highly polymorphic markers. They have a variable number of tandem repeating units which usually comprise a simple sequence consisting of two, three, or four nucleotides. The varying numbers of repeating units are considered alleles. As the number of repeating units can vary between individuals, genotypes within a pedigree can be fully informative such that one can determine which ancestor originated a particular microsatellite marker and subsequently where recombination events occur between markers.

Microsatellite markers are neutral (i.e., free from natural selection) and have an increased rate of mutation compared to other regions of DNA. The increased rate of mutation results in an increased likelihood that the marker has high heterozygosity. Another advantage of using microsatellite markers is that they generally have no physiologic function; they are usually located in the noncoding region of the chromosome. They do not interfere with the phenotypic expression at other loci. There are also limitations of microsatellite markers. Microsatellite marker alleles are recognized as bands on a gel (i.e., Southern blot), but if there are too many alleles, the bands may be so close together that they cannot be distinguished without error.

More recently, SNPs are also being used for linkage analysis. These markers can be processed rapidly on chips, in large quantities, with high accuracy, and for a relatively low cost. These markers only have two alleles at each marker and taken independently have low heterozygosity; however, considered together in the vast number that can be typed genome-wide, they can generate the necessary information with respect to identifying recombination events. The advantages of using SNP markers are that they are much more abundant than microsatellite markers, and they are less prone to error in the laboratory. A disadvantage arises at the analysis stage due to the correlation that occurs between SNP markers on the dense maps available.

The classical statistical algorithms in linkage analysis assume that the alleles at the genetic markers being analyzed are independent of one another, or in other words a genotype at one marker has no effect on a genotype at a neighboring marker. For sparsely spaced genome-wide microsatellite markers, this is the case. Often, however, SNP markers are in close proximity to each other and cannot be assumed to be independent of each other. An allele at one SNP marker may be associated with an allele at another marker, such that two alleles are correlated in the population. For example, an "A" allele at one SNP locus may appear with a "T" allele at a second SNP locus more often than expected by chance across a population. This phenomenon of correlated alleles across loci is known as linkage disequilibrium (LD), or the nonrandom association of alleles at two or more loci.

Applying classical linkage techniques to SNP data can be problematic, because linkage algorithms assume that markers are in linkage equilibrium (i.e., no allelic correlation across markers), and not in LD as is often the case with SNP markers. Linkage analysis results can be artificially inflated if the markers are in LD, and therefore some control for LD must be made. We finish the chapter with a brief discussion of controlling for LD using SNP markers.

5.1. Eliminate SNPs in High LD

One option for dealing with SNPs in high LD is to prune the set of linkage markers to generate a list of SNPs that are in approximate linkage equilibrium. This is often done by calculating measures of LD for pairs of markers that are contiguous (next to each other) along a chromosome. If two markers are in high LD, one of the markers is eliminated. The process continues until a set of markers is generated in which all contiguous pairs of SNPs are in approximate linkage equilibrium. There are other options of eliminating high LD SNPs, including testing for LD between all SNPs (contiguous or not) within a window (e.g., a 50 SNP window). The window can then "slide" a designated number of markers (e.g., slide 5 SNPs) and the process can be repeated until a set of SNPs is selected that is in again approximate linkage equilibrium. At the time of the publication of this chapter, a useful tool for selecting a low LD set of SNPs can be found online in the freely available analysis package PLINK (http://pngu.mgh.harvard.edu/~purcell/plink) (13). For those interested in additional methods for pruning a set of dense SNP data to use for a linkage analysis, we refer readers to other methods proposed by the authors of this book chapter (14). Once a low LD set of SNPs is generated, these SNPs can then be analyzed in the usual way using standard linkage analysis software.

5.2. Incorporate LD into the Linkage Analysis

More advanced techniques for controlling LD attempt to model the LD as part of the linkage analysis. In a two step approach, SNPs in LD can first be grouped such that they are considered simultaneously as a single "multi-allelic" locus (assuming no recombination between loci), and second, be used in the linkage analysis in these groupings that minimize between group correlation. This approach has been incorporated in the linkage software MERLIN (15). A more integrated approach uses graphical modeling to establish an optimal model of LD for all the markers which is then incorporated into the linkage algorithm. This approach is incorporated in the linkage package McLink (14, 16). In both approaches, SNP markers are not discarded, rather LD between markers is modeled, and the model is incorporated into a linkage analysis.

6. Conclusion

Linkage analysis can only be successful for identifying disease genes if appropriate maps and markers are used. In this chapter, we have discussed criteria for what makes a good genetic map and what types of markers or subsets of markers are most informative for linkage analysis. We stress that the information provided here only briefly covers the topics, and we encourage interested individuals to read additional sources.

References

1. Sturtevant AH. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. J Exp Zool, 1913; 14:43–59.
2. Ott J. *Analysis of Human Genetic Linkage*, 3rd edition, 1999, Johns Hopkins University Press, Baltimore, MD.
3. Haldane JBS. The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet, 1919; 8:299–309.
4. Kosambi D. The estimation of map distances from recombination values. Ann Eugen, 1944; 12:172–175.
5. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel J-M, White R. Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative genetic mapping of the human genome. Genomics, 1990; 6:575–577.
6. Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, et al. A genetic linkage map of the human genome. Cell, 1987; 51:319–337.
7. Matise TC, Chen F, Chen W, De La Vega FM, Hansen M, et al. A second-generation combined linkage physical map of the human genome. Genome Res, 2007; 17(12):1783–1786.
8. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. Am J Hum Genet, 1998; 63: 861–689.
9. Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M. A second-generation linkage map of the human genome. Nature, 1992; 359:794–801.
10. Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, et al. The 1993–1994 Genethon human genetic linkage map. Nat Genet, 1994; 7:246–339.
11. Dib C, Faure S, Fizames C, Samson D, Drouot N, et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature, 1996; 380:152–154.
12. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, et al. A high-resolution recombination map of the human genome. Nat Genet, 2002; 31:225–226.
13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: A toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet, 2007; 81:559–575.
14. Allen-Brady K, Horne BD, Malhotra A, Teerlink C, Camp NJ, Thomas A. Analysis of high-density single-nucleotide polymorphisms data: Three novel methods that control for linkage disequilibrium between markers in a linkage analysis. BMC Proc, 2007; 1(Suppl 1):S160.
15. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet, 2002; 30:97–101.
16. Thomas A. Towards linkage analysis with markers in linkage disequilibrium by graphical modelling. Hum Hered, 2007; 64(1):16–26.

Chapter 5

Approaches to Genetic Linkage Analysis

M. Dawn Teare

Abstract

Genetic linkage analysis concerns the estimation of genetic distance between two or more genetic loci. In genetic epidemiology, it is predominantly used to identify, or map, a genetic locus that is associated with quantitative trait variation or, in the case of binary or discrete traits, modification of the risk of being affected with a disease or phenotype. Linkage analysis uses a panel of reference genetic markers to track the segregation of genomic segments within families or sets of relatives. Individuals within the families must be measured for the trait, and often the families have been selected because they segregate the phenotype of interest.

Key words: Human genetic linkage, Recombination fraction, Allelic segregation, Model-based linkage analysis, Model-free linkage analysis

1. Introduction

When a phenotype or trait is influenced by a gene, then correlation is seen between the phenotype and the genetic variation within the gene. When the gene is known and variation directly observed, this correlation can be precisely measured; however, genetic linkage analysis uses this expected or anticipated correlation to actually map or identify the location of the gene. So the objective of the analysis is to infer how frequently marker alleles are co-inherited with the disease alleles, and thereby estimate the recombination fraction between them (see Chapter 4).

We illustrate genetic linkage analysis for a binary trait with two hypothetical pedigrees drawn in Fig. 1. These two families have been selected for study due to the occurrence and segregation of a rare dominant disorder in some of the family members. In this illustrative example, four polymorphic markers have been genotyped at 5 cM intervals and the measured genotype at each

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_5, © Springer Science+Business Media, LLC 2011

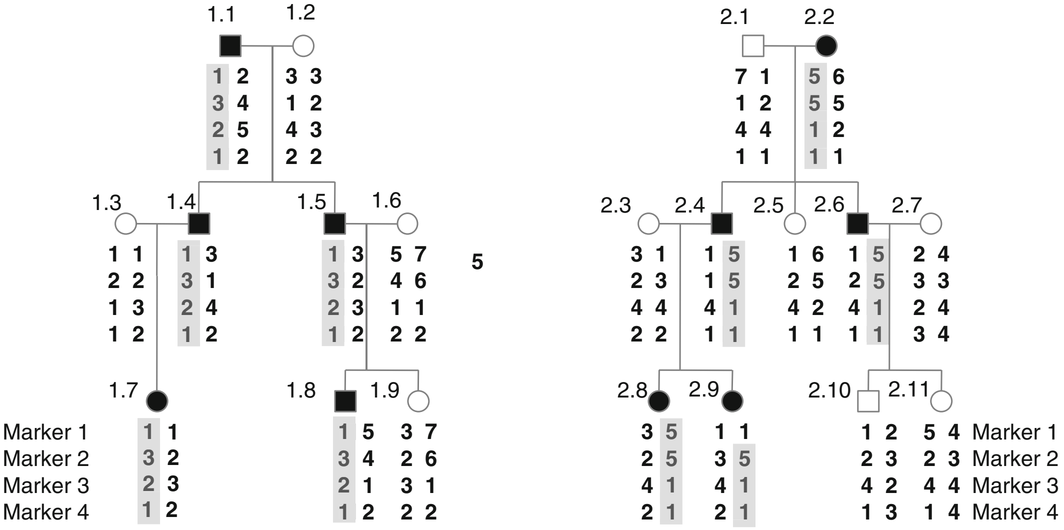


Fig. 1. *Numbers above the symbol on the left* identify the person in the pedigree, e.g. 1.7, person 7 in family 1. *Shaded bold numbers below each symbol* indicate the pair of alleles that make up the genotype for the person above.

locus is displayed in the figure. For example, person 2.8 is heterozygous at all four markers and carries alleles 2 and 5 for Marker 2. From the pedigree you can see that person 2.8 has inherited allele 2 from her mother and allele 5 from her father. As we are assuming a simple dominant disease, the pattern of phenotypes tells us the disease locus genotypes for most people in the pedigree. Anyone unaffected must carry two copies of the wild type ("d") allele, whereas affected individuals may carry one or two copies of the mutant or "disease" ("D") allele. In the absence of new mutation, the affected offspring must have one inherited one wild type allele from their unaffected parent, and the disease allele from the affected parent. The only affected individuals who may carry two copies of the disease allele are the founders 1.1 and 2.2. The probability that they are each homozygous for the mutant is a function of the disease allele population frequency. For purposes of illustration, we shall assume that the parents of 1.1 and 2.2 were observed for phenotype only. Provided one of both pairs of the ancestral parents were unaffected, we can further assume that 1.1 and 1.2 each carry only one risk copy of the disease allele D.

2. LOD Scores

Evidence for genetic linkage is traditionally presented in terms of a logarithm of the odds (LOD) score (1). These are directly related to the classic likelihood ratio methodology widely used in statistical methods. For historical reasons (1) the likelihood ratio is reported as a logarithm to the base 10, in place of a natural logarithm. Maximum likelihood methods enable efficient estimation of parameters within a formal statistical hypothesis framework. A parametric model is proposed such that the null and alternative hypothesis is both specified within the model. In the case of linkage analysis, the null hypothesis assumes that the gene associated with variable risk of disease is not linked to this genetic location. In terms of the model, this is equivalent to saying that the recombination fraction, denoted as θ, is equal to exactly 0.5. The alternative hypothesis states that the gene is linked to this region, i.e. $\theta < 0.5$. Recall from Chapter 4, that θ represents the probability of a recombination between two loci at meiosis. The likelihood of the hypothesis given the data is found by computing the probability of the data conditional on the hypothesis or model. The *likelihood function* is then examined to find the parameter value which maximises the likelihood. This is then termed the maximum likelihood estimate. The maximised likelihood is compared to the likelihood under the null, and twice the natural logarithm of this ratio is then compared with a chi-square distribution with one degree of freedom. Genetic linkage analysis focuses on the maximisation of the LOD score function.

In our carefully constructed hypothetical example, there are potentially five informative meioses in pedigree 1 and 7 informative meioses in pedigree 2. As we assume a dominant phenotype, the only informative meioses are those involving allele transmissions to offspring from an affected parent. Person 1.1 transmits alleles to 1.4 and 1.5, 1.4 transmits to 1.7, and 1.5 transmits to 1.8 and 1.9. Hence, pedigree 1 consists of 5 potentially informative meioses. Pedigree 2 consists of 7. We say potentially because in some cases it is not possible to distinguish whether a recombination has occurred or not. For example, person 2.4 is homozygous at marker 4 so we cannot track the segregation at this meiosis to person 2.8. Though we appear to have potentially 12 informative meioses, this is not strictly true. As we cannot observe the gametic phase (where the parental origin of each allele is known) of the disease locus alleles and the marker alleles directly we do not know the phase in the ancestral genotyped affected (here denoted by 1.1 and 1.2). We have to allow for both possible phases in the likelihoods, and this means we effectively lose one informative meiosis per family (2). The LOD score is computed by comparing the likelihood for a range of values of θ and comparing to the likelihood when $\theta = 0.5$.

Two point LOD scores have been computed for these two families and are listed in Table 1. In genetic epidemiology, two point LOD scores evaluate the evidence for linkage between the disease locus and only a single marker. As families are assumed to be independent, family-specific LOD scores can be added (since the probabilities of independent events can be multiplied). A positive LOD score is evidence for linkage at the specific value of θ.

Table 1
LOD scores for each marker and family computed at seven distinct values of the recombination fraction θ

Marker	$\theta = 0.0$	$\theta = 0.01$	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$
1 (Total)	–Infinity	–1.05	0.18	0.55	0.67	0.52	0.27
Family 1	1.19	1.17	1.08	0.97	0.74	0.50	0.25
Family 2	–Infinity	–2.22	–0.91	–0.42	–0.07	0.02	0.01
2 (Total)	2.40	2.36	2.20	2.46	1.56	1.08	0.57
Family 1	1.19	1.17	1.08	0.97	0.74	0.50	0.25
Family 2	1.20	1.19	1.12	1.02	0.82	0.58	0.32
3 (Total)	3.00	2.95	2.73	2.46	1.87	1.25	0.62
Family 1	1.19	1.17	1.08	0.97	0.74	0.50	0.25
Family 2	1.81	1.78	1.65	1.49	1.13	0.75	0.37
4 (Total)	0.89	0.87	0.80	0.72	0.54	0.35	0.17
Family 1	0.89	0.87	0.80	0.72	0.54	0.35	0.17
Family 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00

A negative LOD score is evidence against linkage at the specified θ. You will notice that one of the entries in the table is "–Infinity", this represents negative infinity which results from taking a logarithm of 0. This means that for marker 1 the observed data in family 2 is impossible (i.e. has probability 0) if the two loci are completely linked (i.e. when θ = 0.00). This is because it is clear that two recombinations have taken place in the transmission from 2.4 to 2.9 and 2.6 to 2.11. This result is interpreted as very strong evidence against complete or tight linkage to marker 1, but the maximum LOD score for marker 1 corresponds to a (maximum likelihood) estimate of 0.2, in other words in ten meioses we see two recombinants.

You may notice that family 2 gives an LOD score of 0 for every value of θ reported for marker 4. This means that the likelihood of the model with linkage is equal to the likelihood with no linkage. If you examine family 2 at marker 4, while all individuals are genotyped it is not possible to track the segregation of alleles due to 2.2, 2.4, and 2.6 being homozygous. We would say that family 2 is uninformative for this marker. The highest LOD scores are seen for markers 2 and 3. In each case, the maximum LOD score occurs at θ = 0.00, so what is the difference between the two markers? The difference again is due to the occurrence of homozygosity in a key individual 2.2 at this marker. This results in a loss of information due to the inability to track the alleles.

3. Multipoint LOD Scores

Figure 1 has represented the genotypes as most probable haplotypes (alleles on the left inherited from the father, and those on the right inherited from the mother) and shading indicates how the genetic material has been inherited through the family from the original affected ancestor. So although some individuals are homozygous at some of the markers, all are heterozygous for at least one marker. If we use the extended haplotype information, we are able to see more clearly which haplotype is transmitted and where any recombinations have occurred. When more than one marker is considered, multipoint LOD scores are reported. Multipoint LOD scores for the two families are shown in Table 2.

Table 2
Multipoint LOD scores for two families

	Distance of disease	LOD_scores			NPL_ score		
Marker	**Locus from Marker 1**	**Family 1**	**Family 2**	**Total**	**Total**	***p*-value**	**Information**
Marker 1	0.00	1.204120	–Infinity	–Infinity	1.52074	0.082520	0.851789
	1.00	1.202906	0.406956	1.609862	1.63727	0.077637	0.744463
	2.00	1.202300	1.007582	2.209882	1.77544	0.063477	0.709101
	3.00	1.202300	1.358906	2.561206	1.93561	0.052246	0.713407
	4.00	1.202906	1.608500	2.811406	2.11821	0.039062	0.757690
Marker 2	5.00	1.204120	1.802611	3.006731	2.32370	0.027344	0.875791
	6.00	1.202906	1.802059	3.004966	2.32562	0.027344	0.844485
	7.00	1.202300	1.802141	3.004440	2.33015	0.027344	0.824344
	8.00	1.202300	1.802855	3.005154	2.33730	0.027344	0.809935
	9.00	1.202906	1.804201	3.007108	2.34708	0.027344	0.800280
Marker 3	10.00	1.204120	1.806180	3.010300	2.35949	0.025879	0.795989
	11.00	1.198759	1.775932	2.974691	2.29878	0.029785	0.741715
	12.00	1.193961	1.745990	2.939951	2.24218	0.029785	0.701580
	13.00	1.189727	1.716358	2.906084	2.18953	0.035156	0.668154
	14.00	1.186056	1.687037	2.873092	2.14066	0.039062	0.639786
Marker 4	15.00	1.182948	1.658030	2.840977	2.09543	0.041504	0.616051

Unlike the two point case, the likelihood for linkage now jointly considers all of the fixed markers and the potential location of the disease locus within the region covered by the fixed set. In Table 2, the multipoint LOD scores have been calculated (using the program genehunter (7)) at 16 equally spaced locations beginning by placing the disease locus coincident with marker 1 then moving along in 1 cM steps. The table reports the family-specific LODs and the Total. Now, Family 2 is informative for linkage at MARKER 4 because the neighbouring linked markers provide some information about which chromosomal haplotype is more likely to have been transmitted at each meiosis.

The column of total LOD scores contains the summary evidence for linkage. In this example, the maximum LOD score is obtained when the disease locus is placed coincident with MARKER 3. The LOD score just exceeds the threshold of 3 which was historically regarded as the minimum score required to declare linkage (3). When the LOD score approach was first proposed by Morton in 1955, he envisaged many independent research groups working on genetic linkage in relative isolation and then pooling their results. Researchers would compute LOD scores at several standard values of θ. As the likelihoods are probabilities, they can be added on the log scale enabling easy sharing of results. One of the most important aspects of his proposed method was his recognition of the problem of multiple testing. Morton proposed a sequential testing framework so that linkage would be declared if the total LOD score at a specified theta exceeded the threshold of 3 and linkage to a region or locus would be declared excluded if the total LOD score fell below –2. Though these thresholds were declared using a sequential testing argument, later it was shown (3) that a threshold of 3 was equivalent to a genome-wide *p*-value of around 9%. The stringent threshold of 3 helped to ensure that most significant linkage reports for single gene disorders were possible to replicate (i.e. the type 1 error rate was well controlled).

We have illustrated the LOD score method here for a simple dominant disease example as the properties of the disease locus itself makes the method easier to explain. If we had instead used a recessive disease example, the phenotype of affected tells you that the person must have two copies of the disease alleles at the disease locus, but unaffected individuals may be heterozygous or homozygous for the wild type. This makes the computation of the likelihoods more complex, though the principle is the same. Parametric LOD score analysis requires the full model at the disease locus to be specified so that only the recombination fraction (in two-point analysis) or the location (in multipoint analysis) is estimated.

4. Computational Approaches

As the previous text has stated, the classic focus of parametric analysis estimates the recombination fraction θ when considering the observed marker and the inferred disease locus genotypes. In the 1980s, computing power increased enabling many more than one genetic marker to be considered at a time. The early linkage programmes used the statistical approach known as "peeling" (4, 5). This method was suited to the analysis of large extended pedigrees with few genetic markers. However, in the 1990s very dense linkage markers were commonly used with as many as 50 markers on some chromosome arms. Further advances in statistical methods took advantage of the inheritance vector approach (6) enabled the multipoint analysis of the full chromosome though this methodology was limited in use to pedigrees with a small number of founders. Programmes using the inheritance vectors include genehunter (7) and merlin (8).

The application of parametric genetic linkage methods to complex disease presents a problem as the genetic component or familial aggregation could arise through a number of different genetic models. For example, if the model allows for incomplete (or partial) penetrance a rare dominant locus would to similar familial clustering as a common recessive. Hence, it is common to see linkage reports consisting of LOD scores computed under a variety of models (9). However, if the genetic component of a disease is heterogeneous the power to detect several distinct causative loci with a parametric (sometimes referred to as model-based) approach is weak. In the early 1990s, non-parametric linkage (NPL) approaches became popular as these methods limited the analysis to affected members only, and hence did not require the genetic model to be specified.

5. Model-Based Versus Model-Free Methods

The objective of genetic linkage analysis is to identify the location of the gene or genes which are assumed to influence the phenotype under consideration. In parametric or model-based linkage analysis, the manner in which the gene acts upon the phenotype is assumed to be known. This seems rather odd at first sight that you should have to declare what the characteristics of the gene are before you have characterised it. However, parametric linkage analysis is a powerful strategy for mapping genes with a simple Mendelian form of inheritance, when the mode of inheritance was fairly easy to infer from clinic-based experience. By contrast, families presenting at clinic concerned about their family history with respect to non-Mendelian disease are not a random sample but rather families with exceptionally high incidence of disease and are not representative of patterns seen in the population. So a clinic-based series is not a good resource for estimating a likely genetic model. However, in many cases, the most likely genetic model generating the familial aggregation of phenotypes can be found from prior segregation analysis of population based or systematically collected families. In complex disease, such major gene models usually assume that there is risk of the disease associated with each genotype but that carriers of the gene are at substantially higher risk of disease than non-carriers.

6. Linkage Analysis and Complex Disease

In the early 1990s, it became clear that though there had been some success in identifying major genes in diseases, such as breast cancer, the parametric approach suffered dramatically in power if there was any locus heterogeneity. Though parametric analysis can allow for locus heterogeneity through the heterogeneity LOD score, an extension to the model in which the alternative hypothesis is that only a proportion of families are linked to the disease locus. A framework which more naturally allowed for locus heterogeneity became more popular partially because this also reflected the way that many studies of complex diseases were conducted. In adult onset disease, it is difficult to recruit extended families; affected sibling pair studies were more easily collected. When considering only pairs of affected relatives, the "identical-by-descent" allele sharing methods offer the advantage that the mode of inheritance does not need to be specified. Instead the observed allele sharing at a genetic region can be compared with that expected based on the relationship between the pairs of relatives.

In our pedigrees in Fig. 1, you can see that there are three pairs of affected siblings. We focus on the pair 2.4 and 2.6 for illustration. If we consider the states of their alleles, we see that they share 2 alleles identical by state (IBS) at markers 1, 3, and 4. At marker 2, they share only allele labelled "5" so here they share 1 allele IBS. If you look at the parents of these two affecteds, you can distinguish between those alleles that are IBS and those that are identical due to having descended from the common ancestor, identical by descent (IBD). At marker 1, they share 2 alleles IBD, and at marker 2 they share 1 IBD; however, at the other two markers, they may share 1 or 2 alleles IBD. It seems that IBD = 1 is more likely because they appear to have inherited different haplotypes from person 2.1. The estimated IBD distribution in a large number of such sibling pairs can be compared to the expected distribution. In full siblings, the proportions should be 0.25, 0.5, and 0.25 for sharing 0, 1, or 2 alleles IBD, respectively. The observed distribution can be compared with the expected distribution and if the result is found to be statistically significant, then it can be inferred that a predisposing locus may be present at this location.

In the Lander and Green article in 1994 (6), a generalisation to this method was proposed to study affected relative pairs only, compare their IBD sharing with their expected sharing under the null and report this as an NPL score. We can illustrate how this is done with our two families in Fig. 1. In family 1, we have five genotyped affecteds, this makes a total of ten possible affected pairs. We then estimate their IBD sharing at a specific location with that expected if there was no predisposing locus present (i.e. what sharing would be expected for this pair of relatives). Four of these relative pairs consist of a parent child pair. These are not informative for linkage in the absence of inbreeding, as a parent always shares exactly one allele IBD with an offspring. However, we can see that for the pair of siblings 1.4 and 1.5 they share 1 allele IBD at markers 2 and 3, and may share 1 or 2 alleles IBD for the markers 1 and 4. The pair of first cousins 1.7 and 1.8 shares exactly 1 allele IBD at each of the four markers.

The results of applying this methodology to the two families reported here are also presented in Table 2. The NPL statistic is a normalised score so in theory you can compare it to a standard normal distribution. However, as the IBD distribution is estimated rather than known the exact *p*-value can be computed corresponding to the amount of statistical information contained in the analysed pedigrees.

Though the principal is simple to understand, it is actually not so straightforward to compute the most likely (maximum likelihood) estimate of the IBD distribution especially when only the affected pairs are genotyped. Though with the recent availability of high density SNP chips offering ten 1,000 SNPs per chromosome the observed IBS converges to the IBD distribution (10).

Though the affected pair approach or model-free method is attractive and does not require the knowledge of the genetic model, its performance in terms of localising predisposing genes has been disappointing. This was one of the reasons behind the move to genetic association studies when it was realised that for many complex diseases the lack of identification of major genes by linkage meant one of two things, either the major gene component was due to many private mutations and each family could be assumed to be linked to a different locus, or the genetic components tended to be due to a polygenic model, where very many loci would have additive and accumulated effects on an individual's risk (11).

7. Quantitative Trait Linkage Analysis

Linkage mapping in quantitative traits relies on the assumption that pairs of relative with similar trait values tend to share more alleles IBD at loci influencing those trait values. When pairs of relative are considered, the IBD distribution can be estimated as in the examples shown above. There are three distinctive approaches, using variance components, a regression model, and extreme sampling. Haseman and Elston (12) proposed a regression method for use on sibling pairs. They suggested that the square of the difference between the sibling pair of trait values could be modelled as the dependent variable in a regression taking estimated IBD for the pair, as the independent variable. Under the null hypothesis of no linkage the slope of the regression (or the coefficient for estimated IBD) is 0.

This simple and appealing approach was extended to work with other pairs of relatives but a considerable improvement was found in 1997 by Wright (13) observed that limiting attention to the similarity only between the pairs lost a lot of the linkage information. He showed that the trait sum also contained useful data for detecting linkage. This observation was carried forward by Drigalenko (14) who proposed to use both the trait sum and the trait difference to gain power to detect linkage. A simple regression, where the mean corrected trait product (or covariance) was taken as the dependent variable, was shown to offer a significant robust advantage over the trait difference approach only. This development was termed trait-product regression, and further improvements have addressed some emergent problems, such as loss of power, when there is very high sibling correlation.

The extreme sampling methods rely on selecting pairs who show extreme differences (discordant) or show extreme and similar trait values (concordant). In discordant pairs, you would expect less sharing at those loci influencing trait values than under no linkage. While in concordant pairs, you would expect more sharing. This approach essentially maximises power while minimising cost, in recognition that most statistical information comes from the extremes of the distribution. Risch and Zhang (15) compared these sampling strategies and found that discordant siblings generally are more powerful when it is likely that common environmental factors induce correlation. As the two forms of sampling are associated with different alternative hypotheses, a number of methods have proposed a testing framework to handle both designs jointly (16).

The variance components approach applies the model outlined in Chapter 2. In this application, the covariance between pairs of relatives can be modelled by partitioning the total genetic variance into that due to IBD sharing at a specific genetic location or marker and the variance due to other unmeasured genes. The component due to the specific region is isolated by estimating the actual IBD sharing at the candidate locus, and the residual covariance is that defined by the expected sharing based on kinship. This approach can be applied to extended families, making it potentially more powerful that methods limited only to sibling pairs. This application relies on fitting a multivariate normal model to the measured trait values on relatives (17, 18), and therefore, assumes that the trait is normally distributed at the population level. As we have seen in Chapter 2, a major gene component (which is indeed the reason for conducting the linkage analysis) results in a deviation from normality. This poses a difficulty when deriving robust statistical tests associated with the approach. A number of solutions have been offered to handle the non-normal distributions, and these are extensively reviewed by Feingold (16).

8. Genetic Heterogeneity

Genetic heterogeneity in this context means that more than one genetic risk variant accounts for the genetic component. For example, cystic fibrosis occurs when individuals receive two defective copies of the cystic fibrosis conductance regulator gene (CFTR) gene. However, a very large number of distinct mutant copies of this gene exist [OMiM 602421]. Linkage analysis is robust to this form of heterogeneity as it only considers co-inheritance of marker and disease alleles within pedigrees. The main requirement of the marker is to be highly polymorphic and hence is as informative as possible. In the example families in Fig. 1, the strongest evidence of linkage is coincident with marker 3. In family 1, allele “2” is in phase with the disease allele, whereas in family 2 allele “1” is in phase. This is quite different to association studies, where the actual allelic state of the marker is important.

The form of heterogeneity that makes gene mapping difficult is locus heterogeneity, where a phenotype can arise due to mutations occurring at one of several distinct genetic loci. Parametric linkage allows for this by assuming the heterogeneity is between families and maximises the heterogeneity LOD score (2). However, power reduces sharply in parametric models as soon as locus heterogeneity exists. Simulation studies strongly suggest that model free methods are much more robust to heterogeneity (19).

9. Missing Data

The families in Fig. 1 are atypical of those encountered in practice, as everyone is measured for the phenotype and marker genotypes. When studying adult onset disease, it can be difficult to recruit parents and extended family members. Relatives in earlier generations may be deceased, they may be too ill to participate, they may refuse to consent or have lost touch with the rest of the family. So the information on some members of the family is incomplete. Missing data inevitably makes an impact and reduces power as it becomes more difficult to track the genetic material though the pedigree.

10. Following up Linkage Signals

As in any research, positive signals need replication. In the study of single gene disorders, crossover events can be identified and the region of interest refined. Replication in linkage can be difficult as the interval associated with the strongest signal in similar studies can be several tens of centimorgan (cM). In complex disease, distinct linkage studies frequently report strong signals to different loci. This can be due to genuine genetic differences between populations or to type 1 errors and lack of power. One way to handle the diverse nature of results is to perform a meta-analysis. The GMSA method is frequently used to bring together the summary analyses of genome-wide linkage studies (20). They allow for different linkage marker sets by dividing the genome up into approximately equal-sized chromosome location bins, taking the "best" result for each bin in each study, then ranking these results. The distribution of these clustered ranks is then compared to that under random assortment. In this way, bins that receive more high ranks that would be expected under the null can be identified.

11. Future Prospects

Genetic linkage analysis has somewhat fallen out of fashion in recent years, partly due to an accumulation of disappointing results in many well-powered linkage searches in complex disease and quantitative traits. This has to some extent helped to drive the enthusiasm for genome-wide association studies, where unrelated subjects are studied. However, as molecular genetic technology provides ever denser means to measure and observe individual genetic data, the association and linkage methods inevitably overlap. Kinship estimation (21) and hence potential pedigree reconstruction (22) from genome-wide genotyping may arise directly from the genetic analysis and may no longer require the family-based sampling and recruitment that made studies so difficult in the past.

References

1. Morton NE. (1955) Sequential tests for the detection of linkage. Am J Hum Genet. 7:277–318.
2. Teare MD, Barrett JH. (2005) Genetic epidemiology 2 – Genetic linkage studies. Lancet. 366:1036–1044.
3. Lander E, Kruglyak L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet. 11:241–247.
4. Cannings C, Thompson EA. (1977) Ascertainment in the sequential sampling of pedigrees. Clin Genet. 12:208–212.
5. Elston RC, Stewart J. (1971) A general model for the genetic analysis of pedigree data. Hum Hered. 21:523–542.
6. Lander ES, Green P. (1987) Construction of multilocus genetic-linkage maps in humans. Proc Nat Acad Sci. 84:2363–2367.
7. Markianos K, Daly MJ, Kruglyak L. (2001) Efficient multipoint linkage analysis through reduction of inheritance space. Am J Hum Genet. 68:963–977.
8. Abecasis GR, Cherny SS, Cookson WO, et al. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 30:97–101.
9. Rapley EA, Crockford GP, Teare D, et al. (2000) Localization to Xq27 of a susceptibility gene for testicular germ-cell tumours. Nat Genet. 24:197–200.
10. Thompson EA. (2008) The IBD process along four chromosomes. Theor Pop Biol. 73:369–373.
11. Risch N, Merikangas K. (1996) The future of genetic studies of complex diseases. Science. 273:1516–1517.
12. Haseman JK, Elston RC. (1972) The investigation between a quantitative trait and a marker locus. Behav Genet. 2:3–19.
13. Wright FA. (1997) The phenotypic difference discards sib-pair QTL linkage information. Am J Hum Genet. 60:740–742.
14. Drigalenko E. (1998) How sib pairs reveal linkage. Am J Hum Genet. 63:1242–1245.
15. Risch N, Zhang H. (1996) Mapping quantitative trait loci with extreme discordant sib pairs: Sampling considerations. Am J Hum Genet. 58:836–843.
16. Feingold E. (2001) Methods for linkage analysis of quantitative trait loci in humans. Theor Pop Biol. 60:167–180.
17. Amos CI. (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet. 54: 535–543.
18. Almasy L, Blangero J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 62: 1198–1211.
19. Clerget-Darpoux F, Elston RC. (2007) Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. Hum Hered. 64:91–96.
20. Wise LH, Lanchbury JS, Lewis CN. (1999) Meta-analysis of genome searches. Ann Hum Genet. 63:263–272.
21. Skare O, Sheehan N, Egeland T. (2009) Identification of distant family relationships. Bioinformatics. 25:2376–2382.
22. Gusev A, Lowe JK, Stoffel M. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19:318–326.

Part III

Genetic Mapping by Association

Chapter 6

Fine-Scale Structure of the Genome and Markers Used in Association Mapping

Karen Curtin and Nicola J. Camp

Abstract

In this chapter, mutation (specifically single-nucleotide polymorphisms, SNPs) and recombination will be covered in more detail, and the concepts of genotype and haplotype will be reviewed. Linkage disequilibrium (LD) describes the strength of a relationship between alleles at different loci. The definition for LD, its visual representation, and the calculation of statistics that measure LD will be presented. The power of genetic association studies to identify disease susceptibility alleles fundamentally relies on the genetic variants studied. A standard approach is to determine a set of tagging-SNPs (tSNPs) that capture the majority of genomic variation in regions of interest by exploiting local correlation structures. The concept of LD and how it is used to select tSNPs will be addressed, as well as specific procedures and algorithms that are practiced by researchers to determine these variants.

Key words: Linkage disequilibrium, Haplotype blocks, Mutation, Recombination, Tagging-SNPs

1. Introduction

1.1. Genome Structure

A genome is an organism's complete set of DNA. The human genome is made up of three billion bases of DNA across 23 distinct chromosomes and contains about 30,000 genes, which are the basic physical and functional units of heredity. Genes comprise only about 2% of the human genome, with the remainder consisting of noncoding regions, whose functions may include regulating proteins encoded by genes and providing chromosomal structural integrity (1). An understanding of genomic structure (the relationship between genetic variants at different positions in the human genome) and genetic architecture (the genetic model that underlies a trait) is knowledge used by researchers in their selection of genetic variants to study the inherited basis of a disease, or other traits of interest (2).

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_6, © Springer Science+Business Media, LLC 2011

Evolutionarily, genomic structure is dynamic, changing as a result of recombination, mutation, selection, or genetic drift. Mutation events introduce new genetic variants that initially have strong relationships with the alleles at other variant positions close by. Recombinant crossover events physically break the genome and thus reduce the relationship between alleles on either side of the crossover. Initial studies using population data have clearly indicated areas of the genome that are more likely to experience recombination (recombination "hotspots") and those less likely to experience recombination events, thus defining a variety of patterns for genomic structure across the human genome (*3, 4*).

1.2. Mutation

A mutation is a permanent change in DNA. Hereditary mutations are those in the germline, that is, mutations that are present in egg and sperm cells and therefore they can be transmitted from parent to offspring and be perpetuated in a population (see Fig. 1). DNA changes that cause harm may be removed from the population via selection, and tend to be very rare. If the changes are not of direct harm, then these genetic changes can become common in the population. Genetic changes that occur in more than 1% of the population are often called polymorphisms. They are common enough to be considered a normal variation in the DNA. Polymorphisms are responsible for many of the phenotypic differences observed between people for traits such as eye color, hair color, and blood type. Most polymorphisms influence personal characteristics and have no adverse effects on a person's health; however, some may influence the risk of developing certain disorders (5).

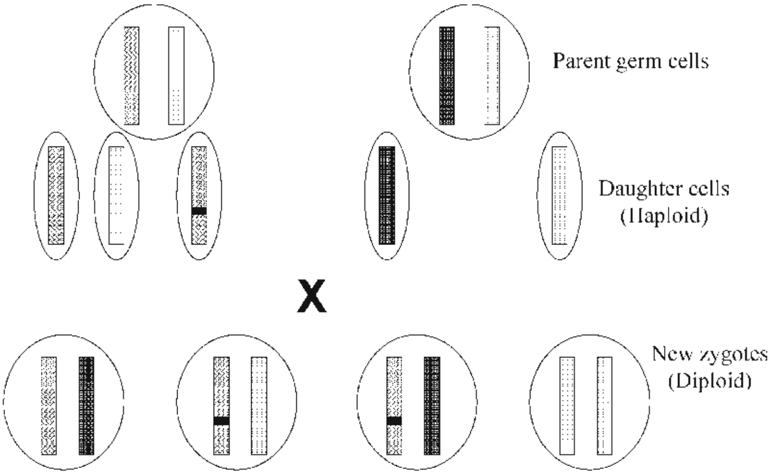


Fig. 1. *Hereditary mutation*. Daughter cells are produced from parent germ cells during the process of meiosis. Daughter cells are haploid, each containing one set of chromosomes. New zygotes are the diploid cells resulting from fertilization, and thus contain two copies of each chromosome. A mutation (denoted as a *black bar*) that occurs in a germline chromosome may be inherited in new zygotes.

Almost all (>99%) nucleotide bases are exactly the same in all humans. In the less than 1% of bases that differ, the most common form of variation (90%) is the single-nucleotide polymorphism (SNP). There are approximately ten million SNPs in the human genome. A SNP is a nucleotide site where different bases can reside. Most SNPs are biallelic, having only two alleles, for example, G or A, or, C or T. Other variants are single or multiple insertions or deletions of one or several bases.

1.3. Recombination

In addition to mutation, genetic variation evolves by meiotic recombination (see Fig. 2). In germ cells which produce eggs or sperm, the chromosome pairs match up and may exchange segments of DNA, a process called recombination. After recombination, the chromosome pairs separate and produce haploid cells that contain only a single chromosome (6).

Over the course of many generations, segments of the ancestral chromosomes in a population are shuffled through repeated recombination events, or *ancestral recombination*. In general, recombination occurs more frequently between positions that are a long way apart, and rarely between DNA sequences that are close together (7). However, it has been shown that there are regions in the genome that are more preferential to recombination events, and conversely those where recombination is suppressed. This leads to segments of ancestral chromosomes, that is, regions of DNA sequences, that are shared by multiple individuals in a population, and represent regions of chromosomes that have not been broken up by recombination (“haplotype blocks”), separated by places where recombination has occurred (“recombination hotspots”).

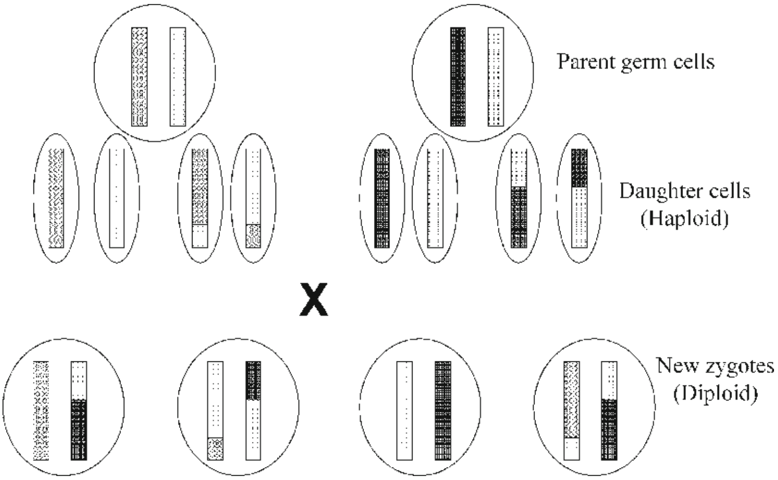


Fig. 2. *Recombination*. During meiosis, the pairing process in the parent germ cells generates haploid daughter cells that contain exchanged DNA segments due to recombination. New zygotes after fertilization may contain a mixture of the two chromosomes from each parent.

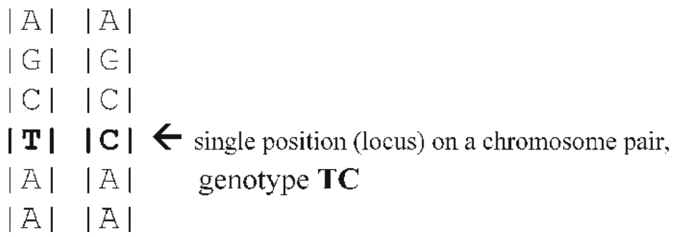


Fig. 3. *Genotype*. At each SNP locus there are two alleles, read across a chromosome pair, which describe the genotype at that locus.

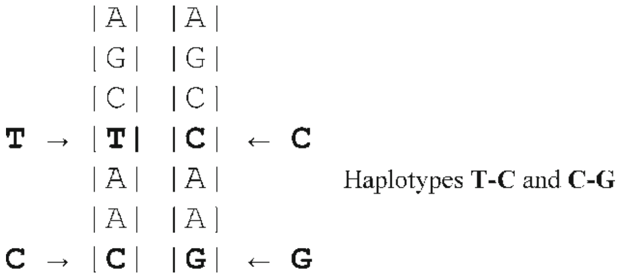


Fig. 4. *Haplotype*. A pair of haplotypes are illustrated across seven base pair positions. The haplotype pair is {A-G-C-**T**-A-A-**C**, A-G-C-**C**-A-A-**G**}. If in this sequence of DNA only two positions are variant in the population (say, positions 4 and 7), then the haplotypes may be referred to more succinctly as **T-C** and **C-G**.

1.4. Genotypes and Haplotypes

Humans are diploid organisms. A *genotype* refers to the "type" seen at a single position (locus) on a chromosome pair in which the base is read across both chromosomes (see Fig. 3).

In contrast to a genotype, a *haplotype* consists of multiple bases that are read along a chromosome (see Fig. 4). Humans are diploid; hence, these haplotypes exist in pairs. Often haplotypes are written to only note the positions that are polymorphic in the population; that is, the bases that are nonvariant are ignored.

The International HapMap Project is a consortium whose goal is to catalogue and compare the genetic sequences of different individuals to identify haplotypes that are shared and not shared both within and across different populations (8). Such information is valuable to investigate the genetics behind common diseases, as discussed in this chapter.

2. Linkage Disequilibrium

2.1. Concept of Linkage Disequilibrium

Linkage disequilibrium (LD) is the association between alleles at two or more loci in a population (see Fig. 5). More specifically, LD describes a situation in which a haplotype occurs more (or less) frequently in a population than would be expected by chance.

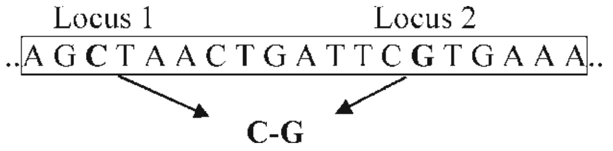


Fig. 5. Two loci on a chromosome.

The concept of LD can also be thought of in terms of prediction. If knowledge of an allele at one locus can predict the allele that will reside at a second locus, then linkage disequilibrium exists between the alleles. However, if knowledge of an allele at the first locus cannot help predict the allele that is at the second locus, then linkage equilibrium exists.

2.2. Common Measures of LD

There are many statistical measures for LD, the more common metrics are discussed here. In mathematical terms, if there is no association or dependence between alleles C and G at two loci, then:

$$\mathrm{P(haplotype\ C-G) = P(allele\ C) \times P(allele\ G)},$$

where P denotes the probability of an event.

If alleles C and G are associated (in LD), then:

$$\mathrm{P(haplotype\ C-G) = P(allele\ C) \times P(allele\ G)} + \delta,$$

where δ is the *raw disequilibrium coefficient.*

This equation can be rearranged to indicate the raw disequilibrium as equal to the probability of a haplotype less the product of probabilities of the two alleles at each locus separately. If two alleles are in linkage equilibrium, $\delta = 0$.

$$\delta = \mathrm{P(haplotype\ C-G) - [P(allele\ C) \times P(allele\ G)]}.$$

The raw disequilibrium coefficient, δ, can be difficult to interpret because its range can vary dependent on allele frequencies at the two loci. Two popular measures of LD that have consistent ranges and that are widely used are D' (Lewontin's D-prime) and r^2 (8, 9). D' is a scaled version of δ that measures LD as a proportion of the maximum amount of LD possible for the specific allele frequencies at the two loci. It can take values $-1 \le D' \le +1$ and can be expressed as follows:

$$D' = \delta/\delta_{\max},$$

where $\delta_{\max} = \min[p(1-q),\ (1-p)q]$; p denotes the frequency of the allele at the first locus and q the frequency of the allele at the second locus.

The linear correlation of alleles, r^2 has a range of $0 \le r^2 \le 1$ and can be calculated as follows:

$$r^2 = \frac{\delta^2}{p(1-p)q(1-q)}.$$

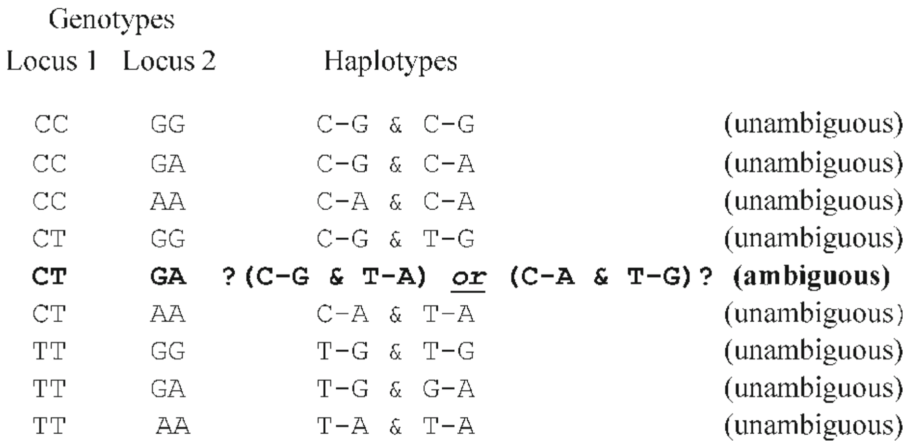
The interpretation of these two LD measurements differs and each has respective advantages and disadvantages. An extreme value for D' (close to 1 or −1) indicates there is no evidence for recombination between the markers, since the initial mutation occurred. Because it is a strong indicator of recombination, it is useful for measuring phylogeny (evolutionary relatedness) and for selecting haplotypes to test in association analyses. On the negative side, D' is more likely to take on extreme values when allele frequencies are rare, and estimates are inflated in small samples. If allele frequencies are similar, a high D' value means the markers are good surrogates for each other; however, if allele frequencies are not similar, they will not be good surrogates. Hence, D' is not considered a good metric for SNP selection (see Section 3). Two positive characteristics of the r^2 LD measure are that its value indicates how well an allele at one locus can predict the other allele at a second locus. Hence, it is a more useful metric for selecting tagging-SNPs, and it measures the loss of efficiency when one marker is replaced with another that has consequences for power in a subsequent association study (10). A drawback of r^2, however, is that it has no direct relationship to recombination.

A list of internet sites and URLs that offer analyses of LD are available at: http://www.nslij-genetics.org/ld/. A list of genetic software packages and a description of their capabilities is available at: http://linkage.rockefeller.edu/soft/.

