Metadata of the chapter that will be visualized online

Chapter Title	Key Concepts in Genetic Epidemiology	
Copyright Year	2018	
Copyright Holder	Springer Science+Business Media, LLC, part of Springer Nature	
Corresponding Author	Family Name	**Panoutsopoulou**
	Particle	
	Given Name	**Kalliope**
	Suffix	
	Organization	Wellcome Trust Sanger Institute
	Address	Hinxton, Cambridgeshire, UK
	Email	kp6@sanger.ac.uk
Author	Family Name	**Wheeler**
	Particle	
	Given Name	**Eleanor**
	Suffix	
	Organization	Wellcome Trust Sanger Institute
	Address	Hinxton, Cambridgeshire, UK
Abstract	Genetic epidemiology is a discipline closely allied to traditional epidemiology that deals with the analysis of the familial distribution of traits. It emerged in the mid-1980s bringing together approaches and techniques developed in mathematical and quantitative genetics, medical and population genetics, statistics and epidemiology. The purpose of this chapter is to familiarize the reader with key concepts in genetic epidemiology as applied at present to unveil the familial and in particular genetic determinants of disease and the joint effects of genes and environmental exposures.	
Keywords (separated by '-')	Genetic epidemiology - Mendelian genetics - Genes - Recombination - Linkage disequilibrium - Population genetics - Kinship - Identity-by-descent - Identity-by-state - Hardy-Weinberg equilibrium - Heritability - Association - Odds ratio	

AU1

Chapter 2

Key Concepts in Genetic Epidemiology

Kalliope Panoutsopoulou and Eleanor Wheeler

Abstract

Genetic epidemiology is a discipline closely allied to traditional epidemiology that deals with the analysis of the familial distribution of traits. It emerged in the mid-1980s bringing together approaches and techniques developed in mathematical and quantitative genetics, medical and population genetics, statistics and epidemiology. The purpose of this chapter is to familiarize the reader with key concepts in genetic epidemiology as applied at present to unveil the familial and in particular genetic determinants of disease and the joint effects of genes and environmental exposures.

Key words Genetic epidemiology, Mendelian genetics, Genes, Recombination, Linkage disequilibrium, Population genetics, Kinship, Identity-by-descent, Identity-by-state, Hardy-Weinberg equilibrium, Heritability, Association, Odds ratio

1 Introduction to Genetic Epidemiology

Genetic epidemiology is the scientific discipline that aims to unravel the role of the genetic determinants in health and disease and their complex interplay with environmental factors. In the past, genetic epidemiology has been particularly successful in mapping genes with large effect sizes at the individual level, for example in monogenic disorders where familial recurrence follows the laws of mendelian inheritance. With the advent of more high-throughput genotyping technologies and the development of more sophisticated statistical genetics methodologies, the field of genetic epidemiology has recently focused its attention on dissecting the genetic architecture of common complex diseases. Unlike monogenic diseases, common complex diseases are caused by a large number of genes with small to modest effect sizes and their complex interplay with environmental factors. Large-scale genome-wide and whole-genome sequencing association studies (GWAS and WGS) have catalogued a large number of genetic variations that are implicated in complex traits and diseases. It is anticipated that subsequent translational efforts will transform the way medicine

Evangelos Evangelou (ed.), *Genetic Epidemiology: Methods and Protocols*, Methods in Molecular Biology, vol. 1793,
https://doi.org/10.1007/978-1-4939-7868-7_2,

will be practised in the near future. This chapter introduces the reader to key concepts in molecular genetics, mendelian genetics, population genetics, and the fusion of these disciplines with epidemiology that has led to what is known today as genetic epidemiology.

2 Molecular Genetics and Variation

Genetics is the study of genes and heredity, the process by which characteristics are passed on from one generation to another. The carrier molecule of an organism's genetic information is called deoxyribonucleic acid (DNA). In this section, we describe the central dogma of biology to explain the flow of genetic information from DNA sequence to protein product and introduce the consequences of DNA variation in health and disease.

2.1 From DNA to RNA to Proteins

DNA is a large molecule consisting of two single strands, and each strand is composed of smaller molecules called nucleotides (Fig. 1). The nucleotides are composed of a sugar residue (deoxyribose),

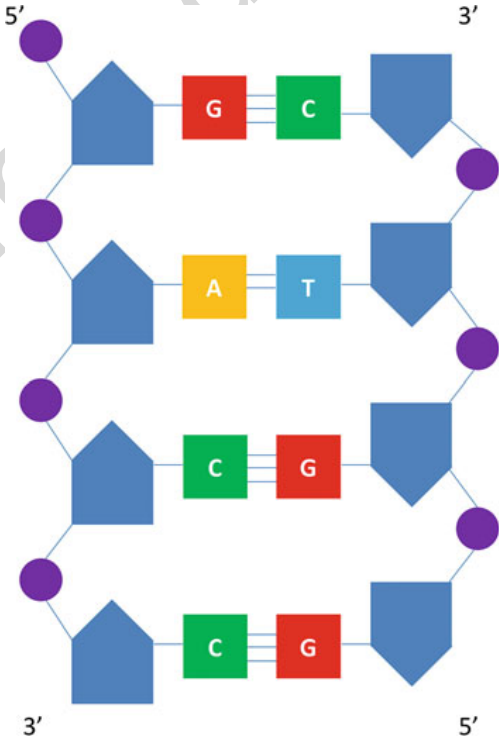

Fig. 1 Structure of a DNA molecule. Basic representation of an unwound DNA double helix segment depicting the phosphate group (purple circle), the sugar residue (blue pentagon), and the four different chemical bases (differentially colored squares). Complementary base pairing occurs between guanine (G) and cytosine (C) and between adenine (A) and thymine (T)

a phosphate group and a nitrogenous base which can be any of four types: adenine (A), cytosine (C), guanine (G), and thymine (T). The sugar residue and the phosphate group together form the nucleoside and alternating nucleosides form the DNA backbone. Covalent bonds bind bases to the nucleoside in one single strand. Weaker hydrogen bonds bind specifically A with T and G with C (also known as complementary bases) between the two single DNA strands resulting in the formation of the DNA double helix. Each single strand has different ends oriented in opposite directions termed five primed (5′) and three primed (3′) ends. The DNA sequence is essentially the order of the four bases across the genome and it is written down as letters for one strand only in the 5′ to 3′ direction, in this example GACC. This linear sequence of DNA is also known as its primary structure. The complementary strand in this case, written in the 3′ to 5′ direction, would be CTGG (Fig. 1). The length of the DNA is measured in base pairs (bp) so the DNA fragment in the example shown is 4 bp long. As we will describe below it is the order of these four chemical bases in the DNA that determines the proteins that are synthesized and carry out all the important functions in human organisms.

The process of protein synthesis can be summarized in two steps: transcription of a DNA sequence into ribonucleic acid (RNA) and translation of RNA into amino acids which form proteins. During the process of transcription the DNA double helix is unzipped into single strands. A single DNA strand acts as a template for the synthesis of a complementary strand of RNA in the 5′ to 3′ direction which is catalyzed by the RNA polymerase enzyme. The structure of RNA is similar to the single stranded DNA except that its backbone is composed of a sugar residue called ribose and the chemical base uracil (U) is present instead of T. RNA transcription that leads to proteins occurs in certain regions of the DNA which are transcribed to messenger RNA (mRNA). These regions are known as genes and typically contain alternating segments of sequence called exons, the protein coding sequences, separated by segments of noncoding DNA called introns. mRNA is further edited to make mature mRNA where introns are cut out and exons are spliced. Differential or alternative splicing of exons gives rise to different gene transcripts ensuring that multiple proteins can be coded by one gene.

The genetic information that is now contained in mRNA is translated into proteins according to the genetic code (Table 1). The genetic code defines how specific base triplets known as codons are combined to form amino acids, the building blocks of proteins. The combination of the four different bases (A, G, C, U) into triplets can make 4^3=64 different codons which encode 20 different amino acids. Because several amino acids can be encoded by more than one codon the code is said to be degenerate and codons that correspond to the same amino acid are called synonymous. Start

t.1 **Table 1**
The genetic code

t.2		U		C		A		G		
t.3	**U**	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	**U**
t.4		UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	**C**
t.5		UUA	Leu	UCA	Ser	UAA	'Stop'	UGA	'Stop'	**A**
t.6		UUG	Leu	UCG	Ser	UAG	'Stop'	UGG	Trp	**G**
t.7	**C**	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	**U**
t.8		CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	**C**
t.9		CUA	Leu	CCA	Pro	CAA	Gin	CGA	Arg	**A**
t.10		CUG	Leu	CCG	Pro	CAG	Gin	CGG	Arg	**G**
t.11	**A**	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	**U**
t.12		AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	**C**
t.13		AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	**A**
t.14		AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	**G**
t.15	**G**	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	**U**
t.16		GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	**C**
t.17		GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	**A**
t.18		GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	**G**

t.19 Table of codons showing the corresponding amino acid encoded by each base triplet
Ala Alanine, *Arg* Arginine, *Asp* Aspartate, *Asn* Asparagine, *Cys* Cysteine, *Gln* Glutamine, *Glu* Glutamate, *Gly* Glycine, *His* Histidine, *Ile* Isoleucine, *Leu* Leucine, *Lys* Lysine, *Met* Methionine, *Phe* Phenylalanine, *Pro* Proline, *Ser* Serine, *The* Threonine, *Trp* Tryptophan, *Tyr* Tyrosine, *Val* Valine

(methionine codon) and stop codons signal the initiation and termination of the RNA translation into long chains of amino acid residues (polypeptides) a process that occurs in the cell plasma, at organelles called ribosomes. This process involves two more classes of RNA molecules: ribosomal RNA (rRNA) molecules that form the core of a cell's ribosome and transfer RNA (tRNA) molecules that carry amino acids to the ribosomes during protein synthesis. Proteins consist of one or more chains of amino acid residues folded into a 3D structure that determines their function and activity. It is the changes in the DNA sequence, either inherited or spontaneously induced, that can result in alterations of the level or structure and function of proteins that can affect human health and disease.

2.2 Human Genome and Variation

Nuclear DNA (nDNA) is found in the nucleus of almost every human cell (except for red blood cells) tightly packed in structures called chromosomes. Mitochondrial DNA (mtDNA) which is found in the cell structures known as mitochondria is responsible for providing the energy that the cell needs to function. nDNA encodes for the majority of the genome in eukaryotes; in humans it is 3.3 billion bp long and contains approximately 20,000 genes [1].

nDNA is distributed in 22 pairs of autosomes and in one pair of sex chromosomes which is XY in males and XX in females. One of

the pair is derived from the mother and one from the father. All human cells contain two copies of each chromosome and are thus called diploid, except for gametes (sperm and ova) which are haploid. Because autosomal chromosome pairs contain the same genes at the same position they are called homologous chromosomes. However, because each chromosome from a homologous pair is derived from a different individual (mother or father) variations at certain DNA locations can be present. There are several classes of variation but the most frequent are single nucleotide polymorphisms (SNPs) which are variations in a single DNA base. Thus, at a given locus (region) in a homologous pair of chromosomes an individual can have either the same DNA base between the members of the pair (i.e., AA), or a different base, i.e., (AT). At that same position another individual may have TT. AA, AT, or TT denote the genotype of an individual at this particular site. Because of this variation, the site is said to be polymorphic and A and T are called alleles. The individual who carries AA at that locus is said to be homozygous for the A allele, AT heterozygous, and TT is homozygous for the T allele. The series of alleles along a single chromosome is called haplotype.