2.3. Estimating LD

To estimate LD, haplotype probabilities are required. While genotypes are observed directly from standard experimental assays, haplotypes are not. If there are two or more heterozygous loci in a segment of chromosome, then haplotype assignment based on genotype is ambiguous (see Fig. 6).

Haplotypes can be determined using experimental methods, such as, somatic cell hybrids or allele-specific PCR, or alternatively using family data to follow inheritance. However, these options are costly and therefore more often statistical inference is used to estimate haplotypes and their frequencies, which are then used to determine LD. There are two general approaches to statistical inference of haplotypes, expectation maximization (EM) algorithms and Bayesian Monte Carlo Markov Chain (MCMC) methods. These methods estimate the population haplotype frequencies as estimated from a group of independent individuals, and often use these haplotype frequencies to also represent the



Genotypes			
Locus 1	Locus 2	Haplotypes	
CC	GG	C-G & C-G	(unambiguous)
CC	GA	C-G & C-A	(unambiguous)
CC	AA	C-A & C-A	(unambiguous)
CT	GG	C-G & T-G	(unambiguous)
CT	**GA**	**?(C-G & T-A) *or* (C-A & T-G)?**	**(ambiguous)**
CT	AA	C-A & T-A	(unambiguous)
TT	GG	T-G & T-G	(unambiguous)
TT	GA	T-G & G-A	(unambiguous)
TT	AA	T-A & T-A	(unambiguous)

Fig. 6. Ambiguity of haplotypes.

possible haplotype pairs (and associated probabilities) for each individual.

Briefly, an EM algorithm (11–13) uses a two-step iterative process to reach the maximum likelihood estimates for haplotype frequencies. The iterations begin with an initial "guess" or set of starting values for the haplotype frequencies from the observed genotype data (an expectation, E step). For example, this could be the frequencies calculated only from the unambiguous haplotypes. In the maximization step (M step), possible haplotype combinations with respective probabilities are assigned to the ambiguous based on the frequencies from the previous E step. Using the assigned haplotypes and frequencies, haplotype frequencies are recalculated across all individuals (a new E step), followed by another M step, and the process continues until the frequencies converge. Software packages that perform EM haplotype estimation include *SNPHAP* (14) and *GCHap* (15). For large numbers of loci, however, the time to find the solution becomes excessively large and local, rather than global, maxima may result. Techniques that have been offered as a solution include sampling techniques such as Gibbs sampling using multiple loci (16), for example, in the *Haplotyper* software.

Bayesian MCMC approaches are statistically more sophisticated methods that can use prior expectations of genome structure to inform the haplotype reconstructions. Because they can accommodate more complex models of the evolutionary process, Bayesian MCMC methods have been shown to be slightly more accurate than EM; however, they generally take longer in terms of computing time (17). A commonly used software package for Bayesian MCMC haplotype estimation is *PHASE* (18). A list of genetic software packages is available at: http://linkage.rockefeller.edu/soft/.

2.4. Significance of LD

Linkage disequilibrium metrics are estimated and therefore are prone to sampling error. It may be important to determine whether LD is significant, that is, whether it is significantly different than zero (linkage equilibrium). This can be statistically tested by using a likelihood ratio test of the null hypothesis of no LD and alternative hypothesis of LD between markers (19). A significant *p*-value indicates that the LD exists.

2.5. Patterns of LD: Haplotype Blocks and Recombination Hot Spots

Eventually, LD decays over time; random mating and recombination enable mutations from an ancestral haplotype to spread throughout a population. If there were no recombination between the loci, then LD remains the same in a population and does not decay. Recombination between the loci leads to reduced LD, and when sufficient recombination has occurred over time, a state of linkage equilibrium returns. Generally, if two loci are physically distant, recombination between them is common, and equilibrium returns quickly. If two loci are physically close, recombination occurs rarely, and LD is maintained. If recombination rates were consistent across the genome, varying only with physical distance, then a regular pattern of genome structure would be expected across the human genome. In 2002, Gabriel et al. (20) suggested that recombination rates were not consistent across the human genome and that, instead, the human genome was made up of *haplotype blocks* (areas of suppressed historical recombination), interspersed with areas of preferential recombination, known as *recombination hotspots*. Other studies confirmed the unexpected extent of correlation and structure in haplotype patterns (21–25). Marked differences in correlation structure are seen when comparing LD between populations. LD tends to extend over longer distances in those populations that have passed through a recent bottleneck or arise from a small number of founders (22).

Linkage disequilibrium is commonly described using pairwise measures between two SNPs. However, the number of pairwise LD statistics increases exponentially with the number of markers, so that their interpretation becomes unwieldy and difficult to summarize in tabular form. This has led to the development of software to provide visualization of LD. *Haploview* is one such visualization software that was adopted by the HapMap initiative to illustrate their data (26). Another graphical tool for LD is *GOLD* (http://www.sph.umich.edu/csg/abecasis/GOLD) (27). These graphical summaries are called *heat plots*, and are well-suited for summarizing LD in densely genotyped map data. An example of a heat plot using *Haploview* is shown in Fig. 7, and an example of a *GOLD* heat plot is shown in Fig. 8.

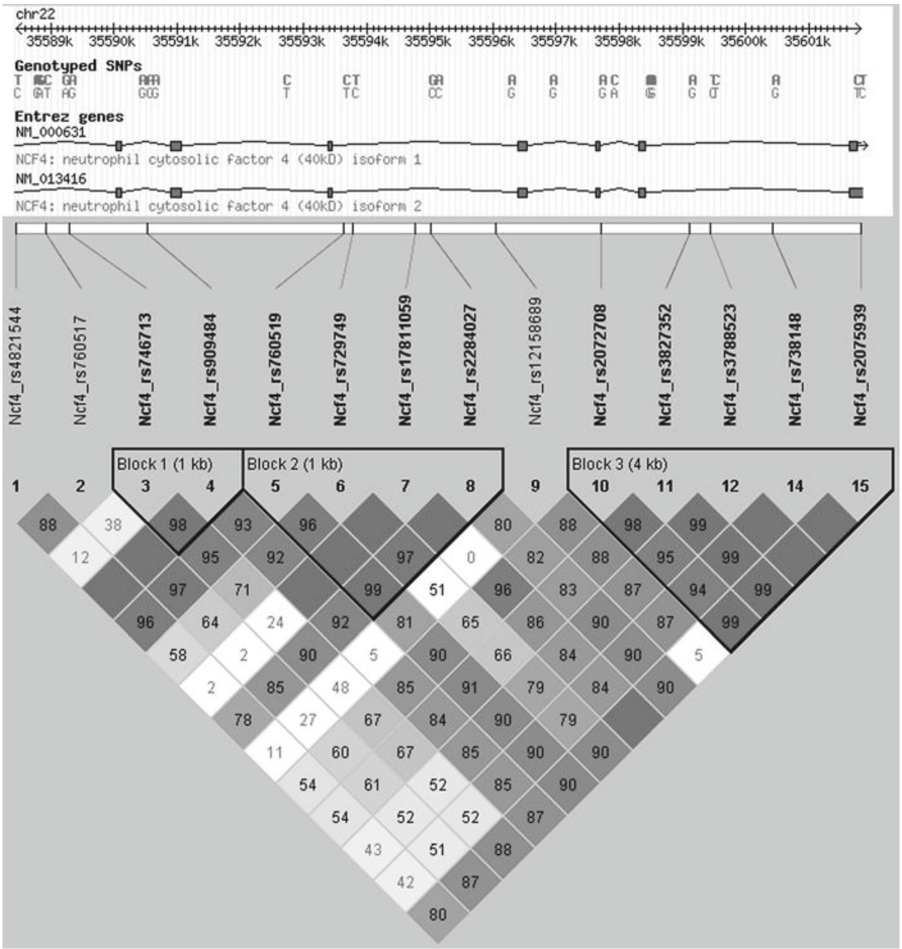


Fig. 7. *Example of a heat plot generated using Haploview*. The darker the shade of *gray*, the stronger the LD. This graph illustrates the haplotype blocks in NCF4 (Olsson et al. Arthritis Res Ther, 9: R98, 2007).

3. Selection of Genetic Variants for Association Studies

Linkage disequilibrium may exist between loci as far apart as 300 kb or more, or may only stretch a few kilobases (20, 23, 28). If LD exists between alleles at two loci (e.g., high r^2), this indicates that the alleles are correlated and, to some extent, the two loci contain redundant information. When selecting SNPs to study in an association analysis, this redundancy can be exploited, such that a smaller set of SNPs can be chosen that adequately represent the majority of the genetic variation in a region. These subsets of SNPs are called tagging-SNPs (tSNPs) (14) because they are selected to "tag" all other variants within the region. The smaller set of tSNPs is selected to be genotyped, which is a much

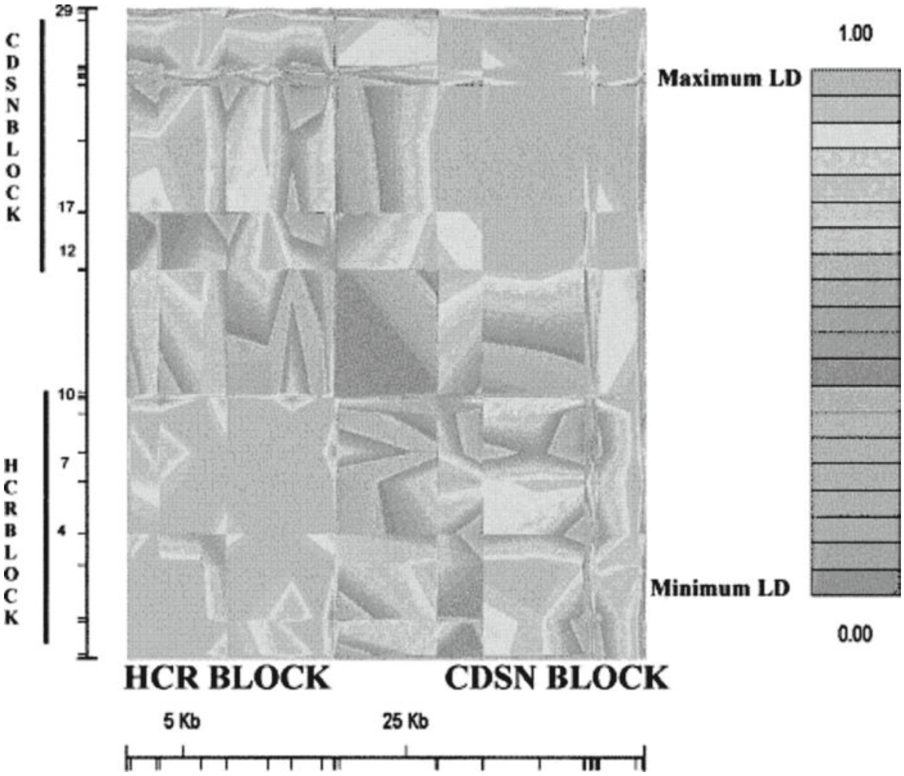


Fig. 8. *Example of a heat plot generated using GOLD*. LD in the 46 kb psoriasis candidate gene (29). Shades of *gray* represent the degree of LD between markers, as measured by *D′*.

more cost effective way to study genetic association than genotyping every SNP in a region of interest. Selecting SNPs using LD is, therefore, a widely used strategy in genetic association studies.

3.1. Discovery Panels for Variant Selection

Before tSNPs can be selected by any method, "full" data (genotypes for "all" variants in a region) must be available for a sample of individuals such that LD can be estimated and used for the selection process. A *discovery panel* is such a sample of individuals. The data available on the discovery panel may vary from full sequencing, which will identify all variants in the individuals, to partial sequencing (such as across exons and regulatory regions in a candidate gene), to dense map data (1 SNP/500 bp). Clearly, dense map data and partial sequencing may miss some SNPs and all methods will be sensitive to the size of the discovery panel for picking up rarer SNPs. Extensive coverage of human genetic variation in discovery panels from diverse populations are now readily available to researchers (29). Publicly available, downloadable sequence data is available for specific genes or genomic regions from the NIEHS SNPs Program (30), SeattleSNPs (31), and

HapMap ENCODE (32). In addition, dense map data is available from HapMap (33) for the entire human genome. These have provided investigators with the means to select variants for subsequent association study. See Section 5, "Web Resources," for a list of URLs.

The resources outlined above were designed to target common population variants and sequencing and genotyping map data are based on discovery panels of limited size. For a single ethnic/racial group, the maximum discovery panel size with sequence data is 24 unrelated individuals (NIEHS SNPs Program, HapMap ENCODE), and the maximum panel size with map data is 60 unrelated individuals (SeattleSNPs, HapMap). The individuals in the discovery panels for these resources are "neutral"; that is, population-based and chosen without regard to a disease or trait. The advantage of neutral discovery panels is that they are universally applicable. The disadvantage of small, neutral panels, however, is that they are inadequate for the detection and characterization of genomic variation surrounding less common alleles (particularly those with minor allele frequencies, MAF, in the range of 0.01–0.05), and may lead to suboptimal tSNPs (34, 35). This problem is worsened if tSNP selection procedures are used that prescreen variants by considering only variants over a predefined MAF or that use map instead of sequence. It has been shown that researchers can supplement data from existing sources for tSNP selection by sequencing additional population-based samples, or by sequencing a set of diseased individuals to increase their power to detect rarer susceptibility variants in subsequent association study (36). However, this is costly. The completion of the production phase of the 1000 Genomes Project (37), anticipated in 2011, will provide a new generation of universally useful tagging sets by sequencing of up to 200 people in each population-specific panel, which will adequately tag common susceptibility alleles and also rare variants that may have low- to moderate-effect sizes.

3.2. Tagging-SNP Selection

A number of algorithms have been proposed to define groups of SNPs that are in high LD and perform tSNP selection. There is no consensus as to which algorithm is best, although a comparison of three widely used algorithms indicated that even when distinct tSNPs were selected using different approaches, the area tagged was highly concordant between methods (38). Here we will briefly describe three different approaches, one that uses the pairwise-SNP LD measure r^2 (33), one that uses SNP-haplotype r^2 (39) and one that uses principal component analysis (PCA) (40).

The simplest and one of the most widely used tagging methods uses pairwise r^2 between pairs of SNPs. In this method, a minimal set of tSNPs is selected such that each SNP that is not selected is in high pairwise LD with a selected SNP (based on a user-defined r^2 threshold, often 0.8). Basically, values for r^2 between pairs of loci are calculated for all SNP pairs, and those values are used to assign SNPs in to "bins." Within a bin, the algorithm identifies those SNPs that best represent the entire bin, that is, those SNPs that surpass the user-defined r^2 threshold with all other SNPs in the bin. Each of these SNPs are identified as a potential tSNP for the bin; if not, the SNPs are designated as an "other" SNP in that bin. A user then simply selects one tSNP from each bin. This approach is implemented in *ldSelect* (31) (http://droog.gs.washington.edu/ldSelect.html). An advantage is that it is quick and intuitively simple. Additional options are that specific SNPs can be specifically forced to be included (perhaps already genotyped, or putatively functional from previous studies) or excluded (assay doesn't work well) in the final tSNP set, which adds to the flexibility of the software. A disadvantage is that the frequencies of the two-locus haplotypes from a pair of SNPs that are used to determine the r^2 values use only data from the two loci to estimate haplotype frequencies, and ignore data at all other loci.

Another widely used tagging method is a haplotype r^2 method. In essence, this differs from the pairwise r^2 method described above, only in that the correlation used to identify tSNPs includes inspection of multiple SNPs together in haplotypes. This procedure will allow that an unselected SNP is represented by a haplotype of selected tSNPs. Thus either single tSNPs, or haplotypes of tSNPs, can serve as proxies for the unselected SNPs. The advantage of a haplotype r^2 method is that fewer SNPs may be required, and hence, the method can be more cost-efficient for genotyping the tSNP set in a subsequent association analysis. However, there are disadvantages. Although theoretically this could be done for all haplotypes of any length, practically only two- and three-locus haplotypes are investigated, which may limit its advantage. A specific restriction of tSNPs selected in this way is that it requires the user to perform specific haplotype analyses to represent the unselected SNPs, which can overcomplicate the analysis planned for subsequent association study. Both pairwise and haplotype r^2 tagging can be performed using the online resource *Tagger* (39), which can be easily implemented locally using Haploview (http://www.broad.mit.edu/mpg/tagger/). Similarly to *ldselect*, SNPs can be specifically included and excluded. In addition, there is an option to identify the "best N" tSNPs for a specified N. Tagger can also evaluate an existing list of tSNPs to see how well they capture unselected SNPs.

Several algorithms based on PCA or other matrix decomposition methods have been proposed (40–43). The basic concept is that SNPs in high LD will load on to the same factors in the PCA procedure. The methods are similar to the r^2 methods in that the factor loadings that define factor-membership in PCA are closely related to multivariate r^2. An advantage of PCA-based approaches is that the tSNPs are selected based on multivariate measures that assess each SNPs relationship to all others simultaneously and may therefore be able to select superior tSNPs. A disadvantage is that the method is more complicated, and hence less intuitive. An example of a PCA method is implemented in the software *PCAtag* (http://www-genepi.med.utah.edu/PCAtag/index.html).

3.3. Candidate Gene tSNP Selection

Candidate gene association studies are undertaken based on functional studies, biological hypotheses, or as a follow-up to regions of interest identified in genome-wide association. The aim of a candidate gene study is to thoroughly interrogate a gene or region, usually consisting of approximately 10,000–200,000 DNA base pairs, either with an aim of finding novel evidence, or to pinpoint a putative disease variant if initial evidence has already been suggested. In general, candidate gene/region studies tend to study SNPs of both common and rarer MAFs and tSNP selection may be performed with a higher r^2 and without a minimum frequency threshold. If variants are prescreened to consider only those with frequencies above a certain level (say 1 or 5%), then fewer tSNPs will be necessary; however, rare, potentially causal alleles will be missed. In addition, researchers will often supplement their tSNPs sets with other SNPs that have been associated with a disease of interest in the literature, particularly if the SNP is in a coding region, and also may perform sequencing to discover additional variants that are not evident in the publicly available resources.

3.4. Genome-Wide tSNP Selection

Power to detect associations in a genome-wide association (GWA) study depends on a high level of LD between an underlying causal variant and the SNPs that are studied (40). Major technological advances in high-throughput genotyping, with very low error rates, has resulted in low per-SNP costs and it is now possible to genotype 500,000 or more variants across the genome in thousands of DNA samples. Hence, the GWA study has become possible and is currently a popular approach for identifying genes that influence common diseases or traits.

To maximize cost efficiency by minimizing the number of required tSNPs, most tagging methods (including all those detailed above) use greedy or exhaustive algorithms. Although useful for candidate genes or small regions, these implementations do not scale well to deal with large genome- or chromosome-wide datasets (41). Research has recently focused on the development of methods for choosing tSNPs for GWA study and has lead to the development of feature selection algorithms that do not involve computation-intensive searches for tSNPs. One such feature selection SNP selection tool that uses hierarchical minimax clustering is *CLUSTAG* (42) (http://www.math.hkbu.edu.hk/~mng/CLUSTAG/CLUSTAG.html).

The commercially available genome-wide gene chip arrays that are currently available can feature one million or more SNPs. These chip arrays are genome-wide tSNP sets designed to tag common variation, focused mostly on SNPs with MAF over 5% and which are most powerful to tag those with MAF over 20%. These gene chip arrays are often used as the first part of a multistage association strategy. The first stage is the genome-wide tSNP set which is usually genotyped on moderate sample sizes (less than 3,000 individuals). At the second stage, selected markers from the first stage are genotyped in a larger number of individuals (sometimes 10,000 or more); and later stages involve more thorough investigation of SNPs identified as potentially important in earlier stages, in a candidate gene or region.

4. Summary

The Human Genome Project and the International HapMap Project have led to the cataloguing of millions of SNPs across the human genome, the majority of which have been genotyped on multiple, small discovery panels of different ethnic/racial backgrounds. Focused gene/region initiatives such as NIEHS SNPs Program, SeattleSNPs, and the HapMap-ENCODE project have led to sequencing data for genes and regions in the same, or similar, discovery panels. All the data from these resources are publicly and freely available and their existence has revolutionized genetic association studies. The characterization of genetic structure using linkage disequilibrium has enabled the selection of informative SNPs resulting in more powerful and comprehensive candidate gene association studies. In addition, and due to parallel technological advances in high-throughput genotyping, genome-wide association studies have become not only a reality, but an increasingly common association study design.

Current resources have already begun to impact finding genes associated with disease in both candidate gene and genome-wide studies. A number of genes have been pinpointed and associated with breast cancer, prostate cancer, type II diabetes, idiopathic scoliosis, and age-related macular degeneration (43–47). Additionally, finding the DNA sequences underlying such common diseases as cardiovascular disease, diabetes, arthritis, and cancers is being aided by the human variation maps (SNPs) generated from the Human Genome Project in cooperation with the private sector. These SNPs also provide focused targets for the development of effective new therapies (1). Beyond the current resources, by 2011 the 1000 Genomes Project initiative aims to provide publicly available genome-wide sequence data for over 1,000 individuals (37). It will therefore provide the larger panels necessary to select tSNPs that will have superior power to identify both common and rarer susceptibility alleles in association studies (36). As technological advances continue, it is expected that full genomic sequencing for all individuals may eventually eliminate the need for selecting tSNPs. However, until full sequencing of large numbers of individuals is technically and financially viable, the tSNP selection process remains an important part in the design of genetic association studies.

5. Additional Web Resources

Abecasis, G. Linkage Disequilibrium Lecture Notes, available at: http://www.sph.umich.edu/csg/abecasis/class/666.03.pdf

Nickerson, D. SNP Discovery and Genotyping Workshop presentation, available at: http://pga.gs.washington.edu/presentations/SNPDiscovery&Genotyping_Sep.ppt

U.S. National Library of Medicine. Genetics Home Reference: Your Guide to Understanding Genetic Conditions, available at: http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/genemutation

Wellcome Trust. The Human Genome, available at: http://genome.wellcome.ac.uk/

References

1. U.S. Department of Energy Office of Science. Human Genome Program. Human Genome Project Information. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml. U.S. Department of Energy Office of Science: Oak ridge, TN, 2008.
2. National Institutes of Health. NIH Guide: Genetic Architecture of Complex Phenotypes. http://grants.nih.gov/grants/guide/pa-files/PA-98-078.html. National Institutes of Health, Office of Extramural Research: Bethesda, MD, 1998.
3. Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., and Stephens, M. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet, *36*: 700–6, 2004.
4. McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. The fine-scale structure of recombination rate variation in the human genome. Science, *304*: 581–4, 2004.
5. U.S. National Library of Medicine. Genetics Home Reference: Your Guide to Understanding Genetic Conditions. http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/genemutation. National Institutes of Health: Bethesda, MD, 2008.
6. National Center for Human Genome Research, National Institutes of Health. Crossing-Over: Genetic Recombination. New Tools for Tomorrow's Health Research, Access Excellence Resource Center: Bethesda, MD, 1992.
7. Wellcome Trust. The Human Genome. http://genome.wellcome.ac.uk/. Wellcome Trust: London, UK, 2008.
8. Devlin, B. and Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics, *29*: 311–22, 1995.
9. Lewontin, R. C. On measures of gametic disequilibrium. Genetics, *120*: 849–52, 1988.
10. Pritchard, J. K. and Przeworski, M. Linkage disequilibrium in humans: models and data. Am J Hum Genet, *69*: 1–14, 2001.

11. Excoffier, L. and Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol, *12*: 921–7, 1995.

12. Hawley, M. E. and Kidd, K. K. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered, *86*: 409–11, 1995.

13. Long, J. C., Williams, R. C., and Urbanek, M. An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet, *56*: 799–810, 1995.

14. Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., and Todd, J. A. Haplotype tagging for the identification of common disease genes. Nat Genet, *29*: 233–7, 2001.

15. Thomas, A. GCHap: fast MLEs for haplotype frequencies by gene counting. Bioinformatics, *19*: 2002–3, 2003.

16. Niu, T. Algorithms for inferring haplotypes. Genet Epidemiol, *27*: 334–47, 2004.

17. Li, S. S., Cheng, J. J., and Zhao, L. P. Empirical vs Bayesian approach for estimating haplotypes from genotypes of unrelated individuals. BMC Genet, *8*: 2, 2007.

18. Stephens, M., Smith, N. J., and Donnelly, P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet, *68*: 978–89, 2001.

19. Zhao, J. H., Curtis, D., and Sham, P. C. Model-free analysis and permutation tests for allelic associations. Hum Hered, *50*: 133–9, 2000.

20. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. The structure of haplotype blocks in the human genome. Science, *296*: 2225–9, 2002.

21. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. High-resolution haplotype structure in the human genome. Nat Genet, *29*: 229–32, 2001.

22. Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. Linkage disequilibrium in the human genome. Nature, *411*: 199–204, 2001.

23. Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., Studebaker, J. F., Ankener, W. M., Alfisi, S. V., Kuo, F. S., Camisa, A. L., Pazorov, V., Scott, K. E., Carey, B. J., Faith, J., Katari, G., Bhatti, H. A., Cyr, J. M., Derohannessian, V., Elosua, C., Forman, A. M., Grecco, N. M., Hock, C. R., Kuebler, J. M., Lathrop, J. A., Mockler, M. A., Nachtman, E. P., Restine, S. L., Varde, S. A., Hozza, M. J., Gelfand, C. A., Broxholme, J., Abecasis, G. R., Boyce-Jacino, M. T., and Cardon, L. R. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet, *33*: 382–7, 2003.

24. Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R., and Dunham, I. A first-generation linkage disequilibrium map of human chromosome 22. Nature, *418*: 544–8, 2002.

25. Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science, *294*: 1719–23, 2001.

26. Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, *21*: 263–5, 2005.

27. Abecasis, G. R. and Cookson, W. O. GOLD – graphical overview of linkage disequilibrium. Bioinformatics, *16*: 182–3, 2000.

28. Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., Whittaker, P., Collins, A., Morris, A. P., Bentley, D., Cardon, L. R., and Deloukas, P. The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet, *13*: 577–88, 2004.

29. Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H.,

Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., et al. A second generation human haplotype map of over 3.1 million SNPs. Nature, *449*: 851–61, 2007.

30. Livingston, R. J., von Niederhausern, A., Jegga, A. G., Crawford, D. C., Carlson, C. S., Rieder, M. J., Gowrisankar, S., Aronow, B. J., Weiss, R. B., and Nickerson, D. A. Pattern of sequence variation across 213 environmental response genes. Genome Res, *14*: 1821–31, 2004.

31. Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet, *74*: 106–20, 2004.

32. The International HapMap Consortium. A haplotype map of the human genome. Nature, *437*: 1299–320, 2005.

33. The International HapMap Consortium. The International HapMap Project. Nature, *426*: 789–96, 2003.

34. Iles, M. M. Quantification and correction of bias in tagging SNPs caused by insufficient sample size and marker density by means of haplotype-dropping. Genet Epidemiol, *32*: 20–8, 2008.

35. Zeggini, E., Rayner, W., Morris, A. P., Hattersley, A. T., Walker, M., Hitman, G. A., Deloukas, P., Cardon, L. R., and McCarthy, M. I. An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat Genet, *37*: 1320–2, 2005.

36. Curtin, K., Iles, M. M., and Camp, N. J. Identifying rarer genetic variants for common complex diseases: diseased versus neutral discovery panels. Ann Hum Genet, *73*: 54–60, 2009.

37. National Institutes of Health, N. H. G. R. I. International Consortium Announces the 1000 Genomes Project. pp. News Release, 2008.

38. Ke, X., Miretti, M. M., Broxholme, J., Hunt, S., Beck, S., Bentley, D. R., Deloukas, P., and Cardon, L. R. A comparison of tagging methods and their tagging space. Hum Mol Genet, *14*: 2757–67, 2005.

39. de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. Efficiency and power in genetic association studies. Nat Genet, *37*: 1217–23, 2005.

40. Amos, C. I. Successful design and conduct of genome-wide association studies. Hum Mol Genet, *16* (*Spec No. 2*): R220–5, 2007.

41. Halldorsson, B. V., Istrail, S., and De La Vega, F. M. Optimal selection of SNP markers for disease association studies. Hum Hered, *58*: 190–202, 2004.

42. Ao, S. I., Yip, K., Ng, M., Cheung, D., Fong, P. Y., Melhado, I., and Sham, P. C. CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. Bioinformatics, *21*: 1735–6, 2005.

43. Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C. S., Bowman, R., Meyer, K. B., Haiman, C. A., Kolonel, L. K., Henderson, B. E., Le Marchand, L., Brennan, P., Sangrajrang, S., Gaborieau, V., Odefrey, F., Shen, C. Y., Wu, P. E., Wang, H. C., Eccles, D., Evans, D. G., Peto, J., Fletcher, O., Johnson, N., Seal, S., Stratton, M. R., Rahman, N., Chenevix-Trench, G., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Garcia-Closas, M., Brinton, L., Chanock, S., Lissowska, J., Peplonska, B., Nevanlinna, H., Fagerholm, R., Eerola, H., Kang, D., Yoo, K. Y., Noh, D. Y., Ahn, S. H., Hunter, D. J., Hankinson, S. E., Cox, D. G., Hall, P., Wedren, S., Liu, J., Low, Y. L., Bogdanova, N., Schurmann, P., Dork, T., Tollenaar, R. A., Jacobi, C. E., Devilee, P., Klijn, J. G., Sigurdson, A. J., Doody, M. M., Alexander, B. H., Zhang, J., Cox, A., Brock, I. W., MacPherson, G., Reed, M. W., Couch, F. J., Goode, E. L., Olson, J. E., Meijers-Heijboer, H., van den Ouweland, A., Uitterlinden, A., Rivadeneira, F., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Hopper, J. L., McCredie, M., Southey, M., Giles, G. G., Schroen, C., Justenhoven, C., Brauch, H., Hamann, U., Ko, Y. D., Spurdle, A. B., Beesley, J., Chen, X., Mannermaa, A., Kosma, V. M., Kataja, V., Hartikainen, J., Day, N. E., et al. Genome-wide association

study identifies novel breast cancer susceptibility loci. Nature, *447*: 1087–93, 2007.

44. Eeles, R. A., Kote-Jarai, Z., Giles, G. G., Olama, A. A., Guy, M., Jugurnauth, S. K., Mulholland, S., Leongamornlert, D. A., Edwards, S. M., Morrison, J., Field, H. I., Southey, M. C., Severi, G., Donovan, J. L., Hamdy, F. C., Dearnaley, D. P., Muir, K. R., Smith, C., Bagnato, M., Ardern-Jones, A. T., Hall, A. L., O'Brien, L. T., Gehr-Swain, B. N., Wilkinson, R. A., Cox, A., Lewis, S., Brown, P. M., Jhavar, S. G., Tymrakiewicz, M., Lophatananon, A., Bryant, S. L., Horwich, A., Huddart, R. A., Khoo, V. S., Parker, C. C., Woodhouse, C. J., Thompson, A., Christmas, T., Ogden, C., Fisher, C., Jamieson, C., Cooper, C. S., English, D. R., Hopper, J. L., Neal, D. E., and Easton, D. F. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet, *40*: 316–21, 2008.
45. Gao, X., Gordon, D., Zhang, D., Browne, R., Helms, C., Gillum, J., Weber, S., Devroy, S., Swaney, S., Dobbs, M., Morcuende, J., Sheffield, V., Lovett, M., Bowcock, A., Herring, J., and Wise, C. CHD7 gene polymorphisms are associated with susceptibility to idiopathic scoliosis. Am J Hum Genet, *80*: 957–65, 2007.
46. Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. Complement factor H polymorphism in age-related macular degeneration. Science, *308*: 385–9, 2005.
47. Salonen, J. T., Uimari, P., Aalto, J. M., Pirskanen, M., Kaikkonen, J., Todorova, B., Hypponen, J., Korhonen, V. P., Asikainen, J., Devine, C., Tuomainen, T. P., Luedemann, J., Nauck, M., Kerner, W., Stephens, R. H., New, J. P., Ollier, W. E., Gibson, J. M., Payton, A., Horan, M. A., Pendleton, N., Mahoney, W., Meyre, D., Delplanque, J., Froguel, P., Luzzatto, O., Yakir, B., and Darvasi, A. Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium. Am J Hum Genet, *81*: 338–45, 2007.

Chapter 7

Genome-Wide Association Studies

Mark M. Iles

Abstract

Genome-wide association (GWA) studies are best understood as an extension of candidate gene association studies, scaled up to cover hundreds of thousands of markers across the genome in samples usually of several thousand cases and controls. The GWA approach allows the detection of much smaller effect sizes than with previous linkage-based genome-wide studies. However, this sensitivity makes them vulnerable to false positive findings caused by subtle differences between cases and controls that may arise as a result of issues, such as genotyping errors, population stratification, and sample mix-ups as well as the more obvious issue of multiple testing.

After some background and an introduction to GWA, studies are considered stage-by-stage with particular focus on quality control as this is by far the most time-consuming and complex issue related to GWA.

Key words: Genetics, Epidemiology, Genome-wide, Statistics, Association

1. Introduction

There has been a huge increase in the number of genome-wide association (GWA) studies conducted in recent years. Ten were published in 2006, 90 in 2007, 144 in 2008, and 48 up to the end of March 2009. In part, the prevalence of such studies reflects technological developments in genotyping technology, but the driving force behind these can be traced back to two key papers from 1996 (1, 2). In these papers, the authors proposed that common genetic variants may underlie many common traits or diseases and that these would be best found using population-based association studies rather than family-based linkage analysis even though this may require the testing of every gene in the genome (1). This in turn would require identification of all common variants in human genes (2). These proposals gained credence

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713, DOI 10.1007/978-1-60327-416-6_7, © Springer Science+Business Media, LLC 2011

and led to the International HapMap Project (3), with the aim of cataloguing common human genetic variation in a range of ethnic groups. Combined with the latest single nucleotide polymorphism (SNP) chip genotyping technologies, allowing the simultaneous genotyping of hundreds of thousands of markers, HapMap has enabled GWA studies to be conducted, leading to the recent discovery of common genetic variants associated with diseases, such as cardiovascular disease (4–8), breast cancer (9–11), and type II diabetes (6, 12–18).

Fundamentally, GWA studies differ little from traditional case-control studies conducted on candidate genetic regions. However, given that the common genetic variants that cause disease may have quite small effect sizes and the level of multiple testing inherent in genotyping many markers genome-wide, GWA studies require the collection of large numbers of cases with a particular disease and controls. As a result of the association (linkage disequilibrium or LD) between nearby loci not all loci in a region need be typed for most common variation to be captured. Marker (usually SNP) spacing should be dense enough to capture the variation at those loci that have not been genotyped. SNPs may be chosen randomly across the genome or may be chosen specifically for their coverage (using a pilot sample or existing data such as HapMap) in which case they are known as tagging SNPs (19). Studies should be designed in terms of both sample size and marker coverage to have sufficient power to detect common disease susceptibility alleles of modest effect. Genotype data may be analysed in various ways, but the simplest is a comparison of frequencies between cases and controls. These issues are discussed in more depth below.

The strength of GWA is that it represents a method for capturing a new class of disease-associated genetic variants. Pedigree-based association studies utilise families in which disease clusters, so are well powered to find rare variants of large effect. GWA relies on population-based samples, so requires common variants (as rare variants are infrequently observed) of more modest effect, which could not be found using traditional linkage-based approaches.

2. Sample Collection: Sample Size, Stratification, Case Enrichment

The first stage of a GWA study is the collection of suitable samples. Most studies simply compare genotype frequencies in cases and controls for a particular disease. However, if the trait were quantitative, researchers would usually collect either a random sample of individuals or those with extreme values of the trait.

As in any study, it is important that sufficient numbers of individuals are collected to ensure good power to detect an effect and that the cases and controls are reasonably matched. The power of a study depends on the frequency and effect size of the functional genetic variant(s), both of which are unknown. Furthermore, it is likely that despite the large number of SNPs that are now available on commercial platforms the true functional variant is not genotyped; rather we depend on one of the SNPs that has been genotyped being in LD with it and therefore being able to pick up the effect. This results in a further reduction in power, inflating the required sample size proportional to the inverse of the correlation (measured by r^2) between the genotyped and causative markers. How strong this correlation is depends on the coverage of the platform being used. The latest 1,000,000 SNP chips (http://www.illumina.com) capture at least 80% of the variation in more than 90% of the genome in a European sample. As to effect sizes and frequencies, our best indicator is to look at the results of studies that have been conducted. These suggest (20) that the median effect size detected is an odds ratio (OR) of 1.25 (with very few having an $OR > 1.5$) with a median allele frequency of 0.4. Thus, while common variants influencing disease do exist, they have quite small effect sizes requiring large samples. It is highly likely that there are genetic variants that have an even smaller effect on disease than those seen so far, but studies have simply not been large enough to detect these. If we assume we require a p-value of 5×10^{-7} (see Subheading 5 for a discussion of this) and are looking for a variant with an OR of 1.25 and a frequency of 0.4, where we have a SNP in LD with it ($r^2 = 0.8$) we require a sample size of 3,125 cases and 3,125 controls for 80% power (for a $GRR = 1.5$ and frequency of 0.2, we require 1,290 cases and controls for 80% power). Thus, it is rare for studies to be conducted with fewer than 1,000 cases and 1,000 controls.

The next question is how such a sample should be obtained. Ideally, samples should always be collected with the study design in mind. However, this is usually not feasible for such a large study and instead existing collections must be utilised. Since these will not have been collected with GWA in mind, any design issues that are particular to GWA will not have been considered. Foremost among these is the problem of population stratification. Population stratification usually arises when a sample consists of distinct subpopulations between which there is little mating. Differences in allele frequency may then occur by chance between the subpopulations. If the subpopulations are not sampled equally frequently in cases and controls (for instance, one subpopulation is overrepresented in cases but underrepresented in controls), any loci that differ in frequency between these subpopulations may appear to be associated with disease risk. Differential sampling may occur as a result of bad design, by chance or because one subpopulation has a higher incidence of disease (for cultural, environmental, or genetic reasons). For instance, while Europe does not consist of completely distinct subpopulations, there is likely to be more mating between individuals from the same geographic region, thus alleles at many loci vary in frequency across Europe (21). These are too small to have affected previous studies, designed to find large effect sizes, but may give rise to false positives in a GWA study that is designed to find smaller effect sizes. Thus, while in previous studies it may have been enough to ensure that cases and controls were from the same continent or very broad geographic region, in a GWA study matching should ideally be tighter, for instance within country, if this represents a narrow enough ethnic origin. Existing data sets may not record such information, or the definition of ethnicity may not be precise enough, for example, samples from the USA of "European origin". Fortunately, methods exist for detecting population stratification and for correcting this. These are discussed below, but it should be remembered that such approaches are never as satisfactory as a well-matched case-control set. The extreme of this is to use a ready-genotyped set of controls, which can be used against any set of cases. While this has been shown to be relatively unproblematic for a UK set of controls against various sets of UK cases (6), stratification is always a potential hazard. There are also issues regarding cases and controls genotyped at different times (see Subheading 3 below).

The last issue to be discussed with regards to sample collection is "genetic enrichment". Individuals who not only have the disease in question but also have other characteristics may be considered to be more likely to have a genetic basis to their disease. Such characteristics may include an early age of onset, a family history or a greater severity of disease. The main advantage to this approach is that such "genetic enrichment" may improve power by oversampling those individuals who are likely to have more genetic risk factors. However, there are various disadvantages. Firstly, such information may be unavailable, or unreliable, in which case such an approach is not possible. Secondly, there is no guarantee that the factors chosen do indeed "genetically enrich" the sample, at least not for the common low penetrance variants that GWA is well powered to detect. Thirdly, choosing only those individuals who are "genetically enriched" may lead to a smaller sample size, so there is a trade off here in power. Finally, there is no guarantee that the results from a genetically-enriched sample are applicable to the general population. Thus, the usefulness of "genetic enrichment" depends on the disease being studied and the samples available.

3. Genotyping: Interpretation of Scatter Plots, Calling Algorithms, Indicators of QC Problems

Once the samples have been collected, the next stage is to genotype them. Genome-wide SNP chips output the results of each genotype in terms of the intensity of the two alleles at the SNP. Those who are homozygous have a high intensity for that allele and low for the other, while those who are heterozygous have an intermediate intensity for both. If the genotype intensities are plotted for all individuals at a particular SNP, these appear as three clusters of points (a cluster plot). Genotypes are called based on being within one of these clusters. The boundaries of each cluster may be declared based on either the current genotype data or on previously genotyped data and any genotypes lying outside of these boundaries is declared as "uncalled" and so is treated as missing data (see Fig. 1). Usually, calling of genotypes

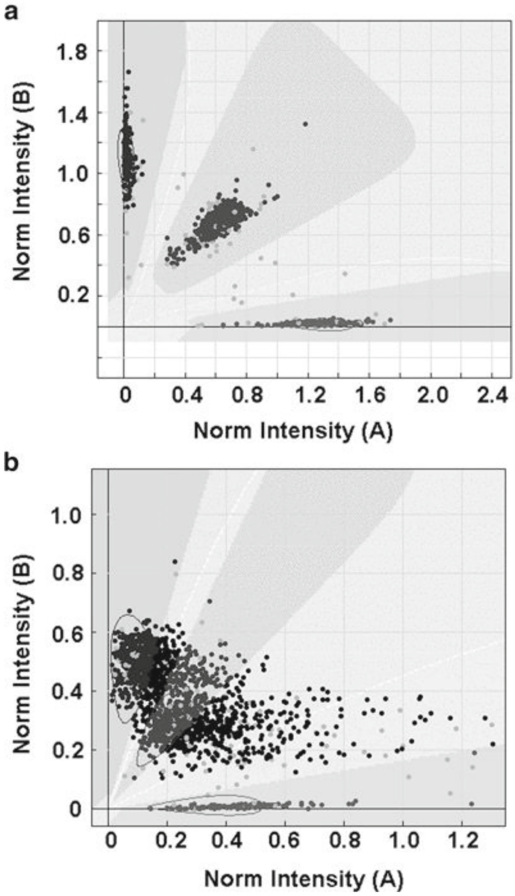


Fig. 1. Two examples of cluster plots. The *x* and *y* axes show the intensities for the two alleles so the three genotypes should form three separate clusters. (**a**) Is an example of good genotyping: the clusters are well-defined and most genotypes are called. (**b**) Is an example of poor genotyping: the clusters overlap to the extent that it is not even clear how many clusters there are.

is performed automatically by software provided by the genotyping company. Problems may arise when minor allele frequency is low, in which case one homozygote may not be observed, and so identification of clusters becomes harder. Other problems may be caused by low sample quality or particular problems in genotyping some SNPs in which case clusters may be difficult to distinguish. A further issue that has been identified (6, 22) occurs when samples come from various sources and have been handled in different ways or were collected at different times. This can lead to a consistent difference in intensities between samples producing, in extreme cases, six clusters rather than three. Suggestions for dealing with this are to either call genotypes for different groups (even cases and controls) separately (22) or to account for these potential strata within the data in the calling algorithm itself, as in CHIAMO (http://www.stats.ox.ac.uk/~marchini/software/gwas/chiamo.html; (6)).

Problems in calling genotypes may not be identified by the calling algorithm itself, but may give rise to false positive results, particularly when there are differences between the calling of case and control samples. These are often readily apparent when the cluster plots are examined, but with one cluster plot for each SNP, this would require individual examination of hundreds of thousands of plots. Instead other indicators of genotyping problems are usually employed and the cluster plots only examined for those SNPs that show strong evidence of association after statistical analysis as a final quality check.

Such indicators of genotyping problems may be low SNP call rates, deviation from Hardy–Weinberg equilibrium (HWE) or low minor allele frequency. There are no strict rules for what exact values constitute low quality genotyping. For instance, out of 600,000 SNPs you would expect 600 to reach a HWE *p*-value of 0.001 simply by chance, even in the absence of any genotyping problems. Thus, we would recommend not excluding SNPs on the basis of such measures until after statistical analysis for association has been conducted. At this stage, those SNPs that appear to be strongly associated can then be considered more carefully in terms of their various quality measures. More often than not those SNPs that do poorly in terms of one quality measure do poorly in terms of others.

It is important, however, to exclude individuals on the basis of the quality of the sample before statistical analysis is conducted. So, for example, those samples missing more than 5%, or even 1% of the genotypes when considering all the tested SNPs may be excluded. This may seem quite stringent but it should be remembered that this is not because the data is missing “at random” but because this indicates potential genotyping problems that might affect the genotyping of all SNPs in that sample. Only a very small proportion of samples is excluded even with such strict QC.

Another check that ought to be conducted on samples is that none are duplicates or closely related. Any pair of samples that are closely related shares far more alleles identical by state than an unrelated pair. This is easily implemented using software, such as PLINK (http://pngu.mgh.harvard.edu/purcell/plink/; (23)). Duplicates may arise because individuals have accidentally been recruited more than once into the study or because samples have been mixed-up. Another simple check for sample mix-up is to compare individuals' sex as recorded in their phenotype and their sex according to heterozygosity of the X chromosome (this is also implemented in PLINK, http://pngu.mgh.harvard.edu/purcell/plink/; (23)). Further checks for genotyping problems can be applied when the sample being analysed consists of subsets defined either by case/control status, geographic origin, stratified genotyping or sample collection. The various quality measures discussed may be applied to each subset, in case one subset is of low quality, and this is not identifiable when all groups are considered together. Alternatively, there may be heterogeneity between subsets in terms of quality measures or even minor allele frequency, also indicating possible quality problems.

4. Identifying and Correcting for Population Stratification

As mentioned in Subheading 1, population stratification may give rise to false positive associations between disease phenotype and genotype. While careful study design (in terms of well-matched cases and controls) is the best approach to dealing with this, it is possible that ethnic/geographic matching is either not possible or that even if this is done there remains so-called cryptic stratification, which is not identifiable from phenotype data. It is possible, however, to test for stratification and either to exclude population outliers or to some extent correct for such stratification. By far the simplest method for identifying population stratification is to check for deviation from HWE, which occurs as a result of population stratification. However, inflation of false positive association rates may be caused by quite modest levels of stratification, which are not detectable as significant deviations from HWE.

Another simple approach is that of the Q–Q (quantile–quantile) plot. Here, the test of association is conducted and the ordered test statistics for all SNPs (having excluded those of low quality) are plotted against their quantiles. While some SNPs may genuinely be associated with disease, the vast majority should not. Thus, a deviation from the $x=y$ line indicates that there is some inflation of the test statistics above that expected under the null hypothesis of no association. Such inflation is usually measured by

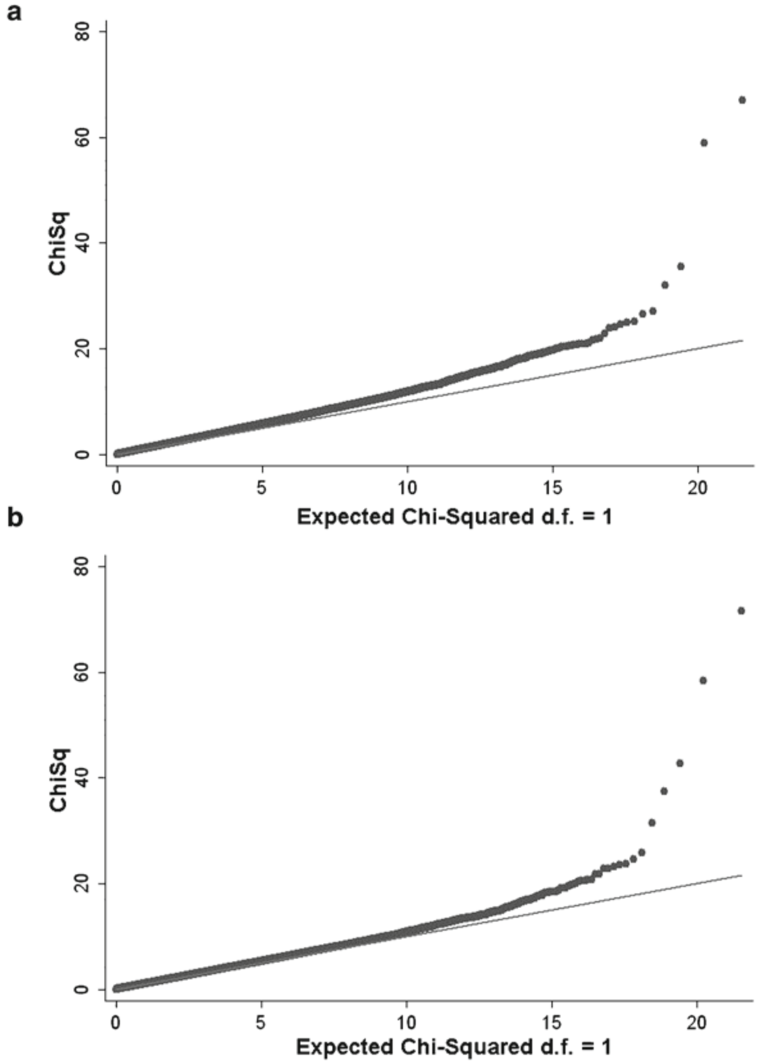


Fig. 2. Examples of Q–Q plots. The data come from a genome-wide study of melanoma, with samples from eight different centres in Europe and Australia (30). Observed chi-squared values are plotted against expected values. (**a**) Shows the results when an unstratified analysis is conducted, not taking account of the potential differences in allele frequencies between centres. Here, λ = 1.14. (**b**) Shows the results when a stratified analysis is conducted, adjusting for potential differences in allele frequencies between centres. Here, λ = 1.06. The results suggest that there is some stratification in the sample, but that this is mostly accounted for by differences between centres.

dividing the median of the test statistics by its expectation under the null hypothesis, denoting this ratio λ (see Fig. 2). This gives rise to the genomic control method for correcting for such inflation, whereby each test statistic is divided by λ (24, 25). However, it has been shown that, in part because the inflationary effect of stratification varies from marker to marker and λ is an averaging of this, genomic control overcorrects those markers that are not very stratified and undercorrects those that are more stratified. Thus, the approach results in the test both losing power and having an inflated false positive rate (26).