One of the two alleles will be present at a lower frequency in the population than the other allele; the less frequent is called the minor allele and the most frequent is called the major allele. A DNA variation is said to be rare when the minor allele frequency (MAF) is less than 0.01 meaning that the minor allele is observed in 10 or less individuals out of 1000. For rare variants the term single nucleotide variation (SNV) is used instead of SNP.

If $f(\mathrm{AA})$, $f(\mathrm{AB})$, and $f(\mathrm{BB})$ are the frequencies of the three genotypes at a bi-allelic locus, then the frequency p of the A-allele and the frequency q of the B-allele in the population are obtained by counting alleles.

$$p = f(\mathrm{AA}) + 1/2f(\mathrm{AB}) = \text{frequency of A}$$
$$q = f(\mathrm{BB}) + 1/2f(\mathrm{AT}) = \text{frequency of B}$$

Because p and q are the frequencies of the only two alleles present at that locus, they must sum to 1.

$$p + q = f(\mathrm{AA}) + f(\mathrm{BB}) + f(\mathrm{AB}) = 1$$
$$q = 1 - p \textit{ and } p = 1 - q$$

SNPs are the simplest form of DNA variation among individuals and are the focus of current research to unravel the genetic aetiology of common, complex diseases. There are several other forms of genetic variation such as microsatellites (typically nucleotide repeats that exist in variable numbers), insertions/deletions (one or several bases are duplicated/lost), duplications and translocations (usually large sequences that are cut from one site in the genome and inserted in another site). These are called structural variations and are covered in more detail elsewhere in this book (*see* Chapter 12).

2.3 The Impact of DNA Variation in Health and Disease

DNA sequence variations are the result of genetic mutation that may be introduced during DNA replication or due to DNA exposure to damaging agents. Hereditary mutations are passed on from parent to offspring. Mutations are essential for our evolution and our long-term survival. However, a very small percentage of all mutations can also lead to medical conditions of various severities.

Variants that fall in protein-coding genes are the best understood because it is easier to make predictions about the effect that these have on gene function, known as functional consequences. There is a wide range of databases that describe these such as the Ensembl [2] and UCSC [3] databases. For example, non-synonymous variants, i.e., those that cause amino acid changes may introduce a premature stop codon leading to a shortened transcript; small insertion/deletions (indels) can change the translational reading frame. These belong to the category of loss of function (LoF) variants that comprise highly deleterious variants responsible for severe diseases. Non-synonymous, missense variants where the length is preserved can sometimes, but not always, affect the structure or function of the protein. A very well-known example is sickle-cell anaemia, caused by a missense mutation, A to T, in the gene coding for the beta-globin chain constituent of hemoglobin. This mutation results in the substitution of glutamic acid to valine (GAG codon changes to GTG); the disease is manifested in homozygous individuals and is caused by aggregation and precipitation of hemoglobin. In heterozygous individuals (known as carriers) 50% of the hemoglobin is still produced so the symptoms are far less severe. Interestingly, the mutation has thought to have arisen because it provides protection to malaria.

The protein-coding part of the genome represents approximately 1% of the genome. Base variations outside gene regions are typically implicated in common complex diseases. The exact mechanisms by which changes in the DNA sequence outside genes and their complex interplay with environmental factors can cause disease are the subject of extensive research in the current era [4]. We will briefly introduce some terms in order to understand how variations at the DNA outside of protein-coding regions can affect tightly controlled dynamic processes that govern transcription and translation of the primary sequence to genes and proteins respectively.

Transcription and translation are complex processes regulated by many factors [5, 6]. Briefly, the initiation of transcription is controlled by promoters, which are DNA elements upstream of the gene where different forms of RNA polymerase and other associated transcription factors bind. Transcription factors are broadly divided into activators and repressors that bind to enhancers (noncoding DNA sequences 200–1000 bp long containing multiple activator and repressor binding sites) and can activate and/or repress a wide repertoire of target genes. Enhancers can

be found near the regulated gene (5′ upstream of the promoter or within the first intron of the gene they affect) or they can be distal, found in introns of neighboring genes or intergenic regions, i.e., between genes. The configuration of the genome called DNA looping brings together promoters, enhancers, activators, repressors, and other RNA processing factors to achieve the tight regulation at the gene expression level. The process of translation involves several components of the translational machinery and is also tightly regulated by several factors for example short oligonucleotides called microRNAs (miRNAs). Therefore, variants falling outside protein-coding regions that affect the tight regulation or alter the dynamics of these processes can increase susceptibility to a certain disease.

Transcriptional regulation also occurs at the level of chromatin structure by controlling the accessibility of the DNA to polymerase and other complexes. Histone modification, DNA methylation, and noncoding RNAs are epigenetic changes (heritable changes in gene expression not involving changes in the underlying DNA sequence). Epigenetic change is a natural process that can silence genes but can also be influenced by age, lifestyle, other environmental factors, and disease state. The crosstalk between genetics and epigenetics may also explain the impact of variants outside promoters or protein-coding sequence in health and disease.

3 DNA Transmission

The first step in the process by which genetic information is transmitted from generation to generation is called meiosis. During this process a single cell divides to produce four cells containing half the original amount of genetic information. This section gives an overview of the process of meiosis and describes the patterns of DNA transmission first introduced by Mendel and how these relate to modern genetics.

3.1 Meiosis and Recombination

Meiosis is the process of cell division that leads to gametes, sperm, and ovum. A simplistic description of this process is depicted in Fig. 2 for one homologous chromosome.