Other approaches include assigning individuals to theoretical subpopulations and testing conditional on these (27, 28), but such approaches assume that the sample consists of distinct subpopulations (which may not be true) and, crucially, are highly computationally intensive.

By far the most popular approach for detecting and adjusting for population stratification is the application of principal components analysis (PCA) (26). The idea is that individuals who are geographically close are likely to be more correlated in terms of genotypes (i.e. closely related) than those who are far apart. Even if there is only a slight correlation on a SNP-by-SNP basis, when this is considered genome-wide, it may be great enough to distinguish between subpopulations (when these are distinct) or reveal gradients in SNP frequencies when there are no distinct subpopulations. The first principal component gives the linear combination of genotypes that best captures the variation in the data. The second principal component is the orthogonal combination that best captures the remaining variation in the data and so forth.

Thus, the sample may be combined with data from across the world (for example, the HapMap data (http://www.hapmap.org) which includes samples from Europe, Asia, and Africa) and PCA applied to the combined sample. This identifies individuals that are ethnically distinct from the rest of the sample. For example, in a study of a European population, combining with the HapMap data and applying PCA produces principal components 1 and 2 that separate out the continents into three distinct clusters. Anyone who is not of European origin appears in one of the non-European clusters (6). Ethnic outliers from the sample can then be excluded from further analyses.

Once outliers have been removed to give a more homogeneous sample, PCA can be reapplied to detect more subtle stratification. It has been shown (21, 29) that the first and second principal components are likely to correspond to orthogonal two-dimensional geographical axes, such as latitude and longitude. Further principal components may correspond to further orthogonal axes, depending on the structure of the population. It is important when applying PCA that the SNPs are first thinned so that LD is minimised, otherwise the principal components may just pick out regions of strong LD. The weightings within the components are also interpretable as representing regions that are particularly stratified, perhaps even having undergone selection. For instance, within both the UK (6) and Europe (30) it has been shown that there is particularly high population stratification around the lactase gene (which affects lactose intolerance and is well known to have undergone ancestral selection), and HLA, which has also undergone selection. In Europe, PCA distinguishes individuals from different countries extremely well (21, 30). The principal components may then be used as covariates in a logistic regression on disease status, for instance, to adjust for potential stratification.

5. Analysis: Testing for Association and Interactions Estimating Effect Size

Once any problematic samples have been removed, association analysis can begin. Various complex association analyses may be applied to candidate gene studies but, given the number of SNPs that are tested and the number of samples involved in a GWA study, only very basic fast tests are currently computationally feasible. The simplest of these is to test for an association between genotype and case-control status using Pearson's chi-squared or equivalent. However, since this has two degrees of freedom in most cases (where all three genotypes are observed), the Cochran–Armitage trend test is more commonly applied. Here, a log additive mode of inheritance is assumed, so the test is asymptotically equivalent to a logistic regression of case-control status on genotype measured as a continuous trait (taking the value 0, 1, or 2). While this makes the model less flexible and therefore likely a worse fit to the data, this is traded off against increased power because it has only one degree of freedom. Various authors have suggested simple alternatives to be more powerful (31, 32), such as applying tests assuming additive, recessive, and dominant modes of inheritance and picking the best of these, but the Cochran–Armitage trend test remains by far the most popular. Other possible approaches would be to apply multilocus tests, under the hypothesis that some disease risk is due to interaction effects that are undetectable when analysing one SNP at a time (33). Again, these are not widely applied and are unlikely to be used in a primary analysis.

The next question is the interpretation of such results. The usual approach would be to look at those SNPs with the most significant *p*-values. Given that hundreds of thousands of SNPs have been tested, the cut-off for significance must be quite stringent. A simple Bonferroni correction is conservative because SNP genotypes are not independent (due to linkage disequilibrium). There is no agreed cut-off for GWA studies, although the values used are generally of the order of 10^{-5}–10^{-8} (6, 7, 11) based on both frequentist and Bayesian arguments.

At this stage, the various SNP quality measures may be checked along with cluster plots (see above) to ensure as far as possible that significance is not a result of genotyping error.

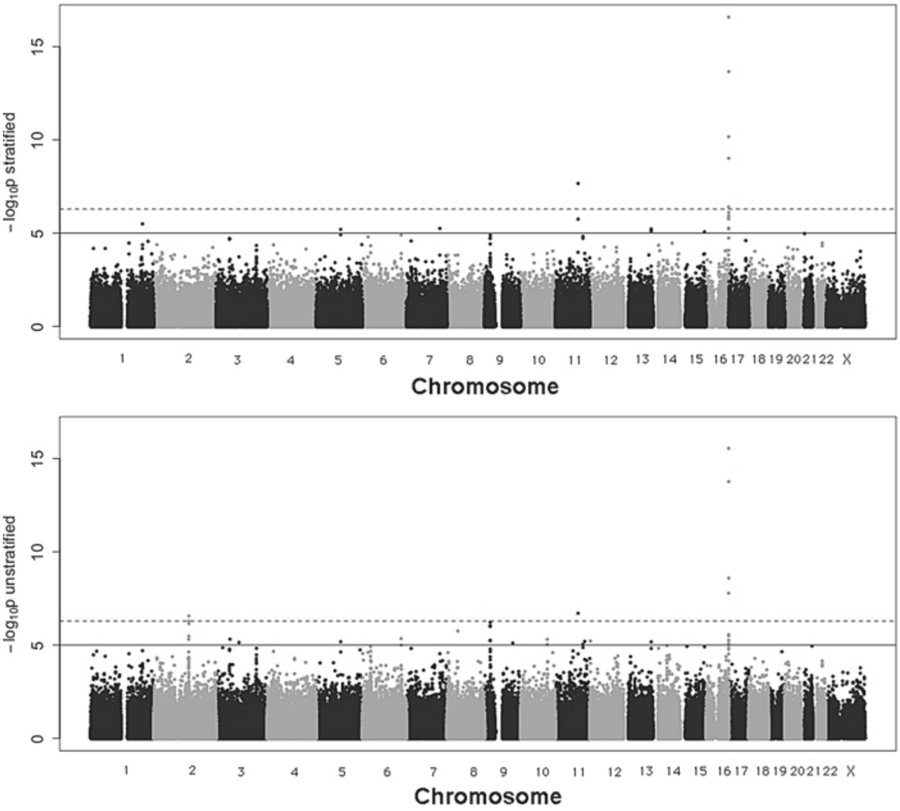


Fig. 3. Manhattan plot showing-$\log_{10}$ *p*-values plotted against genomic location. Again the data are from a melanoma genome-wide association study (30), with results adjusted for stratification. Alternating colours are used simply to distinguish chromosomes more easily. The more convincing findings are those where there are multiple SNPs with high significance. Lone SNPs with high significance are more likely to be due to genotyping error.

Having several significant SNPs in a region is further evidence that genotyping error is not the cause, but is no guarantee that population stratification or a simple statistical false positive is not the reason. Such results can be displayed as a Manhattan plot (Fig. 3).

It is also of interest to estimate the contribution of any replicable genetic variant to disease risk. Most commonly this would be by OR or relative risk, either for each genotype or assuming an additive risk model. Many authors also estimate in some way the proportion of overall disease risk that is explained by the variant using a measure such as heritability (or sibling relative risk) or population attributable risk. Although easily stated these are not always easy to interpret and may vary widely. For instance, a rare variant with a large effect size contributes greatly to heritability but little to attributable risk due to its rarity. A common variant with small-to-moderate effect size may have a high attributable risk because it is common but a very small effect on heritability because of its small effect size. Population attributable risks (PAR) reported from GWA studies tend to be high as the variants found are common: 0.54 for Restless Legs Syndrome (34), 0.38 for Coronary Artery Disease (7), and 0.13 for Prostate Cancer (35); while measures of the proportion of the genetic risk are lower: excess familial risk of 0.036 for Breast Cancer (11) and 0.002 of the variance in risk for Multiple Sclerosis (36). Tellingly, estimates of PAR for the replicated SNP found for colorectal cancer vary between 0.11 and 0.42 (because of differences in frequency between populations), while explaining only 0.009–0.018 of the increased risk to siblings of cases (37). It should be remembered that the ORs estimated in the initial studies that find the causative locus are likely to overestimate: the so-called winner's curse (38, 39). Estimates from replication data are less biased.

Further analyses may include subgroup analysis if, say, certain phenotypic subtypes are suspected to have a different genetic basis and checking of heterogeneity of ORs (by Mantel–Haenszel test) across either disease subtypes or geographic regions. Adjustments may be made, both for *p*-values and ORs for phenotypic covariates, such as sex, age, etc., and for principal components by logistic regression; the latter to ensure that stratification does not inflate false-positive rates.

6. Replication and Imputation

It is common to require independent replication of any significant results, giving further confirmation that the result is not due to either chance or error and allowing independent estimation of effect size (40). In terms of classical epidemiology, this may seem odd, as it could be seen as equivalent to splitting a dataset into "testing" and "replication" sets which, while on the surface attractive, have less power than simply analysing everything together. But this ignores the special nature of GWA data – a replication dataset allows confirmation that genotyping was not at error or that subtle stratification has not occurred. Furthermore, the replication data may be ethnically distinct from the rest of the data and so not well suited for combining. Finally, genome-wide genotyping is, of course, far more expensive than genotyping a single SNP, so that there may be more samples available to genotype genome-wide than can be afforded. In this instance, the remainder may be genotyped at the few most significant SNPs after the GWA. For these reasons, it is common for replication in one or more independent data sets to be published along with the initial study.

It is at this point that the initial study may finish, but this is only really the beginning of understanding the results of GWA.

While one or more hits in a region at a suitable significance level and replication in independent datasets indicate genuine findings, it is unlikely that the causative locus has been identified. It is more likely that the most significant loci are simply in strong LD with the causative locus (or loci). The obvious next step would be to conduct further dense genotyping or even sequencing (or perhaps sequencing a small sample to discover all common SNPs in the region followed by genotyping these in the full sample) to narrow down the location of the true causative locus. Such approaches may be prohibitively expensive, particularly if the region in question is large.

One alternative is to impute the genotypes at those SNPs that have not been typed. By using a dataset, such as HapMap (http://www.hapmap.org) or the 1,000 genomes project (http://www.1000genomes.org) which contain far denser genotype data than is currently available on commercial SNP chips, the pattern of LD between nearby SNPs can be established and then applied to the sample data such that SNPs that have not been genotyped may be estimated ((41); http://www.stats.ox.ac.uk/~marchini/software/gwas/impute.html; http://www.sph.umich.edu/csg/abecasis/MACH/). Those SNPs that are estimated with suitable confidence in sufficient samples can then be treated as though genotyped and the data analysed as usual, although it is more correct to account for this uncertainty. Such an approach offers a way of "genotyping" extra SNPs within a dataset and hopefully facilitating further narrowing down of the location of the causative locus. Researchers should of course be more cautious about the results at SNPs that are imputed rather than genotyped. They are only as good as the dataset from which they have been imputed which may either be quite small relative to the researchers' sample or may not ethnically be a good match. Thus, the estimates of SNP frequencies and the LD patterns between them suffer. Furthermore, care should be taken in the sample to remove low quality SNPs before imputation – imputation based on a SNP that has been badly genotyped is unreliable. Imputation also allows the combination of samples that have been genotyped on different sets of SNPs (i.e. on different SNP chips). This allows analysis of more of the data than a simple meta-analysis based on *p*-values at common SNPs.

7. The Future

Clearly, GWA has been a success. Many new genetic variants have been identified that are associated with a wide range of diseases (20). The obvious extension to this is to continue to apply the method to further diseases and to increase sample sizes as far as possible such that ever smaller effect sizes and rarer variants may be identified. But eventually there must be some limit to this – samples cannot increase in size indefinitely. Then, we need to either use other sources of information (such as bioinformatic tools) or approach the analysis of such data in a different way focusing, for instance, on the discovery of rarer variants. But it is likely that GWA will reveal more about the underlying genetic cause of diseases for some time.

References

1. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281): 1516–1517.
2. Lander ES (1996) The new genomics: global views of biology. Science 274(5287): 536–539.
3. International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796.
4. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R et al (2007) A common allele on chromosome 9 associated with coronary heart disease. Science 316(5830): 1488–1491.
5. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T et al (2007) A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316(5830): 1491–1493.
6. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
7. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M et al (2007) Genomewide association analysis of coronary artery disease. N Engl J Med 357: 443–453.
8. Matarin M, Brown WM, Scholz S, Simon-Sanchez J, Fung HC et al (2007) A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. Lancet Neurol 6(5): 414–420.
9. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M et al (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39(7): 870–874.
10. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J et al (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39(7): 865–869.
11. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087–1093.
12. Sladek R, Rocheleau G, Rung J, Dina C, Shen L et al (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445(7130): 881–885.
13. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI et al (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316(5829): 1331–1336.
14. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316(5829): 1336–1341.
15. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y et al (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316(5829): 1341–1345.
16. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316(5826): 889–894.
17. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T et al (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39(6): 770–775.
18. Salonen JT, Uimari P, Aalto JM, Pirskanen M, Kaikkonen J et al (2007) Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium. Am J Hum Genet 81(2): 338–345.
19. Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype

tags: a class of tests and the determinants of statistical power. Hum Hered 56: 18–31.

20. Iles MM (2008) What can genome-wide association studies tell us about the genetics of common disease? PLoS Genet 4(2): e33.
21. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456(7218): 98–101.
22. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JMM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37: 1243–1246.
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559–575.
24. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
25. Devlin B, Roeder K (2001) Genomic control: a new approach to genetic-based association studies. Theor Pop Biol 60: 155–166.
26. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8): 904–909.
27. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170–181.
28. Satten G, Flanders WD, Yang O (2001) Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68: 466–477.
29. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40(5): 646–649.
30. Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, Randerson-Moor J, Aitken JF, Avril MF, Azizi E, Bakker B, Bianchi-Scarrà G, Bressac-de Paillerets B, Calista D, Cannon-Albright LA, Chin-A-Woeng T, Dębniak T, Galore-Haskel G, Ghiorzo P, Gut I, Hansson J, Hočevar M, Höiom V, Hopper JL, Ingvar C, Kanetsky PA, Kefford RF, Landi MT, Lang J, Lubiński J, Mackie R, Malvehy J, Mann GJ, Martin NG, Montgomery GW, van Nieuwpoort FA, Novakovic S, Olsson H, Puig S, Weiss M, van Workum W, Zelenika D, Brown KM, Goldstein AM, Gillanders EM, Boland A, Galan P, Elder DE, Gruis NA, Hayward NK, Lathrop GM, Barrett JH, Newton Bishop JA (2009) Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet 41(8): 920–925.
31. Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V (2008) Maximising association statistics over genetic models. Genet Epidem 32: 246–254.
32. Bacanu S-A, Nelson MR, Ehm MG (2008) Comparison of association methods for dense marker data. Genet Epidem 32: 791–799.
33. Lunetta KL, Hayward BL, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 5: 32.
34. Stefansson H, Rye DB, Hicks A, Petursson H, Ingason A (2007) A genetic risk factor for periodic limb movements in sleep. N Engl J Med 357(7): 639–647.
35. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D et al (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39: 631–637.
36. International Multiple Sclerosis Genetics Consortium (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357(9): 851–862.
37. Haiman CA, Le Marchand L, Yamamato J, Stram DO, Sheng X et al (2007) A common genetic risk factor for colorectal and prostate cancer. Nat Genet 39: 954–956.
38. Zöllner S, Pritchard JK (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80(4): 605–615.
39. Garner C (2007) Upward bias in odds ratio estimates from genome-wide association studies. Genet Epidemiol 31: 288–295.
40. NCI-NHGRI Working Group on Replication in Association Studies (2007) Replicating genotype-phenotype associations. Nature 447: 655–660.
41. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nat Genet 39: 906–913.

Chapter 8

Candidate Gene Association Studies

M. Dawn Teare

Abstract

Candidate gene association studies aim to establish or characterise association between the genetic variation occurring within a specific gene or locus and a phenotype. If the phenotype is quantitative, then the effect size is often measured as the difference between the genotype specific means or a per allele effect. When the phenotype is binary and the disease is either present or absent, the effect is summarised as a genotype specific risk or relative risk. This chapter focuses on methodology employed when a single or small number of genetic loci are being investigated for an association with a specific phenotype.

Key words: Odds ratio, Relative risk, Genotype specific relative risk, Case–control study, Haplotype risk

1. Introduction

Candidate gene association studies aim to establish or characterise association between the genetic variation occurring within a specific gene or locus and a phenotype. Historically, genetic association studies became popular because it was clear that the genetic variants underlying risk of complex disease were likely to have individually weak effects, making each effect difficult to detect through the linkage approach. Weak effects due to common variants are possible to detect through genetic association studies; however, as the variant becomes rarer, this is also difficult for association studies (1). Many of the early positive reports of association studies suffered from being underpowered resulting in few convincing replications (2). Exactly what constituted a candidate gene was open to argument, and therefore how to control for the potential problem of multiple testing was also unclear as the size of the set of candidate genes was not defined. Since the onset of genome wide association studies (GWAS) the "prior interest"

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_8, © Springer Science+Business Media, LLC 2011

can now be justified and to some extent quantified. The further study of regions of the genome associated with risk of disease is now required to see which of the GWAS "top hits" are replicable and hence more likely to be genuine associations.

GWAS are designed or powered to detect weak effects, and hence the samples used are frequently not representative of the population. By contrast, the candidate gene study generally requires representative population-based samples, either cohort samples for quantitative traits or disease-based sampling for case–control studies. In Chapter 2, we have introduced a number of terms which define a genetic component. When considering major genes, we present the effect of one or two copies of the risk allele by estimating the penetrances of specific genotypes for a binary trait and estimating the mean increase (or decrease) in trait value for a quantitative trait. Many phenotypes can be studied as a quantitative or qualitative trait. For example, obesity is frequently measured on the body mass index (BMI) scale; however, when the BMI exceeds a particular threshold, the individual can be classed as obese. The approaches to study the two forms of phenotype are considered separately as the methods employed are slightly different. For simplicity in the following examples, we assume that a candidate gene or locus has only two specific alleles, though all methods can be extended to multiple risk allele systems.

2. Quantitative Traits

In the genetic analysis of quantitative traits, we may be interested to identify the proportion of total trait variance attributed to the variation at the candidate risk locus. This would be equivalent to an estimate of heritability. Put another way, we may want to estimate the genotype specific mean values of the trait. This question can be addressed through linear regression. The measured trait in individual i is denoted by y_i and the genotype by g_i, which takes values 0, 1, or 2 reflecting how many risk alleles are present. The linear model takes the form

$$y_i = \alpha + \beta_1(g_i = 1) + \beta_2(g_i = 2) + \beta_c x_i + \cdots + \varepsilon_i$$

where β_1 is the effect on the trait value when an individual carries one copy of the candidate gene risk allele, β_2 is the effect of two copies of the risk allele on the trait, and $\beta_c x_i$ represents the term for the effect on the phenotype of another known covariate, such as age or sex. Further covariates can be added to the model as required. The final term ε_i is the random error term, assumed to be normally distributed. This parameterisation enables us to specify the genotypic means of the trait directly. While this form of analysis could be done by simple analysis of variance, the linear regression model allows additional covariates to be taken into account. The linear model has the further advantage that the trait distribution is not required to follow a normal distribution, only the residual error term is required to be normally distributed (3). An alternative parameterisation is to use additive and dominance coefficients.

$$y_i = \alpha + \beta_1 g_i + \beta_2 (g_i = 2) + \beta_c x_i + \cdots + \varepsilon_i$$

Now β_1 represents the additive allelic effect and β_2 is the dominance effect (4).

3. Case–Control Studies

While many quantitative traits have been studied, the vast majority of these are of interest because they are regarded as a surrogate marker for disease risk, or that a trait value above a specific threshold is strongly associated with clinical disease. So the more frequently encountered study design in genetic epidemiology (as with classic epidemiology) is the case–control study design. This form of study uses targeted sampling, setting the number of cases and controls to optimise statistical power. In Chapter 2, we discussed how the penetrance function was used in the study of binary traits. This is the conditional probability of disease in an individual with a specified genotype, an absolute risk. Some studies do report penetrance probabilities, for example, the impact of the BRCA1 gene on an individual's risk is quite significant. The probability of a woman developing breast cancer by age 70 if she carries one high risk allele for BRCA1 is estimated to be 65% with a 95% confidence interval (CI) of 44–78% (5). This is a large risk compared to the average lifetime risk of breast cancer for women in the UK of 11% (6). If we compare the report of the effect of BRCA1 on disease risk with the more recently established association of the CASPASE-8 gene genetic variant CASP8 D302H (7), we find that these results are reported in a slightly different way. A comparative effect is reported rather than an absolute effect. In this example, the summary effect is reported as an odds ratio (OR) of 0.89 (95% CI: 0.85–0.94) for carriers of one allele and an OR of 0.74 (95% CI: 0.62–0.87) for carriers of two alleles, when compared with common homozygote genotype. The results are reported in this comparative manner because the case–control design prevents the estimation of absolute risk. The reason for this is outlined in the next section, which gives a concise overview of the methods and terminology used in case–control studies.

4. Terminology and Methodology Used in Case–Control Studies

4.1. Disease Risk

The term *risk* is widely used in a number of disciplines though in epidemiology (and genetic epidemiology) it has a precise meaning. *Disease risk* is generally used in the context of binary traits to mean the probability of new or incident disease occurring within a specified time period in a defined group or subgroup of a population, all of whom are disease free and therefore "at risk" at the beginning of the time interval. From this definition it is clear that *risk* is a *cumulative probability* and hence a dimensionless parameter which must lie between 0 and 1. An alternative way of characterising risk is to consider the *incidence rate* of new cases of disease per unit of time, often reported in units of *person-years at risk*. Models that take account of incidence (or mortality) rates are based on probability density models. When the incidence rate is considered over shorter and shorter time intervals this converges to the *hazard function* (a function of time) which can be thought of as the instantaneous probability of developing disease, conditional upon having survived disease free up to the instant before. This model allows the hazard rate to vary with time which makes it very suitable for the analysis of disease with variable age at onset. Statistical survival models can be used to derive the cumulative disease risk from the hazard function.

When population based sampling is used the risk or hazard can be directly estimated. However, as argued in Chapter 3, population-based sampling is costly as a means to study factors associated with disease risk when the disease is rare. A more efficient design is to sample cases with disease and cases without, then compare the frequencies of exposures in the two groups to infer association. By sampling in this way, you can estimate the frequency of the exposure in cases and controls, but as you have fixed your numbers of cases and controls you cannot directly estimate the absolute risk of disease conditional on exposure. To illustrate the sampling, Table 1 shows genotype counts in a UK case–control study for lung cancer (8). In this example, their cases are restricted to "ever smokers". From these data, we can estimate the probability of each of the three genotypes in lung cancer patients, we can report the estimated probability or risk that an "ever smoker" lung cancer case has genotype AA is 0.37. This knowledge appears to be not very useful as we would really like to know the risk of disease in those individuals with genotype AA. This is how penetrance is defined in Chapter 2. At first sight we are not even able to estimate the relative risk of disease in one exposure group vs. another (for example AA vs. AG), as the relative risk is a ratio of two penetrance probabilities. This apparent difficulty is resolved by reporting the Odds Ratio (OR). The OR is defined as the odds of disease in exposed divided by the odds of

Table 1
Genotype-specific odds ratios in ever smoking lung cancer cases compared to population-based controls. The cases and controls are sampled from a UK population (8)

SNP genotypes rs8034191	Lung cancer cases (ever smokers)	Controls	Odds ratio (95% confidence interval)	*p* Value
AA	670	448	1.00 reference	
AG	858	415	1.38 (1.17–1.63)	
GG	303	97	2.09 (1.61–2.70)	1.5×10^{-8}

disease in unexposed. In Table 1, we define genotype AG as the exposure and genotype AA as not exposed, the estimated OR is $(448/415)\div(670/858)$, or 1.38.

While the OR appears to be difficult to interpret, its common utility comes from three features. Firstly, that it approximates the relative risk (or risk ratio) when the disease is rare, secondly in case-cohort designs it estimates the relative hazard or rate ratio without the rare disease assumption, and thirdly, the OR is the natural parameter in the logistic regression model which is discussed in the next section. Case-cohort designs are very common in genetic epidemiology where the "control" group are disease free at the beginning of the risk period only (9).

4.2. Logistic Regression

Candidate gene case–control analyses can be reported as shown in Table 1 with statistical significance computed through a chi-squared test on two degrees of freedom. However, if the disease risk is known to be modified by other factors, such as age and sex, it is common to see "adjusted" ORs reported. This is frequently achieved by using a general linear model with a logit link, otherwise known as logistic regression. In this model, the natural logarithm of the odds of disease can be modelled by a linear function of the independent variables. If we denote the genotype specific probability (or risk) of disease as π_i, where subscript i indicates the genotype class, the expression takes the form:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\alpha+\beta_1(g_i=1)+\beta_2(g_i=2)+\beta_c x_i$$

In this framework, the error terms are binomially distributed rather than normally distributed (10). You can see that the right side of this equation takes a similar form to that used by linear regression in quantitative traits. However, in the logistic regression analysis the coefficient terms are the natural logarithm of the ORs. The general linear model framework means that binary and quantitative traits can be analysed in an essentially similar way. These statistical models can estimate the effect of the genetic variant on phenotype while adjusting for other known modifiers or covariates. More importantly, several independently conducted replication studies can be jointly analysed to produce a combined single estimate of the impact, or relative impact, of the genetic variant on phenotype.

The breast cancer CASP8 report referred to earlier (7) is a good example of how logistic regression can be used to jointly consider the evidence from many studies (in this case up to 15) to study the relative effect of extensive covariate measurements and investigate potential statistical interactions. However, among the 15 studies they included there were many differences in the type of covariate data collected so to report their summary evidence they used the meta-analysis technique.

5. Meta-Analysis

Statistical meta-analysis is used to bring together the summary results of similar studies to find an overall more precise estimate of the effect size in question. Meta-analysis techniques require only summary data, such as the point estimate and standard error, for each study and then a weighted analysis can be performed to estimate the overall effect (11). This can be a useful approach in many circumstances. In collaborative replication studies, it allows a weighted analysis of all the evidence, and the summary analysis produces a "forest" plot similar to that shown in Fig. 1. Here, 15 hypothetical studies are presented. The summary statistics and 95% confidence intervals for each study are visually presented in a columnar format. The study point estimate is displayed as a darkened square, and the size of the square is inversely related to the standard error associated with the point estimate. The line extending from each square covers the numerical values included in the study specific 95% confidence interval. Hence, larger squares tend to have narrower confidence intervals. Studies 4, 10, and 15 are relatively small, whereas 3, 7, and 13 are relatively large. In this example, the 15 studies are in random order, but it is common to see studies ordered by sample size or date of publication. The figure shows that though all of the point estimates are above the null value of zero, the log of the OR when equal to 1, eight of the studies would not reject the null hypothesis of no association. Taking the 15 studies together, a weighted analysis results in a summary point estimate (log (OR) = 0.089) and 95% CI (0.064–0.118). The overall summary estimate is represented as a diamond and the width of the diamond covers the range of values in the 95% confidence interval. Taking all 15 studies together, there

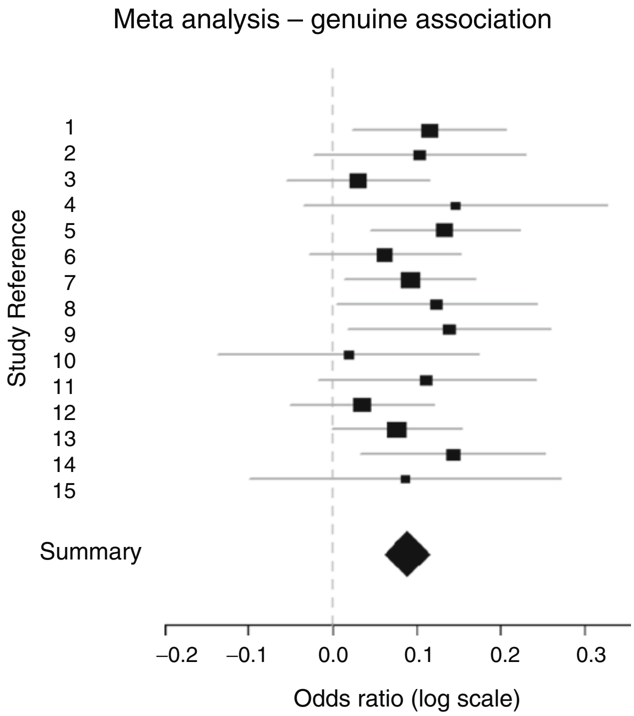


Fig. 1. Results of 15 hypothetical similar studies presented in a forest plot. Each study is represented by a *horizontal line* and *central square*. The studies are of varying sizes, the larger the sample size, the larger the *central square.*

is strong evidence against the null hypothesis of no association. By contrast, Fig. 2 shows an example analysis of 15 studies when there is no underlying association.

Meta-analysis methods have developed from the need to quantitatively synthesise a number of study results. These techniques have frequently been used to summarise published evidence from similar studies or clinical trials. Relying only on published associations can lead to a concern of publication bias. Publication bias arises when many groups may study a particular hypothesis but only those reporting a statistically significant effect are accepted for publication. Figure 3 shows a rather exaggerated example of publication bias. In this example, the studies are listed in publication date order with study 1 the earliest and study 15 the latest. From the figure you can see that the first five studies all show a statistically significant effect, though their small sample size leads to wide confidence intervals. Once several positive reports are published, larger studies attempt to replicate the association for a more precise estimate of the effect. In this figure, the analysis of the 15 studies appears to show a marginally significant effect, as the diamond alongside the "Summary" appears to exclude zero. In this example, the result is likely to be strongly influenced by the

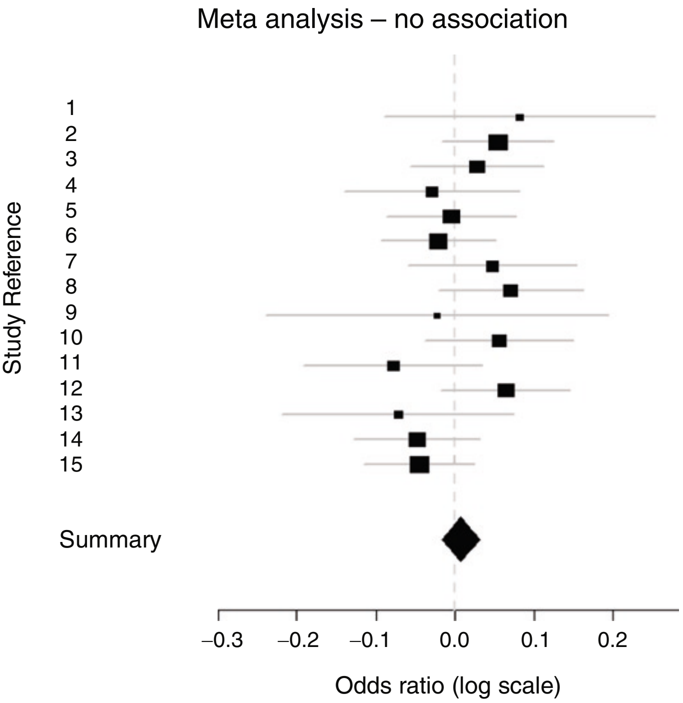


Fig. 2. Results of 15 hypothetical studies presented in a forest plot. Studies have been generated under the null model of no association.

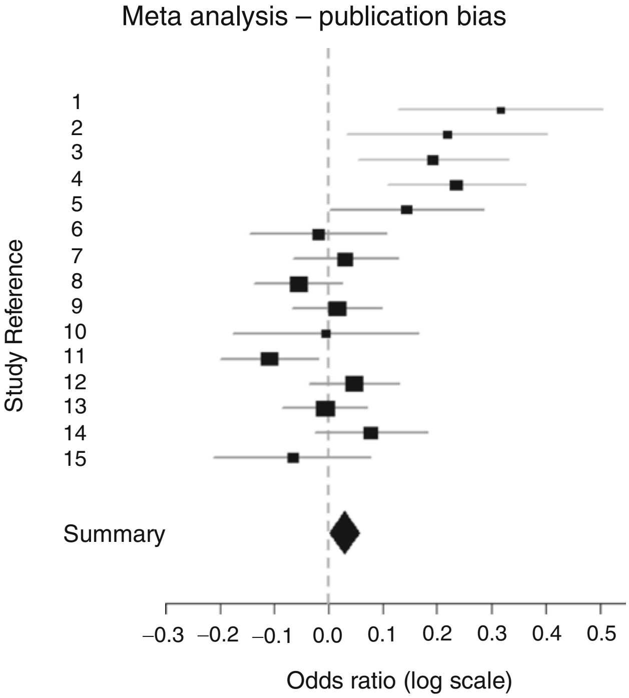


Fig. 3. Results of 15 hypothetical studies presented in a forest plot. The studies are listed in publication date order.

first five reports. Formal techniques for testing for publication biases include rank correlation tests and visual funnel plots (12).

Before the onset of GWAS, meta-analyses of candidate gene studies were also potentially prone to publication bias. However, one major advantage of the GWAS approach has been the general shift towards two-stage, collaborative designs. The first stage involves dense genotyping in a small highly selected sample. The strongest signals are then followed up in many independently conducted but essentially collaborative studies through Consortia (13). These multi-centre studies are then published jointly and frequently use the meta-analysis technique to allow for the very different study designs employed.

6. Association with Specific Genotypes and Haplotypes

Candidate gene association studies the variation within a candidate region captured by one or more polymorphic loci. Though genetic variation can take many forms, the most frequently studied variants are currently of the SNP class. Chapter 6 has described how to select SNPs that collectively tag or capture as much of the variation present within a region. These SNPs can then be analysed separately or jointly as haplotype blocks suggest appropriate groupings. The consideration of haplotypes made up of the allelic states of several SNPs within a candidate region can be handled in two different ways. The haplotypes may be considered equivalent to a multi-allelic polymorphic marker, such as a micro-satellite, and risk is then examined with respect to each of these specific haplotypes (14), or an LD mapping approach can be used which considers all the variation seen in the gene but allows either a sliding window (15) or variable length haplotype method (16) to identify the locus showing the highest evidence for retention of a putative ancestral haplotype. It is generally assumed that the variant is either the causal variant or is in very strong LD with something that is causal.

Humans are diploid, and have two copies of each of the autosomal chromosomes. The alleles at each locus are therefore linearly arranged on the two chromosomes though the phase cannot easily be observed directly. When considering the functional consequences of genetic variation for example, in gene expression, *cis* effects (i.e. alleles in phase over short genetic distances) are known to be important (17). The phase effect can be thought of as an interaction, and therefore it may be important to consider haplotype phase as well as compound genotype in candidate region analysis (18).

However, as the haplotypes are not directly observed and most association studies do not include genotypes relatives the “phased” haplotypes must be estimated. Various computational statistical methods are employed to estimate the haplotype phase. The simplest of these uses the Expectation–Maximisation algorithm. Rather as before when the counts of genotypes in each group are compared by a chi-squared test, now the estimated frequencies of each haplotype are compared between cases and controls. When considering only one SNP, the frequencies between cases and controls can be compared with a chi-squared test for two degrees of freedom. If we just compared compound genotypes between cases and controls, we would have nine possible compound genotypes but only four haplotypes, so we appear to have more power to detect a difference when considering only haplotypes. This is the same argument as when allele frequencies are compared in cases and controls, this type of comparison assumes that there are only *cis* effects and no *trans* (opposite phase) type of interaction. Once more than two SNPs are considered, the number of degrees freedom involved in such tests becomes prohibitively large so it seems reasonable to restrict the model to *cis* effects.

One further complication of an analysis comparing estimated haplotype frequencies between cases and controls is that the estimation of the frequencies must now be taken into account in the analysis. These frequencies cannot be compared with a simple chi-squared test but require a likelihood framework to compare the estimates. Several methods exist to both estimate the haplotype frequencies and perform a likelihood based test to test for association (18).

A statistically significant result in a haplotype-based study tells you there is evidence of differences between cases and controls, but it can be difficult to interpret the resulting association. Earlier we listed the quantities of interest when embarking upon these studies. In epidemiological terms, it is the risk or relative risk to the individual when falling into an exposure risk group that is of interest. If we compare estimates of haplotype frequencies, we have a haplotype relative risk. The only way to interpret this for a diploid individual is to assume no interaction between haplotypes and assume the individual genotype relative risk is the product of the two haplotype relative risks.

7. Lung Cancer: A Case–Control Example

We illustrate the application of the case–control methods in a recent publication. Amos et al. (8) reported the follow-up of candidate SNPs identified through a GWAS in lung cancer cases and controls. The first stage examined the genotype distributions of over 300,000 SNPs in 1,154 lung cancer cases and 1,137 controls. From this stage they selected the top ten (statistically significant) SNPs from the GWAS study. These ten SNPs were then genotyped in two larger case–control collections, a Texan series of 711 case and 632 controls and a UK series of 2,013 cases and 3,062 controls. Of the ten selected SNPs, two showed very strong evidence of association. The authors reported the *p* values resulting from a combined study allele-based test which were less than 10^{-17}, though methods allowing for an additive allelic effect (and without assuming Hardy–Weinberg equilibrium) also found highly significant signals ($p < 10^{-12}$).

Although the set of first stage SNPs had been selected to be tagging SNPS (i.e. not be in strong linkage disequilibrium) two of the SNPs in the top ten set were in the same region or block of strong linkage disequilibrium. Within this region on 15q25.1 there are several genes known to play a role in nicotine dependence (CHRNA3 and CHRNA5). Amos et al. speculated that the statistically significant association to two SNPs in this region may be due to an individual's propensity to be exposed to smoking through stronger addiction to nicotine. They were able to test this hypothesis by adjusting the analysis by pack years of smoking, number of cigarettes consumed per day and examining the distribution of smoking patterns by SNP genotype. They found suggestive evidence that the association with lung cancer is limited to ever-smokers but that the effect of the genotype and of smoking act independently on risk. Clearly, this hypothesis needs to be examined in further independent follow-up studies. The group was also able to explore the evidence for one vs. two causal loci, i.e. were the two SNPs capturing the same causal variant or two distinct causal variants. They estimated extended haplotypes in the original discovery set at seven further SNPs located within this candidate region on 15q25.1 and found that a single extended haplotype was significantly associated with lung cancer risk ($p = 0.00007$). This result would support the hypothesis that the two SNPs are detecting evidence for a single causal variant through LD. This hypothesis can also be further investigated through independently conducted follow-up studies.

8. Osteoporosis: A Quantitative Trait Example

Osteoporosis is defined as weakening or thinning of bone predisposing the individual to osteoporotic fracture (OF). The weakening and thinning of bone is measured by a number of quantitative markers, such as bone mineral density (BMD), bone size, and bone turnover markers (19). Studies investigating the genetic epidemiology of osteoporosis frequently choose to study surrogate phenotype markers such as BMD and the WHO Working Group defines osteoporosis according to measurements of BMD using dual-energy X-ray absorptiometry (DEXA) (20). The justification for the use of BMD alone as a surrogate marker of osteoporosis phenotype has been brought into question as clinical trials have shown only a weak correspondence between BMD levels and fracture risk (19).

The vitamin D receptor gene (VDR) has been extensively studied as a candidate gene for osteoporosis risk. The keen interest was generated by an early report suggesting that 75% of the heritability of BMD could be accounted for by allelic variation within the VDR gene (21). Although this report was based on a small sample and subsequently declared genotyping errors weakened the results, variation within the VDR gene has been the focus of many follow-up studies of BMD and OF. Though several follow-up studies reported significant association with variation at this locus, many of the single studies suffered from small sample size. Three subsequent meta-analyses of published reports found evidence for only modest effects (22–24). Within the VDR gene there are five polymorphic loci that have been investigated for association with BMD, measured at femoral neck and lumbar spine, and OF. In 2006, the GENOMOS consortium published a comprehensive participant level meta-analysis analysis of the association between these five VDR variants (*Cdx2*-promotor, *Fok1*, *Bsm1*, *Apa1*, and *Taq1*). The multicentre study was able to jointly consider results on over 25,000 participants. This represented a mix of longitudinal and cross-sectional studies. Some earlier reports suggested haplotype-specific effects.

As the three loci Bsm1, Apa1, and Taq1 are known to be in strong LD, these were analysed as a multi-allelic locus, and common haplotype-specific effects were studied. The effect of the locus (genotype or haplotype) polymorphism on BMD variation was analysed with mixed model analysis of variance. This enabled a full joint analysis of all nine studies while allowing for variation between studies. Two quantitative outcomes were considered, lumbar spine, and femoral neck BMD. They found no statistically significant evidence of consistent differences in mean BMD between individuals from distinct VDR genotype or haplotype groups. When considering the binary outcome OF, they used the study reported measure classed as “all fractures” and then also repeated the analysis limiting the outcome to “vertebral fracture”, this was to overcome the problem that each study had collected the fracture data in different ways. The binary outcome analysis was summarised by reporting per allele or per haplotype ORs arising from adjusted logistic regression.

In conclusion, they found no significant association between any of these considered polymorphisms with lumbar spine or femoral neck BMD. They reported a statistically weak association between Cdx2 alleles and vertebral fracture risk (a per allele relative risk reduction of 9%, p value = 0.039). It therefore appears that variation within the VDR gene has little or no effect on osteoporosis risk.

References

1. Risch, N., Merikangas, K. (1996) The future of genetic studies of complex diseases. *Science* 273:1516–1517.
2. Cordell, H.J., Clayton, D.G. (2005) Genetic epidemiology 3 – Genetic association studies. *Lancet* 366:1121–1131.
3. Dobson, A.J. (2002) Normal Linear Models. In: *An Introduction to Generalised Linear Models*, 2nd Edition. Chapman and Hall/CRC: Boca Raton, FL, pp. 85–114.
4. Vignal, C., Bansal, A.T., Balding, D.J., et al. (2009) Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-DRB1 loci. *Arthritis Rheum* 60:53–62.
5. Antoniou, A., Pharoah, P.D.P., Narod, S., et al. (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. *Am J Hum Genet* 72:1117–1130.
6. Cancer Research UK FactSheet. http://info.cancerresearchuk.org/cancerstats/types/breast/incidence/
7. Cox, A., Dunning, A.M., Garcia-Closas, M., et al. (2007) A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* 39:352–358.
8. Amos, C.I., Wu, X., Broderick, P., Gorlov, I.P., et al. (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 40:616–622.
9. Jewell, N.P. (2004) Study Designs. In: *Statistics for Epidemiology*. Chapman and Hall/CRC: Boca Raton, FL, pp. 43–56.
10. Dobson, A.J. (2002) Binary Variables and Logistic Regression. In: *An Introduction to Generalised Linear Models*, 2nd Edition. Chapman and Hall/CRC: Boca Raton, FL, pp. 115–134.
11. Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., Song, F. (2004) *Methods for Meta-Analysis in Medical Research*. Wiley: New York.
12. Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., Song, F. (2004) Publication Bias. In: *Methods for Meta-Analysis in Medical Research*. Wiley: London, pp. 109–132.
13. Iles, M.M. (2010) Genome Wide Association. In: M.D. Teare (ed.) *Genetic Epidemiology*. Springer: New York.
14. Clayton, D., Chapman, J., Cooper, J. (2004) The use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* 27:415–428.
15. Guo, Y., Li, J., Bonham, A.J., Wang, Y., Deng, H. (2009) Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: A comparison of association-mapping strategies. *Eur J Hum Genet* 17: 785–792.
16. Browning, B.L., Browning, S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84:210–223.
17. Price, A.L., Patterson, N., Hancks, D.C., Myers, S., Reich, D., et al. (2008) Effects of *cis* and *trans* genetic ancestry on gene expression in African Americans. PLoS Genet 4:e1000294. doi:10.1371/journal.pgen.1000294.
18. Balding, D.J. (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7:781–791.
19. Chen, Y., Shen, H., Yang, F., Liu, P.-Y., et al. (2009) Choice of phenotype in osteoporosis genetic research. *J Bone Miner Metab* 27:121–126.
20. NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy (2001) Osteoporosis prevention, diagnosis, and therapy. *JAMA* 285:785–795.
21. Morrison, N.A., Qi, J.C., Tokita, A., Kelly, P.J., et al. (1994) Prediction of bone density from vitamin D receptor alleles. *Nature* 367:284–287.
22. Cooper, G.S., Umbach, D.M. (1996) Are vitamin D receptor polymorphsims associated with bone mineral density? *J Bone Miner Res* 11:1241–1248.
23. Thakkinstian, A., D'Este, C., Eisman, J., Nguyen, T., Attia, J. (2004) Meta-analysis of molecular association studies: Vitamin D receptor gene polymorphisms and BMD as a case study. *J Bone Miner Res* 19: 419–428.
24. Gong, G., Stern, H.S., Cheng, S.C., Fong, N., Mordeson, J., Deng, H.W., et al. (1999) The association of bone mineral density with vitamin D receptor gene polymorphisms. *Osteoporos Int* 9:55–64.

Chapter 9

Family-Based Association Studies

Frank Dudbridge

Abstract

Family-based association methods are useful because they offer improved matching of controls to cases, with the result that they are not susceptible to confounding by population stratification. They also allow analysis of parent-of-origin effects and maternal–fetal interactions. The transmission/disequilibrium test (TDT) is a test of linkage and association that is equivalent to a matched case/control analysis, from which various extensions are possible. A logistic regression formulation leads to modifications for multiallelic markers, haplotypes, and quantitative traits. Some pitfalls are described, for the situations in which one parent is missing, genotyping errors have occurred, and haplotype phase is uncertain. The problem of testing association in general pedigrees is discussed, with particular reference to sib pairs without parents.

Key words: TDT, Matched case/control, Population stratification, Linkage, Within-family test

1. Introduction

Family-based association studies use the transmission information within families to infer association between genetic markers and disease or quantitative phenotypes. Typically, the family units are small, but a large number of them are collected for analysis. Most often, nuclear families are used that consist of two parents and some full siblings, but extended pedigrees may also be used for association analysis, as may subsets of nuclear families, such as sib pairs or single parent families.

Family-based association is appealing for several reasons. Firstly, in common with all epidemiological studies, there is the need to match cases to controls at the population and, ideally, the individual level. A situation of particular concern is population stratification, in which the study population actually consists of a mixture of subpopulations having different gene frequency and

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_9, © Springer Science+Business Media, LLC 2011

disease prevalence. When collecting cases and controls, relatively more cases might come from the subpopulations with higher prevalence, which would lead to a difference in the gene frequency between cases and controls even if it has no influence on disease. This is an example of *confounding* in epidemiology. Family-based methods avoid this problem by essentially matching each case to a control from the same family. Since families are by definition in the same subpopulation, the problem of confounding by stratification is avoided.

Members of the same family are likely to share a high proportion of environmental exposures so that differences in phenotype are more likely to be due to genetic differences. This means that more evidence for association should be present in a family-based sample than a population-based sample of similar size, so the family-based design is more powerful. However, this advantage is offset by the fact that family members share up to one-half of their genome in common so that a certain proportion of the information in a family sample is redundant.

Another important advantage of family-based studies is that they allow parent of origin studies that allow for imprinting effects and interactions between maternal and fetal genotypes.

While family-based studies have compelling advantages, their primary limitation is the difficulty and cost of recruiting suitable family members into the study. For each case that is identified, its parents or siblings must also be located, consent must be obtained and their DNA extracted. This is particularly difficult for late-onset disease, as family members may be dispersed or even deceased. Consequently, it is common for a family-based study to contain a mixture of family structures, which creates difficulties in analysis. For these reasons, family-based studies are more often applied to early-onset diseases, and there is a preference for smaller-scale candidate gene studies rather than genome-wide scans for which many thousands of cases are required.

2. Transmission/ Disequilibrium Test

The most commonly used family-based association test is the transmission/disequilibrium test (TDT) (1). In its original form, the TDT considers the transmission of the variant allele of a biallelic marker from heterozygous parents to affected children. Table 1 shows the four cell counts relating to the transmissions from a parent to an affected child, arranged as a contingency table. The TDT treats the untransmitted allele as a matched control to the transmitted allele, in which case only the heterozygous parents are informative. Intuitively, if the marker does not affect the disease, then we should observe Mendelian transmission

Table 1
Counts of transmissions from *n* parents of affected children, used in calculating the transmission/disequilibrium test

	Non-transmitted allele		
Transmitted allele	**Variant**	**Common**	**Total**
Variant	n_{11}	n_{12}	$n_{1\cdot}$
Common	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

in the family. Formally, we test the null hypothesis that heterozygous parents transmit each allele with equal probability. Let $T = 1$ when the parent transmits the variant allele, $T = 0$ otherwise; then under equally likely transmission $E(T) = 1/2,$ $\mathrm{var}(T) = 1/4,$ and by applying the central limit theorem over the $n_{12} + n_{21}$ heterozygous parents, we have the TDT statistic

$$\mathrm{TDT} = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})},$$

which is asymptotically distributed as χ^2 with one degree of freedom.

It can be shown that equal transmission probability occurs either when there is no linkage between marker and disease, or when there is no association (2). Figure 1 shows the four situations that are possible when a parent carries a penetrant allele and is heterozygous at the marker. When there is no linkage, either marker allele is transmitted with the disease allele with probability ½ regardless of the haplotype distribution in parental chromosomes: scenarios (a) and (b) are equally likely, as are (c) and (d). But when there is no association, there is no information on which marker allele occurs on the same haplotype as the disease allele, given that the parent is heterozygous: scenarios (a) and (c) are equally likely, as are (b) and (d). Therefore, either marker allele occurs on the disease haplotype with probability ½ and is then transmitted with probability ½ regardless of the recombination fraction. This means that the null hypothesis of the TDT may be taken to be either no linkage, or no association.

By regarding the TDT as a matched case/control design, in which the matched control is the nontransmitted allele, standard

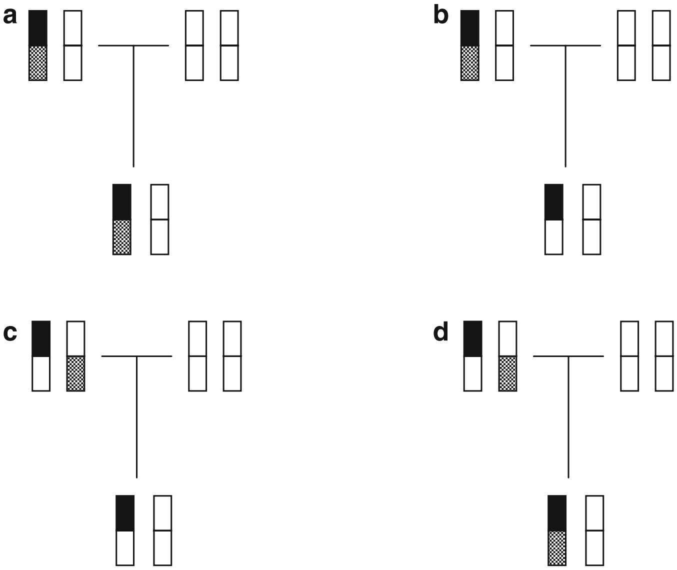


Fig. 1. Four transmission scenarios for a parent heterozygous at a test marker. The disease locus is shown above the marker locus. Disease risk allele is shown in *black*, variant marker allele in *gray*: (**a**) marker allele on disease chromosome, no recombination; (**b**) marker allele on disease chromosome, with recombination; (**c**) marker allele on normal chromosome, no recombination; (**d**) marker allele on normal chromosome, with recombination.

methods from epidemiology can be used to extend the test in various ways. The usual model for matched designs in epidemiology is conditional logistic regression, for which here transmission is the random outcome and alleles are the predictors. For a biallelic marker, we model the log odds ratio β for transmission of the variant allele from a heterozygous parent:

$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta.$$

The likelihood for n heterozygous parents is

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(\delta_i \beta)}{1+\exp(\beta)},$$

where $\delta_i = 1$when parent i transmits the variant allele, 0 otherwise. Statistical theory shows that the TDT is the same as a score test for $\beta = 0$ from this likelihood.

This model can now be extended to allow for multiple alleles, multiple loci and environmental covariates. A useful approach is to regard the analysis not in terms of transmissions from parents, but rather of comparing cases, the affected children, to controls formed from the other combinations of parental alleles. The likelihood contribution from a full case-parent trio is equivalent to that obtained by matching the case to three controls corresponding to each of the other three child genotypes that could be formed from the parental genotypes. A wide variety of analyses is possible in this framework, including parent of origin effects (3).

3. Quantitative Traits

For quantitative traits, two types of regression models are in common use. The first uses logistic regression in a similar way as above, treating transmission as the random outcome. The second uses linear regression with the genotype as the independent variable and the quantitative trait as the dependent variable, with adjustment to respect the family-based design. Generally, the linear regression approach is more sensitive to the assumption of trait normality, but it can be more flexible and powerful when this assumption is met.

The logistic regression approach treats transmission as the random outcome, as for binary traits, and treats the quantitative trait as an effect modifier for alleles acting as predictors (4). For a biallelic marker, the transmission probability of the variant allele from a heterozygous parent is given by

$$\text{logit}(p) = \alpha + \Upsilon\beta,$$

where Υ is the trait value of the child, regression coefficient β is the transmission parameter of the variant allele, and the intercept α is included to account for possible association to a different phenotype which determines inclusion in the study. In an unselected sample, the intercept could be omitted. As for discrete traits, this model can be readily extended to allow for multiple alleles, multiple loci, and covariates (3).

The linear regression approach decomposes the total population association into two components: between-family and within-family (5). The within-family association is robust to population stratification, and a difference between the between- and within-family association parameters can be taken as evidence of stratification. The model for the trait mean of child j in family i is

$$\Upsilon_{ij} = \alpha + \beta_{\text{b}} B_i + \beta_{\text{w}}(G_{ij} - B_i) + e_{ij}$$

where β_{b} and β_w are between- and within-family coefficients for association, G_{ij} codes for the genotype of child j in family i, and B_i is an expected value of B_{ij} for family i. To allow for multiple children, B_i is defined as

$$B_i = \frac{1}{2}(G_{i\mathrm{F}} + G_{i\mathrm{M}}),$$

where $G_{i\mathrm{F}}$ and $G_{i\mathrm{M}}$ code for the father's and mother's genotypes, respectively, or when parents are not available

$$B_i = \frac{1}{n_i}\sum_j G_{ij}$$

where n_i is the number of children in family i. The likelihood for a nuclear family is the multivariate normal density with the mean vector specified by this model and the variance–covariance matrix constructed to allow for linkage and shared environment among sibs (5). Likelihood ratios are used to test for within-family association ($\beta_\mathrm{w} = 0$), total association ($\beta_\mathrm{w} = 0$ and $\beta_\mathrm{b} = 0$, with two degrees of freedom) and population stratification ($\beta_\mathrm{w} = \beta_\mathrm{b}$).

4. Some Pitfalls

In spite of its simplicity, the TDT has some hidden pitfalls that arise from its use of a family structure. These can give misleading results, and various modifications of the TDT are available to allow for these problems.

If a parent is missing so that the data consist of one parent and the child, a bias can arise because the transmitted allele cannot always be determined (6). Consider a biallelic marker and recall that only heterozygous parents are informative for the TDT. If the child is homozygous, then the transmitted allele is obvious, but if it is heterozygous, we cannot say which allele was transmitted by the observed parent. We might then naively restrict analysis to the homozygous children, but they are likely to be homozygous for the more common allele, whether or not the marker is associated with disease. The result is that we count more transmissions of the common allele than the variant allele, leading to a transmission probability different from one-half even when there is no association. This bias can only be corrected with specialized methods (7, 8).

A similar bias arises when there is genotyping error. If a truly heterozygous parent is misclassified as homozygous, then the transmission is disregarded and no information is gained. If, however, a truly homozygous parent is misclassified as heterozygous, then an apparently informative transmission of the homozygous allele will be counted. Again, this would lead to a preferential scoring for the more common allele, assuming that misclassification rates are similar for the three genotypes (9).

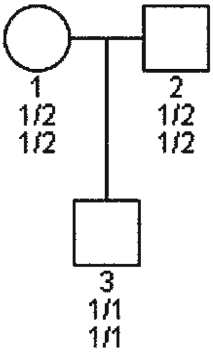


Fig. 2. Family in which haplotypes can be deduced, but would not be deducible from other children of the same parents.

Sometimes, we may wish to perform a haplotype analysis. While haplotypes can usually be inferred from family data, there is one situation in which ambiguity occurs. This is when both the parents have the same heterozygous genotype and the child is also heterozygous. Then, we cannot say which parent transmitted which allele, and the transmitted haplotypes become unknown. If families with ambiguous haplotypes are discarded, a bias can sometimes arise. Figure 2 shows a family in which the haplotypes can be deduced so that we can score two transmissions of the 1-1 haplotype. However, for these parents the haplotypes can only be deduced in this case and in the case in which the 1-1 haplotype is twice not transmitted, so the expected transmission count is 1, with variance 1. We have seen that the usual TDT assumes an expected transmission count of 1 for two parents, with variance ½, so scoring just the certain haplotype transmissions would lead to an underestimate of the true variance and an inflated test statistic. Methods are available for haplotype analysis in the presence of ambiguity (8, 10–12).

We have seen that the TDT is both a test of linkage and of association, but this is only true when there is one child per family. When there are several affected siblings, the standard TDT is not a valid test of association, if linkage is present. To see this, note that one definition of linkage is an increase of alleles shared identical by descent (IBD) among affected sibs. Thus, while lack of association means that we cannot predict the marker on a parental chromosome (see Fig. 1), the transmissions from that chromosome are correlated so that we cannot treat them as independent. This is potentially an important issue when following up a linkage scan, although in practice it seems to have a minor impact because the increase in IBD is quite small in a complex disease. Again, special methods have been developed for this situation (7, 8, 13).

5. Summary

Family-based methods are useful because they eliminate the possibility of confounding by population stratification, and increase power by matching on shared environmental variables. Their main limitation is the difficulty in recruiting all the suitable family members, and for this reason these methods are preferred for early onset disease, and tend to be applied in candidate gene or replication studies rather than genome-wide scans.

Although the analysis is similar to that of a matched case/control design, we have seen that the family design introduces some subtleties, which can give rise to misleading results. Because of these complications, family-based analysis often uses specialized software: some commonly-used programs include QTDT (5), FBAT (7), UNPHASED (8), and PLINK (14).

The most widely used method, the TDT, is appropriate for nuclear family data, in particular trios consisting of a case and its two parents. In a general pedigree, case-parent trios can be extracted and the TDT applied to each one. This is always a valid test for linkage, but can lead to problems when testing for association in the presence of linkage. We have seen that transmissions to affected siblings are correlated, and similar issues of correlation arise in general pedigree settings. Special conditioning approaches and adjustments are available for general pedigrees and are implemented in the FBAT software.

When parents are unavailable, which is often the case for a late-onset disease, the standard TDT cannot be applied. Variants of the TDT for sibships have been proposed (15), and again these have been generalized in FBAT. Another approach, implemented in UNPHASED, is to average over all possible parents that are compatible with the observed children. This has the advantage of easily combining sibships with trio data, and it can also be combined with case/control data. The main disadvantage is that a probability model has to be assumed for the missing parents, which may not be correct under population stratification, although this problem is not thought to be severe.

A very large number of variations on the TDT have been proposed, covering applications, such as multiple phenotypes, sex chromosomes, and gene–environment interaction, as well as further developments on the topics covered here. The programs referenced here are adequate for most basic analyses, but the reader is encouraged to study some recent review articles (16, 17) for a more wide ranging survey of this area.

References

1. Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52,** 506–516.
2. Ewens, W.J. and Spielman, R.S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* **57,** 455–464.
3. Cordell, H.J., Barratt, B.J. and Clayton, D.G. (2004). Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene–gene and gene–environment interactions, and parent-of-origin effects. *Genet Epidemiol* **26,** 167–185.
4. Waldman, I.D., Robinson, B.F. and Rowe, D.C. (1999). A logistic regression based extension of the TDT for continuous and categorical traits. *Ann Hum Genet* **63,** 329–340.
5. Abecasis, G.R., Cardon, L.R. and Cookson, W.O. (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66,** 279–292.
6. Curtis, D. and Sham P.C. (1995). A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet* **56,** 811–812.
7. Lake, S.L., Blacker, D. and Laird, N.M. (2000). Family-based tests of association in the presence of linkage. *Am J Hum Genet* **67,** 1515–1525.
8. Dudbridge, F. (2008). Likelihood-based association analysis in nuclear families and unrelated subjects with missing genotype data. *Hum Hered* **66,** 89–98.
9. Mitchell, A.A., Cutler, D.J. and Chakravarti, A. (2003). Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet* **72,** 598–610.
10. Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65,** 1170–1177.
11. Dudbridge, F., Koeleman, B.P., Todd, J.A. and Clayton, D.G. (2000). Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *Am J Hum Genet* **66,** 2009–2012.
12. Horvath, S., Xu, X., Lake, S.L., Silverman, E.K., Weiss, S.T. and Laird, N.M. (2004). Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol* **26,** 61–69.
13. Martin, E.R., Bass, M.P., Hauser, E.R. and Kaplan, N.L. (2003). Accounting for linkage in family-based tests of association with missing parental genotypes. *Am J Hum Genet* **73,** 1016–1026.
14. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81,** 559–575.
15. Spielman, R.S. and Ewens, W.J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* **62,** 450–458.
16. Laird, N.M. and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* **7,** 385–394.
17. Tiwari, H.K., Barnholtz-Sloan, J., Wineinger, N., Padilla, M.A., Vaughan, L.K. and Allison, D.B. (2008). Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Hum Hered* **66,** 67–86.

Chapter 10

Genome Variation: A Review of Web Resources

Andrew Collins and William J. Tapper

Abstract

An enormous number of high-quality Web-based resources are now available to facilitate research into genome variation. Although identification of the most appropriate and informative resources can be challenging, a number of key sites provide links to more specialized resources that may be useful to follow up. Given ongoing research, focussing on the sequencing of many different genomes, we can expect sequence databases and their associated polymorphism-based resources to greatly increase in depth and complexity in a relatively short period of time. However, databases and tools developed to date, and described here, provide a sound basis for accommodating this next generation of genomic data. As well as sequence-oriented resources this review presents databases providing genotypic and common disease phenotype data, copy number variation, genetic maps, cytogenetic data, and gives an overview of key software tools, with the emphasis on analysis of the genetic basis of common disease.

Key words: Genome sequence, Single nucleotide polymorphism, Copy-number variation, Linkage maps, Linkage disequilibrium, Common diseases

1. Introduction

The volume and quality of web-based resources for the analysis and interpretation of genomic variation has increased exponentially in recent years. In many cases, the development of one resource has been essential for progress with another. An obvious example is the sequencing and analysis of the human genome (1) which was necessary for the development of single nucleotide polymorphism (SNP) genotyping panels, which are essential for the genome-wide association studies that are currently underway. SNP-based research has been greatly boosted by the International HapMap project (2) which presents SNP genotypes at very high density in a relatively small number of individuals. These data provided the raw material for detailed characterization of the

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_10, © Springer Science+Business Media, LLC 2011

linkage disequilibrium (LD) structure of the genome and reliable panels of "tag" SNPs which capture most of the common variation with minimal genotyping effort and cost. Although arguably conflicting with this strategy, a further popular application of the HapMap data is the imputation of untyped genotypes in disease case-control association studies (3) which facilitates combining information across samples (meta-analysis) and modestly increases power.

The analysis of genomic variation is an extremely fast moving field and the study of copy number variation, for which Web resources are available, is gaining increasing prominence. Furthermore, the realization that common genes only explain a small fraction of the genetic variation related to disease has encouraged an ambitious undertaking to sequence many more genomes (http://www.1000genomes.org/), which is expected to provide a view of DNA variation at much higher resolution. The low resolution of current studies is evident because the genome is thought to contain at least 10 million SNPs (4), and therefore only about 1 in 20 is tested in most genome-wide association studies. One of the goals of the 1,000 genomes project is not only to identify rarer genetic variants which have a frequency of 1% or more, but also to include variants with 0.5% or lower frequency when within genes. The database presumably becomes a valuable resource for future studies which aims to establish which of the variants are functionally implicated in disease.

This review considers some of the more important Web-based resources that facilitate studies of genome variation. In addition to Web-based databases and browsing tools, a number of software packages for the analysis of data are highlighted. The review is divided into sections with wide overlaps because many of the Web-based resources are useful across multiple categories. We initially consider resources that provide views of the genome sequence (sequence and physical map resources), where the emphasis is on browsing the genome structure at different levels of resolution. Secondly, we examine resources that emphasize polymorphic variation in the sequence, predominantly SNPs and sources of SNP genotype data, but, increasingly, copy-number variation. This is followed by resources which provide genetic maps, both linkage and linkage disequilibrium, which are valuable for multilocus analysis in linkage studies and association mapping of disease. Fourthly, we consider resources that provide phenotypic information and include cytogenetic variation (which may or may not influence phenotype) and also some of the sources for disease case-control and related data. We also consider here some of the sites that review current knowledge of genes and disease. Finally, we briefly describe some of the software tools that are valuable for the analysis of genome variation emphasizing those that focus on aspects of association mapping. To illustrate some searches with real examples, we refer to the FGFR2 gene and, where relevant, its association with breast cancer.

2. Sequence and Physical Map Resources

2.1. Sequence Databases, Browsers, and Tools

The three main resources for sequence browsing are provided by the University of California Santa Cruz (UCSC), the National Center for Biotechnology Information (NCBI), and the Ensembl resources in the UK. A user's guide to all three is provided at: http://www.nature.com/ng/journal/v35/n1s/full/ng1189.html. The UCSC and NCBI resources are very similar as both integrate data from a variety of maps, including genes, polymorphisms, repetitive elements, and sequence conservation across species to form comprehensive annotations of the human genome that can be interrogated and presented as figures or tables. For conservation queries, UCSC takes preference as it presents histograms derived from multiple alignments of 28 vertebrate species or a subset of 17 placental mammals. As a result, each nucleotide is given a score that is readily accessible from the histogram or table browser. In comparison, NCBI aligns mRNA and EST sequences from five other organisms to the assembled human genomic sequence so that information on conservation is limited to a smaller number of species in coding regions. Unlike NCBI, the UCSC Web site also presents information on structural variation originating from the Database of Genomic Variants (DGV). Although Ensembl also presents data on conservation, polymorphisms, and repeat sequences, it focuses on the annotation of coding sequences and differs in this respect with UCSC and NCBI. Further differences between these three sites include the annotation of different mRNA sequences and using different symbols for the same genes which can complicate comparisons between them, despite links between the three sites.

The UCSC genome browser ((5), http://genome.ucsc.edu/, http://genome.ucsc.edu/cgi-bin/hgTracks?org=human) enables scrolling over chromosomes at different levels of resolution and represents the state-of-the-art in genome annotation. The huge volume of information presented here is provided through powerful graphical interfaces and a large number of "tracks" which can be selected for display across many categories. These categories include mapping and sequencing data (for example, recombination rates, clones, GC content, chromosome band), phenotype and disease (quantitative trait loci, case-control associations), genes and gene prediction, mRNA and expressed sequence tag (EST) data, expression and regulation, comparative genomics, variation and repeats and detailed analysis in the ENCODE regions (6). Other UCSC resources include the Human Gene Sorter (http://genome.ucsc.edu/cgi-bin/hgNear) which is useful for examining gene families and the ways in which genes are interrelated. These relationships include those at the level of protein product, expression profiles, or map location. It is possible to determine a set of genes with shared properties for further analysis. A search for FGFR2 on protein homology unsurprisingly yields not only the FGFR genes 1, 3, and 4, but also less immediately obvious matches with RET, KDR, and FLT1. This type of information may be useful in selecting further candidates for association studies.

The NCBI mapviewer (http://www.ncbi.nlm.nih.gov/projects/mapview/) provides powerful genome browsing tools for the human and many other genomes (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome). The browser enables viewing of genomes by organism, displays chromosome maps and provides zoomed views with increasing resolution down to sequence level. Multiple maps on different scales can be aligned based on shared marker and gene names and a common coordinate system for sequences. Within this suite of programs is GenomeView (http://www.ncbi.nlm.nih.gov/Tour/3a.html) which displays the genome graphically through scaled chromosome ideograms and presents the results of searches as marked locations on the ideograms. The SequenceView tool displays sequence data for a specific chromosomal region and graphically depicts any biological features annotated in that region.

Ensembl (http://www.ensembl.org/index.html) provides gene-oriented information, including ContigView which presents specific genes in genomic context (sequence position, gene length, cytogenetic band, microsatellite markers, protein coding regions) and a more detailed view which illustrates conservation, clones, and transcripts. For example, the link for the gene FGFR2 (http://www.ensembl.org/Homo_sapiens/contigview?gene=ENSG00000066468) contains much information which identifies three microsatellite markers in the gene, the sequence position, protein coding status, and the closest neighboring gene (ATE1). Also useful is a graphical display of syntenic regions for a wide range of organisms. Due to its focus on coding regions, Ensembl typically aligns more full transcripts, including some manual annotations and provides easier access to exon and intron information than the NCBI and UCSC sites.

3. Polymorphism Databases

3.1. SNP Databases

The dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) is regularly updated to reflect clustering of SNPs which map to the same location, thereby reducing redundancy. A search for human gene FGFR2 yields 1,274 SNPs listed by "rs" number and each with links to MapView, GeneView, SeqView, and protein databases. The GeneView link gives a categorical breakdown of coding SNPs (synonymous, missense, etc.) within each exon. Searches for individual SNPs by rs number yields sequence, SNP function (intronic, coding, etc.) and information on population diversity (genotype and allele frequencies in HapMap and Perlegen samples).

SNPper (http://snpper.chip.org/bio/snpper-enter/) is a similar database which combines information from dbSNP and the UCSC Human Genome Browser. SNPs can be searched for using dbSNP "rs" names, the SNP Consortium's "TSC" names, or by position on the chromosome. Alternatively, one or more gene names can be submitted to find all relevant SNPs classified by promoter, 5 UTR, intron, exon, intron boundaries, 3 UTR, and downstream. The basepair location for each SNP, alleles, links to corresponding public databases and sequence surrounding the SNP are also displayed. Automatic primer design is also provided through the Primer3 program from the Whitehead Institute.

The SNPbrowser™ software (http://marketing.appliedbiosystems.com/mk/get/snpb_landing) is a free commercial tool which offers knowledge-guided selection of genotyping assays for association studies. The graphical display presents SNPs and LD maps together with haplotype blocks derived from the analysis of HapMap and Applied Biosystems SNPs. The tool enables searches for tagging SNPs using a wide variety of inputs, including SNP and gene names, chromosomal or microsatellite locations, and allows browsing of detailed SNP and gene information.

3.2. SNP Genotypes

The International HapMap Project (http://www.hapmap.org/) started in 2002 with a goal to compare genetic sequences of different individuals and identify chromosomal regions with shared genetic variants to determine panels of tag SNPs across the genome. This is an international effort between Japan, the UK, Canada, China, Nigeria, and the USA. The database currently holds ~4 million SNP genotypes for populations of northern and western European ancestry (CEPH samples from Utah, abbreviation CEU), Han Chinese from Beijing (CHB), Japanese individuals from Tokyo (JPT), and Yoruba in Ibadan, Nigeria (YRI). The genotypes are subject to rigorous quality control and the data are available for 270 individuals from the four populations. The search facility enables, for example, a search for the gene FGFR2 which yields 75 tag SNPs capturing the variation of 270 alleles (from the CEU population) with an r-squared cut-off of 0.8.

The HGDP-CEPH Diversity Panel Database (http://www.cephb.fr/hgdp-cephdb/) receives, stores, and provides marker genotypes generated by users of the DNA samples from the HGDP-CEPH Diversity Panel. The DNA samples are derived from 164 individuals representing 51 populations worldwide. Along with the genotypic data, information on geographic and population origin, along with gender, are presented.

3.3. Copy-Number and Other Sources of Structural Variation

The Database of Genomic Variants (DGV, http://projects.tcag.ca/variation/) provides data on structural variation defined as alterations in DNA segments of at least 1 kb in length, along with insertion–deletion polymorphisms (InDels) in the range 100 bp–1 kb. The database only considers variants that are found in healthy patient samples. These data are also available as a track on the UCSC genome browser.

The human genome structural variation project (http://hgsv.washington.edu/) characterizes the extent of structural variation at the sequence level covering deletions, insertions, inversions, and other rearrangements. Eight individuals from HapMap are currently analyzed. This project maps clones against the reference sequence to systematically identify and sequence structural variants genome-wide (7). The database contains the data reported in Kidd et al. (8) presented through a modified mirror of the UCSC genome browser and as a track on the browser.

4. Genetic Map Databases

4.1. Linkage Maps

Genetic linkage maps give the relative positions of genes along a chromosome with distances that depend on the frequency with which two loci recombine during meiosis. Accurate linkage maps are essential for multilocus mapping of disease genes by linkage in family pedigrees and other studies related to recombination. The Rutgers combined linkage-physical maps http://compgen.rutgers.edu/maps/ (9) combine genotype data from the Centre d'Etude du Polymorphisme Humain (CEPH, http://www.cephb.fr/) and deCODE pedigrees (http://www.decode.com/) with genetic distances conditional on sequence-based positional information. Build 36 of the Rutgers map comprises 28,121 markers. The maps include regression-based smoothing to separate all markers by nonzero distances, which facilitates interpolation of locations for markers not already in the map (including the enormous number of SNPs not yet typed in these pedigrees). An online tool is provided to locate additional markers in the map given their sequence position.

The CEPH genotype database ((10), http://www.cephb.fr/cephdb/), which has provided a resource for the development of linkage maps for many years, contains genotypes for 32,356 markers typed on the set of reference families and maps which describe all the recombination breakpoints in the families. Other linkage map sources include the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway) and the NCBI mapviewer (http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606) which both give recombination map locations from deCODE, Marshfield, and Genethon linkage maps.

The HapMap resource provides sex-averaged genome-wide "recombination rate" data computed through a coalescent model from LD using the HapMap SNP data (2). The maps have very high resolution but recombination patterns are distorted because LD patterns also depend on selection, mutation, and population history (11).

4.2. Linkage Disequilibrium Maps

LD maps ((12), http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/LDMAP/map2.htm) provide an LD analog of the linkage map in linkage disequilibrium units (LDUs) for which one LDU represents the extent of LD that is useful for mapping (and spanning widely differing distances on the kilobase scale). The LDU scale has a simple relationship to time since a major population bottleneck. Recent applications include LD structure of isolated populations (13) and multilocus disease gene mapping (14).

5. Phenotypic Variation

5.1. Cytogenetic

The DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER, https://decipher.sanger.ac.uk/) collects clinical information about chromosomal microdeletions, duplications, insertions, translocations, and inversions and displays this information alongside the genome map. The project provides a research tool to increase knowledge about chromosomal rearrangements with a focus of improving medical care and providing genetic advice for the relevant families and individuals. Chromosome analysis remains an important tool in the diagnosis of children with developmental delay, learning disability, and/or multiple congenital anomalies. However, the resolution of Giemsa banding is limited and many rearrangements are missed, hence the usefulness of array-CGH to detect imbalances genome-wide. The database provides the necessary tools to manage the voluminous data from more than 30,000 clones in each experiment. The Ensembl interface of DECIPHER allows the visualization of location of clones that are deleted or duplicated and matches them, where possible, against known microdeletion or microduplication syndromes. The database provides detailed information of known syndromes with a chromosome ideogram-based interface, together with comprehensive lists of references.

The chromosome anomaly collection (http://www.som.soton.ac.uk/research/geneticsdiv/Anomaly%20Register/) is rather different in that it contains examples of unbalanced chromosome abnormalities (UBCAs) that lack detectable phenotypic effect. These are anomalous in the sense that cytogenetically visible UBCAs usually do have phenotypic consequences which would come to medical attention. The collection also includes euchromatic variants that form part of the continuum of copy number variation in the human genome (15). This resource enables the characterization of genomic regions with unbalanced anomalies which are consistently free of phenotypic consequences.

5.2. SNP Genotype and Phenotype Resources

The Wellcome Trust Case Control Consortium (WTCCC, http://www.wtccc.org.uk/) is analyzing case and control samples to identify common disease variants. Burton et al. (16) describe the genome-wide association study in the British population which contrasted 2,000 individuals for each of seven major diseases with a common set of 3,000 controls. The study determined 24 independent association signals with genome-wide significance: 1 for bipolar disorder, 1 for coronary artery disease, 9 for Crohn's disease, 3 for rheumatoid arthritis, 7 for type 1 diabetes and 3 for type 2. A second arm to the study has analyzed 15,000 polymorphic markers that alter protein sequence to look for genetic variation relating to four diseases – breast cancer, autoimmune thyroid disease, multiple sclerosis, and ankylosing spondylitis. The SNP genotypes and the case-control status information are available to researchers wishing to pursue studies on SNP variation and disease, along with population genetics analyses.

The Cancer Genetic Markers of Susceptibility (CGEMS, http://cgems.cancer.gov/) is a 3-year initiative of the National Cancer Institute focused on identifying genetic variants involved in susceptibility to prostate and breast cancer. The project is a genome-wide case-control study with more than 500,000 SNPs embracing five large prostate and breast cancer samples. For prostate cancer, the study includes 1,177 individuals who developed prostate cancer during the observational period contrasted with 1,105 individuals who did not. For breast cancer, CGEMS has a genome scan in a sample of 1,200 cases and 1,200 controls. The CGEMS study will follow up markers identified as promising in the first phase through further epidemiologic and case-control analyses.

5.3. Genes and Disease

Online Mendelian inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim), is a database of human genes and genetic disorders. The database contains textual information, references and copious links to MEDLINE, genome sequence, and related resources at NCBI and elsewhere. Searching is simple and the literature referenced is comprehensive. OMIM is probably the first resource to access for specific gene and phenotype-based queries. As an example, a search of OMIM for FGFR2 yields a detailed text derived from 108 references. The breast cancer-specific text lists Easton et al. (17) as identifying a G/A SNP rs2981582 strongly associated with familial breast cancer and identifies further studies which find associations with sporadic postmenopausal breast cancer and in BRCA2 gene carriers, respectively.

The genetic association database: http://geneticassociationdb.nih.gov/ concisely lists seven references and five diseases associated with FGFR2, including breast cancer. GeneCards http://www.genecards.org/ presents data and generates gene-specific links to numerous online databases, including comprehensive lists of gene symbol aliases, genomic location data, protein/expression information, gene ontology, mutation phenotypes, splice variants, orthologs, and details of SNPs (935 NCBI SNPs listed for FGFR2).

6. Software for the Analysis of Genome Variation

Although this review is directed at online database and related genomic resources, we briefly outline here some of the key software tools, and sites which supply these tools, with the analysis and interpretation of genomic variation in mind.

The Rockefeller Web site (http://linkage.rockefeller.edu/) provides a huge number of links to key references, software packages, and other resources, with a particular emphasis on linkage analysis and the LINKAGE suite of programs.

There are a large number of sequence analysis tools of which the best known is the Basic Local Alignment Search Tool (BLAST http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) which finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

The construction of a multilocus genetic map is usually a prerequisite for accurate fine mapping of disease genes. For constructing a linkage map, the CRI-MAP program (http://compgen.rutgers.edu/multimap/crimap/) is efficient for constructing multipoint linkage maps in family data. The LDMAP program (http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/LDMAP/default.htm) constructs metric LD maps for SNP data, from unrelated individuals, which have distances in LDUs. The maps are useful in determining panels of SNPs for coverage of a genomic region, for fine mapping, and the analysis of population structure and history.

For planning an association or other gene mapping study, the appropriate power calculations can be achieved effectively through specialized software packages. The genetic power calculator ((18), http://pngu.mgh.harvard.edu/~purcell/gpc/) provides power calculations for variance components, quantitative trait locus and linkage and association tests in sibships. The program CaTS ((19), http://www.sph.umich.edu/csg/abecasis/CaTS/) is designed for power calculations for large genetic association studies focusing on two-stage designs in genome-wide association.

There are many programs which determine haplotypes for testing association with disease and for population-based studies. One of the best known is PHASE ((20), http://stephenslab.uchicago.edu/software.html), which estimates haplotypes from population genotype data and incorporates methods for estimating recombination rates and identifying recombination hotspots.

For disease association mapping, particularly genome-wide association, the CHROMSCAN program (http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/chromscan/) implements a composite likelihood approach for fine mapping, where locations are determined on an underlying LD map. The PLINK program (21) is a comprehensive and whole genome analysis suite (http://pngu.mgh.harvard.edu/~purcell/plink/) which embraces many tools useful for data management, quality control, examining population stratification, performing case-control and quantitative association tests, haplotype analysis, genotype imputation, and testing epistasis.

7. Discussion

Inevitably, it is possible to include only a small fraction of the potentially useful resources that are available in such a review. However, the power of modern Web-searching tools should make it possible to identify other potential resources, where a specific and well-defined application is envisaged. The ongoing developments, which include the sequencing of multiple genomes, create many challenges for the development of tools which provide access and interpretation of these data, particularly as we have become accustomed to having a single reference genome. It is possible to anticipate extensive and profound changes to some of the key sites, particularly sequence browsers, in the next few years as these data are made available. However, the powerful and flexible Web-based tools developed thus far for the presentation of genomic variation suggest that a smooth transition to embrace the new set of reference genomes can be achieved.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
2. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
3. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78(4):629–644.
4. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet* 33(Suppl):228–237.
5. Thomas DJ, Trumbower H, Kern AD, Rhead BL, Kuhn RM et al (2007) Variation resources at UC Santa Cruz. *Nucleic Acids Res* 35:D716–D720.
6. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816.
7. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM et al (2005) Fine-scale structural variation of the human genome. *Nat Genet* 37:727–732.
8. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N et al (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453:56–64.
9. Matise TC, Chen F, Chen W, De La Vega F, Hansen M et al (2007) A second-generation combined linkage physical map of the human genome. *Genome Res* 17:1783–1786.
10. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM et al (1990). Centre d'Etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* 6(3):575–577.
11. Tapper W, Gibson J, Morton NE, Collins A (2008) A comparison of methods to detect recombination hotspots. *Hum Hered* 66(3): 157–169.
12. Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR et al (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci* USA 99:2228–2233.
13. Service S, DeYoung J, Karayiogou M, Roos JL, Pretorious H et al (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet* 38(5):556–560.
14. Collins A, Lau W (2008) CHROMSCAN: genome-wide association using a linkage disequilibrium map. *J Hum Genet* 53(2):121–126.
15. Barber JCK (2005). Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. *J Med Genet* 42(8): 609–629.
16. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P et al (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678.
17. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447:1087–1093.
18. Purcell S, Cherny SS, Sham PC (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19(1):149–150.
19. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 38:209–213.
20. Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73:1162–1169.
21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR et al (2007) PLINK: a tool-set for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 81(3):559–575.

Part IV

Emerging Themes

Chapter 11

Advanced Methods in Twin Studies

Jaakko Kaprio and Karri Silventoinen

Abstract

While twin studies have been used to estimate the heritability of different traits and disorders since the beginning of the twentieth century, statistical developments over the past 20 years and more extensive and systematic data collection have greatly expanded the scope of twin studies. This chapter reviews selected possibilities of twin study designs to address specific hypotheses regarding the role of both genetic and environmental factors in the development of traits and diseases. In addition to modelling latent genetic influences, current models permit inclusion of information on specific genetic variants, measured environmental factors and their interactive effects. Examples from studies of anthropometric traits are used to illustrate such approaches.

Key words: Quantitative genetics, Twins, Models, Behaviour genetics, Obesity, Longitudinal studies

1. Moderation of Variance Components by Measured Environmental Factors for Gene–Environment Interaction Analyses

When the data permit, the twin model can be extended to analyse more detailed questions about the variance and covariance structures than only heritability estimates (1). A simple example is to estimate heritability of height in different birth cohorts to investigate whether the magnitude of genetic and environmental variances of height have changed over time (2). Likewise, sex differences in the magnitude of the genetic and environmental variance components can be estimated in classical twin models. If the data are available also on opposite-sex dizygotic (OSDZ) twin pairs, the models of sex differences permit tests of whether different sets of genes (or environmental factors) influence phenotypic variation in males and females.

To study sex differences in the genetic architecture of body mass index (BMI, kg/m^2), data from approximately 37,000 twin

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_11, © Springer Science+Business Media, LLC 2011

pairs aged 20–39 years from eight countries based on the GenomEUtwin consortium were analysed. The variance structures were mostly very similar across countries, and genetic differences explained 45–85% of the variation in BMI. There was greater variation in BMI among women than among men across the countries, and overall the analyses also indicated that sets of genes influencing variation in BMI at this age are not fully identical for males and females (3). Such twin findings prompted researchers to probe further the Xq22-24 region of the X-chromosome for genes associated with obesity. Linkage results suggested a sex-specific effect and association with haplotypes of the SLC6A14 gene, which encodes an amino acid transporter speculated to affect appetite regulation (4).

Although heritability for BMI is substantial, the recent increases in the prevalence of overweight cannot be attributed to changes in the gene pool. However gene–environment interactions, whereby genetic differences between individuals modulate responses to obesity-promoting behaviours and environments, may account for a substantial part of the apparent heritability. Advances in statistical modelling allow tests of the effects of moderators on heritability and estimate the role of gene–environment interactions (5).

The moderation model tests whether the magnitude of additive genetic variance (a), common environmental variance (c), and unique environmental variance (e) are changing as a function of a measured environmental factor M. In such structural equation models, a classic univariate twin model is modified to include a moderation component (6). Beta terms are added to the standard paths a, c, and e, which indicate the magnitude of the effect of additive genetic influences, common environmental influences, and unique environmental influences as a function of the environmental factor. The extra terms thus indicate the significance of a potential moderator variable M on each of these sources of variance. The value of M can assume different values in each subject, unlike earlier models of gene–environment interaction, where only pair-specific factors (such as age or sex or urban/rural residence) could be used. In the moderation model, the additive genetic value is a linear function of the moderator M, represented by the equation $a + \beta x M$, where βx is an unknown parameter to be estimated from the data. When βx is significantly different from zero, there is evidence for a moderating effect. Likewise, the βy and βz terms represent the extent to which the moderator changes the impact of common and unique environmental influences. The pathway $\mu + \beta_m M$ models the effect of the moderator variable on the mean of the outcome variable as in a standard linear regression model. Any gene–environment correlation effects between the moderator variable and outcome are also included in this pathway. Figure 1a

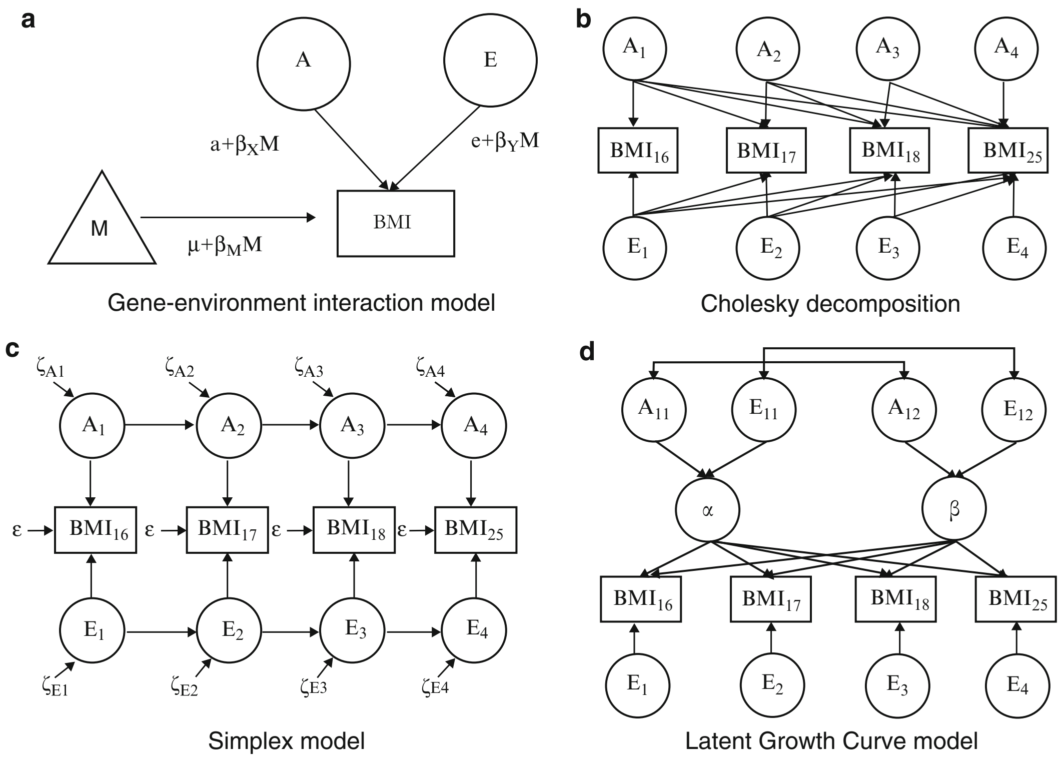


Fig. 1. Conceptual figures for multivariate twin models, see text for details for the models. (**a**) The moderator model allows for modification of the genetic and environmental effects on the trait being studied by values of a moderator variable (*M*), thus permitting testing hypotheses on gene–environment interactions using twin data. Longitudinal data, for example of BMI at four ages (16, 17, 18, and 25) can be analysed with different approaches illustrated as (**b**) a Cholesky decomposition, (**c**) a Simplex model and (**d**) a longitudinal growth model with first order latent variables of level (*intercept*) and rate of change (*slope*), that are further decomposed into genetic and environmental components. As represent latent additive genetic variables and Es represent latent environmental variables. These models can also permit the inclusion of genetic effects due to dominance or environmental effects shared by family members (such as twins in twin pair), but these are not shown in the figures.

shows the conceptual components of the model for additive genetic and unique environmental effects.

Using such moderator models, we have demonstrated that physical activity moderates the heritability of BMI and waist circumference (7), likewise that parenting variables modifies the relative contribution of genes and environment on adolescent smoking and drinking (8, 9).

2. Multivariate Analyses of Twin Data

Multivariate modelling of twin data can assess the existence of environmental and genetic correlations between traits. The question asked is whether the phenotypic correlation between two or more traits is due to genetic effects that are in common, i.e. pleiotropic effects. This might then guide phenotype construction and analysis approaches in linkage and association analyses. Alternatively, the association may be due to environmental factors affecting both traits or due to a causal relationship between them, in which case the nature of the research to investigate the more detailed aspects of the association is very different from attempting to find the common, underlying genes.

Analysis of longitudinal data to study causes of phenotypic stability and tracking over time uses many of the same models as used for cross-sectional multivariate studies. Longitudinal analysis of twin data naturally improves statistical power as multiple observations from the same individuals are available. When we are asking whether there are changes in the magnitude of genetic and environmental influences over time, a cross-sectional study design can be used assuming that there are no cohort effects. For example, we found in Finnish twin data that the effect of genetic factors on coffee consumption was higher in young than in middle-aged adults but increased again in old age (10). In order to investigate whether the same genetic and environmental factors operate over time, a longitudinal approach is needed. There are three commonly used methods to analyse this kind of longitudinal data.

The Cholesky decomposition approach has a useful conceptual interpretation as all factors are constrained to impact current and later (but not earlier) time points and is easy to apply to data sets if measurement points are reasonably few (Fig. 1b). The figure illustrates the concept of a Cholesky model for BMI measures at four time points (ages 16, 17, 18, and 25). The full model posits four latent genetic effects acting on each time point and each later point, with an equivalent number of latent environment effects, again at each time point and each later time point. The model can be reduced by dropping statistically non-significant paths, and for example end up with a model with only one latent genetic effect acting equally at all time points. This would imply that no new genes are activated in weight development as subjects age from 16 to 25 years of age. On the other hand, it may turn out that age-specific genetic effects are required implying that new genes are expressed at each age. As a model-free method, Cholesky decomposition can accommodate any pattern of change but is not falsifiable. It is a suitable method if the nature of the growth process or the relationships of the constituent variables are not well known in advance. For longitudinal analyses, Cholesky decomposition has the disadvantage that it makes no prediction about future time points. In longitudinal contexts, it provides information about the relative contribution of genetic and environmental factors to the tracking of trait but is not directly informative about the rate of change in a trait.

The Simplex model makes more restrictive predictions about covariance pattern and hence is falsifiable (Fig. 1c). The degree to which genetic (or environmental) effects are transmitted from one time point to the next can be estimated from the model, while change in BMI is seen as a dynamic process in which new genetic and environmental factors start to affect at each age. This new part of genetic or environmental variation is denoted as innovation parameters ($\zeta_{A_1}-\zeta_{A_4}$ and $\zeta_{E_1}-\zeta_{E_4}$, respectively). Simultaneously some, or all, of the factors affecting at the previous age can also be of importance. This genetic or environmental persistence is denoted as transmission paths in the model (from A_1 to A_2 or E_1 to E_2, etc.). Further an error variance term (ε) is expected to affect each measure. It has the disadvantage that the future (next measurement point) depends on current state (most recent measurement) only (11). As with the Cholesky decomposition, the number of parameters increases with the number of measurements. In a Swedish twin study, a simplex model was used to investigate longitudinal growth in height from age 3 to 18 years of age (12).

The third useful set of longitudinal models is growth curve models (13). They construct two growth variables – initial level (“intercept, denoted as α”) and rate of change (“slope” or β) – to predict level at a series of time points (Fig. 1d). These growth variables are modelled as (first-order) latent factors which load on the observed longitudinal measures. Using twin data, second order latent variables (A_{11} and A_{12} for additive genetic effects, and E_{11} and E_{12} for environmental effects) are used to estimate the contribution of genes and environment to the growth variables (13). In addition, specific environmental effects on each measurement point of BMI are specified (E_1 through E_4). The paths between the additive genetic effects on initial level and rate of change, and the paths between the environmental effects on initial level and rate of change can be used to estimate the genetic and environmental correlations of initial level with rate of change.

The simplest form of individual trajectories is one of constant rate of change over time, but with sufficient data points, such linear growth models can be extended to non-linear processes. Using linear growth models of twin data, we can ask what is the contribution of genetic factors to inter-individual variation in initial level and rate of change. Moreover, we can also conduct a bivariate analysis to see whether the same or different genes influencing initial level and rate of change. This is done formally by estimating the genetic correlation between intercept and slope. Growth models are very efficient and the number of parameters does not increase with number of measurements. It also provides prediction about time points that have not yet been measured.

We have used a latent growth model to estimate genetic effects on BMI level at baseline and the rate of change in BMI over a 15-year study period based on a longitudinal cohort of Finnish twins. They were aged 20–46 years at baseline and provided data on weight and height from mailed surveys in 1975, 1981, and 1990. We found a substantial genetic influence on rate of change in BMI (h^2 for men = 58% (95% CI 0.50–0.69), h^2 for women = 64% (95% CI 0.58–0.69)) (14). The genetic correlation between BMI level and rate of change was virtually zero, suggesting that the genes affecting BMI are different from those involved in weight changes in adults.

Other longitudinal models can be applied to estimate the heritability of traits with variable age of onset, but multivariate models to estimate whether the same genes contribute to say risk factors of disease and the variation in age of onset of the disease are not yet easily available.

3. Co-twin Control Study of Either Disease Discordant or Exposure Discordant Pairs

The co-twin control design was first put forward by Gesell (15) over 60 years ago but was proposed by Fisher (16) in the late 1950s as a design to test the causal versus constitutional hypotheses of smoking as a cause of chronic diseases. Differences between monozygotic (MZ) pairs are taken to arise from environmental differences between them. It is a powerful design for investigating either causes or consequences of a disease by detailed investigations into the differences between twins from discordant pairs. Over time, it has been realised that the environmental differences can arise from a very broad spectrum of effects and over a long time span. Thus, the effective genotype of MZ twins may begin to diverge over time as epigenetic and various environmental effects modify gene expression in the twins, even though their genomic DNA remains unchanged, except for possible somatic mutations. Recently, differences in copy number variants (CNVs) were found in MZ pairs discordant for Parkinson's disease (17), illustrating the power of the discordant MZ twin pair design to investigate the importance of novel genetic mechanisms for disease.