In a diploid cell the maternally derived and paternally derived dsDNA of a chromosome undergoes DNA replication (it is duplicated) to produce two identical dsDNA molecules, the sister chromatids, held together by the centromere. The resulting homologous chromosomes pair up. At this stage it is possible to exchange different segments of genetic material between homologous chromosomes leading to the formation of recombinant chromosomes. In the first meiotic division event that follows non-sister chromatids are separated and distributed in two diploid cells. In the second meiotic division the sister chromatids are separated and

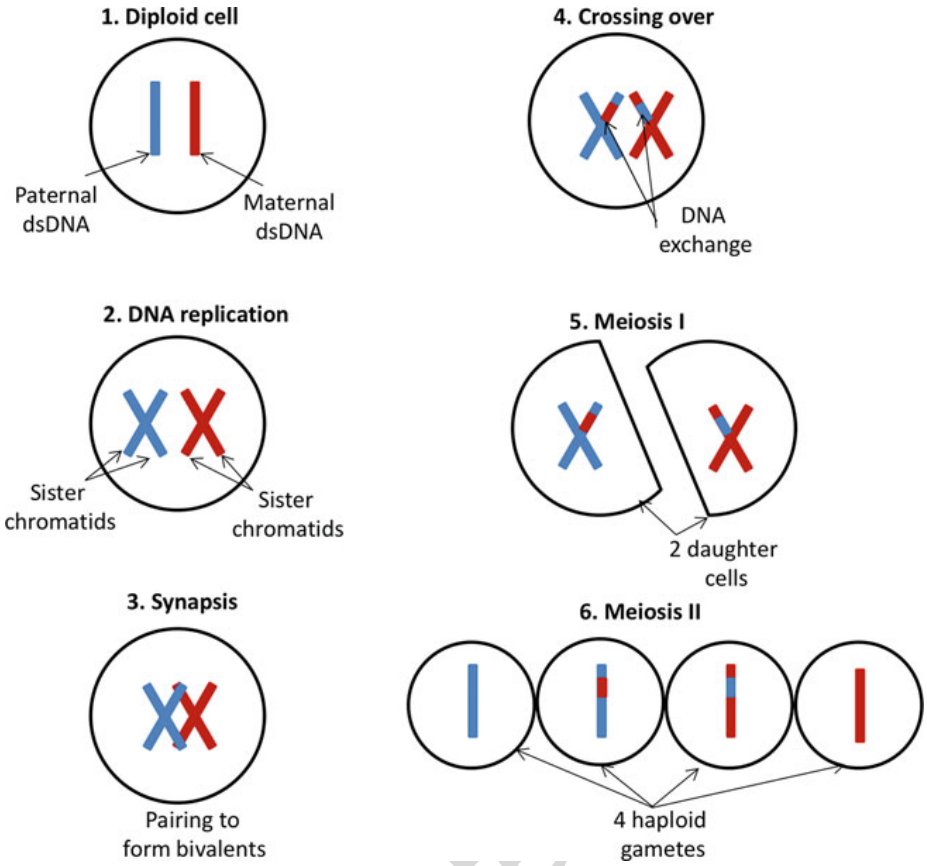

Fig. 2 An overview of meiosis. (1) A homologous chromosome of a diploid cell which contains the maternally derived and paternally derived double-stranded DNA (dsDNA). (2) DNA replication to produce two identical dsDNA molecules, the sister chromatids. (3) Pairing up of homologous chromosomes. (4) Crossing over and exchange of DNA segments between homologous chromosomes. (5) First meiotic division—separation of non-sister chromatids to two diploid cells. (6) Second meiotic division—separation of sister chromatids to four haploid gametes

distributed in four haploid gametes. Gametes (sperm and ova) fuse together during reproduction to form a zygote diploid cell.

An important aspect of meiosis is that homologous chromosomes are distributed randomly and independently to the gametes. So there is a 50% probability that a gamete will receive one chromosome from the mother rather than from the father and there are 2^{23} distinct gametes that a mother or father will produce.

Furthermore, crossing over accounts for further shuffling of genetic material because the sister chromatids held together by the centromere are not identical. Figure 2 shows one recombination event between two chromosomal segments but in reality the mean number per cell is ~55 in males and double as much in females. The further apart 2 genes are, the more likely it is that there will be recombination between DNA segments. The probability of recombination is termed the recombination fraction (θ) and forms the key to linkage analysis as discussed in Subheading 4.1.

3.2 Mendelian Genetics and Relevance to Modern Genetics

Modern genetics originated with Gregor Mendel, an Augustinian monk living in Czech Republic in the mid-eighteenth century, long before it was known that genes are the basic units of heredity. Mendel carried out a lengthy series of cross-breeding experiments with garden peas and observed the transmission of seven easily distinguishable traits including stem height (tall versus short), flower color (violet versus white), and pea morphology (round versus wrinkled). By describing the inheritance patterns mathematically he was able to demonstrate that heredity was transmitted by what he described as "factors" in a predictable and logical manner that could be studied by experimental means. He proposed three laws that are known today as Mendel laws: The law of uniformity, the law of segregation, and the law of independent assortment.