For example, the effects of obesity can be studied in the absence of confounding due to genetic effects using MZ twin pairs discordant for obesity. The obese and the non-obese co-twins share the same genes and differ only by environmental exposures (considered in the broadest sense including epigenetic effects) and the resultant acquired obesity. With this co-twin control design, we were able to identify and study up to 15 healthy MZ pairs with 10–25 kg differences in weight from the FinnTwin16 study. A control group of normal-weight or obesity concordant MZ pairs was also studied. These studies show that acquired obesity is associated with increased liver fat content, insulin resistance, various vascular abnormalities, and multiple changes in adipose tissue metabolism and lipid profiles using lipidomics (18, 19). This observational approach of long-term obesity discordance complements the classic short-term experimental studies of overfeeding MZ pairs (20).

Why then are these MZ twin pairs so strikingly discordant for obesity? They represent only a small fraction of the MZ pairs in the base population and showed some differences in birth weight; however, their growth and weight development was normal in childhood and adolescence until after puberty when intra-pair weight differences began to appear (21). There are differences in their physical activity at age 16–17 years preceding their weight change (22). This suggests that physical activity may be a proximal causal factor for future obesity, consistent with much but not all of the epidemiological literature (23). Experimental evidence indicates a strong inverse link between fitness and fatness; in rats bred for low aerobic capacity, risk factors comprising the metabolic syndrome were much more common than in high capacity rats (24). In mid-adolescence, there may also be differences in dietary patterns and ingestion of specific foods or in other environmental factors. It is also possible that the manifestation of physical inactivity in adolescence and obesity in adulthood is preceded by much earlier events. One mechanism may be through epigenetic modification of gene expression in these MZ pairs. Fraga et al. (25) suggested that epigenetic methylation changes increase with increasing age in trait discordant MZ pairs, but the epigenetic effects relevant to obesity could possibly develop in childhood or even prenatally. Tsankova et al. (26) showed that social stress induced lasting downregulation of brain-derived neurotrophic factor (BDNF) mRNA transcripts and repressive histone methylation of their promoters, suggesting a role for long-term histone remodelling in depression.

An interesting extension of the MZ discordant pair design is the children of twins design. Children of MZ twins share 50% of their genes both with their parents and with their parent's co-twin, i.e. their aunt or uncle. This means that all of the children of a discordant twin pair have the same average genetic risk for disease, while their environment during their upbringing may differ because of their parent's phenotype. This has been used to confirm the importance of genetic liability not only in studies of offspring of MZ pairs discordant for schizophrenia (27), but also in studies of the role of smoking during pregnancy (28). The children to a DZ twin pair may be described as cousins both genetically and socially while the children to an MZ twin pair are social cousins but genetic half-siblings.

4. Genotyping of Twin Data

Genotyping of DZ twins, like siblings, can contribute to estimate associations within and between families and has increased power when additional siblings are added. Including MZ pairs in such data sets, the role of residual polygenic effects can be better defined. In principle, specific genotypic data from DZ pairs can be included in any of the aforementioned twin models. Such models may be estimated by either incorporating the individual genotype data as a moderator variable or by estimating variance components using DZ twin pairs subdivided into groups by identity-by-descent (IBD) status.

Genetic factors play an important role in the responsiveness to changing environmental conditions. Some 20 years ago, Kåre Berg (29) put forward the variability gene concept and indicated empirical evidence in its favour based on studies of intra-pair differences of MZ twin pairs differing in phenotype to study whether specific genetic "variability" loci are associated with differences in lipid levels between the two members of a pair. An association may indicate that some alleles increase or restrict the effect of environmental factors or that environmental factors affect the gene expression differently for different alleles, and this is seen as elevated or diminished levels of variability in trait values. Such a result was seen in a recent study of genetic variants of the Ghrelin gene in MZ pairs discordant and concordant for obesity (30). In the largest study of its kind, genome-wide association (GWA) data for 1,754 MZ female twin pairs from GenomEUtwin consortium (http://www.genomeutwin.org) was used to identify loci affecting serum lipid levels specifically in females (Ripatti et al., unpublished data). In addition to MZ twins, the variability gene effects may potentially also be tested in general population samples using repeated measures of the traits for each individual (however, then assessed at different ages), or even more simply by comparing the trait variances between different genotype classes in general population studies. However, MZ twins offer the additional benefit of controlling for possible confounding due to possible epistatic effects and more control over a range of unmeasured environmental influences in both members of the twin pair.

5. Twin Resources: How to Find and Recruit Twins, Twin Registries Worldwide

For a long time, much of the twin research conducted in the world emanated mostly from countries of Anglo-Saxon origin or Northern Europe. In the past 10–15 years, twin cohort studies and twin registries have expanded to many South and East-Asian countries, thus providing information about the contribution of a much wider range of cultural and environmental influences and in population with different genetic inheritance. A compilation of such studies and registers can be found in two theme issues of *Twin Research and Human Genetics* from October 2002 and December 2006. In addition to such larger and more permanent studies, many smaller twin studies are continuously being carried out using a wide range of study designs (31). Approximately every 50th person is a twin, and twins are born into all classes of society. While twins have some pregnancy and infancy characteristics, such as low birth weight, that distinguish them from singletons, twins are highly representative of the general population for nearly all traits after pregnancy and early childhood.

Research incorporating new methodologies (i.e. transcriptomics, metabolomics and proteomics, neuroimaging) that target endophenotypes of the disease or trait under study is uniquely powerful when used in twin studies, such as the discordant MZ design. In order to take into account developmental aspects underlying the complexity of most phenotypes, it is essential to use longitudinal studies with repeated measures of biomarkers and environmental influences. Twin studies provide a cost-effective approach to such studies. Within the framework of genetically informative data sets, we might need detailed phenotyping starting from gene expression studies all the way to neurobiological and social correlates of human traits and diseases in relevant cultural contexts.

References

1. Posthuma D, Beem AL, De Geus EJ, van Baal GC, von Hjelmborg JB, Iachine I et al. Theory and practice in quantitative genetics. Twin Res 2003; 6(5):361–376.
2. Silventoinen K, Kaprio J, Lahelma E, Koskenvuo M. Relative effect of genetic and environmental factors on body height: differences across birth cohorts among Finnish men and women. Am J Public Health 2000; 90(4):627–630.
3. Schousboe K, Willemsen G, Kyvik KO, Mortensen J, Boomsma DI, Cornes BK et al. Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. Twin Res 2003; 6(5): 409–421.
4. Suviolahti E, Oksanen LJ, Ohman M, Cantor RM, Ridderstrale M, Tuomi T et al. The SLC6A14 gene shows evidence of association with obesity. J Clin Invest 2003; 112(11): 1762–1772.
5. Purcell S. Variance components models for gene-environment interaction in twin analysis. Twin Res 2002; 5(6):554–571.
6. Purcell S, Sham P. Variance components models for gene-environment interaction in quantitative trait locus linkage analysis. Twin Res 2002; 5(6):572–576.
7. Mustelin L, Silventoinen K, Pietiläinen K, Rissanen A, Kaprio J. Physical activity reduces the influence of genetic effects on BMI and waist circumference: a study in young adult twins. Int J Obes (Lond) 2009; 33(1):29–36.
8. Dick DM, Pagan JL, Viken R, Purcell S, Kaprio J, Pulkkinen L et al. Changing environmental influences on substance use across development. Twin Res Hum Genet 2007; 10(2):315–326.
9. Dick DM, Viken R, Purcell S, Kaprio J, Pulkkinen L, Rose RJ. Parental monitoring moderates the importance of genetic and environmental influences on adolescent

smoking. J Abnorm Psychol 2007; 116(1):213–218.

10. Laitala V, Kaprio J, Silventoinen K. Genetics of coffee consumption and its stability. Addiction 2008; 103(12):2054–2061.
11. Boomsma DI, Molenaar PCM. Constrained maximum likelihood analysis of familial resemblance of twins and their parents. Acta Genet Med Gemellol 1987; 36:29–39.
12. Silventoinen K, Pietiläinen KH, Tynelius P, Sorensen TI, Kaprio J, Rasmussen F. Genetic regulation of growth from birth to 18 years of age: the Swedish young male twins study. Am J Hum Biol 2008; 20(3):292–298.
13. Neale MC, Mcardle JJ. Structured latent growth curves for twin data. Twin Res 2000; 3(3):165–177.
14. Hjelmborg JB, Fagnani C, Silventoinen K, Mcgue M, Korkeila M, Christensen K et al. Genetic influences on growth traits of BMI: a longitudinal study of adult twins. Obesity (Silver Spring) 2008; 16(4):847–852.
15. Gesell A. The methods of co-twin control. Science 1942; 95(2470):446–448.
16. Fisher RA. Cancer and smoking. Nature 1958; 182:596.
17. Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, de Stahl TD et al. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 2008; 82(3):763–771.
18. Pietiläinen KH, Sysi-Aho M, Rissanen A, Seppanen-Laakso T, Yki-Järvinen H, Kaprio J et al. Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects – a monozygotic twin study. PLoS One 2007; 2(2):e218.
19. Pietiläinen KH, Naukkarinen J, Rissanen A, Saharinen J, Ellonen P, Keränen H et al. Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity. PLoS Med 2008; 5(3):e51.
20. Bouchard C, Tremblay A, Desprès JP, Nadeau A, Lupien PJ, Thèriault G et al. The response to long-term overfeeding in identical twins. N Engl J Med 1990; 322:1477–1482.
21. Pietiläinen KH, Rissanen A, Laamanen M, Lindholm AK, Markkula H, Yki-Jarvinen H et al. Growth patterns in young adult monozygotic twin pairs discordant and concordant for obesity. Twin Res 2004; 7(5):421–429.
22. Pietiläinen KH, Kaprio J, Borg P, Plasqui G, Yki-Järvinen H, Kujala UM et al. Physical inactivity and obesity: a vicious circle. Obesity (Silver Spring) 2008; 16(2):409–414.
23. Hill JO, Wyatt HR. Role of physical activity in preventing and treating obesity. J Appl Physiol 2005; 99(2):765–770.
24. Wisloff U, Najjar SM, Ellingsen O, Haram PM, Swoap S, Al Share Q et al. Cardiovascular risk factors emerge after artificial selection for low aerobic capacity. Science 2005; 307(5708):418–420.
25. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci USA 2005; 102(30):10604–10609.
26. Tsankova NM, Berton O, Renthal W, Kumar A, Neve RL, Nestler EJ. Sustained hippocampal chromatin regulation in a mouse model of depression and antidepressant action. Nat Neurosci 2006; 9(4):519–525.
27. Gottesman II, Bertelsen A. Confirming unexpressed genotypes for schizophrenia. Risks in the offspring of Fischer's Danish identical and fraternal discordant twins. Arch Gen Psychiatry 1989; 46(10):867–872.
28. D'Onofrio BM, Turkheimer EN, Eaves LJ, Corey LA, Berg K, Solaas MH et al. The role of the children of twins design in elucidating causal relations between parent characteristics and child outcomes. J Child Psychol Psychiatry 2003; 44(8):1130–1144.
29. Berg K. Variability gene effect on cholesterol at the Kidd blood group locus. Clin Genet 1988; 33:102–107.
30. Leskela P, Ukkola O, Vartiainen J, Rönnemaa T, Kaprio J, Bouchard C et al. Fasting plasma total ghrelin concentrations in monozygotic twins discordant for obesity. Metabolism 2009; 58(2):174–179.
31. Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. Nat Rev Genet 2002; 3(11):872–882.

Chapter 12

Mendelian Randomisation: A Tool for Assessing Causality in Observational Epidemiology

Nuala A. Sheehan, Sha Meng, and Vanessa Didelez

Abstract

Detection and assessment of the effect of a modifiable risk factor on a disease with view to informing public health intervention policies are of fundamental concern in aetiological epidemiology. In order to have solid evidence that such a public health intervention has the desired effect, it is necessary to ascertain that an observed association or correlation between a risk factor and a disease means that the risk factor is *causal* for the disease. Inferring causality from observational data is difficult, typically due to confounding by social, behavioural, or physiological factors which are difficult to control for and particularly difficult to measure accurately. A possible approach to inferring causality when confounding is believed to be present but unobservable, as it may not even be fully understood, is based on the method of instrumental variables and is known under the name of *Mendelian randomisation* if the instrument is a genetic variant. While testing for the presence of a causal effect using this method is generally straightforward, point estimates of such an effect are only obtainable under additional parametric assumptions. This chapter introduces the concept and illustrates the method and its assumptions with simple real-life examples. It concludes with a brief discussion on pitfalls and limitations.

Key words: Causal inference, Instrumental variable, Confounding

1. Introduction

The study of risk factors for disease is central to epidemiological research. Here, we distinguish between prognostic and aetiological research by considering the notion of risk in its original context of studying conditions thought to be *caused* by a particular factor and not in the broader sense of *predicting* the probability of a condition for prognostic purposes. For the latter, all factors associated with the outcome are of interest, regardless of whether they are *causal* or not. For aetiological research, the focus is on assessing the effects of modifiable exposures on disease with view

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_12, © Springer Science+Business Media, LLC 2011

to informing health intervention policies. It is hence important to verify that an observed association between the exposure and disease of interest indicates a *causal* relationship between the two. Inferring causality from observed associations is problematic as it is not clear which of two correlated variables is the cause, which the effect, or whether the association is due to another unmeasured factor, or *confounder*. Randomised controlled trials (RCTs) provide the accepted solution since they render reverse causation and confounding implausible. However, RCTs are neither ethical nor practical for many exposures of epidemiological interest, such as exercise, alcohol consumption, and diet regimes. From the practical viewpoint, many exposures develop over years so people cannot be randomised easily to "lifetime" exposure and trials attempting to do so are very costly. Moreover, the population of volunteers in a trial is likely to differ considerably from the general population. Thus, we often have to make causal inferences from observational epidemiological studies and, arguably, we actually need to do so when public health interventions are of interest as we require a representative study population (1).

There have been many success stories where evidence from epidemiological studies has informed public health policy and led to health improvements in the general population. These include the links between smoking and increased risk of lung cancer (2), and between maternal folate supplementation and reduced risk of neural tube defects (3, 4) leading to widespread banning of smoking in public places and the mandatory fortification of cereal flour with folic acid in the USA, Canada, and Chile, for example. There have also been many high-profile failures, where reported associations failed to be replicated in follow-up RCTs. For example, the observation that increased beta-carotene intake reduces the risk of smoking-related cancers was not replicated in the subsequent large-scale RCTs (5–7). More recent failures to replicate observational findings in RCTs include the associations between hormone replacement therapy and cardiovascular disease and between oestrogen levels and Alzheimer's disease or dementia.

There are many reasons why an observational study and an RCT could provide contradictory results. Different dose levels, different durations of follow-up and interactions with other risk factors are usually proposed, but they do not fully explain such discrepancies. The most likely reasons are *confounding* by unobserved lifestyle, socioeconomic factors or baseline health status, *reverse causation*, where the presence of disease influences what is thought to be exposure rather than vice versa, and the usual problems of selection or reporting bias. Since only those associations with high observational support are ever likely to be verified in an RCT, we can only presume that many other reported associations are likely to be non-causal (8). Given the tendency of high-profile findings to persist in the medical literature and thus influence public health and clinical policy long after they have been refuted by RCT evidence (9), it is important to have alternative methods for assessing causality from observational data. Here, we particularly address the case where we have unobserved confounding factors and so cannot adjust our analyses in the usual way.

Mendelian randomisation is an *instrumental variable* (IV) approach to the problem of inferring causality when unobserved confounding is believed to be likely, possibly because the underlying biological processes are not fully understood (1, 10–16). It uses a well-understood genetic variant, known to be associated with the exposure but without direct effect on the disease, as an instrument. The exposure may itself be a phenotype or a genetically influenced behaviour. The fact that genes are assigned randomly at meiosis (given the parental genes) implies that the instrument should be independent of any unobserved confounding between the exposure and the disease, and so we can think of Mendelian randomisation as a natural imitation of a randomised trial although the randomisation is not, of course, perfect. Reverse causation is not an issue here since genes are determined before birth. The basic idea is that there should be no association between the genetic variant and the disease *unless* the considered exposure or phenotype is actually causal for the disease.

2. Mendelian Randomisation

In this section, we introduce a formal framework for causal inference and the core conditions for IV methods with a brief discussion of some of the implications for testing and estimating causal effects in epidemiological applications. We then illustrate the method with some examples.

2.1. Causal Concepts

We first need to formalise how we distinguish between *association* and *causation*. If we say that a variable X is associated with another variable Y, we mean that observing X is informative, or predictive, for Y. The usual conditional probability notation $P(Y=y|X=x)$ describes the distribution of Y, given that we happen to know that $X=x$ has occurred.

We regard causal inference to be about studying the effect of *intervening* in a particular system (17–20). Other causal frameworks are based on counterfactual or potential outcome variables (21) or structural equation models (15, 22). It can be argued that the notion of intervention is implicit to all these formal approaches to causality (23, 24). Specifically, when we say that X *causes* Y, we mean that *manipulating* or *intervening on* X is informative for Y. Ordinary conditional probability notation does not reflect the changes in the distribution of Y when X is *set* to a particular value. We need some extra notation in order to formally distinguish between association and causation. We use the notation do($X=x$) to represent the intervention of setting X to a value x, as suggested by Pearl (22). The two conditional distributions $P(Y=y|\text{do}(X=x))$ and $P(Y=y|X=x)$ are not necessarily the same. The former depends on *x only* if X is causal for Y and corresponds to what we observe in a randomised study. The latter also depends on x, for instance when there is confounding or reverse causation, and this is what we observe in an observational study. As a simple example, let X be a binary variable indicating whether an individual's fingers are stained yellow or not, and let Y be a binary outcome for lung cancer. Since we know that stained fingers are due to smoking and smoking causes lung cancer, $p(y|x)$ describes how someone's risk of lung cancer can be predicted from inspection of their fingers. However, if we could intervene on X by staining or removing the stain from everyone's fingers, for example, $p(y|\text{do}(x))$ would no longer depend on x since finger stain in its own right does not affect lung cancer risk.

A *causal effect* is some contrast of two different interventions on X (x_1 and x_2) on the outcome Y. For continuous outcomes, the *average causal effect* (ACE), describing the average change in Y induced by setting X to be some value x_2 compared with a baseline value x_1, is an obvious causal effect parameter to consider and is the parameter that we focus on for illustrative purposes. It is defined as

$$\text{ACE}(x_1, x_2) = E(Y \mid \text{do}(X = x_2)) - E(Y \mid \text{do}(X = x_1)). \qquad (1)$$

When Y is binary, the *causal relative risk* (CRR), given by

$$\text{CRR} = \frac{P(Y=1 \mid \text{do}(X=x_2))}{P(Y=1 \mid \text{do}(X=x_1))},$$

or the *causal odds ratio* (COR), defined analogously, are both relevant parameters. A causal effect is identifiable if we can show mathematically, under the model assumptions and given the observable data, that the expression of the ACE in Eq. 1 – or equivalent expression for other parameters – is equal to an expression without the "do()" notation that depends purely on observational terms. Sometimes, this can be achieved by adjusting for a sufficient set of observed confounders in the usual way (18, 22). IV methods provide an alternative approach when unobserved confounding is present.

The ACE, CRR, and COR are all *population* parameters in that they are defined in terms of changes across the whole population of interest. There are other *local* causal effect parameters defined on specific subgroups of the population that we might wish to target, depending on the focus of the analysis. One well-known local effect is the "effect of the exposure on the exposed" or, the "effect of treatment received" as it is sometimes described in a clinical trials context. Different causal parameters are identifiable under subtly different assumptions which in turn need to be justified in any given case. The interpretation of the various parameters in epidemiological applications is also open to some debate. We do not go into details here but some discussion of these issues can be found in (15, 25–27).

2.2. Instrumental Variables

Let the variable X represent the modifiable phenotype or exposure of interest, and let Y be the outcome or disease indicator, as before. We let G denote the known genetic variant associated with X which plays the role of instrument in our approach. We are interested in the causal effect of X on Y when we believe that unobserved confounding is present. Denote the unobserved confounder(s) by U. Causal inference using IV methods falls into two main categories. In order to *test* that an association is causal, we need to make certain (in)dependence assumptions concerning the four variables above. In addition, we have to make some structural assumption describing how any proposed intervention affects their joint distribution. For *estimating* the causal effect, should it seem likely that one is present, we need further parametric assumptions.

There are some *core conditions* that must be satisfied in order for the genetic variant, G, to qualify as an instrument (14, 16, 20). Using the notation $A \perp\!\!\!\perp B|C$ to mean "A is independent of B given C", these can be stated as follows:

1. $G \perp\!\!\!\perp U$ – the genetic variant is unrelated to the confounding between X and Y.
2. $G \not\!\perp\!\!\!\perp X$ – the genetic variant is associated with the exposure and the stronger this association, the better.
3. $G \perp\!\!\!\perp Y|(X, U)$ – given the exposure status and the confounders (if the confounders were observable), the genetic variant does not provide any additional information for the outcome, i.e. there is "no direct effect" of G on Y and no other indirect effect other than through X.

These three conditional independence assumptions define a unique directed acyclic graph (DAG) connecting the variables G, X, Y, and U as shown in Fig. 1. An equivalent statement is that

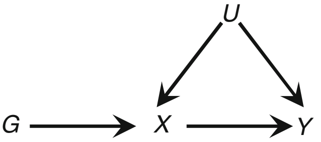


Fig. 1. The unique DAG connecting G, X, Y, and U described by the core IV conditions.

the joint distribution of the four variables factorises in the following way:

$$p(y,u,g,x) = p(y \mid x,u)p(x \mid u,g)p(u)p(g).$$

Note that these assumptions do *not* imply that $G \perp\!\!\!\perp Y|X$ or $G \perp\!\!\!\perp Y$, as has been sometimes misunderstood. Furthermore, assumptions 1 and 3 cannot be easily tested from data as they depend on U, which is unobserved, and hence have to be justified from background knowledge. In our case, the first assumption means that you must be reasonably satisfied that G is not associated with the sort of confounding you might typically expect for any particular X–Y relationship. However, Mendelian randomisation is based on the idea that genes are randomly assigned at meiosis and this implies that, across the population, genetic effects are relatively robust although not completely immune to confounding (28). Assumption 3 demands a comprehensive understanding of the underlying biological and clinical science and may be appropriately considered in a sensitivity analysis of alternative pathways.

So far, we have made assumptions about how our four variables are related "naturally". The additional structural assumption concerns what happens to the joint distribution when we intervene on X, and demands that the distributions $p(y|x, u)$, $p(g)$, and $p(u)$ are not changed by the particular intervention in X, i.e. are not changed when conditioning on do(X=x). This implies that the joint distribution under intervention is given by

$$p(y,u,g,x \mid \mathrm{do}(X = x^*)) = p(y \mid x^*,u)\mathbf{1}\{x = x^*\}p(u)p(g),$$

where $\mathbf{1}\{x=x^*\}$ is the indicator function taking the value 1 if $x=x^*$ and 0 otherwise. The plausibility of this assumption depends, of course, on the type of intervention being considered and needs to be justified based on background knowledge. For instance, a drug that adjusts homocysteine level might plausibly be judged to leave an individual's lifestyle behaviour unchanged. There could, however, be a placebo effect that changes the distribution of Y more than is warranted by the new value of X (homocysteine level), or the drug could affect other relevant biological processes in the body. On the other hand, if people are prevented from drinking alcohol by some change in the law, then their other health and lifestyle behaviours might change in order to compensate. Graphically, as shown in Fig. 2, intervening on X removes all

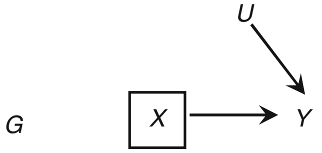


Fig. 2. The DAG representing the core IV conditions under intervention in *X*.

directed edges into X. In particular, we can see from the graph that, under intervention, we get what is often called the *exclusion restriction* (15): $G \perp\!\!\!\perp Y | \mathrm{do}(X=x)$.

Note that by deleting either the edge $U \rightarrow Y$ or $U \rightarrow X$ from Fig. 1, a test for independence of Y and G given X ($Y \perp\!\!\!\perp G|X$) is tantamount to a test for no confounding between X and Y, but we are not aware of this having been used in practice.

2.3. Testing for a Causal Effect

The three core IV conditions in Subheading 2.2, together with the structural assumption, are sufficient to test for a causal effect of X on Y *regardless* of the distributional form of the factors of the joint distribution. We do require that the joint probability distribution is *faithful* to the relevant DAG in that there are no conditional independence relationships, perhaps due to interactions that cannot be read from the graphs. Consequently, any appropriate statistical test of association between the instrument G and outcome Y amounts to a test for a causal effect of X on Y. (See ref. 16 for a detailed discussion.)

2.3.1. Homocysteine and Stroke

Is there a causal relationship between high plasma homocysteine concentrations and risk of stroke (29)? The T allele of the MTHFR C677T polymorphism is known to be associated with homocysteine levels with TT homozygotes having higher levels than CC homozygotes, in particular. A summary estimate of the association between homocysteine levels and risk of stroke (X–Y relationship) from a meta-analysis gave an odds ratio of 1.59 corresponding to a 5 μmol/L observed increase in homocysteine. Dichotomising MTHFR into TT and CC carriers, the odds ratio for the genotype–stroke association (G–Y) was 1.26 and found to be significant. The conclusion is that a causal effect of homocysteine level on risk of stroke is plausible. However, *none* of the reported values give any indication of the *size* of this effect as the homocysteine–stroke association could be confounded.

2.3.2. Plasma Fibrinogen and CHD

Do higher plasma fibrinogen levels increase the risk of coronary heart disease (CHD) (30)? Various observational studies have reported increased risk associated with higher fibrinogen levels for various cardiovascular outcomes. The G-455 → A polymorphism in the promoter region of the β-fibrinogen gene is consistently associated with differences in fibrinogen levels and plays the role of the instrument. A meta-analysis of 16 studies produced a “per allele” odds ratio of 0.96 with associated 95% confidence interval of (0.89, 1.04). The conclusion is that there is no support for a causal effect of fibrinogen levels on CHD; or in other words, if fibrinogen has a causal effect then it is too small to be detected in this meta-analysis of 16 studies.

2.3.3. Alcohol Consumption and Blood pressure

The previous examples concerned the effect of some intermediate phenotype on a disease. We can also use the idea of Mendelian randomisation when we have a modifiable exposure, such as alcohol consumption, for which positive (e.g. CHD) and negative (e.g. liver cirrhosis and some cancers) effects have been reported in observational studies. Besides being difficult to measure due to reporting bias, alcohol consumption is strongly associated with all kinds of confounding factors, and so there are doubts about the causal nature of any of the above associations (31). We consider the issue of whether there is a causal effect of alcohol consumption on blood pressure.

The ALDH2 gene determines blood acetaldehyde, the principal metabolite for alcohol, and is known to be associated with alcohol consumption. In particular, individuals homozygous for the "null" variant *2 suffer unpleasant symptoms, such as facial flushing, nausea, drowsiness, and headache after alcohol consumption. Heterozygotes have a limited ability to metabolise acetaldehyde but have a less severe reaction than *2*2 homozygotes. Consequently, *2*2 homozygotes have lower alcohol consumption than the "wild type" *1*1 homozygotes *regardless* of their other lifestyle behaviours while heterozygotes tend to drink intermediate amounts. In the meta-analysis of Chen et al. (31), there was no apparent association between ALDH2 and typical confounding factors that one would expect for the alcohol–blood pressure relationship. This, together with the random allocation of genes at conception makes us fairly confident about core IV assumption 1. Current knowledge of the biochemical function of ALDH2 excludes the possibility that it could be associated with blood pressure via another pathway besides alcohol consumption (core IV assumption 3).

Blood pressure was found to be 7.44 mmHg higher on average for *1*1 homozygotes than for *2*2 homozygotes with 95% CI (5.39, 9.49) yielding a p value of $p = 1.1 \times 10^{-12}$ for high versus low consumption. Blood pressure was 4.24 mmHg higher on average for *1*2 heterozygotes than for *2*2 homozygotes with 95% CI (2.18, 6.31) giving a p value of $p = 0.00005$ for moderate versus low consumption. Most of the studies were on Japanese populations (where ALDH2*2*2 is common) so these results are for males as Japanese women drink very little alcohol in general. The fact that there was no observed relationship between genotype and blood pressure for women indicated that the above association is indeed due to alcohol consumption, for which ALDH2 is a proxy, and not due to the gene itself or some alternative pathway by which ALDH2 might predict blood pressure. The highly significant association between the ALDH2 variant and blood pressure is strong evidence of a causal effect. In fact, contrary to reported observational claims, it would appear that even moderate drinking can be harmful.

2.4. Estimating a Causal Effect

Once a test indicates that a causal effect is likely, we would typically want to know the size of this effect. This is more difficult. When all variables are binary, or categorical, only upper and lower *bounds* on the causal effect can be calculated without any extra assumptions (32). The width – and hence usefulness – of these bounds depend on the strength of the IV and the amount of confounding, but they do give an idea of how informative the data are. Hence, the IV core conditions and structural assumption are not sufficient for point identification of causal parameters and extra parametric assumptions are required.

When the ACE of Eq. 1 is of interest and *Y* is continuous (possibly suitably transformed), it is popular to assume *linearity* of all relationships and *no interactions.* The structural (causal) assumption only appears in the regression of *Y* on *X* and *U*:

$$E(Y \mid X = x, U = u) = E(Y \mid \mathrm{do}(X = x), U = u) = \mu + \beta x + \delta u$$

This yields β=ACE(x+1, x) as the relevant causal parameter. Note that the above assumes that there is no effect modification of the effect of *X* on *Y* by *U* on the linear scale, i.e. people in various subpopulations, like men/women or older/younger people, all react in the same way to exposure. The parameter β cannot be estimated from the above regression as *U* is unobserved. Likewise, we cannot ignore *U* and estimate it from a regression of *Y* on *X* as this would give a biased estimate due to the collinearity $U \not\perp\!\!\!\perp X$. From the regression of *X* on *G* and *U*

$$E(X \mid G = g, U = u) = \eta + \alpha g + \zeta u,$$

we *can* estimate α by ignoring *U* since $G \perp\!\!\!\perp U$. It is easy to show (16) that

$$E(Y \mid G = g) = \tilde{\mu} + \alpha\beta \cdot g,$$

so $\alpha\beta$ can be estimated from a regression of *Y* on *G*. Hence, a consistent estimator for $ACE(x+1, x) = \beta$ is given by the ratio of the estimated coefficients, $\hat{\beta}_{Y|G}$ and $\hat{\beta}_{X|G}$ from the regressions of *Y* on *G* and of *X* on *G*, respectively. It is useful that these could even be estimated from separate studies, one where *X*, *G* are observed and another one where *Y*, *G* are observed. In this situation, the above ratio estimator is equivalent to the popular "two-stage least squares" (2SLS) estimator which regresses Y on values of *X* predicted from the "first-stage" regression of *X* on *G* and the terms are often used interchangeably.

In an investigation into the causal effect of circulating C-reactive protein (CRP) and the metabolic syndrome, three-SNP haplotypes from the CRP gene were used as instruments for associations between serum CRP levels and various metabolic syndrome phenotypes (33). For one particular outcome – insulin resistance measured by homoeostasis model assessment (HOMA-R) – a clear observational association was reported with a doubling of CRP levels leading to a significant increase of about 8% in HOMA-R ($p < 0.0001$). But CRP is known to be associated with a wide range of lifestyle and socioeconomic characteristics. Moreover, it could be elevated as a result of atherosclerosis or insulin resistance, so confounding and reverse causation cannot be excluded. The core IV assumptions appear to be reasonably satisfied, although not enough is known about the biological pathways involving CRP to be fully sure about assumption 3. Standard checks for linearity on log(CRP) and log(HOMA-R) looked reasonable. It is, of course, impossible to check any parametric assumptions about the unobserved confounders, especially the one of no effect modification. This has to be justified with subject matter background knowledge, instead. The 2SLS approach, using the regressions of log(HOMA-R) on the CRP haplotypes and of log(CRP) on the CRP haplotypes, estimated that doubling CRP levels *reduces* the HOMA-R score by 6% ($p > 0.1$). Since the result is non-significant, we conclude that the data do not support a causal effect of circulating levels of CRP on insulin resistance. This result appears to contradict the naive analysis, which may indeed be due to confounding and reverse causation.

3. Further Issues and Complications

There are well-known problems with the 2SLS-estimator. The standard deviation of the estimator is typically much larger than that of the estimator obtained from a naive regression of Y on X. This is especially so when we have a weak instrument, i.e. when $Corr(G, X) \approx 0$ so that IV is not very informative for X. Note that it is impossible to find a strong instrument when there is a lot of confounding. One notable problem is that the assumption of linearity cannot be true when Y is binary, although it could be a good approximation over a particular range of exposure levels in some cases. This is an issue for epidemiological applications since many outcomes of interest are naturally binary.

The main problem for the non-linear case is that the relationship between the two regressions (Y on G, and X on G) and the relevant causal parameter, i.e. CRR or COR, is no longer straightforward and any estimators derived from these are biased (16, 27). There are other IV methods that can yield estimates of certain causal effects for binary outcomes, but they all require strong additional assumptions (15, 34–36). It is important to note that different approaches target different causal parameters in the sense that they estimate individual, local, or population effects. Some estimators, such as those derived from structural mean modelling approaches, also require joint observation of all three variables (G, X, and Y) for all individuals, whereas the "Wald-type" estimators based on ratios of differences (of which

2SLS is an example), do not (27). This has implications for meta-analyses as not all studies typically supply joint observations. Structural mean models make weaker assumptions than the other approaches in that a parametric model for the regression of X on G and U is not required. However, these approaches target the local effect of exposure on the exposed and are only unbiased for a population effect with additional assumptions.

Violations of the core IV conditions are also possible and these have implications even for testing for a causal relationship. The most important and likely violation occurs when there is population stratification, where we have different allele frequencies in subpopulations which may in turn also differ in their lifestyles (giving rise to an association between G and U) or in their disease risk (giving rise to an association between G and Y not screened off by X, U). A sensible study design should take this possibility into account. The chosen instrument could also be in linkage disequilibrium (LD) with another variant which is associated with the disease via a route other than through its effect on X, the exposure of interest. Likewise, problems can be caused by pleiotropic effects and canalisation or developmental compensation (1, 10, 37). If only insufficient prior knowledge about the genetic or confounding mechanisms is available to justify the core conditions, results that seem to indicate a causal effect may very well have an alternative, non-causal explanation that we are not aware of. DAGs can be used to represent what is believed about the biology and then be queried regarding the validity of our assumptions (12, 16). For example, the genetic variant chosen as instrument may not be *the* causal gene for the exposure of interest but is in LD with a causal gene which is unobserved. This could be thought of as *measurement error* in the genetic data. However, as illustrated in Fig. 3, this does not necessarily imply any violations of the core IV conditions. G_1 might not be as good an instrument as G_2 in the sense that its association with X is weaker, but it is (1) independent of U, (2) associated with X, and (3) conditionally independent of Y given X and U.

Finding a genetic variant that is a suitable IV is also problematic, and there are currently not very many well-studied variants for the typical exposures of interest in epidemiology. Genetic variants that arise from genome-wide association studies could be problematic in that the gene–phenotype associations are often weak

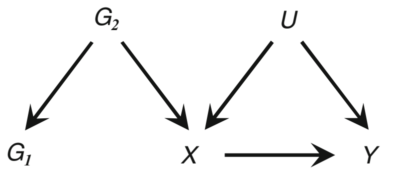


Fig. 3. The chosen instrument G_1 is not causal for X but is associated with another genetic variant, G_2, which is driving all the association. All IV core conditions are satisfied for G_1.

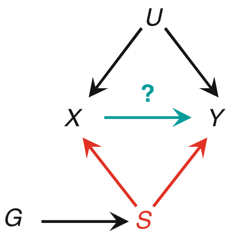


Fig. 4. $G \not\perp\!\!\!\perp X$ and $G \perp\!\!\!\perp U$ but the IV core conditions are not satisfied.

and may not even be reproducible. Even when strong, reproducible associations are found, we then have to be convinced that enough is known about the functionality of the gene in order to claim that the core conditions are satisfied for an IV analysis. Such knowledge does not derive from an association study. On the positive side, thanks to the current rapid advances in functional genomics, the required information on such variants is gradually being accrued. Figure 4 depicts a situation, where we have a clear association between G and X, we can argue the independence between G and U but without understanding the functionality of the gene, there is no way of knowing that the third condition is violated. Since an association between G and Y is evident, we would incorrectly deduce that X causally affects Y, whereas an alternative explanation for this association is that the gene causes an unobservable health problem S which then affects both X and Y.

4. Conclusion

A Mendelian randomisation analysis is *not* aimed at identifying genetic factors that are causal for disease risk. On the contrary, the method requires a known and well-understood genetic variant in order to facilitate causal inference about the effect of an exposure on the disease of interest. One of the limitations for IV methods is finding valid instruments. This is also an issue with genetic instruments in our applications but is hopefully becoming less so with the recent rapid advances in genetic epidemiology (8). Inferring causality from observational data is problematic, but we would argue that some of the confusion about misleading results from observational studies stems from the lack of clear delineation between the notions of association and causation, at a conceptual as well as formal level (24). Only when this distinction is made explicit, can we identify and understand the crucial assumptions that permit a causal interpretation. Only then are we able to critically scrutinise these assumptions, justify or reject them, and hence assess the practical impact of any results. Solid background knowledge is essential for causal analysis. With Mendelian randomisation, we have an advantage over many other areas of application of IV methods in that genetics provide a rich source of information.

Acknowledgement

We acknowledge research support for all authors from the Medical Research Council through a collaborative project grant (G0601625) addressing causal inference in observational epidemiology using Mendelian randomisation.

References

1. D.A. Lawlor, R.M. Harbord, J.A.C. Sterne, N. Timpson, and G.D. Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27:1133–1328, 2008.
2. R. Doll, R. Peto, J. Boreham, and I. Sutherland. Mortality from cancer in relation to smoking: 50 years observations on British doctors. *British Journal of Cancer*, 92:426–429, 2005.
3. MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council vitamin study. *Lancet*, 338:131–137, 1991.
4. T.O. Scholl and W.G. Johnson. Folic acid: influence on the outcome of pregnancy. *American Journal of Clinical Nutrition*, 71 (Suppl.):12955–13035, 2000.
5. W.C. Willett. Vitamin A and lung cancer. *Nutritional Review*, 48:201–211, 1990.
6. R. Peto, R. Doll, J.D. Buckley, and M.B. Sporn. Can dietary beta-carotene materially reduce human cancer rates? *Nature*, 290:201–208, 1981.
7. Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine*, 330:1029–1035, 1994.
8. G.D. Smith, S. Ebrahim, S. Lewis, A.L. Hansell, L.J. Palmer, and P.R. Burton. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet*, 366:1484–1498, 2005.
9. A. Tatsioni, N.G. Bonitis, and J.P.A. Ioannidis. Persistence of contradicted claims in the literature. *Journal of the American Medical Association*, 298:2517–2526, 2007.
10. G.D. Smith and S. Ebrahim. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32:1–22, 2003.
11. M.B. Katan. Commentary: Mendelian randomization, 18 years on. *International Journal of Epidemiology*, 33:10–11, 2004.
12. N.A. Sheehan, V. Didelez, P.R. Burton, and M.D. Tobin. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Medicine*, 5:e177, 2008.
13. R.J. Bowden and D.A. Turkington. *Instrumental Variables.* Cambridge University Press, Cambridge, 1984.
14. S. Greenland. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29:722–729, 2000.
15. M.A. Hernán and J.M. Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17:360–372, 2006.
16. V. Didelez and N.A. Sheehan. Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16:309–330, 2007.
17. J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–710, 1995.
18. S.L. Lauritzen. Causal inference from graphical models. In O.E. Barndorff-Nielsen, D.R. Cox, and C. Kluppelberg, editors, *Complex Stochastic Systems*, Chapter 2, 63–107. Chapman & Hall, Boca Raton, 2000.
19. A.P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002.

20. A.P. Dawid. Causal inference using influence diagrams: the problem of partial compliance. In P.J. Green, N.L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, 45–81. Oxford University Press, Oxford, 2003.
21. D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.
22. J. Pearl. *Causality*. Cambridge University Press, Cambridge, 2000.
23. M.A. Hernán. A definition of causal effect for epidemiologic research. *Journal of Epidemiology and Community Health*, 58:265–271, 2004.
24. V. Didelez and N.A. Sheehan. Mendelian randomisation: why epidemiology needs a formal language for causality. In F. Russo and J. Williamson, editors, *Causality and Probability in the Sciences*, volume 5, *Texts in Philosophy*, 263–292. London College Publications, London, 2007.
25. J.M. Robins, T.J. VanderWeele, and T.S. Richardson. Comment on: Causal effects in the presence of non compliance: a latent variable interpretation. *Metron*, 64:288–298, 2006.
26. S. Geneletti and A.P. Dawid. The effect of treatment on the treated: a decision theoretic perspective. In P. McKay Illari, F. Russo and J. Williamson, editors, *Causality in the Sciences*,Oxford University Press, 2010.
27. V. Didelez, S. Meng, and N.A. Sheehan. Assumptions of IV methods for observational epidemiology. Statistical Science, 25: 22-40, 2010.
28. G.D. Smith, D.A. Lawlor, R. Harbord, N. Timpson, I. Day, and S. Ebrahim. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine*, 4:e352, 2007.
29. J.P. Casas, L.E. Bautista, L. Smeeth, P. Sharma, and A.D. Hingorani. Homocysteine and stroke: evidence on a causal link from Mendelian randomisation. *Lancet*, 365:224–232, 2005.
30. G.D. Smith, R. Harbord, J. Milton, S. Ebrahim, and J. Sterne. Does elevated plasma fibrinogen increase the risk of coronary heart disease? *Arteriosclerosis, Thrombosis and Vascular Biology*, 25:2228–2233, 2005.
31. L. Chen, G.D. Smith, R. Harbord, and S.J. Lewis. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS Medicine*, 5:e52, 2008.
32. A.A. Balke and J. Pearl. Counterfactual probabilities: computational methods, bounds and applications. In R.L. Mantaras and D. Poole, editors, *Proceedings of the 10th Conference on Uncertainty in Artificial Inteligence*, 46–54, 1994.
33. N.J. Timpson, D.A. Lawlor, R.M. Harbord, T.R. Gaunt, I.N.M. Day, L.J. Palmer, A.T. Hattersley, S. Ebrahim, G.D.O. Lowe, A. Rumpley, and G.D. Smith. C-reactive protein and its role in metabolic syndrome: a Mendelian randomisation study. *Lancet*, 366:1954–1959, 2005.
34. J.M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics*, 23:2379–2412, 1994.
35. S. Vansteelandt and E. Goetghebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society, Series B*, 65:817–835, 2003.
36. P. Clarke and F. Windmeijer. Instrumental variable estimators for binary outcomes. Working Paper 10/239, Centre for Market and Public Organisation, University of Bristol, 2010.
37. D. Nitsch, M. Molokhia, L. Smeeth, B.L. DeStavola, J.C. Whittaker, and D.A. Leon. Limits to causal inference based on Mendelian randomization: a comparison with randomised controlled trials. *American Journal of Epidemiology*, 163:397–403, 2006.

Chapter 13

Copy Number Variation

Louise V. Wain and Martin D. Tobin

Abstract

Recent genetic epidemiology studies have been dominated by genome-wide association (GWA) studies using single nucleotide polymorphisms (SNPs). However, a form of structural genomic variation, termed copy number variation (CNV), is also widespread throughout the human genome, and can be highly polymorphic between individuals. Such variation has long been shown, through candidate gene studies using low-throughput molecular biology techniques, to have direct consequences on human health and variation. Many studies have now sought to extensively characterise this variation on a genome-wide scale and, increasingly, attempts are being made to identify associations between CNV and human disease. Although many of the study design issues that have been described for SNP GWA studies are also relevant for CNV GWA studies, CNV studies also present their own unique set of challenges and considerations. New microarray-based technologies are enabling more accurate mapping of CNVs, and CNV maps of the human genome are being regularly refined with increasing resolution. The study of CNV and its effects on human health and disease therefore present a dynamic and exciting challenge for researchers in the field of genetic epidemiology.

Key words: Copy number variation, Structural variation, Genome-wide association studies, Genetic epidemiology, Human disease, Human variation

1. Introduction

The human genome varies between individuals not only at the sequence level but also structurally. Specifically, deletions and duplications can result in alterations in the diploid copy number of affected segments of the human genome. This is referred to as copy number variation (CNV). In recent years, a number of studies have shown that human genomic CNV is widespread (1–12) and this evidence has not only prompted major efforts to further characterise CNVs, but it has also stimulated interest in the extent to which these variants may influence human health and disease.

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_13, © Springer Science+Business Media, LLC 2011

2. Copy Number Variation

2.1. Basic Genetics

The term CNV is used to describe a sub-microscopic region of genomic DNA of up to several megabases, which can be deleted or duplicated (or both) such that its copy number on either chromosome varies from one. This form of variation is analogous to microsatellites and minisatellites which consist of very short repeated units (2–4 nucleotides for microsatellites and 10–100 nucleotides for minisatellites), but the term CNV usually refers to larger repeated regions of more than several hundred nucleotides. CNVs may contain any number of genes or partial genes depending on the size of the variable region and are polymorphic, to varying degrees, both within and between populations (13, 14). Simple deletions and duplications can be assigned to diallelic genotypes (e.g. AA/A–/–– for deletions where "A" is the observed allele and "–" represents a deletion, or AA/A+/++ where "+" represents a duplication of "A" at one copy of the locus) (Fig. 1 i, ii). However, unlike single nucleotide polymorphisms (SNPs), CNVs are not necessarily diallelic. Multiallelic CNVs show diploid copy numbers consistent with both deletions and duplications, or even multiple duplications (amplification) at the same locus (Fig. 1iii).

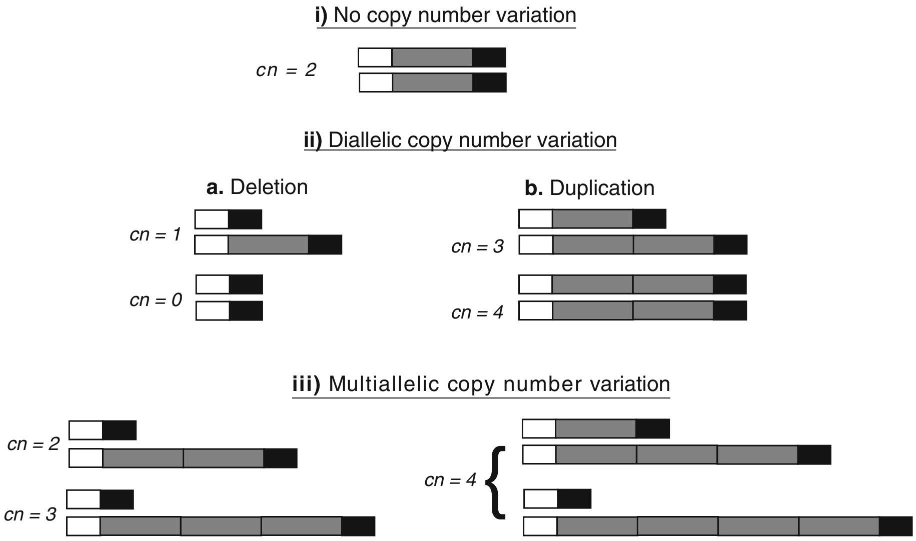

i) No copy number variation
cn = 2
ii) Diallelic copy number variation
a. Deletion
b. Duplication
cn = 1
cn = 0
cn = 3
cn = 4
iii) Multiallelic copy number variation
cn = 2
cn = 3
cn = 4

Fig. 1. Copy number variation. *Grey boxes* represent the presence of a locus. *White* and *black boxes* represent the flanking regions of the locus. (**i**) No copy number variation at the locus: diploid copy number (cn) of two. (**ii**) Diallelic copy number variation. (**a**) Heterozygous and homozygous deletion resulting in copy numbers of one and zero, respectively. (**b**) Heterozygous and homozygous duplication resulting in copy numbers of three and four, respectively. (**iii**) Multiallelic copy number variation. Deletion and duplication at the same locus, and multiplications, can result in overall copy numbers that are indistinguishable from those that result from diallelic situations or from situations where no copy number variation is present (cn = 2).

2.2. Effects of CNV

Deviation from the "normal" diploid (two copies, Fig. 1ii(a)) state has profound consequences when large genomic regions are involved (15). The so-called genomic disorders such as Down syndrome, Cri du chat syndrome, and Charcot Marie Tooth 1A syndrome are examples of where the gain or loss of complete chromosomes or very large segments has very marked phenotypic consequences (16). On a more subtle level, copy number of many loci varies between apparently healthy individuals implying a contributory role for CNV in normal human variation (13, 14, 17–19). In some cases, this phenotypic variation has been associated with susceptibility to diseases such as glomerulonephritis (20), psoriasis (21), systemic lupus erythematosus (SLE) (20, 22, 23), and HIV (13, 24, 25).

The effects of a CNV depend upon its copy number and also on its location relative to genes or their regulatory elements. High copy numbers of the *AMY1* gene, which encodes salivary amylase, are associated with higher protein levels which in turn appear to be associated with dietary starch consumption on a population level (14). This suggests that extra copies of the *AMY1* gene are expressed leading to higher concentrations of the product in saliva. Similarly, an association between the copy number of the *CCL3L1* gene and expression of the chemokine, CCL3L1, has been shown; chemokine expression appears to be proportional to CCL3L1 expression for lower copy numbers, with a plateau of chemokine production for higher copy numbers (13, 26). The copy number of the Fc-gamma receptor gene, *FCGR3B*, correlates with the level of protein expression and the ability of neutrophils to localise to immune complexes (27). On the other hand, duplication of the X chromosome green photopigment gene, *OPN1MW*, does not appear to lead to increased protein expression. Phenotypic variations in colour vision result only from sequence mutations affecting the first copy of *OPN1MW*. Perhaps the second and subsequent copies of *OPN1MW* are too far from the regulatory region to be expressed (28).

A deletion or duplication of one or both copies of a complete gene could result in a proportional reduction of the gene product or its absence. Heterozygous or homozygous deletion of the complement regulator genes *CFHR1* and *CFHR3* is associated with CFHR1 and CFHR3 deficiency with an increased risk of atypical haemolytic uremic syndrome (aHUS) (29). Similarly, the loss or gain of individual exons, the coding regions within a gene, could result in a defective gene product that is either non-functional or has a mutated function (e.g. deletion of exons of the *AAAS* gene results in Triple A syndrome (30)).

Loss, gain, or disruption of regulatory regions, such as promoters, enhancers, and transcription factor binding sites upstream of a gene, may also affect expression (31–33). While deletion of *CFHR1/CFHR3* increases the risk of aHUS, it is possible to carry the deletion with no obvious clinical effects suggesting that the phenotypic consequences of this deletion can be partially or wholly compensated for, possibly by downstream feedback mechanisms (29).

3. CNV Discovery

3.1. Approaches to Detect CNVs

Candidate gene studies using molecular methods, such as fluorescent in situ hybridisation (FISH), Southern blotting, and polymerase chain reaction (PCR)-based methods, for the detection and quantification of CNV at individual loci predate the development of the currently widely used microarray-based screening methods by decades (34–37). These low-throughput methods have continued to develop alongside high-throughput array-based methods, but are not amenable to large-scale multiplexing and remain impractical for screening the entire genome, or even very large genomic segments. However, they are often utilised to validate findings from genome-wide studies (38–43).

Comparative genomic hybridisation (CGH) is used in cancer genetics to detect structural variation between normal cells and tumour cells. The DNA from each source is labelled with a different fluorescent dye, mixed, and hybridised to a target. In traditional CGH, this target would be a chromosome spread, and regions of the genome that differed in copy number between the two cell types could be identified by the fluorescence ratios along the chromosomes. For array-CGH, rather than whole chromosomes, the target genome is represented at a much higher resolution by tens of thousands of probes on a microarray. This technology is now commonly used to identify regions of structural variation between individuals. The probes may be artificially synthesised oligonucleotides of 20–85 nucleotides in length, or clone constructs [such as bacterial artificial chromosome (BAC) clones] containing fragments of genomic DNA, typically 80–200 kb. These cloned regions can be computationally mapped back to specific regions of the genome. Figure 2 illustrates array-CGH. ROMA (representational oligonucleotide microarray analysis) (44) is a variation on array-CGH. Here, a representational subset of the genome based on amplification of short restriction endonuclease fragments that can be easily mapped against the human genome reference sequence is used thereby reducing the complexity of the genome.

SNPs within heterozygous deletions may be called as homozygous during genotyping. Long runs of homozygous calls (loss of heterozygosity) may, therefore, be indicative of deletions (see Fig. 3a). Runs of SNPs deviating from Hardy–Weinberg equilibrium or showing Mendelian inconsistencies can also indicate

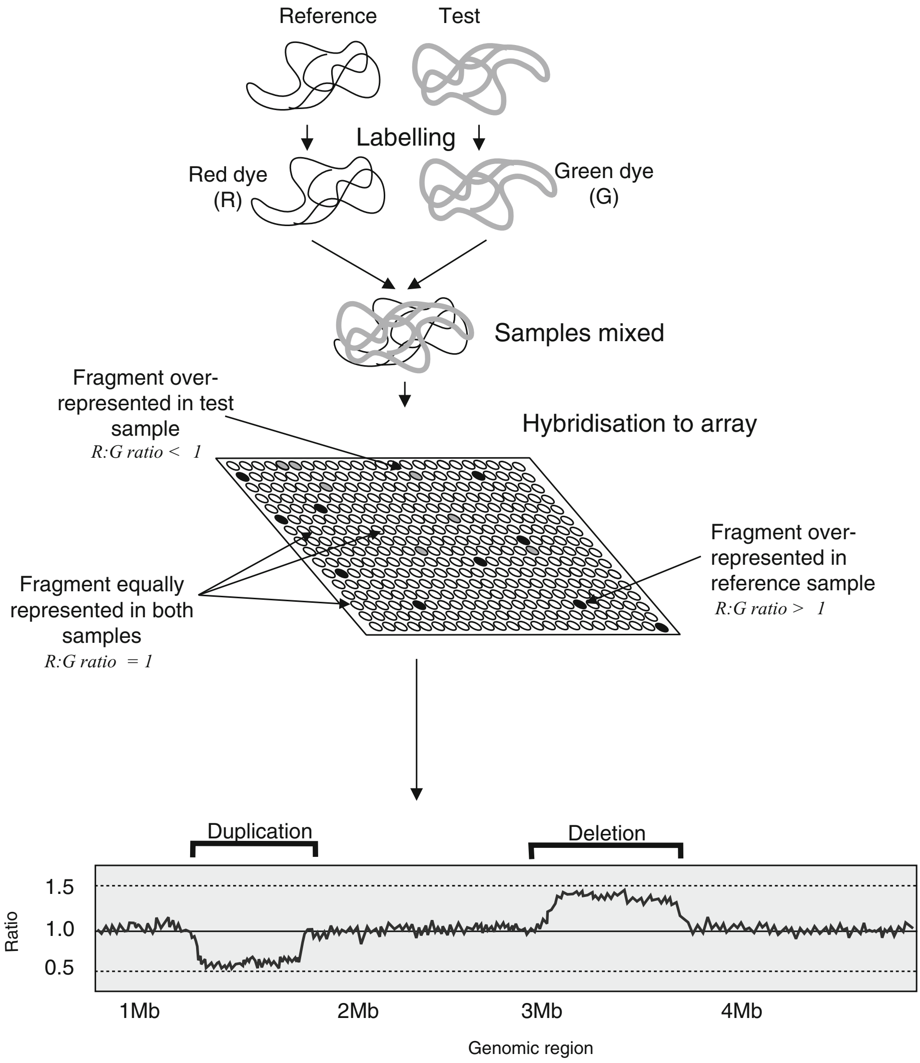


Fig. 2. Array comparative genomic hybridisation (array-CGH). Test and reference DNA samples are differentially labelled, mixed, and hybridised to a target microarray containing probes (e.g. oligonucleotide probes or BAC clone probes). Where a genomic region is equally represented in both the test and reference samples (i.e. where the copy number of the region is the same in both samples), the ratio of fluorescence from the dyes at this spot on the array will be 1. Where one sample is over-represented with respect to the other, a ratio of greater than or less than 1 will be observed. These ratios are then mapped back against the genome to identify regions of variable copy number in the test sample. Reprinted with permission from Elsevier: The Lancet 2009 (74), originally adapted from Feuk et al. (75) with permission from Macmillan Publishers Ltd: Nature Reviews Genetics.

underlying structural variation (Fig. 3b) (1, 6). In addition, derivations of the raw allelic intensities from SNP genotyping arrays have been used to screen for CNV. A number of algorithms have been developed to call CNVs from this information, many of which are based around Hidden Markov Models (45–54). These models assume that the data are generated by a stochastic process

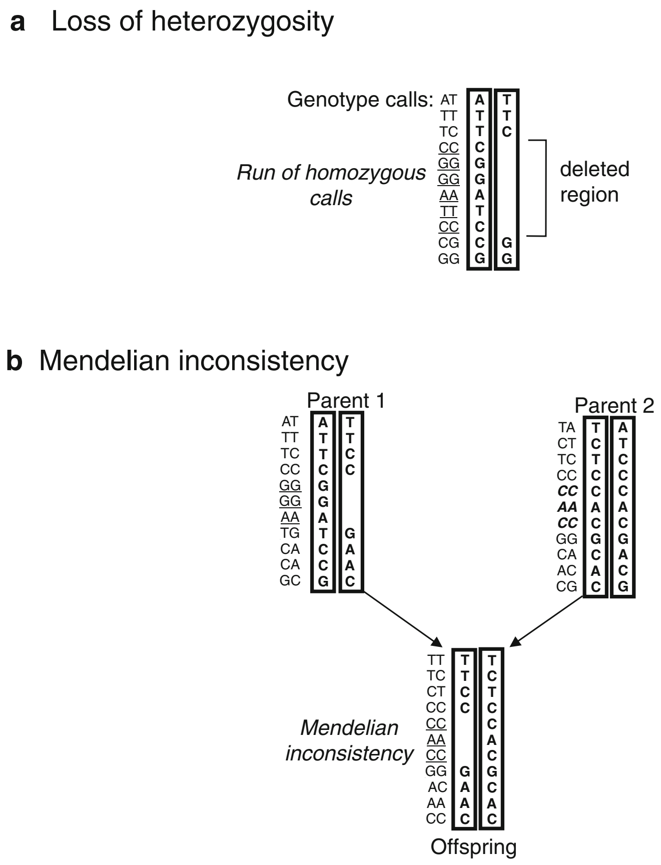


Fig. 3. Effect of CNV on genotyping. (**a**) Loss of heterozygosity. SNPs in regions of heterozygous deletion are called as homozygotes. Unexpectedly, long regions of homozygosity may be indicative of a deletion. (**b**) Mendelian inconsistency. The three SNPs in the heterozygously deleted region of Parent 1 are called as homozygotes. The offspring has inherited the deletion from Parent 1 and a normal (in terms of structural variation) copy from Parent 2 and so is called as homozygous for the alleles from Parent 2 for these three SNPs.

defined by a predetermined number of states representing copy number/genotype combinations (46, 55). More recently, hybrid genotyping arrays have been employed. These contain very large numbers of non-polymorphic probes as well as SNP probes (43, 53, 56, 57). Array-based approaches to CNV discovery and association are considered further in Subheading 4.

Alternatives to the use of array platforms include sequencing-based methods such as end-pair sequencing (11, 58). In this approach, the fragments of test genomic DNA are circularised and then randomly cleaved, and the ends of the resulting fragments are sequenced and computationally mapped to a reference genome sequence. The distance between the two points where the sequenced ends map to the reference genome is used to detect whether there is structural variation between the test genome and the reference genome. In simple terms, if the ends map closer together on the reference sequence than would be expected, then this may be indicative of an insertion in the test sequence relative to the reference sequence. Similarly, if the ends map further apart than expected, this may indicate a deletion in the test sequence. Unlike the array-based methods, this strategy is able to detect inversions which would result in a reversed orientation of the sequenced ends when mapped against the reference sequence (11, 59). The decreasing costs and wide availability of high-throughput sequencing methods should increase the utility of these approaches in genome-wide CNV detection (60).

3.2. CNV Discovery Findings and Their Interpretation

Any reported CNV needs to be interpreted in the light of the strengths and limitations of the approach used to detect the CNV. A principal factor affecting interpretation is the resolution of the detection method, which depends upon the size and spacing of the probes. BAC array-CGH can detect regions of copy number as small as 20 kb, but can only report the whole clone as being copy number variable and so the boundaries of small CNVs tend to be overestimated (recall that BAC clones are large, comprising around 80–200 kb of genomic DNA). SNPs that lie within copy number variable regions have had a greater propensity to be excluded from SNP genotyping arrays, as they are more likely to deviate from Hardy–Weinberg equilibrium and to show Mendelian inconsistencies (see Fig. 3b). This impacts on the SNP coverage of the relevant genomic regions and on inferences about CNV boundaries. The inclusion of CNV probes on hybrid genotyping arrays can circumvent this problem by providing coverage of regions often missed by SNP probes (56).

Further interpretational caveats arise from the use of reference samples. Array-CGH CNV calls are made relative to a reference sample, comprising DNA from a single individual. However, the reference sample is unlikely to have a copy number of two at all loci. Therefore, an apparent duplication may represent a true duplication in the test sample or a deletion at the same locus in the reference sample. Pooled DNA from two or more individuals may be used as a reference sample to minimise problems arising from a rare CNV in a reference sample, but interpretational difficulties may persist in regions of common CNVs and the increased variance of the reference signal may weaken the signal from CNVs in the test sample. Intensity for each SNP on an Illumina SNP genotyping array is derived from the average intensity of the genotype clusters rather than from a reference sample (61).

The Database of Genomic Variants (DGV) (4) is currently the largest repository of CNV data from peer-reviewed screening studies using generally healthy control populations. Within the database, structural variants are sub-classified into indels, CNVs, and inversions. Variants termed CNVs in the database are all greater than 1 kb in length, and the term indel is used to define regions of CNV of between 100 bp and 1 kb. Any CNVs which overlap by at least 70% are merged into a single "CNV locus." These definitions are guided by pragmatism and by the historical resolution of the array-based methods rather than any clear-cut biological significance.

The Redon et al. study of 2006 (8) was a key early paper in the field of CNV discovery. A genome-wide survey of structural variation was undertaken using a whole genome tile path (WGTP) array (CGH), comprising 26,574 large insert clones covering more than 93% of the genome, and an Affymetrix GeneChip Human Mapping 500K early access SNP array. A total of 270 individuals from the HapMap collection (62) were analysed and 12% of the genome was found to be within regions of CNV. A subsequent study which developed and utilised the Affymetrix 6.0 array (containing around 940,000 CNV probes in addition to probes for over 900,000 SNPs) provided a new landmark (56) and challenged conclusions drawn by earlier studies (8) which employed lower resolution approaches. By combining information from both CNV and SNP probes, McCarroll developed a map of CNVs at ~2 kb breakpoint resolution. This indicated that many of the CNVs previously described were between 5 and 15 times smaller than initially reported. Importantly, this not only cast doubt on previous estimates of the proportion of the genome subject to large-scale CNVs (probably <5%), but it also means that inferences drawn about genes and other potentially functional sequence included in known CNVs will need to be revised (56).

Once appropriately characterised, most autosomal CNVs with frequencies >1% (common CNVs) appear to behave like SNPs in that most appear to be diallelic, in Hardy–Weinberg equilibrium, and are stably inherited in a Mendelian fashion (56). This contrasts to evidence from large, rare CNVs which frequently occur de novo and which are known or assumed to be deleterious (40, 63). Emerging evidence also suggests that most common CNVs are well-tagged by SNPs, and although available estimates of the tagging vary, probably at least one-half of common CNVs will be tagged ($r^2 > 0.8$) by the newest genome-wide SNP arrays (56, 64).

4. Genome-Wide CNV Association Studies

Genetic association studies employing SNP genotyping can rely on a very well-developed catalogue of SNP variants, their frequencies, and inter-relationships in different populations (62, 65).

A researcher designing a SNP association study will have a good indication of the likely success of the assay, the allele frequency of an assayed SNP, and the extent to which it should capture (tag) variation at nearby SNPs. Until recently, researchers undertaking CNV association studies had little or no information about the presence and nature of CNVs in a given genomic region. This is beginning to change with the advent of more precise mapping of CNVs and their breakpoints using hybrid arrays and sequencing-based approaches (56, 58). Improving CNV maps is a crucial step forward to more effective and efficient CNV association studies. For example, in planning a genome-wide CNV association study, one can utilise available information about previously discovered CNVs, such as data relating to its breakpoints, in order to target probes for reliably quantifying the copy number of a given CNV rather than refining its breakpoints (56). Thus, resources can be employed to optimise the detection (e.g. replicate probes might improve the signal-to-noise ratio of assays) or to type more CNVs by avoiding the use of wasteful probes which lie outside the known boundaries of the CNV. That said, it will only be possible to rely on CNV maps for studies focused on common CNVs. Different techniques are likely to be required for studies of unique or rare variants which may not occur in the reference populations used to develop CNV maps.

4.1. Rare CNVs

Studies aiming to relate the association of rare CNVs to disease rely first on CNV discovery and, second, on relating the presence or absence of a rare CNV to disease status. Several studies of this kind have been undertaken using data from genome-wide SNP arrays to detect rare CNVs hypothesised to underlie a range of disorders including schizophrenia and autism (Table 1). Deletions may be discovered from SNP genotype data by the detection of runs of contiguous SNPs that show loss of heterozygosity (due to hemizygous genotypes being called as homozygous, Fig. 3a). A more sensitive approach, however, is to utilise the fluorescence intensity signals from the SNP probes. Duplications or deletions cause changes in the normalised fluorescence intensity for neighbouring SNPs. Thus, the approach has been called "SNP-CGH" (61), highlighting the similarities between this and array-CGH.

Association studies of this type have utilised the genome-wide SNP arrays of Affymetrix (40, 53, 61, 63) and Illumina (45, 51, 53), and use a measure of the combined signal (R) from both alleles at a particular SNP. This is often expressed as the $\log_2 R$ ratio; the logarithm to the base 2 of a ratio of the observed intensity to an expected intensity (of the genotype clusters rather than a reference sample). Some studies also use information from the relative ratio of the fluorescence signals from one allelic probe to another, since these ratios would be expected to be 0, 0·5, and 1·0 in the absence of CNV. A range of different normalisation

Table 1
Examples of recent genome-wide CNV association studies including the CNV detection methods used, study sizes, interesting findings, and any validation and replication strategies pursued

Disease	CNV detection platform	Study size	Key association finding(s)	Validation/replication	References
Autism	19K WGTP BAC array-CGH	712 Cases 837 Controls	De novo 16p11.2 microdeletion that is significantly associated with autism (found in 4 cases vs. 0 controls)	FISH, microsatellite analysis, and RT-qPCR. Breakpoints refined using NimbleGen Chr16 oligonucleotide array-CGH	(41)
	ROMA	165 Case families 99 Control families	De novo CNVs significantly associated with autism	ROMA 390K, Agilent 244K array-CGH, and FISH	(39)
	Affymetrix 10K SNP arrays Linkage analysis	1,181 Families containing at least 2 cases	Association of chromosome region 11p12–p13 and neurexins with autism risk	CNV assessment using alternative calling algorithms and comparison with DGV entries	(68)
	Affymetrix 5.0[a] (500,000 SNP probes + 500,000 non-SNP probes including 100,000 targeted to known CNV regions)	1,441 Cases (some related) 4,234 Controls (2,814 with bipolar disorder)	Inherited and de novo 16p11.2 CNV associated with susceptibility to autism	Replication of specific events in 2 further populations: 1. 512 Cases and 434 controls (Agilent 244K array-CGH) 2. 299 Cases and 18,834 controls (Illumina HumanHap300 BeadChip)	(53)
Schizophrenia	ROMA 85K array-CGH for detection	159 Cases	Partially disrupted genes more common in cases suggesting role for mutated gene products	Illumina 550K SNP array and NimbleGen 2.1M feature HD2 CNV array	(51)
		268 Controls	De novo CNVs associated with disease		

	Discovery: Illumina HumanHap300 or HumanCNV370	Discovery: 2,160 trios + 5,558 parent–offspring pairs, all controls	Deletions at 1q21.1, 15q11.2, and 15q13.3 found to be significantly associated with schizophrenia	Replication using 3,285 cases and 7,951 controls	(42)
	Association: Illumina HumanHap300	Association: 1,433 cases 33,250 controls		Illumina HumanHap300, HumanHap550, Affymetrix 6.0, or Taqman RT-qPCR	
	Affymetrix 5.0 and 6.0	3,391 Cases 3,181 Controls	De novo CNVs associated with sporadic cases Expected association with deletions in critical regions identified Association of large deletions at 1q21.1 and 15q13.3	Significantly associated deletions validated using RT-qPCR	(43)
	Affymetrix 5.0[a]	152 Sporadic cases (plus unaffected parents) 48 Familial cases (plus biological parents) 159 Controls (plus biological parents)	De novo CNVs associated with sporadic cases	RT-qPCR	(40)
Ischaemic stroke	Illumina Infinium Human-1 HumanHap300 SNP arrays	263 Cases 275 Controls	No significant association identified	Used DGV as extra control set	(72)

(continued)

Table 1 (continued)

Disease	CNV detection platform	Study size	Key association finding(s)	Validation/replication	References
Subarachnoid aneurismal haemorrhage	Visual inspection of genotype clusters from Illumina HumanHap300 SNP array	203 Cases 294 Controls	One significant association but did not survive Bonferroni correction	Analysis of Hardy–Weinberg equilibrium, Mendelian Inheritance inconsistencies, and comparison with known CNV regions from HapMap database Top hit validated using RT-qPCR	(38)
Coeliac disease	Imputation of triallelic genotypes at SNPs previously identified from raw intensity data to have an extra untyped allele	768 Cases 1,417 Controls	Association with 8 SNPs including 3 within MHC complex (known to be associated). Associations did not survive multiple testing	Resequencing	(73)

The reader is advised to refer directly to the literature for further details of these studies and to evaluate the methods used in light of the design issues introduced here. *WGTP* whole genome tiling path

[a] Approximately 500,000 SNP probes + ~500,000 non-SNP probes including 100,000 targeted to known CNV regions

strategies have been employed to attempt to improve the signal-to-noise ratio, and a wide variety of different CNV calling algorithms have been utilised (45–54, 66, 67). A recently developed algorithm can make use of the combined information from both the SNP probes and the non-polymorphic probes that are available on the most recent genome-wide platforms such as the Affymetrix 5·0 and 6·0 arrays (67).

Since the detection of very rare CNVs often cannot rely on a priori knowledge of the relevant CNV, studies investigating rare CNVs that predispose to disease will rely on strict filtering strategies to limit false-positive findings. Inevitably such strategies impact on the sensitivity to detect CNVs, particularly when using SNP arrays with limited coverage of the relevant genomic regions. These filtering strategies usually include either a minimal CNV size or a minimum number of contiguous SNPs, so that the reported findings of such studies tend to relate to large CNVs (39–43, 51, 53, 63, 68). Even so, validation and replication studies are necessary to exclude spurious findings. Studies of rare CNVs are akin to studies of genomic disorders, particularly those which employ family data in order to focus efforts on the detection of rare de novo variants (40, 42, 43, 51).

4.2. Common CNVs

We appear to be on the threshold of undertaking genome-wide association (GWA) studies of common CNVs. These studies must exploit the new high-resolution CNV maps to develop assays capable of detection of common, generally smaller CNVs with much greater sensitivity and specificity than employed in association studies to date. Analytical approaches may be adopted to deal appropriately with data resulting from sub-optimal CNV assays. For example, where CNV boundaries are uncertain, use of the first principal component from the intensity measures of the different probes assumed to lie within a given CNV (rather than a mean intensity measure of the probes) can downweight the intensities of probes that actually lie outside the CNV boundaries (69).

However, it is still likely that for some CNVs (especially multiallelic CNVs) the data will be noisy and, particularly in case–control studies, differential bias can impact on association tests and lead to false-positive findings. As with differential bias for SNP genotype calling (70), using a strict threshold for CNV calling can actually worsen the differential bias due to differing rates of non-random missingness between cases and controls (69). One potential solution to this is to undertake likelihood ratio testing of quantitative CNV measurements in cases and controls rather than to separate CNV calling and association testing (69).

Notwithstanding the additional challenges of CNV typing and association testing, the epidemiological considerations that apply to genome-wide SNP association studies are highly relevant for CNV association studies. Common CNVs are expected to have an individually modest effect on disease risk (as with SNPs) and therefore similarly large sample sizes are likely to be required, i.e. at least several thousand cases and controls. Accumulating such sample sizes is unlikely to be a major problem for the consortia that have already established collaborative pooling of data across many studies. For GWA studies, stringent significance levels are likely to be required, although the true extent of multiple testing may not be clear without further refinement of CNV maps. As with existing (SNP-based) GWA studies, independent replication of findings will remain pivotal to assessing the robustness of findings of these studies.

5. Concluding Comments

Human genomic CNV is widespread and ongoing studies are attempting to refine the estimates of the proportion of the genome that is copy number variable. Crucially, CNV may contribute to variability in disease risk between individuals. In fact, association signals in some SNPs could be explained by nearby CNVs in strong LD with the tested SNP (56, 71). However, our understanding of CNV is much less well-developed than our understanding of SNPs, and fundamental questions remain unanswered in relation to the likely success and optimal approaches for association studies of common CNVs. Many CNVs remain inaccurately mapped and characterised. Even with rapid improvements in assays, it seems likely that copy number will be very difficult to measure for the more complex CNVs, such as multiallelic CNVs and especially overlapping or nested CNVs (8). How well common CNVs can be tagged by SNPs (thereby avoiding more expensive and possibly inferior assays) requires clarification (56, 64). The extent to which the risk of common diseases is influenced by structural rather than sequence variation is not known, but providing answers to this question seems set to be a major focus of genetic epidemiology research over the coming decade.

References

1. Conrad DF, Andrews TD, Carter NP, Hurles ME, and Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38: 75–81
2. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, et al. (2004) Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36: 861–6
3. Hinds DA, Kloek AP, Jen M, Chen X, and Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38: 82–5
4. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949–51

5. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, et al. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79: 275–90
6. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, et al. (2006) Common deletion polymorphisms in the human genome. Nat Genet 38: 86–92
7. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, et al. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16: 1182–90
8. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–54
9. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525–8
10. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77: 78–88
11. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nat Genet 37: 727–32
12. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, et al. (2007) A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80: 91–104
13. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, et al. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307: 1434–40
14. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256–60
15. Stankiewicz P and Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. Trends Genet 18: 74–82
16. Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14: 417–22
17. Hollox EJ, Armour JA, and Barber JC (2003) Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. Am J Hum Genet 73: 591–600
18. Ingelman-Sundberg M, Sim SC, Gomez A, and Rodriguez-Antona C (2007) Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. Pharmacol Ther 116: 496–526
19. Aldred PM, Hollox EJ, and Armour JA (2005) Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. Hum Mol Genet 14: 2045–52
20. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, et al. (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439: 851–5
21. Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, et al. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40: 23–5
22. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, et al. (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39: 721–3
23. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, et al. (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 80: 1037–54
24. Ahuja SK, Kulkarni H, Catano G, Agan BK, Camargo JF, et al. (2008) CCL3L1-CCR5 genotype influences durability of immune recovery during antiretroviral therapy of HIV-1-infected individuals. Nat Med 14: 413–20
25. Kuhn L, Schramm DB, Donninger S, Meddows-Taylor S, Coovadia AH, et al. (2007) African infants' CCL3 gene copies influence perinatal HIV transmission in the absence of maternal nevirapine. AIDS 21: 1753–61
26. Townson JR, Barcellos LF, and Nibbs RJ (2002) Gene copy number regulates the production of the human chemokine CCL3-L1. Eur J Immunol 32: 3016–26
27. Willcocks LC, Lyons PA, Clatworthy MR, Robinson JI, Yang W, et al. (2008) Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. J Exp Med 205: 1573–82
28. Deeb SS (2005) The molecular basis of variation in human color vision. Clin Genet 67: 369–77

29. Zipfel PF, Edey M, Heinen S, Jozsi M, Richter H, et al. (2007) Deletion of complement factor H-related genes CFHR1 and CFHR3 is associated with atypical hemolytic uremic syndrome. PLoS Genet 3: e41

30. Qin K, Du X, and Rich BH (2007) An Alu-mediated rearrangement causing a 3.2kb deletion and a novel two base pair deletion in AAAS gene as the cause of triple A syndrome. Mol Genet Metab 92: 359–63

31. Beysen D, Raes J, Leroy BP, Lucassen A, Yates JR, et al. (2005) Deletions involving long-range conserved nongenic sequences upstream and downstream of FOXL2 as a novel disease-causing mechanism in blepharophimosis syndrome. Am J Hum Genet 77: 205–18

32. Lee JA, Madrid RE, Sperle K, Ritterson CM, Hobson GM, et al. (2006) Spastic paraplegia type 2 associated with axonal neuropathy and apparent PLP1 position effect. Ann Neurol 59: 398–403

33. Muncke N, Wogatzky BS, Breuning M, Sistermans EA, Endris V, et al. (2004) Position effect on PLP1 may cause a subset of Pelizaeus-Merzbacher disease symptoms. J Med Genet 41: e121

34. Fielder AH, Walport MJ, Batchelor JR, Rynes RI, Black CM, et al. (1983) Family study of the major histocompatibility complex in patients with systemic lupus erythematosus: importance of null alleles of C4A and C4B in determining disease susceptibility. Br Med J (Clin Res Ed) 286: 425–8

35. Howard PF, Hochberg MC, Bias WB, Arnett FC, Jr., and McLean RH (1986) Relationship between C4 null genes, HLA-D region antigens, and genetic susceptibility to systemic lupus erythematosus in Caucasian and black Americans. Am J Med 81: 187–93

36. Dunckley H, Gatenby PA, Hawkins B, Naito S, and Serjeantson SW (1987) Deficiency of C4A is a genetic determinant of systemic lupus erythematosus in three ethnic groups. J Immunogenet 14: 209–18

37. Avent ND, Martin PG, Armstrong-Fisher SS, Liu W, Finning KM, et al. (1997) Evidence of genetic diversity underlying Rh D-, weak D (Du), and partial D phenotypes as determined by multiplex polymerase chain reaction analysis of the RHD gene. Blood 89: 2568–77

38. Bae JS, Cheong HS, Kim JO, Lee SO, Kim EM, et al. (2008) Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population. Biochem Biophys Res Commun 373: 593–6

39. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445–9

40. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, et al. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 30: 30

41. Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, et al. (2008) Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet 17: 628–38

42. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, et al. (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455: 232–6

43. Stone JL, O'Donovan MC, Gurling H, Kirov GK, Blackwood DH, et al. (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455: 237–41

44. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, et al. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res 13: 2291–305

45. Blauw HM, Veldink JH, van Es MA, van Vught PW, Saris CG, et al. (2008) Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. Lancet Neurol 28: 28

46. Colella S, Yau C, Taylor JM, Mirza G, Butler H, et al. (2007) QuantiSNP: an objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013–25

47. Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, et al. (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res 16: 1575–84

48. Laframboise T, Harrington D, and Weir BA (2007) PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. Biostatistics 8: 323–36

49. Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, et al. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 65: 6071–9

50. Partek Genomics Suite. Version 6.3 Copyright © 2008 Partek Inc., St. Louis, MO, USA

51. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539–43

52. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: An integrated hidden

Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 5: 5

53. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, et al. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358: 667–75
54. Zhao X, Li C, Paez JG, Chin K, Janne PA, et al. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 64: 3060–71
55. Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, and Noble WS (2007) Unsupervised segmentation of continuous genomic data. Bioinformatics 23: 1424–6
56. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 7: 7
57. Esteller M (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 8: 286–98
58. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56–64
59. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420–6
60. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–41
61. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136–48
62. The International HapMap Consortium (2003) The international HapMap project. Nature 426: 789–96
63. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, et al. (2008) Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82: 477–88
64. Cooper GM, Zerr T, Kidd JM, Eichler EE, and Nickerson DA (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 7: 7
65. Sherry ST, Ward M, and Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677–9
66. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, et al. (2008) A robust statistical method for case-control association testing with copy number variation. Nat Genet 7: 7
67. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 7: 7
68. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, et al. (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Genet 39: 319–28
69. Barnes C, Plagnol V, Marchini J, Clayton DG, and Hurles ME (2008) A robust and statistical method for case-control association testing with copy number variation. Nat Genet 40: 1245–52
70. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, et al. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37: 1243–6
71. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 24: 24
72. Matarin M, Simon-Sanchez J, Fung HC, Scholz S, Gibbs JR, et al. (2008) Structural genomic variation in ischemic stroke. Neurogenetics 21: 21
73. Franke L, de Kovel CG, Aulchenko YS, Trynka G, Zhernakova A, et al. (2008) Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. Am J Hum Genet 82: 1316–33
74. Wain LV, Armour JA, and Tobin MD (2009) Genomic copy number variation, human health, and disease. Lancet 374: 340–50
75. Feuk L, Carson AR, and Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85–97

Chapter 14

Epigenetic Variation

Kevin Walters

Abstract

Epigenetics is a fast moving field and our understanding of epigenetic mechanisms has dramatically improved in recent decades. We present the role that epigenetics plays in genomic control in humans; the molecular basis of this control and the role that epigenetic aberrations play in the aetiology of human disease. We outline some of the laboratory techniques for characterising epigenetic variation from methylation analysis of a single CpG to characterising histone variation across extensive genomic regions. The fields of computational epigenetics and population epigenetics have recently emerged and we discuss developments in statistical methods that use DNA methylation as biomarker for the prediction of disease. Finally we describe how DNA methylation errors that occur during somatic cell divisions have been used as a molecular clock that allows inferences about cell population histories to be made.

Key words: DNA methylation, Histone modifications, Chromatin remodelling, Transcriptional regulation, Cellular inheritance, Molecular clock

1. Epigenetic Regulation of the Human Genome

DNA is highly condensed into chromatin in the nucleus. The level of DNA packaging defines two contrasting forms of chromatin, tightly packed heterochromatin and (relatively) lightly packed euchromatin. Chromatin structure is strongly associated with the transcriptional activity of genes within that region: heterochromatic genes undergo limited transcriptional activity in contrast to actively transcribed euchromatic genes. The basic building blocks of chromatin are nucleosomes. Each nucleosome consists of 146 bp of DNA wrapped around an octamer of histone proteins (two copies of each of four histone proteins). This extreme condensation of DNA makes it highly inaccessible to regulatory factors. This inaccessibility is resolved by a dynamic process allowing exposure of DNA to regulatory factors. The two main mechanisms involved in this process are DNA methylation

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_14, © Springer Science+Business Media, LLC 2011

and histone modification and these two processes are intimately linked. The modulation of chromatin structure by DNA methylation, histone modifications, and other molecular mechanisms is an example of epigenetic control. The current use of the term "epigenetics" generally relates to the inheritance of cellular states that are not a result of DNA sequence variations. Epigenetic control plays a major role in several biological processes including tissue-specific control of gene expression, control of transposable elements, genomic imprinting, and X-chromosome inactivation. Epigenetic control of gene expression during development ensures transcriptional silencing of those developmental genes whose expression is no longer required. Post-development, the stable inheritance of chromatin states by daughter cells during somatic cell division is facilitated by somatic inheritance of gene-specific patterns of DNA methylation and histone modifications. The stable inheritance of chromatin states ensures that gene-specific transcriptional activity is perpetuated in somatic cell divisions.

In humans, DNA methylation occurs at CpG dinucleotides. CpGs are not evenly distributed throughout the genome and there exist so-called CpG islands; 7% of all CpGs reside within CpG islands (1). These islands are regions of higher than expected density of CpG dinucleotides and current research suggests that approximately 88% of active gene promoters are associated with CpG islands (2). The association with gene promoters implies a major transcriptional role for CpG islands; CpG islands are usually associated with open chromatin structures that are accessible to the transcription machinery. Genes with CpG islands in their promoter regions are very often found in ubiquitously expressed genes whilst tissue-specific genes often have no island. It would appear that the presence of a promoter-linked CpG island implies somatic stability whilst absence of an island makes it easier to reverse the transcriptional activity of a gene.

DNA methylation is maintained by methyltransferases. Five methyltransferases have been discovered to date: DNMT1, 2, 3a, 3b, and 3L. The role of DNMT2 is not clear and DNMT3L is not thought to be capable of methylating cytosines but interacts with other active methyltransferases; the remaining three are involved in maintenance and de novo methylation with DNMT1 being predominantly a maintenance methyltransferase but possessing the capability of de novo methylation (3). These DNA methyltransferases interact with histone deacetylases and histone methyltransferases in regulating transcriptional activity.

There are two identified mechanisms by which DNA methylation inhibits transcription. The first is direct inhibition in which DNA methylation itself inhibits transcription factor binding (4) and the second is via recruitment of an intermediary protein, namely methyl CpG-binding domain proteins (MBDs);

these MBDs bind to the DNA and inhibit transcription. There are several MBDs and they exhibit binding preferences in both the sequence context and number of CpGs. There are a host of complex interactions between the various MBDs and other proteins such as histone deacetylases, histone methyltransferases, and nucleosome remodelling complexes (5).

There are several mechanisms that are involved in histone modification: histone variants, ATP-dependent chromatin remodelling, and various histone modifications. There are many modifications that can occur to histones and these different alterations allow them to take differing roles in a range of biological processes. These modifications include lysine acetylation and methylation, ubiquitination, serine phosphorylation, ADP ribosylation, and sumoylation. The effect of some of these modifications is not well understood, particularly ribosylation. Histone acetyltransferases are transcriptional co-activators (along with ATP-dependent chromatin remodelling proteins) and are involved in regulating gene transcription as part of much larger transcription regulatory complexes (6). Sumoylation is involved in the recruitment of histone deacetylases (HDACs) and has a role in transcriptional silencing. Some of these histone modifications can have opposing outcomes depending on their location; ubiquitination of lysine 123 on Histone 2B has a gene-activating effect while it is thought that ubiquitination of lysine 119 on histone 2A is associated with a repressive state of transcription (7). Histone phosphorylation is another regulator of transcription that is capable of either enhancing or repressing transcription.

The exact way in which these modifications modulate transcription is unclear, but it is thought that they do so through the intermediate recruitment and binding of so-called effector modules (DNA regulatory proteins). Different effector modules are recruited according to the type of methylation (mono/di/tri) of the histone tail lysines and thus different outcomes can result according to which effector modules are recruited. Histone methylation plays a role in X-chromosome inactivation, DNA repair and transcriptional regulation depending upon which lysine residue is methylated and the extent the methylation. Chromatin conformation is controlled by both histone modification and DNA methylation, but the complex interactions between DNA methylation and histone modifications and the temporal aspects are poorly understood. Recent evidence from studies on *Xenopus laevis* suggests that histone acetylation and DNA methylation act in concert to reactivate the *oct*4 promoter (8). Histone deacetylases (HDACs) are recruited following the binding of methylated CpG-binding proteins (MeCPs). The histone deacetylation then allows methylation of the histones themselves. Methylated histones are targets for chromatin condensing proteins which lead to transcriptional inactivation (5). Clearly, much work needs to be done in different cell lines looking at different genes subject to various states of epigenetic silencing to provide a fuller picture of the temporal relationships and interactions between these two mechanisms.

Another class of epigenetic regulatory proteins are the polycomb and trithorax group proteins; the latter being relatively poorly understood. Historically, the primary function of the polycomb group proteins was thought to be maintaining the silenced state of homeotic genes that are involved in specialisation of the various body segments during early development. Recent evidence suggests that they have a much wider regulatory function in cell differentiation, are involved in the global regulation of many genes in different stages of development, and are involved in all the major mammalian developmental pathways (9). Polycomb group proteins play a key role in maintaining the inactive state of facultative heterochromatin (X-chromosome inactivation and imprinting, for example). Three multiprotein polycomb group complexes have been identified that co-operate to maintain the repressed state of genes that they regulate. Their specific functions include binding to certain methylated lysines in histone 3 and involvement in the ubiquitination of histone 2a.

Much of our understanding of how polycomb group proteins are recruited is through our work in Drosophila. It has been shown that genes regulated by polycomb group proteins contain specific sequence elements (polycomb response elements) that recruit polycomb group protein complexes (10). These response elements can act over large distance and can be tens of kilobases from the promoters that they regulate. They are capable of repressing multiple genes in their vicinity and their repressive capability appears to exhibit a dosage effect (11). The stable inheritance of the repressed state of a gene under polycomb group regulation following DNA replication is not known. It has been suggested that histone H3 K27 methylation could be one of the epigenetic marks that mediates the transmission of this epigenetic mechanism (9). The derepressive role of trithorax and AHS1 proteins are poorly characterised; what is known is that they have H3 methyltransferase activity and that they are associated with all polycomb response elements. They may play some part in reversing the epigenetic states of target genes (12).

2. Somatic and Meiotic Inheritance of Epigenetic States

There are several putative mechanisms for the epigenetic inheritance of histone modifications. The first proposes that histone octamers segregate randomly to the daughter strand creating a patchwork of octamers on each stand. The gaps are plugged with naïve octamers and the chromatin modifying machinery is subsequently recruited to modify the naïve octamers. The second scenario sees each octamer split into two heterodimers which are then made into octamers by the addition of four new histones following DNA replication. The final putative mechanism of somatic transmission of the histone marks is one in which the DNA methylation and histone methylation interact in some way; there is evidence of interaction between histone methyltransferases and several DNA methyltransferases (13).

Compared to histone modifications, the mechanism by which methylation patterns are somatically inherited is relatively well understood. Following somatic cell division a new unmethylated DNA strand is synthesised. This process yields hemimethylated CpG sites and it is known that DNMT1 has a preference for hemimethylated CpGs (14). Certain N-terminal amino acids of DNMT1 possess a proliferating cell nuclear antigen (PCNA)-binding domain (15). PCNA is involved in DNA replication and repair so the associated binding of PCNA and DNMT1 helps to target DNMT1 to replication foci. The PCNA-associated DNMT1 then methylates the unmethylated strand in a processive manner (16). Assuming an error-free process, this produces two daughter cells that have the same methylation pattern as the parent cell. However, there is evidence of two types of error that can occur during the remethylation process at DNA replication. These errors are failures to methylate a CpG that should be methylated (failure of maintenance methylation) and methylating a CpG that should not be methylated (de novo methylation). Neither DNMT3A nor DNMT3B localise to the replication loci (17) so that DNMT1, which is generally considered to be a maintenance methyltransferase, must also be capable of de novo methylation activity; this has been demonstrated in vitro (18).

A longstanding question relating to the stable inheritance of DNA methylation patterns is whether demethylation is an active or passive process. In a passive process, the nascent strand does not get methylated by maintenance methyltransferases at the next replication event, possibly as a result of protein–protein interactions that inhibit DNMT1, leading eventually to predominantly unmethylated cells. In this scenario, the methylated CpG is not returned to an unmethylated state but it becomes highly diluted in a cell population. In active demethylation, some process facilitates the transition of a cytosine from a methylated to an unmethylated state. Evidence has recently emerged for a protein that acts as an active demethylating agent. *Gadd45a* is thought to be recruited to sites of demethylation and promotes DNA repair; the methylated cytosines are excised and replaced by unmethylated nucleotides (1).

Transgenerational epigenetic inheritance relates to the passage of epigenetic information through the germ line. There are now several studies showing that in certain genes epimutations are heritable and seen in multiple generations. Mono-allelic expression resulting from imprinting can be viewed as a form of transgenerational epigenetic inheritance where the silencing is sex specific but the specific mechanisms by which epigenetic information could be inherited at certain loci is not understood. The previously accepted view that epigenetic states were not meiotically inherited is supported by the fact that there is genomewide demethylation following fertilisation. How can epigenetic marks be transmitted across generations if the marks are erased and re-established in early development? Whilst the answer to this question is not known there are several studies demonstrating transgenerational epigenetic inheritance of DNA methylation levels (19, 20). The study of Tufarelli (20) also involved the inheritance of a deletion so that unravelling the genetic/epigenetic influences is complex. In human studies, it is always difficult to rule out genetic control of apparent meiotic inheritance of the epigenetic state, and this remains a major obstacle to overcome. Most of the studies to date have involved a single family and usually a single generation. A study by Bjornsson et al. (21) looked at the methylation changes in 21 three-generation families whose DNA methylation was measured at time points an average of 16 years apart. The analysis identified strong familial clustering of changes in DNA methylation levels over time averaged over 807 genes (heritability was 0.743). A recent study using mono- and dizygotic twins found significantly higher methylation differences in dizygotic compared to monozygotic co-twins. Experiments conducted by the authors showed that this was unlikely to be due to DNA sequence differences, strengthening the evidence for epigenetic inheritance (22). Almost all studies have focussed on DNA methylation, but what happens to the epigenetic state of histones in these early stages is not known. Future studies investigating the relationship between DNA methylation and histone modifications in early development may provide some intriguing results.

3. The Role of Epigenetic Errors in Disease

Our knowledge of the role of epigenetic changes in disease is developing rapidly as a result of an increasing number of studies looking for epigenetic variation in a multitude of diseases. The marker of choice is invariably DNA methylation because it is a reliable and easily analysed biomarker but also because it is observed in many epigenetically silenced genes. Environmental exposure is likely to influence the epigenetic landscape and characterising this as well as the between individual variation (in DNA methylation patterns, for example) as a result of random errors in copying epigenetic marks is a major challenge. In studies looking at candidate genes, much of the focus has, understandably, been on genes detected through genetic studies (confirmed or putative oncogenes or tumour suppressor genes in cancer, for example). Many studies focus on a small number of CpGs out of the hundreds contained within CpG islands although this approach is questionable as previous studies have reported both sequence (23) and density-dependent methylation (24).

Our understanding of the contribution of epigenetic variation to disease susceptibility is most advanced in the field of cancer where evidence of epigenetic alterations in just about every type of cancer have been reported (25). Hypermethylation of tumour suppressor genes (TSGs) is a common finding in cancer (26–28) and some TSGs are inactivated in a range of cancers (29). As well as cancer development, promoter hypermethylation is also associated with progression and metastasis (30, 31). The factors influencing aberrant CpG island methylation are not fully understood, but there is some evidence to suggest that it is related to whether the flanking sequence matches the preferred flanking sequences of the de novo methyltransferases (32). Aberrant methylation of CpG islands has also been shown to be influenced by distance from repetitive elements and the local chromatin pattern (33, 34). Recognition that some cancers had a high degree of DNA methylation lead to the controversial proposal of the CpG island methylator phenotype (CIMP). CIMP cancers were considered to be distinct in many molecular and histological ways and were thought to encompass those cancers in which there was a decreased fidelity of methylation maintenance (35). Decreased fidelity of cancer cells to replicate their CpG methylation signatures has certainly been observed in some cancer cell lines (36).

In most cases, epigenetic changes occurring in tumour suppressor genes are accompanied by DNA mutations and unravelling causality is difficult. Clearly, the interplay between genetic and epigenetic factors in disease development will be an area of intense research as the evidence for disease-associated epigenetic modifications accumulates. Screening regions of known loss of heterozygosity for epigenetic changes revealed the TCF21 gene as a possible tumour suppressor gene. CDH1, a known tumour suppressor gene, appears to be silenced by a combination of genetic and epigenetic changes: a genetic mutation at one allele and DNA methylation of the other (37). Epigenetic changes can also be the precursor of genetic changes; for example, promoter hypermethylation of MLH1 leads to microsatellite instability in colon cancer (38). In leukaemia, genetic alterations have been shown to enhance the recruitment of DNA methyltransferases leading to aberrant promoter methylation (39). Much of the research effort has focussed on epigenetic variation in coding genes, but there is emerging evidence of the role of microRNAs in epigenetic control. It has been suggested that some cancers may result from aberrant expression of miRNAs indicating the importance of considering non-coding genes when looking for changes in DNA methylation associated with disease (40).

4. Laboratory-Based Methods to Characterise DNA Methylation Variation

Laboratory methods to quantify epigenetic variation have focussed on DNA methylation as a biomarker. Many methods and many adaptations of these methods have been used to identify, quantify, and characterise the variation in DNA methylation across the human genome. The choice of method depends upon the aim of the experiment. Is the interest in a single CpG, a single CpG island, or genomewide?. The choice also depends upon the anticipated heterogeneity of the methylation patterns likely to be encountered, whether the sequence context is known and how much quantification is required (number of molecules with a specific pattern of methylation or just whether the region can be considered to be methylated as opposed to unmethylated). The choice is also narrowed by the source of the DNA; whether it is from urine, serum, or plasma and also the desired sensitivity (how small a concentration of methylated sites is required to be detected). Demethylating agents can be used on cultured cells if the interest is purely in detecting gene expression changes in the absence of DNA methylation. Methods to detect histone modifications are usually based on mass spectrometry. Chromatin immunoprecipitation followed by microarray analysis (ChIP-on-chip) can be used to investigate histone variation over large genomic regions.