The first two laws were derived as a result of cross-fertilizing plants with distinct traits in monohybrid, reciprocal crosses. For example, Mendel pollinated a male white flower pea plant with pollen from a female violet flower plant and noted that all plants in the first generation (F1) had violet flowers (Fig. 3). He then repeated the cross reciprocally, i.e., with male violet and female white flowered plants and noted the same result. When members of F1 were self-fertilized the second generation (F2) comprised 705 plants with violet petals and 224 plants with white petals. Additional experiments with tall and short-stemmed plants or yellow and green peas demonstrated that the ratio of plants with one characteristic over another in the F2 generation approximated 3:1. Mendel reached his first conclusion: each trait (flower color) is controlled by a unit factor (gene) with each unit factor existing in more than one form (pair of alleles) responsible for the appearance of different characteristics (phenotype). The second conclusion from his experiment was that at the phenotypic level one of the alleles (the V allele) was dominant over the other allele (the v allele that is conversely termed recessive); this explained why plants in F1 appeared violet but were all heterozygotes (Vv). We now know that during sexual reproduction when an organism produces gametes the two alleles of each parent segregate (separate) randomly so that each gamete receives one allele. They then fuse together to produce the pair of alleles that is carried over in the next generation. The resulting genotype ratio in F2 is 1 homozygote for the dominant allele (VV): 2 heterozygotes (Vv): 1 homozygote for the recessive allele (vv). At the phenotypic level this produces a 3:1 ratio of violet versus white flowers (Fig. 3).

The third law of independent assortment was established as a result of Mendel's dihybrid crosses looking at the inheritance pattern of two traits at the same time, for example crosses between plants with round or wrinkled peas that were either yellow or green (Fig. 2). At the phenotypic level the round shape is dominant over the wrinkled shape so we denote the alleles as R and r for each of these distinct traits respectively. Yellow color is dominant over

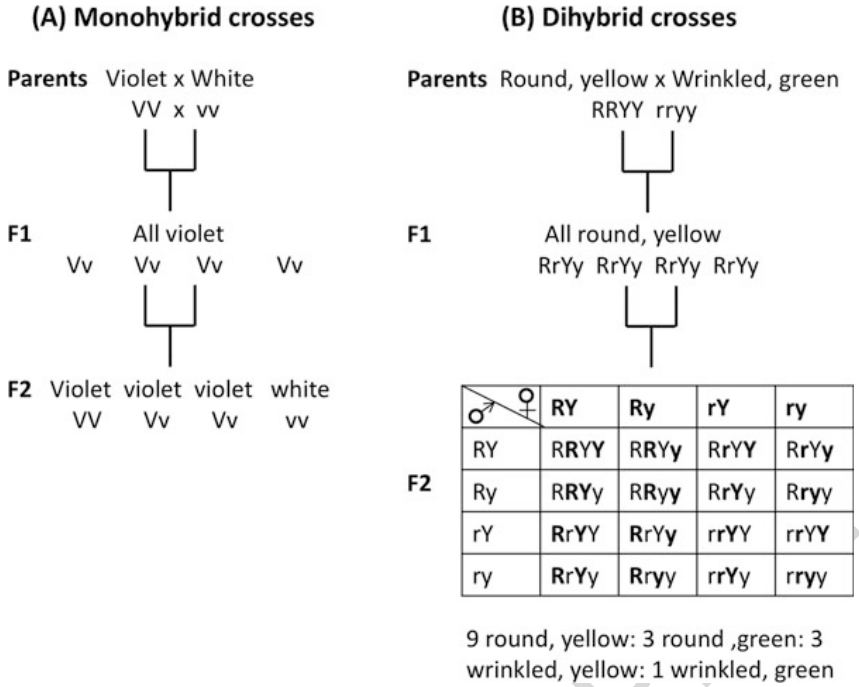

Fig. 3 Mendel's monohybrid and dihybrid crosses. (**a**) An example of monohybrid crosses between peas with violet flowers and white flowers. In the first generation (F1) all heterozygous (Vv) flowers appear violet because of the dominance of the V allele (violet color) over the v allele (white color). In the second generation (F2) the ratio of flowers is 3 violet (VV, Vv, Vv): 1 white (vv). (**b**) An example of dihybrid crosses between round, yellow peas with wrinkled, green peas. In F1 all heterozygous flowers for both characteristics (RrYy) appear round and yellow because of the dominance of the R allele (round shape) over the r allele (wrinkled shape) and the dominance of the Y allele (yellow color) over the (y) allele green color. In F2 several possible genotypes that can arise for these unlinked loci are shown in the Punnett square. Parental and non-parental trait combinations appear a ratio of 9 round yellow peas: 3 round, green peas: 3 wrinkled, yellow peas: 1 wrinkled green pea

green color so we denote the alleles as Y and y respectively. When round, yellow peas (RRYY) were crossed with wrinkled, green peas (rryy) all the plants in F1 were double heterozygotes (RrYy) and appeared as round, yellow peas. In F2 however, parental and non-parental combinations appeared in a regular ratio—9 round yellow peas: 3 round, green peas: 3 wrinkled, yellow peas: 1 wrinkled green pea. The Punnett square table in Fig. 2 shows all the possible genotypes that can arise in F2 that lead to this phenotypic ratio. The first conclusion from this experiment is that the parental traits are not linked; they can be split and give rise to non-parental trait combinations. The second conclusion is that for the 9:3:3:1 ratio to arise different pairs of alleles must segregate independently.

We now know that Mendel studied traits for genes that were in different chromosomes. The third law is generally true for loci that are found in different chromosomes and are thus unlinked.

3.3 Phenotype Transmission in Families

Mendel's monohybrid crosses on pea plants revealed patterns of phenotype transmission that formed the basis of further clinical research unraveling various inheritance patterns in families. Examination of disease transmission in large family pedigrees revealed five basic patterns categorized based on dominant or recessive mode of inheritance and whether the phenotype is transmitted by autosomes or sex chromosomes.

A disease is said to be transmitted in an autosomal dominant fashion if one allele present in autosomal chromosomes is sufficient to cause the affected status. Autosomal recessive inherited disorders require the presence of two disease-causing alleles in autosomes for disease manifestation. Diseases transmitted in a X-chromosomal dominant pattern are infrequent. If the disease-causing allele is inherited from the paternal X chromosome, all daughters will be affected whereas if the disease-causing allele is inherited from the maternal X-chromosome roughly half of the children will be affected irrespective of their gender. Diseases transmitted in a X-chromosomal recessive pattern will almost exclusively affect males if the mutation is passed on by the mother. Females will be affected only if they inherit both disease-causing alleles from each of the parents. Y-chromosomal inheritance affects only males, both fathers and sons. Few diseases follow a straightforward Mendelian inheritance pattern and in most cases this is due to incomplete penetrance (*see* Subheading 5.2 below for more information on penetrance).