Looking for genomewide variation in DNA methylation patterns can involve the use of enzymes to cleave CpGs embedded in specific sequences. Microarray methods (the use of methyl-sensitive restriction enzymes or immunoprecipitation) yield semi-quantitative results (41) where only large differences are easily detectable. As a result of the decades of research into developing DNA technologies and analysis methods, the data available from epigenetic studies is already at the genome level. Like other genomewide methods, genomewide epigenetic studies face many problems, many of which are shared with gene-expression microarray studies: controlling the false discovery rate, removing or allowing for noise, ability to detect complex interactions and amplification bias.

Region-specific methods usually require sodium bisulphite treatment followed by PCR. Treatment with sodium bisulphite deaminates unmethylated cytosines to uracil. Uracils are replaced with thymine following PCR and the methylation status is determined by the C to T ratio. The advantages of using PCR methods are that relatively small amounts of DNA are required and that many of the methods developed for DNA analysis can be used or easily adapted. These region-specific methods indicate the methylation density in a genomic region. Methylation-specific PCR has the advantages of being able to detect very low levels of methylation and of requiring minimal equipment. Drawbacks are that it requires at least two CpG sites with variable methylation and that it is difficult to quantify the level of methylation at the sites amplified. Real-time methylation-specific PCR methods determine the absence or presence of a specific DNA methylation profile and gives information about the number of such cells. A recent technology called Methylight (42) is capable of very high sensitivities, being able to detect a methylated CpG among tens of thousands of unmethylated ones. It is frequently used to detect fully methylated sequences and produces a highly quantitative measure: the percentage of methylated reference (PMR).

At the nucleotide level of resolution, Pyrosequencing, following bisulfite treatment will reveal methylation patterns of an entire sequence of CpGs. This is useful in regions where it is known or suspected that the methylation profiles are likely to be heterogeneous (in tumour samples where there may be a mixture of tumour and non-tumour cells) or where the interest might be in combinations of methylated CpGs. If the interest is in the methylation patterns of complementary strands of the same DNA molecule (looking at methylation error rates or determining the hemimethylated status of a CpG site, for example) then hairpin bisulphite PCR (43) allows this to be ascertained.

5. Statistical Models Incorporating Epigenetic Data

Depending on the technologies used, DNA methylation data can be slightly different to, for example, gene expression data in that there can be an excess of zeros resulting from unmethylated samples. When comparing two groups of such data (tumour vs. normal tissue, for example) a statistical test that treats the zeros separately to the continuous part can be used (44). The test statistic is the sum of two separate statistics: a test statistic for the difference of two independent proportions and a statistic comparing the non-zero values (taking either a parametric or non-parametric form). A two-part permutation test has also been proposed and shown to be more powerful in some situations (45).

Another problem that has received much attention is that of class prediction (classes might be disease onset, severity or progression to a more severe form) using multivariate DNA methylation data profiles of a set of gene promoters. This is a standard statistical classification problem since the number of classes that the data is to be partitioned into is known a priori. If the number of classes are not predefined then clustering, rather than classification, would be used to place each observation into natural groups. The data is partitioned according to some distance metric that depends upon whether the data is categorical or continuous. Hierarchical clustering has been successfully used to differentiate between different histological subtypes of lung cancer using DNA methylation profiles (46). As in microarray analyses, these types of analyses using DNA methylation data are under-determined problems, there may be thousands of genes but only tens of samples. In this situation the first step is usually one of dimensionality reduction (i.e. selecting those promoters that are best able to discriminate between the groups required). There are many methods that could be used including, Fisher's criterion, principal components analysis, *t* tests, or regression methods. Reducing the number of gene promoters in classification problems based on DNA methylation data has been shown to be a key consideration in achieving small classification errors (47).

Machine learning methods are a group of supervised techniques that "learn" from a subset of the data. Support vector machines are a machine learning technique that has been applied to DNA methylation data (47). More black box methods (artificial neural networks and neuro-fuzzy modelling) have also been applied to classification using DNA methylation data and have produced very small classification errors (48). It will be interesting to see whether these methods receive further attention and if so, whether such low classification errors are obtained. More formal statistical clustering methods have also been compared in the context of DNA methylation data including model-based hierarchical clustering and a Bernoulli-lognormal mixture model (49). As well as the Bernoulli-lognormal mixture model, other methods that explicitly allow for the excess of zeros often seen with DNA methylation data produced by the Methylight technology are starting to be developed (50) and promise even better classification error rates than the methods discussed in this review. Developments in bioinformatics and novel statistical methods are needed to prepare for data that current activities, such as the human epigenome project, are generating (51).

DNA methylation is also proving to be a useful molecular marker that can be exploited in population genetics models. Recent works (52, 53) have used DNA methylation copying errors during somatic cell division to begin to unravel colonic stem cell turnover. The advantage of DNA methylation errors over DNA sequence mutations is the approximately 1,000-fold higher error rate. This increased error rate leads to greater diversity in the methylation sequences over the relatively short number of generations considered in these studies. The variation in the methylation sequences allow different models of stem cell replication and differentiation to be compared. DNA methylation is likely to become a popular marker in future research looking at various aspects of somatic cell populations.

References

1. Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH. Large-scale structure of genome methylation patterns. Genome Res. 2006; 16:157–163.
2. Kim TH, Barrera LI, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high resolution map of active promoters in the human genome. Nature 2005; 436:876–880.
3. Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, Li E, Laird PW Jones PA. Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Mol. Cell. Biol. 2002; 22:480–491.
4. Comb M and Goodman HM. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. Nucleic Acids Res. 1990; 18:3975–3982.
5. Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, Kouzarides T. The methyl-CpG- binding protein MeCP2 links DNA methylation to histone methylation. J. Biol. Chem. 2003; 278:4035–4040.
6. Yang XJ and Ogryzko VV, Nishikawa J, Howard BH, Nakatani Y. A p300/CBP-associated factor that competes with the adenoviral oncoprotein E1A. Nature 1996; 382:319–324.
7. Sun ZWand Allis CD. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. Nature 2002; 48:104–108.
8. Barreto G, Schafer A, Marhold J, Stach D, Swaminathan SK, Handa V, Doderlein G, Maltry N, Wu W, Lyko F, Niehrs C. Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. Nature 2007; 445:671–675.
9. Schwarz YB and Pirrotta V. Molecular mechanisms of polycomb silencing. In 'Epigenetics' edited by J. Tost; 2008, Caister Academic Press, Norfolk, England.
10. Chan C-S, Rastelli L, Pirrotta V. A polycomb response element in Ubx gene that determines an epigenetically inherited state of repression. EMBO J. 1994; 13:2553–2564.
11. Kassis JA. Pairing sensitive silencing, polycomb group response elements and transposon homing in Drosophila. Adv. Genet. 2002; 246:421–438.
12. Kahn TG, Schwartz YB, Dellino GI, Pirrotta V. Polycomb complexes and the propagation of the methylation mark at the Drosophila Ubx gene. J. Biol. Chem. 2006; 281:29064–29075.
13. Esteve Po, Chin HG, Smallwood A, Feehery GR, Gangisetty O, Karpf AR, Carey MF, Pradharan S. Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. Genes Dev. 2006; 20:3089–3103.
14. Herman JG, Umar A, Polyak K, Graff JR, Ahuja N, Issa J-P, Markowitz JK, Willson JKV, Hamilton SR, Kinzler KW, Kane MF, Kolodner RD, Vogelstein B, Kunkel TA, Baylin SB. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc. Natl. Acad. Sci. U.S.A. 1998; 95:96870–96875.
15. Iida T, Suetake I, Tajima S, Morioka H, Ohta S, Obuse C, Tsurimoto T. PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. Genes Cells 2002; 7:997–1007.
16. Goyal R, Reinhardt R, Jeltsch A. Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. Nucleic Acids Res. 2006; 34:1183–1188.
17. Margot JB, Cardaso MC, Leonhardt H. Mammalian DNA methyltransferases show different subnuclear distributions. J. Cell. Biochem. 2001; 83:373–379.
18. Jair K-W, Bachman KE, Suzuki H, Ting AH, Rhee I, Yen RW-C, Baylin SB, Schuebel KE. De novo CpG island methylation in human cancer cells. Cancer Res. 2006; 66:682–692.
19. Chan TL, Yuen St, LKong Ck, Chan YW, Chan YS, Ng WF, Tsui WY, Lo MW, Tam WY, Li VS, Leung SY. Heritable germline epimutations of MSH2 in a family with hereditary nonpolyposis colorectal cancer. Nat. Genet. 2006; 38:1178–1183.
20. Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG, Higgs DR. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat. Gent. 2003; 34:157–165.

21. Bjornsson HT, Sigurdsson MI, Fallin MD, Irizarry RA, Aspelund T, Cui H, Yu W, Rongione MA, Ekström TJ, Harris TB, Launer LJ, Eiriksdottir G, Leppert MF, Sapienza C, Gudnason V, Feinberg AP. Intra-individual change over time in DNA methylation with familial clustering. JAMA 2008; 24:2877–2883.
22. Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GHT, Wong AHC, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, McRae AF, Visscher PM, Montgomery GW, Gottesman II, Martin NG, Petronis A. DNA methylation profiles in monozygotic and dizygotic twins. Nat. Genet. 2009; 41:240–245.
23. Chen C, Yang MC, Yang TP. Evidence that silencing of the HPRT promoter by DNA methylation is mediated by critical CpG sites. J. Biol. Chem. 2001; 276:320–328.
24. Cameron EE, Baylin SB, Herman JG. P15iINK4B CpG island methylation in primary acute leukemia is heterogeneous and suggests density as a critical factor for transcriptional silencing. Blood 1999; 94:2445–2451.
25. Jones PA and Baylin SB. The fundamental role of epigenetic events in cancer. Nat. Rev. Genet. 2002; 3:415–428.
26. Yates DR, Rehman I, Abbod MF, Meuth M, Cross SS, Linkens DA, Hamdy FC, Catto JWF. Promoter hypermethylation identifies progression risk in bladder cancer. Clin. Cancer Res. 2007; 13:2046–2053.
27. Jarmalaite S, Jankevicius F, Kurgonaite K, Suziedelis K, Mutanen P, Husgafvel-Pursiainen K. Promoter hypermethylation in tumour suppressor genes shows association with stage, grade and invasiveness of bladder cancer. Oncology 2008; 75:145–151.
28. Marsit CJ, Kim DH, Liu M, Hinds PW, Wiencke JK, Nelson HH, Kelsey KT. Hypermethylation of RASSF1A and BLU tumor suppressor genes in non-small cell lung cancer: implications for tobacco smoking during adolescence. Int J Cancer 2005; 114:219–223.
29. Du Y, Carling T, Fang W, Piao JZ, Sheu JC, Huang S. Hypermethylation in human cancers of the RIZ1 tumour suppressor gene, a member of a histone/protein methyltransferase superfamily. Cancer Res. 2001; 61:8094–8099.
30. Catto JWF, Azzouzi AR, Rehman I, Feeley KM, Cross S, Amira N, Fromont G, Sibnony M, Cussenot O, Meuth M, Hamdy FC. Promoter hypermethylation is associated with tumour location stage and subsequent progression in transitional cell carcinoma. J. Clin. Oncol. 2005; 23:2903–2910.
31. Chen J, Röcken C, Lofton-Day C, Schulz HU, Müller O, Kutzner N, Malfertheiner P, Ebert MPA. Molecular analysis of *APC* promoter methylation and protein expression in colorectal cancer metastasis. Carcinogenesis 2005; 26:37–43.
32. Handa V and Jeltsch A. Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. J. Mol. Biol. 2005; 348:1103–1112.
33. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. DNA motifs associated with aberrant CpG island methylation. Genomics 2006; 87:572–579.
34. Ohm J, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W, Berman DM, Jenuwin T, Pruitt K, Sharkis SJ, Watkins ND, Hermann JG, Baylin SB. A stem cell-like chromatin pattern may predispose tumour suppressor genes to DNA hypermethylation and heritable silencing. Nat. Genet. 2007; 39(2):237–242.
35. Issa JP. CpG island methylator phenotype in cancer. Nat. Rev. Cancer 2004; 4:988–993.
36. Ushijima T, Watanabe N, Shimizu K, Miyamoto K, Sugimura T, Kaneda A. Decreased fidelity in repeating CPG methylation patterns in cancer cells. Cancer Res. 2005; 65:11–17.
37. Grady WM, Willis J, Guildford PJ, Dunbier AK, Toro TT, Lynch H, Wiesner G, Ferguson K, Eng C, Park JG, Kim SJ, Markowitz S. Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. Nat. Genet. 2000; 26:16–17.
38. Hermann JG, Umar A, Polyak K, Graff JR, Ahuja N, Issa J-P, Markowitz S, Wilson JK, Hamilton Sr, Kinzler KW, Kane MF, Kolodner RD, Vogelstein B, Kunkel TA, Baylin SB. Incidence and functional consequences of hMLH1 promoter hypermathylation in colerectal carcinomas. Proc. Natl. Acad. Sci. USA 1998; 95:6870-6875.
39. Di Croce L, Raker VA, Corsaro M, Fazi F, Fanelli M, Faretta M, Fuks F, Lo Coco F, Kouzarides T, Nervi C, Minucci S, Pelicci PG. Methyltransferase recruitment and DNA hypermethylation of target promoters by an oncogenic transcription factor. Science 2002; 295:1079–1082.
40. Meltzer PS. Cancer genomics: small RNAs with big impacts. Nature 2005; 435:745–746.
41. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR. Applications of DNA tiling arrays for whole-genome analysis. Genomics 2005; 85:1–15.

42. Lo YM, Wong IH, Zhang J, Tein MS, Ng MH, Hjelm NM. Quantitative analysis of aberrant p16 methylation using real-time quantitative methylation-specific polymerase chain reaction. Cancer Res. 1999; 59:3899–3903.

43. Laird CD, Pleasant ND, Clark AD, Sneeden JL, Hassan KM, Manley NC, Vary JC Jr, Morgan T, Hansen RS, Stöger R. Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. Proc. Natl. Acad. Sci. U.S.A. 2004; 101:204–209.

44. Lauchenbruch PA. Analysis of data with an excess of zeros. Stat Methods Med Res 2002; 11:297–302.

45. Neuhauser M, Boes T, Jockel K-H. Two-part permutation tests for DNA methylation and microarray data. BMC Bioinformatics 2005; 6:35.

46. Virmani AK, Tsou JA, Siegmund KD, Shen LYC, Long TI, Laird PW, Gazdar AF, Laird-Offringa IA. Hierarchical clustering of lung cancer cell lines using DNA methylation markers. Cancer Epidemiol. Biomarkers Prev. 2002; 11:291–297.

47. Model F, Adorjan P, Olek A, Piepenbrock C. Feature selection for DNA methylation based cancer classification. Bioinformatics 2001; 17:S157–S164.

48. Catto JWF, Linkens D, Abbod MF, Chen M, Burton JL, Feeley KM, Hamdy FC. Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modelling and artificial neural networks. Clin. Cancer Res. 2003; 9:4172–4177.

49. Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. Bioinformatics 2004; 20:1896–1904.

50. Marjoram P, Chang J, Laird PW, Siegmund KD. Cluster analysis for DNA methylation profiles having a detection threshold. BMC Bioinformatics 2006; 7:361.

51. Brena RM, Huang THM, Plass C. Toward a human epigenome. Nat. Genet. 2006; 38:1359–1360.

52. Yatabe Y, Tavare S, Shibata D. Investigating stem cells in human colon by using methylation patterns. Proc. Natl. Acad. Sci. U.S.A. 2001; 98:10839–10844.

53. Walters K. Colonic stem cell data is consistent with the immortal model of stem cell division under non-random strand segregation. Cell Prolif. 2009; 42:339–347.

Part V

Case Studies

Chapter 15

Modeling the Effect of Susceptibility Factors (HLA and PTPN22) in Rheumatoid Arthritis

Hervé Perdry and Françoise Clerget-Darpoux

Abstract

Numerous genome-wide analyses on common multifactorial diseases have been recently published in providing, for each associated Single Nucleotide Polymorphism (SNP), an Odds Ratio (OR), either for one of the susceptibility variant allele versus none, or for two copies of it versus one copy. Besides the poor information attached to these measures, it is a simplistic idea to reduce the effect of a gene to the one of an allele or of an haplotype. It is a far cry from detecting a signal indicating the presence of a causative factor in a genomic region to its identification and the important task of estimating the disease risk due to it. The contrast between cases and controls may be used for the estimation of the genotype relative risks. However, the same population distribution of a marker can be coupled with different modes of inheritance of the trait, and hence different risk estimates. Other sources of information, in particular at familial level must be used and can be crucial in discriminating the genotypes according to the disease risk. Illustration is given on two susceptibility factors to Rheumatoid Arthritis: HLA and PTPN22. In both cases, thanks to the sharing of parental alleles in affected sibs, a refining of the modeling was obtained. Tezenas du Montcel et al. (Arthritis Rheum 52:1063–1068, 2005) show that six HLA genotypes can be distinguished with different RA risks. One HLA genotype confers a risk 6.6-fold higher than another HLA genotype. For PTPN22, Bourgey et al. (BMC Proc 1 (Suppl 1):S37, 2007) show that observed data is not explained by a single variant as initially reported and that using the information on 3 SNPs discriminates the genotypic relative risks (GRRs) from 1 to 4.7.

Key words: Genotype relative risks, Rheumatoid Arthritis, HLA, PTPN22

1. Introduction

One of the greatest current challenges facing human genetics in the post-genome era is the understanding of the pathological pathways leading to multifactorial diseases. Genome-Wide Association (GWA) studies are successfully providing new strings to draw in the huge complexity of gene networks and lead sometimes to the identification of novel unsuspected pathways. This is

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_15, © Springer Science+Business Media, LLC 2011

well illustrated by the huge progress accomplished in Inflammatory Bowel Diseases (IBD) (1) since the localization (2) and identification (3, 4) of NOD2 as the first IBD susceptibility gene.

However, even if GWA studies have been quite productive for some multifactorial diseases in identifying new regions, new genes, new pathways, it is also clear that the limits of this approach will be reached soon and that a huge work will still need to be accomplished. First, we believe that many genes, playing an important role in the pathological pathway, are undetectable by this approach even with very large samples. Indeed, the approach is based on very simplistic assumption of independent effects, whereas interactions are extremely likely in most cases. Besides, when an association signal has been obtained on an SNP, there is still a long way to identify the causal variants, to obtain their mode of action and their connection with the other pieces of the pathway. The length of the region to screen, after an association signal has been obtained, appears to be small when compared to the one of a linkage signal. However, the extent and non-monotony of LD may be such that the region may content many genes which are good candidates and several of them may act in interaction in the disease process. This is of course the difficulty encountered in the HLA region in which associations between antigen alleles and many diseases have been detected more than 30 years ago (5). For the majority of those diseases, the HLA component is not yet elucidated. This is also the case with the 5q31-33 and 4q27 regions, in which linkage and/or association signals have been obtained for several autoimmune diseases.

Even when association with a disease pinpoints a single gene, the estimation of the risk corresponding to the different genotypes is not that simple. The distribution of a marker in cases and controls is too poor for this purpose. They can be coupled with different modes of inheritance of the trait, i.e., different risk estimates. Other sources of information, in particular at familial level must be used and can be crucial in discriminating the genotypes according to the disease risk.

We illustrate this on the modeling of two susceptibility factors to Rheumatoid Arthritis: HLA and PTPN22.

2. Rheumatoid Arthritis

Rheumatoid Arthritis (RA) is one of the most common autoimmune diseases affecting circa 1% of adult population with a majority of women. It is a chronic inflammatory syndrome with a wide clinical spectrum varying from mild to severe disabling disease. Little is known about its etiology. A higher concordance rate in monozygotic (15%) than in dizygotic twins (4%) was reported by

Silman et al. (6), supporting a genetic component in the disease susceptibility.

Besides, the implication of Human Leucocyte Antigens (HLA) in this genetic susceptibility is known for a long time. Indeed, associations between HLA and RA were reported by Statsny (7) three decades ago. However, the biological mechanism underlying these associations remains unknown. In 1987, Gregersen et al. (8) observed that the *HLA-DRB1* alleles reported to be associated, share an RAA (arginine, alanine, alanine) sequence in position 72–74 of the amino acid. They hypothesized that this motif might be functional. Several studies tried to model the role of this shared epitope in RA (9–11), but all concluded that it could not explain *HLA-DRB1* involvement in RA susceptibility.

More recently, the SNP rs2476601 has been repeatedly shown to be associated with RA (12–14). This SNP is located within the hematopoietic-specific protein tyrosine phosphatase gene, PTPN22. Its minor allele T confers 2.1-fold increased risk to heterozygote and 2.7-fold increased risk to homozygote carriers (15) compared to the non-carrier individuals.

We show that using the simultaneous information of association and linkage on these two factors – HLA and PTPN22 – allows a better modeling of their effect in RA susceptibility.

3. Modeling a Gene Effect

When a gene has been identified as involved in the disease susceptibility, we need to better understand the role of this gene, in other words, to go from associated SNPs to the functional variant(s) and to evaluate the differential of risks according to the genotypes.

3.1. Selection of the Most Associated Set of SNPs

When information is available on several SNPs within the gene under consideration, we have to select first the set of SNPs that are the most associated. The association studies usually focused on one marker at a time and compared allele or genotype distributions for the studied markers in cases and controls or in transmitted versus untransmitted alleles in the case of family-based datasets. However, disease susceptibility may be due to the combined effects of multiple sequence variants. The combination test was developed by Jannot et al. (16) to perform a joint analysis of multiple markers. The principle of the method consists in testing all possible subsets of SNPs within a gene. Such a systematic testing of all SNP subsets poses a problem of multiple testing that is solved by the implementation of permutation procedures allowing the estimation of corrected *p*-values. Using simulated data, the combination test was shown to be powerful in many situations, in particular, when several SNPs interact and have small individual effects. H. Perdry extended the combination test to trio families and developed an efficient software[1] for performing this test.

A set of SNPs selected in this step may be considered as a single multiallelic marker, each possible haplotype being considered as an allele. The genotypes are then defined by all combination of two haplotypes (identical or different). For this multiallelic marker, it is then possible to compute the Genotypic Relative Risks (GRRs) by using the genotypic distributions in patients and controls.

Note that several sets of SNPs (several markers) may be indistinguishable in the strength of the association with the disease. When one selected set is nested in another one, we may keep the most parsimonious one. If it is not the case, we must keep in consideration the equivalent sets (equivalent markers).

3.2. Use of Linkage Information in Affected Sibs

The number of parental alleles Identical By Descent (IBD) shared by the affected sibs for the gene under study depends on the gene model in the disease susceptibility (i.e., on the frequencies of the functional genotypes and on the corresponding GRR). Thus, a direct role of a marker selected in the previous step and the estimated GRRs may be tested by comparing the observed IBD to its expectation under this model.

More information may be obtained on the model by using the IBD sharing conditioned on the genotype of one of the affected sibs (17). This information is most often ignored, whereas it may be very useful to confirm or refute a risk hierarchy established in an association study. This is illustrated (Box A) on a VNTR flanking the insulin gene known to be associated to type 1 diabetes (18, 19). The IBD distribution for the insulin gene is close to the expectation under the null (1/4, 1/2, 1/4) when observed on a sample of affected sib pairs but is highly stratified according to proband genotype for this VNTR. This observed IBD stratification fits a direct role of the VNTR in the susceptibility to type I diabetes (18).

3.3. Modeling the Gene Effect While Taking into Account Both Linkage and Association Information

It is possible that none of the gene markers selected in the first step represents what is functional in the disease process. In that case, the observations on the multiallelic marker (M) may be used to infer the modeling of the functional variation in the gene, considered as a multiallelic disease locus (S) (the number of S alleles and M alleles can differ). Additional parameters, which express a

[1] Available on request (herve.perdry@inserm.fr).

Box A. Stratified IBD Sharing

Let us assume the effect of a biallelic variant (a, A) where the probabilities of being affected of the genotypes aa, aA, and AA differ. The proportion of sibs sharing 2, 1, or 0 parental IBD allele(s) depends on the genotype for the variant of the index (one of the two sibs randomly chosen). The proportion of sibs sharing two IBD alleles (IBD = 2) is greater than 0.25 when the index has the highest risk genotype and is smaller than 0.25 when the index has the lowest risk genotype (conversely for the proportion of sibs IBD = 0).

As an example, let us consider the VNTR flanking the insulin gene shown to be associated to type I diabetes (19, 20). This VNTR may be considered as a biallelic variant. The allele "a" represents a number of repeats lower than or equal to 1,000 repeats and the allele "A" more than 1,000 repeats.

In Caucasian populations, the frequency of "a" is 0.70 (the frequency of A = 0.30) and the relative penetrances estimated from cases and controls of the Bell et al. study (20) are (1, 0.33, 0.06).

Under this model, we may compute with the MASC program the IBD expectation conditioned on index genotype.

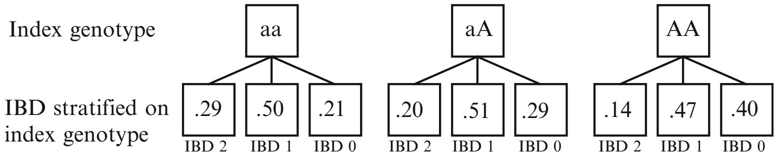


Fig. 1. Expected proportions of sibs sharing 2, 1, or 0 IBD alleles according the index has the genotype aa, aA, or AA, respectively.

The proportion of affected sibs sharing two haplotypes with their index are expected to be 0.29 (greater than 0.25) when the index is homozygote for "a" (less than 1,000 repeats) but 0.14 (lower than 0.25) when the index is homozygote for "A" (more than 1,000 repeats) (see Fig. 1). The observations on 95 affected sib pairs typed for the Genetic Analysis Workshop 5 (19) fit these expectations.

relation between the marker alleles and the disease locus alleles which are the coupling frequencies, have then to be considered. To deal with this high number of parameters (the risks associated to each disease locus genotype, the coupling frequencies between the marker alleles and the disease locus alleles), other sources of

Box B. Principle of the MASC Method

The index cases are classified in three steps, each one being nested within the previous one:

1. First, the index cases are classified following their familial configuration, i.e., the number of affected parents and the presence or absence of at least one affected sib.
2. Each category defined at step 1 is divided in subcategories according to the index genotypes at the marker.
3. Each category defined at step 2 is divided in three subcategories according to the IBD of the index with a given sib.

The distributions expected under a given model at each step of classification can be expressed in term of the marker allele frequencies, the values of the coupling parameters, and the penetrances, which allows computing the likelihood of this model. This likelihood is then maximized using a numeric method. The fit of the model to the observations may then be tested; two nested models may also be compared by a maximum likelihood ratio test.

information that the marker distribution in patient and controls samples have to be used, in particular familial segregation and linkage information.

This was the main idea of the MASC method proposed by Clerget-Darpoux et al. in 1988 (17). The method integrates information on a sample of patients (index cases) at different levels (Box B). Without going into details, it is useful to recall two basic notions underlying this method.

When a marker has been shown to be associated to a disease:

- The marker genotype distribution in a patient sample depends upon the mode of ascertainment of the patients. In particular, the frequency of the more at risk genotype is higher in a sample of patients known to have affected relatives than in a random sample of patients.
- The IBD sharing of a patient with an affected sib depends upon the patient genotype. We expect an excess of sharing when the patient has the high-risk genotype and conversely a decrease of sharing when the patient has a low-risk genotype.

4. Modeling the Effect of HLA-DRB1 in the Susceptibility to Rheumatoid Arthritis

Rheumatoid Arthritis was shown to be associated with a variety of alleles of the class II gene HLA-DRB1. In 1987, Gregersen et al. remarked that the alleles associated with RA (8) share the epitope RAA at position 72–74. They formulated the so-called *Shared Epitope hypothesis*, according to which this specific sequence was responsible for the association between HLA-DRB1 and RA. The risks corresponding to the different associated alleles were not the same and Gao et al. (21) proposed to consider, according to the strength of associations, four different alleles E_1, E_2, E_3, E_X by pooling the *HLA-DRB1* alleles as follows: *HLA-DRB1*0101*, **0405*, **0408*, and **1001* alleles as E_1, the *HLA-DRB1*0401* alleles as E_2, the *HLA-DRB1*0102*, and **0404* alleles as E_3, and the other *HLA-DRB1* alleles as E_X. This classification, which was standard in the literature, was based on alleles known to be associated with RA.

The accuracy of Gao's classification was tested on two different French family samples, which differed in their ascertainment schemes. In the first sample, DNA was collected from one hundred unrelated RA patients and from their two parents (Sample 1). The second sample comprised 132 patients (probands) with at least one affected sib.

The RA patients were recruited in France through the European Consortium on Rheumatoid Arthritis Families (ECRAF). All patients fulfilled the 1987 American College of Rheumatology criteria. Blood samples were collected for DNA extraction and genotyping. *HLA-DRB1* typing (Dynal Classic SSP "DR low resolution") and subtyping (Dynal Classic, "high resolution" for *HLA-DRB1*01*, **04*, **11* and **13*) used the PCR-SSP method (Dynal Biotech, Lake Success, NY). The four alleles still ambiguous after the high resolution subtyping were directly sequenced on exon 2.

The genotype distributions in Samples 1 and 2 (association information) were considered and, for each proband genotype of Sample 2, the proportion of sibs sharing 2, 1, or 0 alleles IBD (stratified IBD distributions). According to the proband genotype, these proportions are expected to differ (11). Tezenas du Montcel et al. (22) showed that this allele classification does not explain all the observations. In particular, 40% of the E_X/E_X patients, rather than the expected less than 25%, shared two identical by descent *HLA-DRB1* alleles with their affected sib.

Consequently, a new classification of the *HLA-DRB1* alleles was proposed (22). It was shown that the risk of developing RA depends on whether the RAA sequence occupies positions 72–74, but is modulated by the amino acid at position 71 (K confers a highest risk, R an intermediate risk, A and E a lower risk) and by the amino acid at position 70 (Q or R confer highest risk than D). The KRAA motif (denoted S_2) at position 71–74 of *HLA-DRB1* confers the highest susceptibility and the RRRAA or QRRAA motif (denoted S_{3P}) an intermediate risk. All other motifs were denoted L (see Table 1). Using the MASC method, Tezenas du Montcel et al. showed that the model which best fit the *HLA-DRB1* genotype distribution in the two samples and the distribution of parental alleles shared by affected sib pairs could be described by three alleles (S_2, S_{3P}, L) and six genotypes with different RA risks. The following risk hierarchy was established (see Table 2): risk was significantly higher for S_2 than for S_{3P} ($p < 0.002$), which in turn was higher than for L ($p < 10^{-11}$). The maximum genotype risk was for the S_2/S_{3P} genotype; its risk was 6.6-fold higher than that for the L/L genotype, followed by S_2/S_2 (GRR = 5.9), S_{3P}/S_{3P} (GRR = 3.3), S_2/L (GRR = 2.7), S_{3P}/L (GRR = 1.9).

In our study, the ERAA sequence is associated with the same level of risk as the alleles without the RAA motif at position 72–74. Glutamic acid (E) at position 71 seems to suppress the effect of the RAA motif and, similarly, for an aspartic acid (D) at

Table 1
Classification of HLA-DRB1 alleles following the amino acid sequence at position 70–74

	Position				
Allele	**70**	**71**	**72**	**73**	**74**
S_2	*	K	R	A	A
S_{3P}	R	R	R	A	A
	Q	R	R	A	A
L	D	R	R	A	A
	*	E	R	A	A
	*	*	Non (R A A)		

* indicates that the risk does not depend on the amino acid at that position

Table 2
Genotype relative risks for rheumatoid arthritis

	L	S_2	S_{3P}
L	1		
S_2	2.7	5.9	
S_{3P}	1.9	6.6	3.3

position 70. In Gao's classification, 50% of the alleles classified as low-risk allele actually shared the RAA motif. A large majority of these alleles share a DRRAA, ARAA, or ERAA motifs, which are indeed low-risk alleles, but 4% share the high-risk motif KRAA.

Our classification and hierarchy of genotypic risks was subsequently confirmed in a new sample of 100 French Caucasian families having one RA patient and both parents genotyped for the HLA-DRB1 gene (23). More recently, support to our modeling was also provided by several studies (24–28) both at statistical and functional levels.

5. Modeling the Role of PTPN22 in Rheumatoid Arthritis

Association between Rheumatoid Arthritis and an SNP of PTPN22 (rs2476601) has been confirmed by numerous studies. However, using data from the North American Rheumatoid Arthritis Consortium (NARAC), Carlton et al. (15) demonstrated that the variant rs2476601 does not fully explain the association between PTPN22 and RA and suggested the existence of at least another variant in PTPN22. Bourgey et al. (29) reanalyzed the NARAC data using both association and linkage information for modeling the role of PTPN22 in RA. Those data were made available though the 15th Genetic Analysis Workshop (GAW 15). A sample of 511 families with affected sib pairs typed for 14 SNPs of PTPN22 and 1,404 unrelated controls from the NARAC data were used. For each affected sib pair, the proband was considered as an index patient. The SNP rs2476601 was among the typed SNPs. An LD analysis leaded to the exclusion of 3 SNPs which were in complete LD with other SNPs.

The Combination Test was applied to the 11 remaining SNPs. Then, a forward procedure using nested chi-square tests was used to select a parsimonious and highly associated subset of SNPs. A set of three SNPs, rs2476601, rs12730735, rs11102685, was retained.

MASC was then used to estimate the GRR of each genotype, for rs2476601 alone, and for the selected three SNPs. The results are displayed in Table 3. It is interesting to note that:

1. When the information is taken on the three SNPs, the GRR varies quite more largely (from 1 to 4.7) than when the information is only taken on the SNP rs2476601 (from 1 to 2.7)
2. Among individuals with the low-risk genotype CC for the SNP rs2476601, the two other SNPs strongly differentiate their GRR which ranges from 1 to 3.60.

In order to avoid cells with small number, some genotypes were pooled together according to their GRR value. This pooling leads to the definition of four different genotype risks for an individual (Table 4).

The number of IBD alleles of PTPN22 shared by the index case and one affected sib were estimated using MERLIN (30). Their distributions conditional on the genotype class of the index case are displayed in Table 5.

Table 3
GRR estimates

rs2476601	GRR	All 3 selected SNPs	GRR
CC	1	CC-AA-AA CC-AA-AG CC-AA-GG CC-AG-AA CC-*G-AG CC-GG-AA	1.60 1.76 3.60 1.73 2.35 1
CT	1.66	CT-AA-AA CT-AA-AG CT-AG-AA	2.88 3.11 2.61
TT	2.7	TT-A*-AA	4.68

*Either the A or the G alleles of rs12730735

Table 4
Genotype risk classes

Class	GRR range	Number of sib pairs in the class
Low (L)	GRR = 1	19
Intermediate 1 (I1)	1 < GRR ≤ 2	295
Intermediate 2 (I2)	2 < GRR ≤ 3	157
High (H)	GRR > 3	34

Table 5
IBD distribution stratified on genotype classes

Genotype class	IBD = 0	IBD = 1	IBD = 2
L	0.47 (9)	0.42 (8)	0.11 (2)
I1	0.29 (85)	0.49 (146)	0.22 (64)
I2	0.26 (41)	0.50 (78)	0.24 (38)
H	0.09 (3)	0.65 (22)	0.26 (9)

In parenthesis, number of sib pairs

Table 6
IBD distribution stratified on rs2477601 genotypes

Genotype	IBD = 0	IBD = 1	IBD = 2
CC	0.27 (96)	0.49 (177)	0.24 (85)
CT	0.25 (35)	0.53 (75)	0.22 (31)
TT	0.14 (2)	0.65 (10)	0.21 (3)

In parenthesis, number of sib pairs

The stratified IBD distributions are consistent with the genotype risk classes. There is a large excess of IBD = 0 when the index has a low-risk genotype. The proportion of sibs that share no haplotype with their index decreases when the GRR increases (from 0.47 to 0.09) and conversely the proportion of sibs sharing two haplotypes increases (from 0.11 to 0.26)

In contrast, stratification of the IBD distribution on rs2477601 genotypes alone is not consistent with the GRR estimates for these genotypes (Table 6).The proportion of sibs sharing 2 IBD alleles being the greatest when the index has the low-risk genotype CC.

The MASC method was applied using the four genotype classes defined above and the IBD stratified on them. The direct effect of rs2476601 is strongly rejected ($p = 0.005$). The hypothesis of the effect of a single untyped SNP was also reject ($p = 0.04$). The data are compatible not only with an interactive effect of the three selected SNPs, but also with an effect of two untyped SNP in LD with those which have been typed.

We showed in this analysis (29) that observed data is not explained by a single variant as initially reported and that using the information on 3 SNPs discriminate the GRRs from 1 to 4.7 quite more largely than reported in the literature (from 1 to 2.7).

6. Discussion

Familial information such as recurrence risks in sibs and parents of affected was, for long, the unique information used by geneticists for modeling the genetic transmission of a disease. The biological revolution of the three last decades offering genetic markers along the genome has provoked a strong shift of the genetic studies with the more ambitious goal of localizing and identifying not only the disease genes involved in monogenic diseases, but also the genetic factors involved in the more complex multifactorial diseases. Geneticists first concentrated their effort in sampling and typing affected relatives, in particular affected sib pairs.

During the last decade of the twentieth century, the most popular approach was genome-wide linkage analysis. Very simple reasoning and observations made clear that the information of marker segregation in affected relatives was too poor in front of the complex etiology of the majority of human diseases. However, instead of enriching and cumulating different sources of information to deal with this complexity, the genetic studies of the last decade promoted the idea that since very dense typing and even the entire sequence of individual human genome were available, it is sufficient to contrast the genome of affected versus unaffected individuals. GWA studies are currently being conducted on samples of unrelated persons in the belief that this is now the design of choice to discover genetic variants underlying relatively common complex diseases. Can we hope to anticipate, understand, and treat the many diseases to which humans are prone, simply by finding genomic locations in which allele frequencies slightly differ between those who have and those who do not have disease?

We firmly believe that the current efforts being put into the construction of huge databases for case-control studies should not be done to the detriment of continued collection of family data (31, 32). We simply show, in this paper, that it is possible to refine the modeling of a disease susceptibility gene by using both association and linkage information. The main motivation behind looking for genetic risk factors for complex diseases is to get a better insight into the pathological process of the diseases. This means that we must go further in looking at the simultaneous effect of the genes and their interaction. We probably have to model in the future the effect of sets of genes involved in the disease susceptibility through complex biological pathways. So many different biological mechanisms are possible that it would be foolhardy to restrict all human genetic research to a single strategy. One of the greatest current challenges facing human genetics is that of how best to gather and synthesize the many lines of evidence possible in order to discover the genetic determinants underlying complex diseases.

References

1. Budarf ML, Labbé C, David G, Rioux JD. GWA studies: rewriting the story of IBD. *Trends Genet* 2009 Mar; 25(3):137–146.

2. Hugot JP, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L, Naom I, Dupas JL, Van Gossum A, Orholm M, Bonaiti-Pellie C, Weissenbach J, Mathew CG, Lennard-Jones JE, Cortot A, Colombel JF, Thomas G. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996 Feb; 29:821–823.

3. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cézard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001 May 31; 411(6837):599–603.

4. Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T,

Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nuñez G, Cho JH. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001; 411(6837): 603–606. Ogura 2001.

5. Dausset J, Svejgard A. (editors) *HLA and Disease*. INSERM, Paris 1976.
6. Silman AJ, MacGregor AJ, Thomson W, Holligan S, Carthy D, Farhan A, Ollier WE. Twin concordance rates for rheumatoid arthritis: results from a genomewide study. *Br J Rheumatol* 1993; 32:903–907.
7. Stastny P. Association of the B-cell alloantigen DRw4 with rheumatoid arthritis. *N Engl J Med* 1978; 298:869–871.
8. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum* 1987; 30:1205–1213.
9. Génin E, Babron MC, McDermott MF, Mulcahy B, Waldron-Lynch F, Adams C et al. Modelling the major histocompatibility complex susceptibility to RA using the MASC method. *Genet Epidemiol* 1998; 15:419–430.
10. Rigby AS, MacGregor AJ, Thomson G. HLA haplotype sharing in rheumatoid arthritis sibships: risk estimates subdivided by proband genotype. *Genet Epidemiol* 1998; 15:403–418.
11. Tezenas-du-Montcel S, Reviron D, Génin E, Roudier J, Mercier P, Clerget-Darpoux F. Modeling the HLA component in rheumatoid arthritis: sensitivity to DRB1 allele frequencies. *Genet Epidemiol* 2000; 19:422–428.
12. Hinks A, Barton A, John S, Bruce I, Hawkins C, Griffiths CE, Donn R, Thomson W, Silman A, Worthington J. Association between the PTPN22 gene and rheumatoid arthritis and juvenile idiopathic arthritis in a UK population: further support that PTPN22 is an autoimmunity gene. *Arthritis Rheum* 2005; 52(6):1694–1699.
13. Criswell LA, Pfeiffer KA, Lum RF, Gonzales B, Novitzke J, Kern M, Moser KL, Begovich AB, Carlton VE, Li W et al. Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *Am J Hum Genet* 2005; 76(4):561–571.
14. Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoerke JM et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* 2004; 75(2):330–337.
15. Carlton VE, Hu X, Chokkalingam AP, Schrodi SJ, Brandon R, Alexander HC, Chang M, Catanese JJ, Leong DU, Ardlie KG et al. PTPN22 genetic variation: evidence for multiple variants associated with rheumatoid arthritis. *Am J Hum Genet* 2005; 77(4): 567–581.
16. Jannot AS, Essioux L, Reese MG, Clerget-Darpoux F. Improved use of SNP information to detect the role of genes. *Genet Epidemiol* 2003; 25:158–167.
17. Clerget-Darpoux F, Babron MC, Prum B, Lathrop GM, Deschamps I, Hors J. A new method to test genetic models in HLA associated diseases: the MASC method. *Ann Hum Genet* 1988; 52:247–258.
18. Dizier MH, Babron MC, Clerget-Darpoux F. Interactive effect of two candidate genes in a disease: extension of the Marker-Association-Segregation χ^2 Method. *Am J Hum Genet* 1994; 55:1042–1049.
19. Spielman RS, Baur MP, Clerget-Darpoux F. Genetic analysis of IDDM: summary of GAW5 IDDM results. *Genet Epidemiol* 1989; 6:43–58.
20. Bell GI, Horita S, Karam JH. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 1984; 33:176–183.
21. Gao X, Gazit E, Livneh A, Stastny P. Rheumatoid arthritis in Israeli Jews: shared sequences in the third hypervariable region of DRB1 alleles are associated with susceptibility. *J Rheumatol* 1991; 18:801–803.
22. Tézenas du Montcel S, Michou L, Petit-Teixeira E, Osorio J, Lemaire I, Lasbleiz S, Pierlot C, Quillet P, Bardin T, Prum B, Cornelis F, Clerget-Darpoux F. New classification of HLA-DRB1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. *Arthritis Rheum* 2005; 52:1063–1068.
23. Michou L, Croiseau P, Petit-Teixeira E, Tezenas du Montcel S, Lemaire I, Pierlot C, Osorio J, Frigui W, Lasbleiz S, Quillet P, Bardin T, Prum B, Clerget-Darpoux F, Cornélis F and The European Consortium on Rheumatoid Arthritis families. Validation of the reshaped shared epitope HLA DRB1 classification in Rheumatoid arthritis. *Arthritis Res Ther* 2006; 8(3):R79.
24. Morgan AW, Haroon-Rashid L, Martin SG, Gooi HC, Worthington J, Thomson W, Barrett JH, Emery P. The shared epitope hypothesis in rheumatoid arthritis: evaluation of alternative classification criteria in a large UK Caucasian cohort. *Arthritis Rheum* 2008; 58(5):1275–1283.

25. Bridges SL Jr, Kelley JM, Hughes LB. The HLA-DRB1 shared epitope in Caucasians with rheumatoid arthritis: a lesson learned from tic-tac-toe. *Arthritis Rheum*, 2008; 58(5):1211–1215.
26. Barnetche T, Constantin A, Cantagrel A, Cambon-Thomsen A, Gourraud PA. New classification of HLA-DRB1 alleles in rheumatoid arthritis susceptibility: a combined analysis of worldwide samples. *Arthritis Res Ther* 2008; 10:R26.
27. Winchester R. Reshaping Cinderella's slipper: the shared epitope hypothesis. *Arthritis Res Ther* 2006; 8:109.
28. Nordang GBN, Gourraud P-A, Viken MK, Constantin A, Cambon-Thomsen A, Thorsby E, Forre O, Kvien TK, Lie B. The Tezenas du Montcel classification of HLA-DRB1 alleles in Norwegian rheumatoid arthritis patients differentiates clinical manifestations: O-30. *Tissue Antigens* 2008; 71(4):276.
29. Bourgey M, Perdry H, Clerget-Darpoux F. Modelling the effect of PTPN22 in rheumatoid arthritis. *BMC Proc* 2007; 1(Suppl 1): S37.
30. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; 30(1):97–101.
31. Bourgain C, Génin E, Cox N, Clerget-Darpoux F. Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases? *Eur J Hum Genet* 2007; 15:260–263.
32. Clerget-Darpoux F, Elston RC. Is linkage and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum Genet* 2007; 64:91–96.

Chapter 16

Coronary Artery Disease: An Example Case Study

Jennifer H. Barrett

Abstract

This chapter illustrates various general issues in genetic epidemiology in relation to coronary artery disease (CAD). This is a disease strongly influenced by environmental/lifestyle factors, such as smoking, but with substantial estimated heritability. Researchers aiming to identify susceptibility genes have used several different definitions of CAD, some focusing on the common presentation of myocardial infarction (MI) and others adopting broader criteria, often imposing an upper limit to age at diagnosis to minimise environmental effects. Many candidate gene association studies and a few large genome-wide linkage studies have been conducted, but with limited success.

Several heritable quantitative traits are strongly related to risk of CAD (e.g. blood pressure and cholesterol levels), and much research has been focussed on identifying genes that influence these traits. Quantitative traits have the advantage of being measurable on any individual, allowing them to be studied in population-based cohorts. However, they also tend to vary considerably over time, and intra-individual variation needs to be taken into account in analyses.

In the last few years, both CAD itself and related quantitative traits have been studied in genome-wide association studies using large sample sizes. Several novel genetic loci influencing CAD have been identified and replicated, in addition to many loci influencing related quantitative traits. However, despite this recent success, only a small fraction of the genetic contribution to risk has been explained.

Key words: Coronary artery disease, Myocardial infarction, Genetic epidemiology, Quantitative trait, Genome-wide association study

1. Introduction

This chapter is not intended to provide a comprehensive guide to the genetic epidemiology of cardiovascular disease (1). Instead, we use this disease to illustrate various general issues in genetic epidemiology.

Cardiovascular disease (disease of the heart and circulatory system) is the main cause of death in the UK, accounting for more than one in three deaths (2). The common occurrence of the disease is not restricted to the developed world; it is also the main

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_16, © Springer Science+Business Media, LLC 2011

cause of death world-wide, and over 80% of such deaths take place in low- and middle-income countries (3). The most common form of cardiovascular disease is coronary artery disease (CAD), which is heart disease associated with underlying atherosclerosis. Atherosclerosis is often undetected until the subject suffers from a myocardial infarction (MI). In other instances, the disease may first give rise to angina. If severe atherosclerosis is diagnosed, this may be treated by angioplasty or coronary artery bypass surgery. Other forms of heart disease contribute importantly to morbidity and mortality, such as cardiomyopathy, which is deterioration of heart muscle potentially leading to heart failure.

Some features of CAD add to the difficulty of studying its genetic aetiology: firstly, it is a complex heterogeneous phenotype, presenting in different ways as discussed above, and secondly disease risk is strongly influenced by lifestyle.

Because of the complexity of the phenotype, researchers have used several different definitions of CAD when studying its genetic epidemiology. For example, the British Heart Foundation (BHF) Family Heart Study (4) used as diagnostic entry criteria any one of MI, angina, angioplasty, or coronary artery bypass surgery before the age of 66. In contrast, the German MI Family Study (5, 6) focussed on subjects who had suffered an MI before their 60th birthday. Disease heterogeneity, if it reflects underlying differences in aetiology, may seriously reduce the power to detect risk factors, a point that argues in favour of a narrower phenotype definition, although any finding may then not be generalisable to other subtypes of CAD. Such differences in phenotype definition hamper comparability between studies, although some recently-identified genetic risk factors, including the single nucleotide polymorphisms (SNPs) on chromosome 9p21.3 identified in recent genome-wide association (GWA) studies, seem to be robust to differences in CAD-related phenotype definitions (6–9).