4 Population Genetics

A basic concept in population genetics is the principle of *Hardy-Weinberg Equilibrium* (HWE), identified independently by Godfrey Hardy and Wilhelm Weinberg in 1908, describing the relationship between allele and genotype frequencies. As above, consider a biallelic autosomal locus (a locus with just two alleles) with alleles A and B whose allele frequencies are p and q (where $q = 1-p$) respectively. If the locus is under Hardy-Weinberg Equilibrium, then in a large, randomly mating population, the genotype frequencies of the genotypes AA, AB, and BB are expected to be in the proportions p^2, $2pq$, and q^2, where $p^2 + 2pq + q^2 = 1$. These proportions do not vary from one generation to the next, and even if the frequencies are not in those proportions in a given generation, they will return to the expected proportions after a single generation. This assumes the absence of selection (occurring through the preferential advantage of a particular genotype over another, or migration of individuals with a particular genotype), mutations or population stratification. The presence of HWE is usually used as a quality check in genetic studies, but significant deviations from HWE can also indicate the presence of selection or inbreeding.

4.1 Linkage and Linkage Disequilibrium

Mendel's third law of independent assortment means that every gene is transmitted from parent to offspring independently from all other genes. However, genes located close to each other on the chromosome are less likely to be separated by a recombination event, and are more likely to be inherited together. This concept is called *linkage*. The probability of recombination is represented by the recombination fraction θ, $0 \leq \theta \leq 0.5$. Alleles at loci on different chromosomes are unlinked and have a 50:50 chance of being inherited together ($\theta = 0.5$), and alleles are linked if θ is less than 0.5. These deviations from independent assortment form the basis of linkage mapping in families [7].

At the population level, we use the term *linkage disequilibrium (LD)* [8] to refer to the residual correlation between specific alleles at SNPs on a chromosome that has not been broken down by historical recombination. For SNPs, the most commonly used measure of LD is r^2 [9], which ranges between 0 and 1, where $r^2 = 1$ implies the SNP alleles are perfectly correlated.

The combination of alleles on a chromosome are called *haplotypes*, and regions of high LD bounded by regions of preferential recombination (*recombination hotspots)* are called *haplotype blocks* [10]. There are typically a limited number of distinct haplotypes in a short segment of the chromosome, so we can select SNPs to represent the haplotypes in the region, and infer the genotypes at the other SNPs which were not directly genotyped. This haplotype-tagging approach has led to the era of whole-genome association studies (*see* Chapter 4).

4.2 Identity by Descent (IBD) and Identity by State (IBS)

Two genes are defined as being *identical by descent* if one is a copy of another, or if they are both copies of the same ancestral gene. Two genes are identical by state if they represent the same allele. For example, if we consider the first simple pedigree (also known as a nuclear family) in Fig. 4, the parents have different alleles at the locus, so both offspring must have inherited the *a* allele from their father and the *c* allele from their mother, meaning they share 2 alleles IBD. In the second pedigree, the parents have the same alleles, although under the assumption of no inbreeding, they will

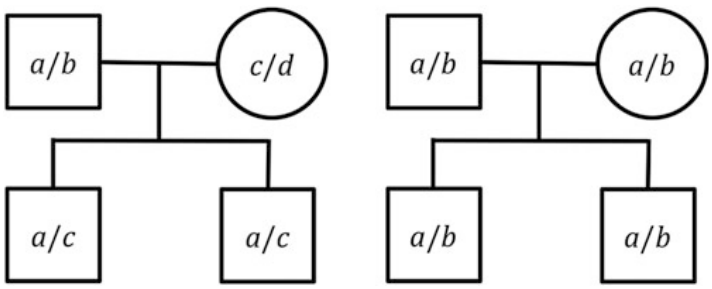

Fig. 4 Example pedigrees with parents and two offspring. Genotypes at a marker (alleles *a,b,c,d*) are shown. The parents are assumed to be unrelated (i.e., no inbreeding)

Table 2
Kinship coefficients and IBD sharing probabilities for relative pairs assuming no inbreeding

Relationship of relative pair	IBD sharing probabilities		Kinship coefficient	
	2 (z_2)	1 (z_1)	0 (z_0)	($\Phi = 1/2\ z_2 + 1/4\ z_1$)
Monozygotic twins	1	0	0	1/2
Parent-offspring	0	1	0	1/4
Full siblings	1/4	1/2	1/4	1/4
Half siblings	0	1/2	1/2	1/8
First cousins	0	1/4	3/4	1/16
Second cousins	0	1/16	15/16	1/64
Uncle-nephew	0	1/2	1/2	1/8

not be from the same common ancestor. Therefore, the offspring will share two alleles IBD if they inherited both alleles from the same parents (e.g., *a* from the father and *b* from the mother, or zero alleles IBD if they inherited their alleles from different parents). If they share zero alleles IBD, we can say that they share two alleles *identical by state* (the same alleles, but not inherited from the same common ancestor). Excess sharing of alleles IBD can be used to identify related individuals, and estimate their relationship. This is also particularly relevant for population-based association studies that assume all individuals are independent (*see* Subheading 5.2).

4.3 Kinship and Inbreeding

If we consider an autosomal locus where each individual carries two copies of a gene, the *kinship coefficient* between two individuals is the probability that genes selected at random, one from each individual, are IBD. The *inbreeding coefficient* is defined as the probability that the two genes carried by an individual are IBD, which is equivalent to the kinship coefficient for the individual's parents. Under the assumption of no inbreeding, Table 2 shows the kinship coefficients and IBD sharing probabilities for relative pairs.

5 Where Genetics Meets Epidemiology

The aggregation of disease phenotypes (such as diabetes status) or similarity in quantitative traits (such as height) in families is an indication that the phenotype may have a genetic component. For binary traits, we use a measure called the *recurrence risk ratio*, λ_R, to define the risk of disease for a relative of an affected family member of type R, compared to the population prevalence of disease [11]. For example, we denote the recurrence risk ratio of disease in siblings as λ_S, and offspring as λ_O. As for traditional

epidemiological studies, the *disease prevalence* is defined as the proportion of cases in the population at a particular time. *The disease incidence* is the rate of new cases in a given time-period.

5.1 Variance Components Models and Heritability

The estimated genetic contribution to a phenotype is called the *heritability*. In 1918, R. A. Fisher introduced the concept of variance, and the *analysis of variance* method [12]. For a trait, X, which we assume to be normally distributed, the total phenotypic variance (V_P) is made up of both environmental (V_E) and genetic components (V_G). We can further subdivide the genetic variance into additive (average effects of loci summed additively across loci), dominance (interactions between alleles at a locus), and epistatic (interactions of alleles between different loci) variances:

$$V_P = V_A + V_D + V_I + V_E$$

Similarly, the environmental variance can be divided into pure environmental variance affecting the individual or the population and variance due to gene-environment interactions.

Broad-sense heritability (H^2) is the proportion of phenotypic variance (V_P) attributable to all genetic effects (V_G) including dominance and epistatic effects:

$$H^2 = V_G/V_P$$

Narrow-sense heritability is most commonly used, and represents the proportion of the phenotypic variance determined by only the additive genetic effects:

$$h^2 = V_G/V_A$$

Although variance component methods were designed for quantitative traits, we can extend the model to binary traits by assuming that a normally distributed quantitative trait called the *liability* underlies the binary trait. A threshold is set such that the proportion of the liability distribution above the threshold is equal to the disease prevalence. For common diseases, estimates of heritability are often used to determine whether genetic approaches will be sufficiently powered to identify genetic variants contributing to risk of disease.

5.2 Genetic Models and Association

At a SNP with two alleles A and B there are three possible unordered genotypes, AA, AB and BB. In epidemiological terms, we can treat the SNP as the exposure variable. *Penetrance* reflects the risk of disease in an individual with respect to the genotype. For a disease trait, there are a number of penetrance models (or *modes of inheritance*) used to define the relationship between genotype and disease, including *multiplicative*, *additive*, *recessive*, and *dominant*.

Table 3
Penetrances under standard genetic models

Genotype	Genetic model			
	Genotype (general)	Recessive	Dominant	Additive
AA (reference)	f_0	0	0	0
AB	f_1	0	1	1
BB	f_2	1	1	2

Table 4
Genotype relative risks for genotypes AB, BB (where B is the risk allele) compared to the baseline genotype AA under standard genetic models

Genotype	Genetic model				
	GRR	Dominant $\gamma_1 = \gamma_2 = \gamma$	Recessive $\gamma_1 = 1$ $\gamma_2 = \gamma$, $\gamma > 1$	Multiplicative $\gamma_1 = \gamma, \gamma > 1$ $\gamma_2 = \gamma_1^2$	Additive $\gamma_1 = \gamma$, $\gamma > 1$ $\gamma_2 = 2\gamma_1$
AB	γ_1	γ	1	γ	γ
BB	γ_2	γ	γ	γ^2	2γ

Under the additive model, γ_2 can also be expressed as $2\gamma_1 - 1$ [17], although $\gamma_2 = 2\gamma_1$ is commonly used [18]

Define, f_0, f_1 and f_2, as the probability of disease given the genotypes AA, AB, BB respectively where the B allele is assumed to be the *risk (increasing) allele*. The penetrances under the models above can be represented as shown in Table 3. For example, under a dominant model, an individual with genotype AB or BB will have disease, whereas under a recessive model, only individuals carrying two copies of the risk allele (BB) will have disease. Well-known examples are Huntington's disease (dominant), and cystic fibrosis (recessive, both parents are required to be "carriers" of the risk allele). The *genotype relative risks* found by comparing the genotypes AB and BB to the reference genotype AA (containing no disease-causing alleles) can be defined as follows:

$$\gamma_1 = \frac{f_1}{f_0}, \quad \gamma_2 = \frac{f_1}{f_0}$$

The relationships between γ_1 and γ_2 under standard genetic models are described in Table 4.

A key concept in genetic epidemiology is that of *association*, the statistical relationship between a genetic variant and a phenotype of interest [13]. In a way that resembles traditional epidemiological

approaches, we test whether a particular allele at a SNP is more frequent in people with disease than people without disease than would be expected by chance. Alleles associated with disease are not necessarily causal for disease (or similarly for influencing a quantitative trait). Due to linkage disequilibrium, it is possible to detect association at a SNP due to linkage disequilibrium between that SNP and the causal SNP, also known as *indirect association*.

The most common design for association analysis of disease traits in the population is a case control study, where a sample of unrelated affected cases and unaffected controls are recruited. The case control design is *retrospective*, given that the individuals are collected and information on their genotype (exposure) is obtained retrospectively. Relative risks (as described above) can only be estimated from the data in *prospective* cohort studies, where individuals are selected into the study on the basis of their exposure (genotype), and followed for a specified time period to see who develops disease. In retrospective studies we can use the *odds ratio* (OR), the ratio of the odds of disease in the exposed group compared to the non-exposed, where exposure is defined by carrying a particular allele at a SNP locus and an odds ratio of one indicates independence between the SNP and disease.