Most genetic studies, like the two family studies mentioned above, impose some upper limit to age at diagnosis. The reasoning is that for many complex diseases early onset is more likely to indicate genetic susceptibility. This is of particular relevance to diseases like CAD, where incidence increases with age and there are strong environmental risk factors. Lifestyle factors that have been established as increasing risk of CAD include: smoking; a diet high in salt and saturated fat and low in fruit, vegetable, and fibre intake; physical inactivity; and psychosocial factors, such as stress and depression. Among these, smoking is the factor that combines relative ease of measurement with a strong effect on risk, with the increase in CAD mortality for smokers compared with lifelong non-smokers estimated at ~60% from 50 years follow-up of a cohort of British male doctors (10). An interesting study design would therefore be to search for genetic risk factors by studying non-smoking cases, a point we return to in Section 4.

Other features of CAD make its genetic epidemiology easier to study. For one thing, it is a common disease so that large studies are feasible, including studies of families with multiple affected members. Secondly, there exist several "intermediate" quantitative phenotypes that influence risk, such as high- and low-density lipoprotein (HDL and LDL) levels, triglyceride levels and blood pressure (BP), which can be measured and investigated using samples of individuals from the population (see Section 3).

There are rare forms of CAD, such as familial hypercholesterolemia, characterised by very high levels of LDL, where rare mutations confer vastly increased risk of disease. Such forms of CAD have been successfully studied by carrying out linkage analysis on large pedigrees with several affected members often showing a clear pattern of inheritance (11). Although risk conferred by the mutations subsequently identified in the LDL receptor gene is very high, because of their rarity the impact on population incidence of CAD is low. The main focus of this chapter is on common forms of CAD with multiple genetic and environmental risk factors.

A first step in studying the genetic epidemiology of any disease is to establish that there is an important genetic contribution to disease risk, and this is often done by estimating the heritability. For a binary trait, this is done by modelling an underlying unobserved continuous variable representing liability (susceptibility to the trait) and estimating the proportion of variation in liability that is due to genetic factors. Because of the strong potential for ascertainment bias in estimating heritability of CAD incidence, more reliable estimates can be expected by studying mortality. A Swedish study of over 20,000 twins compared rates of concordance in death from coronary heart disease in monozygotic and dizygotic twin pairs (12). Heritability was estimated as 57% in males and 38% in females; genetic effects were shown to be important throughout life, but particularly at earlier ages. Another measure of familial contribution to risk is the sibling relative risk, denoted λ_s, which is the risk of disease in siblings of cases compared with the population risk. Ascertainment bias can also lead to difficulties in obtaining unbiased estimates of this measure (13). For CAD, λ_s has been estimated at ~1.6 using the Framingham Heart Study, a large population-based prospective cohort study (14); adjustment for smoking and for risk factors, such as body mass index, systolic BP, and total cholesterol to HDL cholesterol ratio did little to attenuate the estimated risk. It is thus established that genetic susceptibility contributes importantly to risk of CAD, justifying and motivating the search for disease-related genes.

2. Candidate Gene Association Studies and Linkage Analysis

Until recently, the only available gene mapping methods available to researchers were linkage analysis across the whole genome and association analysis applied to candidate genes. As with other complex diseases, these approaches have had only limited success in CAD.

Numerous genome-wide non-parametric linkage studies of CAD have been conducted (1), the largest and most recent of which are the BHF Family Heart Study (4) and the PROCARDIS study (15), both of which collected genetic and phenotypic information from ~2,000 families with at least two affected siblings. The studies had similar inclusion criteria, and both also reported subgroup analyses on the narrower MI phenotype. For the BHF study, the highest LOD score was 1.98 on chromosome 2 for CAD overall, with a pointwise *p*-value of 0.001 and a genome-wide *p*-value estimated by simulation of 0.31. PROCARDIS employed a two-stage design and identified a locus on chromosome 17 showing evidence of linkage in the MI subgroup (LOD score 2.68 from combined analysis of both stages). Using exclusion analysis, 85% of the autosomal genome was excluded for a locus-specific λ_s of 1.24, although interestingly the chromosome 2 region from the BHF study was not excluded. However, despite the relatively large sample size of these two studies, the evidence for linkage is not overwhelming in either study, and in neither case has a disease gene in the linkage region subsequently been identified.

Overall, the regions identified as potentially important from linkage studies show little concordance with SNPs recently found to be associated with CAD from GWA studies (see Subheading 4). Although other factors such as heterogeneity may contribute, the principle factor giving rise to this is almost certainly the lack of power for even the larger linkage studies to detect loci with the magnitude of genetic risk now being identified.

The complexity of CAD and current understanding of its aetiology and related phenotypes have led to very many candidate genes being proposed and investigated through association analysis. Several of these have shown reasonable evidence of association in meta-analyses (e.g. *APOE*, *PAI1*, *ACE*, and *MTHFR*) although results have not been consistent. Morgan and colleagues (16) identified 85 genetic variants in 70 genes previously reported to be associated with atherosclerosis or acute coronary syndrome and tested for association in a case-control study of 811 patients and 650 controls. Only one variant showed nominal evidence of association and the overall results were entirely consistent with the global null hypothesis (i.e. none of the tested variants being associated with disease).

Although several factors may contribute to the inability to replicate findings, including population heterogeneity, the most important factors are likely to be low power and the application of insufficiently stringent significance thresholds to candidate gene studies. A significance level of 0.05, or at most some correction for the number of loci reported within a particular manuscript, has generally been regarded as providing evidence of association. However, a consensus is emerging that, even for candidate loci, more stringent p-values and larger studies are required to avoid an unacceptably high proportion of false positive results (17).

3. Quantitative Traits

Several quantitative traits which are strong predictors of CAD have themselves been shown to be heritable. For example, heritability estimates of LDL and HDL cholesterol have been obtained of 33% and 44%, respectively, from population-based families in the UK (18), with even higher estimates obtained from twin studies (19). Another approach open to genetic epidemiologists is therefore to investigate genetic influences on the traits themselves. This has been the focus of a large number of linkage and candidate gene association studies (see summary in (1)), and more recently GWA studies have been conducted, for example of BP (7, 20) and lipid levels (21–26).

Quantitative traits differ in various respects from binary disease outcomes; most importantly they can be measured on anyone, and they may vary widely over time and be strongly influenced by environmental factors, including treatment. These facts have implications for both design and analysis.

An important advantage of quantitative traits is that they can be analysed in a population-based sample, with no need to ascertain on the basis of disease status, and cohort studies are thus a feasible approach. There are an increasing number of large well-characterised population-based cohorts, where consent for genotyping has been obtained. Aside from the efficiency of being able to study many traits in the same cohort, prospective studies have a particular advantage for the study of gene–environment interaction, since environmental factors can be measured before the onset of disease (27). In contrast, case-control studies may be subject to recall bias, or it may not be possible to distinguish cause from effect.

A pioneering example of this approach is the Framingham Heart Study (http://www.framinghamheartstudy.org), an epidemiological study which was started in 1948 with the objective of identifying common risk factors for cardiovascular disease. Initially ~5,000 men and women were recruited from the town of Framingham, Massachusetts. In addition to questionnaires on lifestyle, various quantitative traits were measured on participants, and they were followed up with subsequent examinations and interviews at frequent intervals. Since then, second and third generations of the original participants' families have been added to the study cohort. To date, nearly 2,000 research articles have been published based on this study, including recent large-scale genetic association studies of 17 groups of phenotypic traits (28).

It is important to minimise the effect of intra-individual variation on the study of a quantitative trait, ideally by measuring the trait on more than one occasion and ensuring that the conditions of measurement are as homogeneous as possible between subjects. With detailed longitudinal data, attempts can be made to allow for treatment effects and other extraneous factors, and summary measures over time can be used. For example, BP was among the traits recently investigated in the Framingham Study (29), and genetic associations were analysed in relation to both measures taken at a single time point and a long-term measure based on average BP over at least 12 years (based on at least three time points), adjusted for age, sex, and body mass index.

Quantitative traits measured on a prospectively followed cohort also lend themselves to more complex statistical modelling fully exploiting the longitudinal nature of the data. Tobin and colleagues (30) recently analysed common variants in the *WNK1* gene in relation to BP in childhood. They used a mixed model that included random effects for individual-specific baseline BP and the gradient of increase in BP with age. The model takes account of the covariance in measures over time within individuals and allows the effects of SNPs on both baseline BP and the rate of increase with age to be modelled without the need to reduce the longitudinal data to summary form.

A potential disadvantage of the cohort approach is that studying the whole range of the distribution of a trait in the population may not be powerful, especially if the genetic influences are more important in the extremes. In addition, from a clinical perspective most interest is generally in the tails of the distribution of the trait. One approach is therefore to select as "affected" individuals whose values exceed some threshold or have been diagnosed with disease on the basis of high values of the trait. An example of this approach is to study individuals diagnosed with hypertension compared with population-based controls as an alternative to studying BP as a quantitative trait. Genetic risk factors identified in this way may or may not have an important effect on the overall distribution of the quantitative trait.

For a relatively common trait like hypertension, population-based controls are likely to include a substantial proportion of individuals who themselves have hypertension, thus reducing the power of the above strategy. In the WTCCC study of hypertensive cases versus population-based controls (7), this was recognised as a potential contributory factor to the lack of significant findings at the genome-wide level. It was suggested that in future studies controls could be screened to exclude those with high BP. In a similar vein, methods have been proposed and investigated based on selecting individuals with extreme values at both ends of the distribution, both in linkage (where Risch and Zhang (31) suggested the selection of extremely discordant sibling pairs) and association studies (31, 32).

4. GWA Studies

In the last few years, an important new development in genetic epidemiology has been the advent of GWA studies, since the whole genome can now feasibly be searched for common genetic variants associated with increased disease susceptibility. These studies have resulted in the identification of many novel disease- or trait-associated loci. One notable success of GWA studies is that, because of the recognition of the need for large sample size and the practice of adopting stringent standards of statistical significance, results reported in the literature have a good record of replication in independent data sets. However, although the associations are robust, usually the mechanism giving rise to the association is still not understood, and indeed some of the disease-associated variants identified do not lie within known genes.

A catalogue of published GWA studies is maintained by the US National Institutes of Health National Human Genome Research Institute (34). The number of published studies is increasing rapidly; at the time of writing, seven papers are listed in the catalogue under disease/trait CAD, "coronary disease" or MI (including MI [early onset]) (6–9, 35–37). Many more studies are listed under CAD-related traits, such as HDL and LDL cholesterol (20–26).

The only locus showing genome-wide evidence of association in the four earliest GWA studies (6–9) is the 9p21.3 chromosomal region. This association has since been replicated many times, for example in a meta-analysis involving a total of 4,655 MI cases and 5,177 controls (38); the per-allele odds ratio for the lead SNP rs1333049 was estimated in this study as 1.29 (95% confidence interval 1.22–1.37) with an additive mode of inheritance. This represents a relatively strong and highly replicable association with CAD, but the underlying mechanism is not understood. The region includes the cyclin-dependent kinase inhibitors *CDKN2A* and *CDKN2B*, but the associated SNPs are some way from these genes.

Several other loci have since been identified through analysing in independent data sets the variants showing association at a level slightly below genome-wide significance. Associations with SNPs in regions 1p13, 1q41, and 10q11 have been confirmed in numerous studies, as have SNPs in 19p13 and 1p32 near the LDL receptor and *PCSK9* genes, respectively, which are understood to affect risk of MI through their effect on LDL cholesterol levels. The MI Genetics Consortium identified three novel loci, giving nine regions in total (including the six mentioned above) that show strong evidence of association (37). Based on nine SNPs from these regions, they constructed a genotypic risk score and estimated that individuals in the top quintile according to this score had a greater than twofold-increased risk of MI compared with those in the bottom quintile. (Such risk estimates tend to be upwardly biased unless estimated using data completely independent of the discovery data set.) However, they also estimate that these loci explain only ~3% of variance in risk of early-onset MI.

Hence, despite the success of GWA studies, it seems likely that only a small proportion of the genetic basis of CAD has been uncovered. GWA studies are only powered to detect common variants, and it should be no surprise that the majority of variants discovered by this approach have had relatively large minor allele frequency combined with a modest effect on risk (39). Progressing further requires novel approaches to detect other possible genetic susceptibility factors, including rare variants, copy number variation, and interactions between genes or between genes and environment.

One study has examined copy number variation across the genome in relation to CAD (37) and found no evidence of an effect, although this is an area still in its infancy and further work will be needed. Almost all GWA studies to date have restricted the genome-wide analysis to separate analyses of each SNP, often then carrying out haplotype or multi-locus analyses in regions showing evidence of association. Recently, two groups carried out genome-wide haplotype analysis of CAD (36, 40) using different analytical approaches; the first group found a novel haplotype on 6q26-q27 to be associated with CAD, with some evidence of replication. Haplotypic effects which are not uncovered by single-SNP analysis could arise from interaction between SNPs or because the haplotype tags another (possibly rare or copy number) variant, suggesting that this approach may be useful in detecting more complex patterns of susceptibility. Research is being conducted into further novel methods of multi-locus analysis of genome-wide data (see (41) for a review), which may be successful in identifying susceptibility not attributable to individual common SNPs.

Where there are strong known environmental risk factors, stratifying by exposure may help to identify disease genes, despite the inherent multiple testing penalty. For example, conducting a GWA study based on non-smoking CAD cases may increase power by removing subjects, the origins of whose disease lies primarily in environmental/lifestyle factors. Conversely, searching for genetic risk factors among exposed cases may be valuable in identifying genes that increase the risk posed by exposure (gene–environment interaction). Now that sufficiently large studies of CAD have been established, it is of interest to stratify cases by smoking history in the search for disease genes.

5. Summary

This short account of some of the history of and current issues in CAD genetic epidemiology research illustrates issues common to the study of many complex diseases. Factors considered here are sometimes more and sometimes less important for other diseases; for example, difficulties arising from differences in phenotype are probably more acute in psychiatric illnesses and less acute in cancers, but always need to be considered when comparing or combining studies. The identification of highly heritable disease-related quantitative traits opens up avenues of research that, although not unique to CAD, distinguish it from many other diseases.

After many years of searching for genetic risk factors using candidate gene association studies and linkage analysis, the recent advent of GWA studies has opened a new era of research in all common complex diseases. Although there have been many successes, it is clear that findings to date only explain a small proportion of genetic risk. In the near future, it is anticipated that further progress will be made, for example, by the systematic study of copy number variation, the analysis of sequence data and the investigate of gene–gene and gene–environment interaction.

References

1. Arnett, D.K., Baird, A.E., Barkley, R.A. et al. (2007). Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: a scientific statement from the American Heart Association Council on epidemiology and prevention, the Stroke Council, and the Functional Genomics and Translational Biology Interdisciplinary Working Group. *Circulation* **115**, 2878–2901.
2. Allender, S., Peto, V., Scarborough, P., Kaur, A. and Rayner, M. (2008). Coronary heart disease statistics, Chapter 1. BHF: London. (Available from http://www.heartstats.org).
3. World Health Organization (2007). Cardiovascular diseases, Fact sheet No. 317. (Available from http://www.who.int).
4. The BHF Family Heart Study Research Group (2005). A genomewide linkage study of 1,933 families affected by premature coronary artery disease: the British Heart Foundation (BHF) Family Heart Study. *American Journal of Human Genetics* 77, 1011–1020.
5. Broeckel, U., Hengstenberg, C., Mayer, B. et al. (2002). A comprehensive linkage analysis for myocardial infarction and its related risk factors. *Nature Genetics* **30**, 210–214.
6. Samani, N.J., Erdmann, J., Hall, A.S. et al. (2007). Genomewide association analysis of coronary artery disease. *The New England Journal of Medicine* **357**, 443–453.
7. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.
8. McPherson, R., Pertsemlidis, A., Kavaslar, N. et al. (2007). A common allele on chromosome

9 associated with coronary heart disease. *Science* **316**, 1488–1491.

9. Helgadottir, A., Thorleifsson, G., Manolescu, A. et al. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491–1493.
10. Doll, R., Peto, R., Boreham, J. and Sutherland, J. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* **328**, 1519–1527.
11. Leppert, M., Hasstedt, S.J., Holm, T. et al. (1986). A DNA probe for the LDL receptor gene is tightly linked to hypercholesterolemia in a pedigree with early coronary disease. *American Journal of Human Genetics* **39**, 300–306.
12. Zdravkovic, S., Wienke, A., Pedersen, N.L. et al. (2002). Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *Journal of Internal Medicine* **252**, 247–254.
13. Guo, S.-W. (1998). Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. *American Journal of Human Genetics* **63**, 252–258.
14. Murabito, J.M., Pencina, M.J., Nam B.-H. et al. (2005). Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. *JAMA* **294**, 3117–3123.
15. Farrall, M., Green, F.R., Peden, J.F. et al. (2006). Genome-wide mapping of susceptibility to coronary artery disease identifies a novel replicated locus on chromosome 17. *PLoS Genetics* **2**, e72.
16. Morgan, T.M., Krumholtz, H.M., Lifton, R.P. and Spertus, J.A. (2007). Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *JAMA* **297**, 1551–1561.
17. Lohmueller, K.E., Pearce, C.L., Pike, M. et al. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* **33**, 177–182.
18. Freeman, M.S., Mansfield, M.W., Barrett, J.H. and Grant, P.J. (2002). Heritability of features of the insulin resistance syndrome in a community-based study of healthy families. *Diabetic Medicine* **19**, 994–999.
19. Austin, M.A., King, M.-C., Bawol, R.D. et al. (1987). Risk factors for coronary heart disease in adult female twins. *American Journal of Epidemiology* **125**, 308–318.
20. Wang, Y., O'Connell, J.R., McArdle, P.F. et al. (2009). Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proceedings of the National Academy of Sciences U S A* **106**, 226–231.
21. Willer, C.J., Sanna, S., Jackson, A.U. et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics* **40**, 161–169.
22. Wallace, C., Newhouse, S.J., Braund, P. et al. (2008). Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *American Journal of Human Genetics* **82**, 139–149.
23. Sandhu, M.S., Waterworth, D.M., Debenham, S.L. et al. (2008). LDL-cholesterol concentrations: a genome-wide association study. *Lancet* **371**, 483–491.
24. Sabatti, C., Service, S.K., Hartikainen, A.L. et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* **41**, 35–46.
25. Kathiresan, S., Willer, C.J., Peloso, G.M. et al. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics* **41**, 56–65.
26. Aulchenko, Y.S., Ripatti, S., Lindqvist, I. et al. (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature Genetics* **41**, 47–55.
27. Manolio, T. (2009). Cohort studies and the genetics of complex disease. *Nature Genetics* **41**, 5–6.
28. Cupples, L.A., Arruda, H.T., Benjamin, E.J. et al. (2007). The Framingham Heart Study 100 K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Medical Genetics* **8**(Suppl 1), S1.
29. Levy, D., Larson, M.G., Benjamin, E.J. et al. (2007). Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Medical Genetics* **8**(Suppl 1), S3.
30. Tobin, M.D., Timpson, N.J., Wain, L.V. et al. (2009). Common variation in the *WNK1* gene and blood pressure in childhood: the Avon Longitudinal Study of Parents and Children. *Hypertension* **52**, 974–979.
31. Risch, N. and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584–1589.
32. Page, G.P. and Amos, C.I. (1999). Comparison of linkage-disequilibrium methods for localization of genes influencing quantitative traits in humans. *American Journal of Human Genetics* **64**, 1194–1205.
33. Abecasis, G.R., Cookson, W.O.C. and Cardon, L.R. (2001). The power to detect linkage disequilibrium with quantitative traits

in selected samples. *American Journal of Human Genetics* **68**, 1463–1474.

34. Hindorff, L.A., Junkins, H.A., Mehta, J.P. and Manolio, T.A. A catalog of published genome-wide association studies. (Available from http://www.genome.gov/26525384). Accessed [16/04/2009].
35. Erdmann, J., Großhennig, A., Braund, P.S. et al. (2009). New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nature Genetics* **41**, 280–282.
36. Trégouët, D.-E., König, I.R., Erdmann, J. et al. (2009). Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nature Genetics* **41**, 283–285.
37. Myocardial Infarction Genetics Consortium (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics* **41**, 334–341.
38. Schunkert, H., Götz, A., Braund, P. et al. (2008). Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. *Circulation* **117**, 1675–1684.
39. Iles, M.M. (2008). What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genetics* **4**(2), e33.
40. Browning, B.L. and Browning, S.R. (2008). Haplotypic analysis of Wellcome Trust Case Control Consortium data. *Human Genetics* **123**, 273–280.
41. Musani, S.K., Shriner, D., Liu, N. et al. (2007). Detection of gene × gene interactions in genome-wide association studies of human population data. *Human Heredity* **63**, 67–84.

Chapter 17

The Genetic Epidemiology of Obesity: A Case Study

Laura M. Johnson

Abstract

Obesity (OMIM #601665) is a disease where excessive stores of body fat impact negatively on health. The first law of thermodynamics dictates that energy cannot be created or destroyed so if energy is taken into the body, but not transformed to ATP for metabolic work or dissipated as heat, it will be stored as fat. Therefore, the ultimate cause of obesity is a long-term positive energy imbalance [energy intake (EI) exceeds energy expenditure (EE)]. Despite this simple explanation, there is no single reason why EI may exceed EE meaning that the proximate causes of obesity are multi-factorial in origin involving a complex interplay of genetic, behavioural, and environmental influences on metabolism, diet, and activity.

Key words: Obesity, BMI, Fat mass, Lean mass, Appetite, Energy intake, Energy expenditure

1. The Obesity Phenotype

The human body is composed of four main components, namely, protein (in muscle), mineral (in bone), water (in and around cells), and fat mass. The best estimate of total body fat mass can be made by combining measurements of total body weight, density, protein, mineral, and water in a five component model. However, assessing each of these components separately involves complex and expensive analytical techniques, which limits their use in large samples (2).

Simpler, two component models of body composition, where body weight is partitioned into fat or fat-free mass, can be used instead. In this model, an assessment of just one component of fat-free mass is used along side assumptions about the density of the other components to estimate total fat-free mass. Fat-free mass is removed from body weight and the remaining mass is assumed to be fat. There are currently no data on which to base a

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6_17, © Springer Science+Business Media, LLC 2011

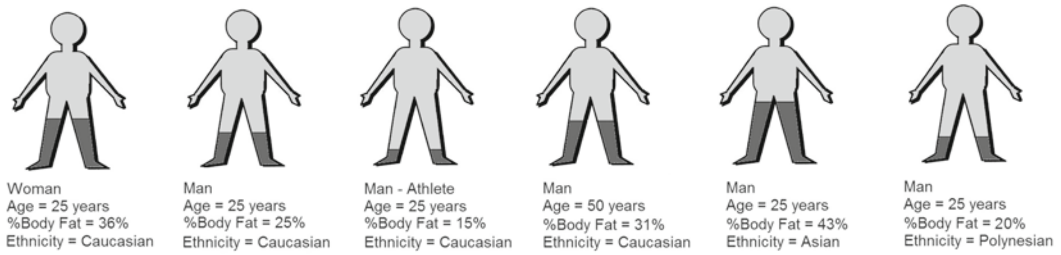


Fig. 1. The average proportion of body weight that is fat mass when BMI = 30 kg/m^2 varies depending on gender, athleticism, age, and ethnicity. *Light grey* = % lean mass; *Dark grey* = % fat mass. Based on data from Prentice and Jebb (8).

cut-off, in terms of body fat, which represents the point that body fat stores become excessive and begin to have an impact on health. In reality, however, any cut-off would be arbitrary as evidence suggests a linear association between fat mass and health outcomes (3, 4). Therefore, any search for the genes for obesity should be done using a continuous measure of body fat mass as a quantitative phenotype.

Traditionally the body mass index (BMI = body weight (kg)/height (m)^2), an indicator of excess body weight relative to height, has been used to define obesity in humans because it is simple and cheap to measure. According to the WHO criteria, overweight in adults is defined as a $25 < \text{BMI} < 30$ kg/m^2 and obesity is defined as a BMI > 30 kg/m^2 (1).

The problem with BMI as a surrogate measure of body fat is that excess body weight could indicate excessive fat mass but could equally suggest high relative hydration or bone or muscle mass. The assumption that a high BMI indicates a large fat mass is not completely flawed, it has been shown that BMI correlates highly with the measures of fat mass (r = 0.5–0.9) (5). However, it also correlates with fat-free mass as well (6, 7). This lack of specificity for indicating the composition of excess weight means that a BMI of 30 kg/m^2 can relate to different levels of actual fatness depending on gender, athleticism, age, and ethnicity (Fig. 1) (8).

If BMI reflects all components of excess body weight, then discovering the genes that effect BMI may lead to the discovery of genes for obesity but genes for body build, proportion, and components of lean mass may be discovered as well. Given this, the use of BMI as an obesity phenotype can be misleading and can add error to estimates of genetic effect. This may help to explain some of the inconsistency that characterises the literature on the genetic epidemiology of obesity.

1.1. Evidence for the Heritability of Obesity

The vast majority of estimates of the heritability of obesity have been made using measurements of BMI to represent fatness. Twin studies have reported that between 50 and 70% of the variation in BMI can be explained by an additive genetic component (9–12). Family studies suggest a more modest effect of genes with heritability estimates in the range of 30–45% (13, 14). When an actual measure of fat mass is used, the proportion of variation attributed to genetic effects has been estimated to be in the range of 50–80% (15–18).

Studies using a twin or family design have specifically modelled genetic variation of different body weight components, i.e. fat, muscle, or bone, and overall BMI to establish the extent to which the same genes (pleiotropy) underlie variation in all components of body composition (14, 16, 18). Separate components of body weight (fat mass, lean mass, and bone density) were estimated using skin-fold thickness (14), bio-impedance analysis (16), or dual-energy X-ray absorptiometry (18). In the classic twin model of heritability, variation in a phenotype, e.g. fat mass is partitioned into additive genetic (A), dominant genetic (D), shared environment (C), and non-shared environmental (E) components based on differences in the proportion of genes shared by family members or monozygotic and dizygotic twin pairs. Whether any genetic effects are shared by two phenotypes can be established by assessing the correlation between the genetic component estimates using a multi-variate genetic model based on structural equation modelling, e.g. the Cholesky factor model. Evidence from these studies suggests that there is a little shared genetic variation between the individual components of body weight. For example, the correlation between the genetic components underlying fat and lean mass was just $r=0.16$, equivalent to an estimated 2% shared heritability (18). In contrast, environmental factors explained a large proportion of shared variation between fat and lean mass ($r=0.51$). This adds further support to the argument that BMI is not the best phenotype for identifying obesity genes as it may be identifying a range of genes that affect body weight that are not specific to variation in fat mass.

2. Mapping Obesity Genes

2.1. Genome-Wide Linkage Scans for Genes for BMI

The last update of the Human Obesity Gene map reported that 317 putative loci linked to obesity have been identified on every human chromosome except Y (19). However, only 15 loci have been replicated in more than three independent studies. A recent meta-analysis of genome-wide linkage scans for BMI and obesity with data from 37 independent samples found no regions with consistently significant evidence of linkage (20). This stark lack of evidence, despite sufficient power to detect an effect, suggests that there is substantial locus heterogeneity underlying variation in BMI. It might be expected that many different genes underlie fatness as fat stores can be altered via multiple pathways, e.g. appetite, energy metabolism, activity levels; but additional heterogeneity may be present in BMI as it characterises all components of body weight which themselves are influenced by an even more diverse set of genes. Segregation analyses suggest that obesity is most likely to be polygenic; the result of interactions between multiple genetic and environmental factors (21). The lack of significance of any one region in the meta-analysis of linkage scans supports this.

2.2. Candidate Genes for Human Obesity

The complex physiology underlying the regulation of energy balance has provided a large number of candidate genes for obesity. Single mutations in genes coding for neurotransmitters, hormones, and their associated receptors in the appetite control pathways of the hypothalamus have been identified as the cause of most cases of extreme or early onset obesity in humans (22).

In most cases, evidence for an effect of these genes (with rare single mutations) on the expression of obesity in the general population is limited. To date association studies of candidate genes for obesity have often experienced problems with replicating observed effects. This may be explained by a large number of false-positive findings, a result of not using an appropriately conservative *p*-value to define a significant effect. Another problem is that replication studies may not include a large enough sample size to detect an effect that is likely to be small given the evidence for a polygenic model of obesity.

One exception is the gene coding for the Melanocortin 4 receptor (MC4R), which has been documented as accounting for 6% of early onset severe obesity (23). Multiple variants in the MC4R gene have been characterised and appear to affect obesity in the general population in opposing ways. For example, the most common mutation of the MC4R gene, the V103I variant, has a minor allele frequency of ~4% in western populations. A protective effect of this allele has been identified in three meta-analyses, which showed that carriers of the minor allele have a reduced risk of being obese (24–26). The largest meta-analysis to date ($n=29,563$) estimated that carriers were 18% less likely to be obese than non-carriers (26).

In contrast, a common variant (rs17782313) near the MC4R gene with a negative effect on BMI was identified by a meta-analysis of genome-wide association data from 16,876 adults and confirmation analyses in 60,352 adults from cohorts of European descent (27). The minor allele frequency was ~24% and the overall per allele effect was estimated to be equivalent to 0.22 kg/m^2 or a 12% increase in the risk of being obese. Interestingly, this variant was highly associated with increased height in adults and modest increases in muscle and bone mass in children as well as increases in fat mass, suggesting that it may not be an obesity gene so much as a gene that enhances body size in general.

2.3. Genome-Wide Association Studies

The first common variant for increased fat mass was recently identified and replicated in 13 cohorts with a total sample size of 38,759 (28). A variant (rs9939609, A allele) of the FTO gene (initially identified from a genome-wide association study of patients with type 2 diabetes) was associated with higher BMI in both adults and children as young as 7 years, which equated to a ~3 kg difference in body weight between homozygous high- and low-risk allele carriers. Further analysis of fat and non-fat mass separately found that the association of the high-risk A allele with weight was mainly attributable to changes in fat mass with a 14% difference across the three genotype groups compared with a 1% difference in non-fat mass across groups.

The effect of FTO has since been confirmed in eight independent cohorts of severely obese adults, children with extreme early onset obesity, as well as samples of healthy adults, children, twins, and newborns of European and American Caucasian descent (29–36). Further studies have failed to replicate the association in samples of African Americans, Chinese, Japanese, and Oceanic Islanders, suggesting that the effect of FTO may be specific to western Caucasians (37–40).

Evidence from a meta-analysis of genome-wide scans shows that the region containing the FTO gene 16q12 was among those with the most consistent evidence for linkage, albeit only nominally significant at the whole genome level (20). The identification of FTO as a high-risk allele for obesity represents a previously unidentified candidate gene and its functional relevance to obesity is currently under investigation. To date gene expression studies have shown that the FTO gene is highly expressed in both the appetite control regions of the hypothalamus (41, 42) and adipose tissue (43).

It has been hypothesised that functional mutation of the FTO gene may lead to a reduction in the sensitivity of the appetite control system, recent analyses of appetite in a sample of young twins genotyped for FTO supports this (30). Using questionnaire-based psychometric measures, the children's sensitivity to internal signals of fullness and enjoyment of food was characterised. Children homozygous for the high-risk A allele were shown to have reduced sensitivity to fullness and a higher enjoyment of food score. Mediation analysis indicated that this variation in appetite, in part, explained the association between FTO and BMI. Much more work is needed to fully characterise the effect of FTO variants on obesity in humans.

3. Interaction in the Aetiology of Obesity

The classic twin studies of Bouchard et al. in the early 1990s demonstrated that the response to disruptions of energy balance, by overfeeding or prescribed exercise, is modified by underlying genetic susceptibility. In these studies, the variation in weight change between monozygotic twin pairs was greater than the variation observed within twin pairs; therefore, when genotype is the same, weight change is more alike. Progress on understanding the details of the impact of interactions between genes and environmental factors on obesity has been slow. This is problematic for studying the genetics of obesity as interactions are often not accounted for in analyses, which can reduce the power to detect a major gene effect. The modest effect of genes for obesity observed to date (44, 45) may be partly explained by the presence of unaccounted-for interactions.

Five models of interaction have been proposed to represent the combined effects of genes and environments on complex disease (Fig. 2) (46). (1) The gene alters the expression of a risk factor; (2) the gene alters the effect of a risk factor; (3) the risk factor alters the effect of a gene; (4) only the combination of gene and the risk factor affect disease; and (5) the gene and the risk factor have independent effects on disease risk. Some examples of each of these models have already been identified for obesity.

3.1. Gene–Gene Interactions

The PPARγ gene is highly expressed in adipose tissue and is associated with the stimulation of adipocyte cell growth and differentiation. The β3-adrenergic receptor is also expressed in adipose tissue and influences fat metabolism and heat production. An interaction between the Pro12Ala and the Trp64Arg

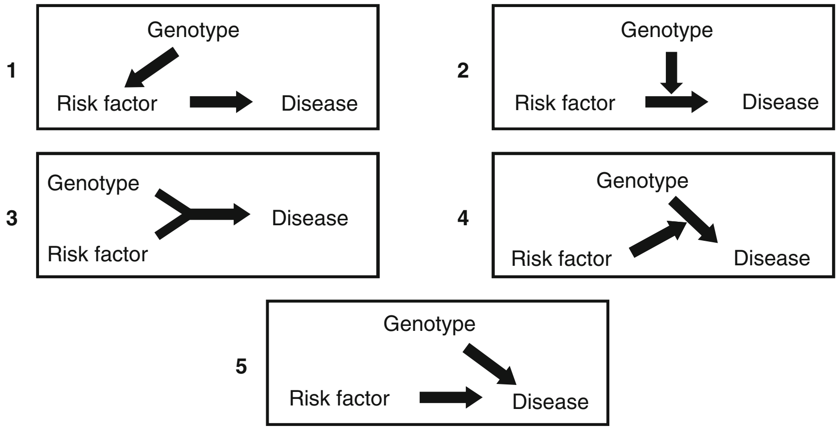


Fig. 2. Models of interaction between genes and other genetic or environmental risk factors in the causation of disease. Adapted from (46).

polymorphisms of the PPARγ and β3-AR genes has been observed in a family study of Mexican Americans and a case control study of obesity in Spanish children (47, 48). In the Spanish case control study, a BMI enhancing effect of the Arg variant of the β3-AR gene is only significant among carriers of the Ala variant of the PPARγ gene representing an example of interaction model 1.

An example of interaction model 5 is the additive independent effects of the FTO gene and the variant near the MC4R gene on BMI in adults and children (27). Adults homozygous for both high-risk variants (1% of sample) had a BMI of ~1.17 kg/m^2 larger than adults carrying no risk alleles (19% of sample).

3.2. Gene–Diet Interactions

A similar example of independent additive effects was observed for an analysis of the combined impact of the FTO gene and diet on later fat mass in early adolescence. Dietary energy density, the amount of energy per gram of food eaten, is an environmental risk factor for obesity that is believed to override physiological indicators of satiety (49). In this study, the effect of both the gene and the diet was significant in a multiple linear regression model so that each high-risk A allele was associated with 330 g more fat mass at age 13 years, and each 1 kJ/g difference in dietary energy density was associated with 150 g more fat mass.

The modifying effect that the Pro12Ala variant has on the impact of dietary fat intake on BMI is an example of the 2nd model of interaction (50–52). In one study, the odds of obesity were no different across quintiles of fat intake for carriers of the rare Ala allele, whereas a positive linear trend was identified in those homozygous for the Pro allele, which suggests that this variant creates resistance to diet-induced obesity (52).

3.3. Gene–Activity Interaction

Tentative evidence that low physical activity accentuates the effect of the FTO gene is an example of interaction model 3 (32). In this analysis of data from a middle-aged Danish population, the difference in BMI of the homozygous AA compared with the TT variant carriers was four times greater among those with low levels of reported activity compared with high levels of physical activity.

In contrast, genes have also been shown to attenuate the impact of activity on obesity risk. The Gln27Glu variant β2-AR gene, involved in fat metabolism in adipocytes, interacts with physical activity and sedentary behaviour (50, 53). In a Spanish case–control study of obesity, adults with the Gln27 allele were protected from obesity regardless of activity, whereas the odds of obesity in adults with the 27Glu allele were linearly related to their physical activity. A similar study in children found that the detrimental effect of TV watching was attenuated in carriers of 27Glu polymorphism, who were at increased risk of obesity compared with carriers of the wild-type variant, which suggests that TV watching has little effect on obesity in children with this genetic predisposition to obesity. More studies are required to replicate these findings and confirm the mechanism of action.

3.4. Methodological Considerations

The power to detect a significant interaction effect is dependent on the hypothesised mode of interaction as well as the precision with which the environment is measured. An important consideration is how to define a purely environmental effect. For example, looking at activity levels, the direct measurement of energy expenditure may in itself represent a genetically determined trait, whereas the type of activity undertaken may be more environmentally controlled by factors such as accessibility or availability of equipment. Alternatively, when assessing interactions with diet, environmental factors such as price and availability may govern food choice but taste preferences and appetite control may influence the amount that is eaten.

In order to detect interaction effects between multiple genetic and environmental factors that themselves have modest effects on disease, large samples are required. A suggestion for maximising power in a limited sample size when a risk factor is easy to measure is to only genotype those people in the extremes of the distribution (54). This approach assumes that there is no gene–environment correlation; however, most methods to assess interactions also fail to address this issue. As with all genetic associations, interactions should be replicated, and establishing large consortia of cohort studies, as has been seen in recent genome-wide association analyses, may help with this (55).

4. Summary

The collaboration of multiple research groups in combining data from a large number of population-based cohort studies has proven to be a successful method for identifying the small genetic effects associated with variation in the obesity phenotype. In addition, the inclusion of a sub-sample with data on body composition has allowed the genetic effect to be allocated to the appropriate component of body weight, i.e. fat mass or lean mass or general body size.

The clinical relevance of the effect of the FTO or MC4R genetic variants on the body weight of individuals is small. But when additive effects of multiple genes and environmental risk factors are combined, this may translate into a substantial effect on health. More importantly, small effects on many individuals can have a significant impact on public health.

References

1. World Health Organization (WHO) (2008) Obesity. In: World Health Organization (WHO), editor. Health topics. Geneva: World Health Organisation (WHO).
2. Ellis KJ (2000) Human body composition: in vivo methods. Physiol Rev 80: 649–680.
3. Heitmann BL, Erikson H, Ellsinger BM, Mikkelsen KL, Larsson B (2000) Mortality associated with body fat, fat-free mass and body mass index among 60-year-old Swedish men-a 22-year follow-up. The study of men born in 1913. Int J Obes Relat Metab Disord 24: 33–37.
4. Allison DB, Zhu SK, Plankey M, Faith MS, Heo M (2002) Differential associations of body mass index and adiposity with all-cause mortality among men in the first and second National Health and Nutrition Examination Surveys (NHANES I and NHANES II) follow-up studies. Int J Obes Relat Metab Disord 26: 410–416.
5. Deurenberg P, Weststrate JA, Seidell JC (1991) Body mass index as a measure of body fatness: age- and sex-specific prediction formulas. Br J Nutr 65: 105–114.
6. Kyle UG, Schutz Y, Dupertuis YM, Pichard C (2003) Body composition interpretation. Contributions of the fat-free mass index and the body fat mass index. Nutrition 19: 597–604.
7. Demerath EW, Schubert CM, Maynard LM, Sun SS, Chumlea WC, et al. (2006) Do changes in body mass index percentile reflect changes in body composition in children? Data from the Fels Longitudinal Study. Pediatrics 117: e487–e495.
8. Prentice AM, Jebb SA (2001) Beyond body mass index. Obes Rev 2: 141–147.
9. Allison DB, Kaprio J, Korkeila M, Koskenvuo M, Neale MC, et al. (1996) The heritability of body mass index among an international sample of monozygotic twins reared apart. Int J Obes Relat Metab Disord 20: 501–506.
10. Korkeila M, Kaprio J, Rissanen A, Koskenvuo M (1991) Effects of gender and age on the heritability of body mass index. Int J Obes 15: 647–654.
11. Price RA, Gottesman II (1991) Body fat in identical twins reared apart: roles for genes and environment. Behav Genet 21: 1–7.
12. Slattery ML, Bishop DT, French TK, Hunt SC, Meikle AW, et al. (1988) Lifestyle and blood pressure levels in male twins in Utah. Genet Epidemiol 5: 277–287.
13. Vogler GP, Sorensen TI, Stunkard AJ, Srinivasan MR, Rao DC (1995) Influences of genes and shared family environment on adult body mass index assessed in an adoption study by a comprehensive path model. Int J Obes Relat Metab Disord 19: 40–45.
14. Rice T, Bouchard C, Perusse L, Rao DC (1995) Familial clustering of multiple measures of adiposity and fat distribution in the Quebec Family Study: a trivariate analysis of percent body fat, body mass index, and trunk-to-extremity skinfold ratio. Int J Obes Relat Metab Disord 19: 902–908.
15. Rice T, Daw EW, Gagnon J, Bouchard C, Leon AS, et al. (1997) Familial resemblance for body composition measures: the HERITAGE Family Study. Obes Res 5: 557–562.
16. Faith MS, Pietrobelli A, Nunez C, Heo M, Heymsfield SB, et al. (1999) Evidence for independent genetic influences on fat mass and body mass index in a pediatric twin sample. Pediatrics 104: 61–67.
17. Rice T, Despres JP, Daw EW, Gagnon J, Borecki IB, et al. (1997) Familial resemblance for abdominal visceral fat: the HERITAGE family study. Int J Obes Relat Metab Disord 21: 1024–1031.
18. Nguyen TV, Howard GM, Kelly PJ, Eisman JA (1998) Bone mass, lean mass, and fat mass: same genes or same environments? Am J Epidemiol 147: 3–16.
19. Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, et al. (2006) The human obesity gene map: the 2005 update. Obesity (Silver Spring) 14: 529–644.
20. Saunders CL, Chiodini BD, Sham P, Lewis CM, Abkevich V, et al. (2007) Meta-analysis of genome-wide linkage studies in BMI and obesity. Obesity 15: 2263–2275.
21. Bouchard C, Perusse L, Rice T, Rao DC (1998) The Genetics of Human Obesity. In: Bray GA, Bouchard C, James WPT, editors. Handbook of obesity. New York: Marcel Dekker. pp. 157–190.
22. Farooqi IS, O'Rahilly S (2005) Monogenic obesity in humans. Annu Rev Med 56: 443–458.
23. Farooqi IS, Keogh JM, Yeo GS, Lank EJ, Cheetham T, et al. (2003) Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. N Engl J Med 348: 1085–1095.
24. Geller F, Reichwald K, Dempfle A, Illig T, Vollmert C, et al. (2004) Melanocortin-4 receptor gene variant I103 is negatively associated with obesity. Am J Hum Genet 74: 572–581.

25. Heid IM, Vollmert C, Hinney A, Doring A, Geller F, et al. (2005) Association of the 103I MC4R allele with decreased body mass in 7937 participants of two population based surveys. J Med Genet 42: e21.

26. Young EH, Wareham NJ, Farooqi S, Hinney A, Hebebrand J, et al. (2007) The V103I polymorphism of the MC4R gene and obesity: population based studies and meta-analysis of 29 563 individuals. Int J Obes (Lond) 31: 1437–1441.

27. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, et al. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40: 768–775.

28. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316(5826): 889–894. DOI: 10.1126/science.1141634.

29. Peeters A, Beckers S, Verrijken A, Roevens P, Peeters P, et al. (2008) Variants in the FTO gene are associated with common obesity in the Belgian population. Mol Genet Metab 93: 481–484.

30. Wardle J, Carnell S, Haworth CM, Farooqi IS, O'Rahilly S, et al. (2008) Obesity-associated genetic variation in FTO is associated with diminished satiety. J Clin Endocrinol Metab 93(9): 3640–3643.

31. Hinney A, Nguyen TT, Scherag A, Friedel S, Bronner G, et al. (2007) Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. PLoS One 2: e1361.

32. Andreasen CH, Stender-Petersen KL, Mogensen MS, Torekov SS, Wegner L, et al. (2008) Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation. Diabetes 57: 95–101.

33. Lopez-Bermejo A, Petry CJ, Diaz M, Sebastiani G, de Zegher F, et al. (2008) The association between the FTO gene and fat mass in humans develops by the postnatal age of two weeks. J Clin Endocrinol Metab 93: 1501–1505.

34. Speakman JR, Rance KA, Johnstone AM (2008) Polymorphisms of the FTO gene are associated with variation in energy intake, but not energy expenditure. Obesity (Silver Spring) 16(8): 1961–1965.

35. Price RA, Li WD, Zhao H (2008) FTO gene SNPs associated with extreme obesity in cases, controls and extremely discordant sister pairs. BMC Med Genet 9: 4.

36. Hunt SC, Stone S, Xin Y, Scherer CA, Magness CL, et al. (2008) Association of the FTO gene with BMI. Obesity (Silver Spring) 16: 902–904.

37. Li H, Wu Y, Loos RJ, Hu FB, Liu Y, et al. (2008) Variants in the fat mass- and obesity-associated (FTO) gene are not associated with obesity in a Chinese Han population. Diabetes 57: 264–268.

38. Scuteri A, Sanna S, Chen WM, Uda M, Albai G, et al. (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet 3: e115.

39. Horikoshi M, Hara K, Ito C, Shojima N, Nagai R, et al. (2007) Variations in the HHEX gene are associated with increased risk of type 2 diabetes in the Japanese population. Diabetologia 50: 2461–2466.

40. Ohashi J, Naka I, Kimura R, Natsuhara K, Yamauchi T, et al. (2007) FTO polymorphisms in oceanic populations. J Hum Genet 52: 1031–1035.

41. Gerken T, Girard CA, Tung YC, Webby CJ, Saudek V, et al. (2007) The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. Science 318: 1469–1472.

42. Fredriksson R, Hagglund M, Olszewski PK, Stephansson O, Jacobsson JA, et al. (2008) The obesity gene, FTO, is of ancient origin, up-regulated during food deprivation and expressed in neurons of feeding-related nuclei of the brain. Endocrinology 149: 2062–2071.

43. Wahlen K, Sjolin E, Hoffstedt J (2008) The common rs9939609 gene variant of the fat mass- and obesity-associated gene FTO is related to fat cell lipolysis. J Lipid Res 49: 607–611.

44. Kurokawa N, Young EH, Oka Y, Satoh H, Wareham NJ, et al. (2008) The ADRB3 Trp64Arg variant and BMI: a meta-analysis of 44 833 individuals. Int J Obes (Lond) 32(8): 1240–1249.

45. Masud S, Ye S (2003) Effect of the peroxisome proliferator activated receptor-gamma gene Pro12Ala variant on body mass index: a meta-analysis. J Med Genet 40: 773–780.

46. Ottman R (1996) Gene–environment interaction: definitions and study designs. Prev Med 25: 764–770.

47. Ochoa MC, Marti A, Azcona C, Chueca M, Oyarzabal M, et al. (2004) Gene-gene interaction between PPAR gamma 2 and ADR beta 3 increases obesity risk in children and adolescents. Int J Obes Relat Metab Disord 28(Suppl 3): S37–S41.

48. Hsueh W-C, Cole SA, Shuldiner AR, Beamer BA, Blangero J, et al. (2001) Interactions between variants in the β3-adrenergic receptor and peroxisome proliferator-activated receptor-γ2 genes and obesity. Diab Care 24(4): 672–677.

49. Johnson L, van Jaarsveld CHM, Emmett PM, Rogers IM, Hattersley AT, et al. (2009) Dietary energy density affects fat mass in early adolescence and is not modified by FTO variants. PLoS One 4: e4594.

50. Marti A, Martinez-Gonzalez MA, Martinez JA (2008) Interaction between genes and lifestyle factors on obesity. Proc Nutr Soc 67: 1–8.

51. Robitaille J, Despres JP, Perusse L, Vohl MC (2003) The PPAR-gamma P12A polymorphism modulates the relationship between dietary fat intake and components of the metabolic syndrome: results from the Quebec Family Study. Clin Genet 63: 109–116.

52. Memisoglu A, Hu FB, Hankinson SE, Manson JE, De Vivo I, et al. (2003) Interaction between a peroxisome proliferator-activated receptor γ gene polymorphism and dietary fat intake in relation to body mass. Hum Mol Genet 12: 2923–2929. DOI: 10.1093/hmg/ddg318.

53. Meirhaeghe A, Helbecque N, Cottel D, Amouyel P (1999) β2-Adrenoceptor gene polymorphism, body weight, and physical activity. Lancet 353: 896.

54. Boks MP, Schipper M, Schubart CD, Sommer IE, Kahn RS, et al. (2007) Investigating gene environment interaction in complex diseases: increasing power by selective sampling for environmental exposure. Int J Epidemiol 36: 1363–1369.

55. Hunter DJ (2005) Gene–environment interactions in human diseases. Nat Rev Genet 6: 287–298.

INDEX

M. Dawn Teare (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, vol. 713,
DOI 10.1007/978-1-60327-416-6, © Springer Science+Business Media, LLC 2011