Conventional X^2 tests of association using contingency tables can be used to test for association between a SNP and disease. Table 5a shows the genotype counts for cases and controls at a SNP with alleles A and B, where allele B is assumed to be the risk allele. The chi-square test statistic, measuring deviation from the expected genotype counts, follows a chi-squared distribution with two degrees of freedom (2 d.f.). This model makes no assumptions on the ordering of the genotypes and each genotype is assumed to have an independent association with disease. The tables can be simplified under standard genetic models described above. For example, under a recessive model, two copies of allele B are required for a γ-fold risk of disease and the contingency table can be summarized as a 2×2 table (1 d.f.) by pooling the AA and AB

Table 5
Contingency tables for the full genotype model and the multiplicative model

(a) Genotype model

Genotype	**AA**	**AB**	**BB**	X^2 (2 d.f.)
Cases	A	b	c	OR (AB relative to AA) $= \frac{bd}{ae}$
Controls	D	e	f	
				OR (BB relative to AA) $= \frac{cd}{af}$

(b) Multiplicative model

Genotype	**A**	**B**	X^2 (1 d.f.)
Cases	$2a + b$	$b + 2c$	Allelic OR $= \frac{(b+2c)(2d+e)}{(2a+b)(e+2f)}$
Controls	$2d + e$	$e + 2f$	

genotypes. The additive model, where there is a γ-fold increased risk of disease for the AB genotype and a 2γ-fold increased risk of disease for the BB genotype, can be tested using the Cochran-Armitage trend test. A commonly used test is the allelic case control test, where the numbers of A and B alleles are pooled ignoring which genotype they came from resulting in a 2×2 table (1 d.f.) as shown in Table 5b. This test is more powerful than the general genotype model under a multiplicative model, but assumes Hardy-Weinberg Equilibrium in the cases and controls. To adjust for covariates such as age and sex or additional SNPs logistic regression in standard statistical software can be used.

In Table 4a), the odds of being a case and having genotype AB is b/e. Similarly, the odds of being a case and having genotype AA is a/d. The odds ratio of genotype AB relative to genotype AA is therefore

$$\frac{b/e}{a/d} = \frac{bd}{ae}$$

The odds ratio for genotype BB relative to AA, and the allelic odds ratio under a multiplicative model can similarly be calculated (Table 5).

For a quantitative trait, tests of association are usually performed in a cohort of unrelated individuals, randomly selected from the population. Assuming additive SNP effects, where the effect of the SNP on the trait increases linearly with the number of copies of the effect allele, the SNP genotypes AA, AB, BB can be coded as 0, 1, 2 and tests of association can be performed using standard linear regression.

It is important to remember that the significance threshold for any test of association needs to be adjusted for the number of independent tests performed. In genome-wide association analyses for example, the number of independent tests in European populations is estimated to be 1 million, and $p = 5 \times 10^{-8}$ ($0.05/1{,}000{,}000$) has become the widely accepted "genome-wide significance" threshold.

Population studies can be susceptible confounding by *population stratification*. This can arise when cases and controls are sampled from populations with different proportions of underlying subpopulations. An extreme example would be when cases and controls are sampled from distinct ethnic groups leading to spurious associations with SNP alleles due to differences in allele frequency between the ethnic groups [14]. Family studies using related controls can control for this problem, in addition to methods designed to deal with known/cryptic relatedness [15, 16].

References

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247
2. Flicek P, Ahmed I, Amode MR et al (2013) Ensembl 2013. Nucleic Acids Res 41: D48–D55. https://doi.org/10.1093/nar/gks1236
3. Meyer LR, Zweig AS, Hinrichs AS et al (2013) The UCSC genome browser database: extensions and updates 2013. Nucleic Acids Res 41: D64–D69. https://doi.org/10.1093/nar/gks1048
4. Ritchie GR, Flicek P (2014) Computational approaches to interpreting genomic sequence variation. Genome Med 6:87. https://doi.org/10.1186/s13073-014-0087-1
5. Carlberg C, Molnár F (2016) Mechanisms of gene regulation. Springer, Netherlands. https://doi.org/10.1007/978-94-007-7905-1
6. Sonenberg N, Hinnebusch AG (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell 136:731–745. https://doi.org/10.1016/j.cell.2009.01.042
7. Teare MD (2011) Genetic epidemiology. Humana Press, New York
8. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14. https://doi.org/10.1086/321275
9. Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231. https://doi.org/10.1007/BF01245622
10. Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19:135–140. https://doi.org/10.1016/S0168-9525(03)00022-2
11. Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228
12. Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinburgh 52:399–433
13. Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366:1121–1131. https://doi.org/10.1016/S0140-6736(05)67424-7
14. Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598–604. https://doi.org/10.1016/S0140-6736(03)12520-2
15. Anderson CA, Pettersson FH, Clarke GM et al (2010) Data quality control in genetic case-control association studies. Nat Protoc 5:1564–1573. https://doi.org/10.1038/nprot.2010.116
16. Price AL, Zaitlen NA, Reich D et al (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463. https://doi.org/10.1038/nrg2813
17. Schaid DJ (1999) Likelihoods and TDT for the case-parents design. Genet Epidemiol 16:250–260. https://doi.org/10.1002/(SICI)1098-2272(1999)16:3<250::AID-GEPI2>3.0.CO;2-T
18. Camp NJ (1997) Genomewide transmission/disequilibrium testing—consideration of the genotypic relative risks at disease loci. Am J Hum Genet 61:1424–1430. https://doi.org/10.1086/301648

Author Queries

Chapter No.: 2 394545_1_En

Query Refs.	Details Required	Author's response
AU1	Please check whether the affiliation and correspondence details are presented correctly.	