PART I

GENOMES

CHAPTER ONE

Identifying the genetic basis of disease

Vineet Bafna

It is all in the DNA. Our genetic code, or *genotype*, influences much about us. Not only are physical attributes (appearance, height, weight, eye color, hair color, etc.) all fair game for genetics, but also possibly more important things such as our susceptibility to diseases, response to a certain drug, and so on. We refer to these “observable physico-chemical traits” as *phenotypes*. Note that “to influence” is not the same as “to determine” – other factors such as the environment one grows up in can play a role. The exact contribution of the genotype in determining a specific phenotype is a subject of much research. The best we can do today is to measure correlations between the two. Even this simpler problem has many challenges. But we are jumping ahead of ourselves. Let us review some biology.

1 Background

Why do we focus on DNA? Recall that our bodies have organs, each with a specific set of functions. The organs in turn are made up of tissues. Tissues are clusters of cells of a similar type that perform similar functions. Thus, it is useful to work with cells because they are simpler than organisms, yet encode enough complexity to function autonomously. Thus, we can extract cells into a Petri dish, and they can grow, divide, communicate, and so on. Indeed, the individual starts life as a single cell, and grows up to full complexity, while inheriting many of its parents’ phenotypes.

Bioinformatics for Biologists, ed. P. Pevzner and R. Shamir. Published by Cambridge University Press.
© Cambridge University Press 2011.

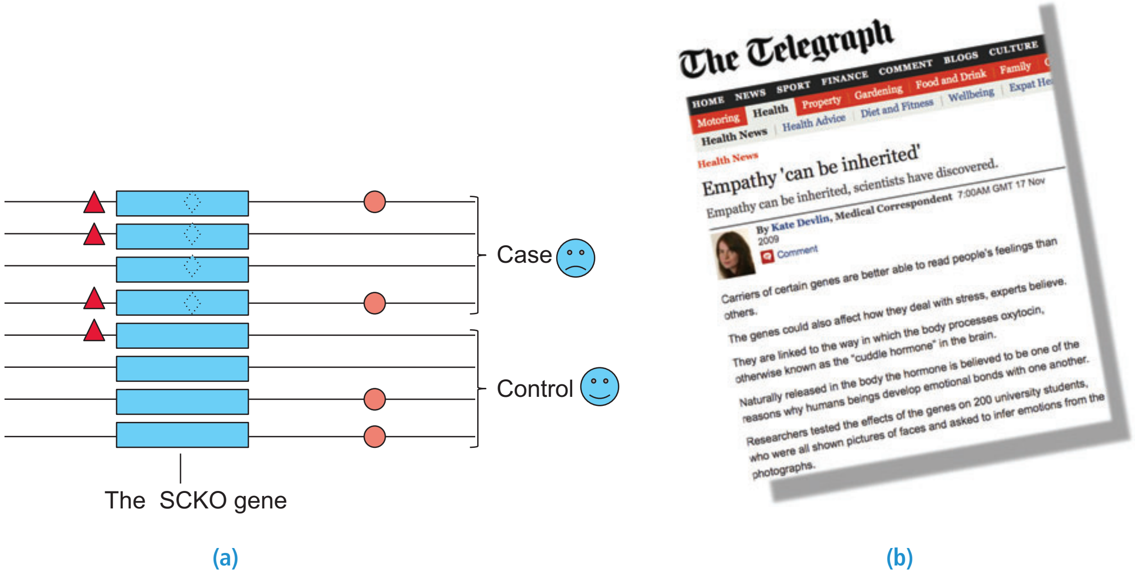
Cambridge University Press
978-1-107-01146-5 - Bioinformatics for Biologists
Edited by Pavel Pevzner and Ron Shamir
Excerpt
More information

There must be molecules that contain the instructions for making the body, and these molecules must be inherited from the parents. The cells have smaller subunits (nucleus, cytoplasm, and other organelles) which contain an abundance of three molecules: DNA, RNA, and proteins. Naturally, these molecules were prime candidates for being the inherited material. Of these, proteins and RNA were known to be the machines in the cellular factories, each performing essential functions of the cell, such as metabolism, reproduction, and signal transduction.

This leaves DNA. The discovery of DNA as the inherited material, followed by an understanding of its structure and the mechanism of inheritance, form the major discoveries of the latter half of the twentieth century. DNA consists of long chains of four nucleotides, which we abbreviate as A, C, G, T. Portions of the nucleotides (*genes*) contain the code for manufacturing specific proteins, as well as the regulatory mechanisms that interpret environmental signals, and switch the production on or off. Interestingly, we have two copies of DNA, one from each of our parents. In this way, we produce a similar set of proteins as our parents, and therefore display similar phenotypes, including susceptibility to some diseases. Of course, as we inherit only a randomly sampled half of the DNA from each parent, we are similar but not identical to them, or to our siblings.

On the other hand, if all DNA were identical, it would not matter where we inherited the DNA from. In fact, DNA mutates away from its parent. Often, these mutations are small changes (insertions, substitutions, and deletions of single nucleotides). There are also many additional forms of variation, which are more complex, and include many large-scale changes that are only now being understood. In this chapter, however, we will focus on small mutations as the only source of variation. If we sample DNA from many individuals at a single location (a *locus*) we often find that it is polymorphic (contains multiple nucleotide variants). Clearly, if these mutations occur in a gene, then the protein encoded by the DNAcan also change, possibly changing some functional trait in the organism. Therefore, different variants at a locus sometimes present different phenotypes, and are often referred to as *alleles*, after Mendel. Loci with multiple alleles are variously called “segregating sites” (they separate the population), “variants”, or “polymorphic markers.” If these variants affect single nucleotides, they are also called single nucleotide polymorphisms or SNPs.

We start with a basic instance of a *Mendelian* mutation: individuals present a phenotype if and only if they carry the specific mutation. Our goal is to identify the mutation (or the corresponding genomic locus) from the set. Figure 1.1a shows this with three candidate variants represented by $\diamond$, $\triangle$, and $\circ$. A simple approach to identifying the causal mutation is as follows: (i) determine the genotypes of a collection of individuals that present the phenotype (*cases*), and those that do not (*controls*); (ii) align the genotypes of all individuals, and identify polymorphic locations; (c) for


Case
Control
The SCKO gene
(a)
(b)
The Telegraph
HOME NEWS SPORT FINANCE COMMENT BLOGS CULTURE
Motoring Health Property Gardening Food and Drink Family
Health News | Health Advice | Diet and Fitness | Wellbeing | Expat He
Health News
Comment

Empathy 'can be inherited'

Empathy can be inherited, scientists have discovered.

By Kate Devlin, Medical Correspondent 7:00AM GMT 17 Nov 2009

Carriers of certain genes are better able to read people's feelings than others.

The genes could also affect how they deal with stress, experts believe.

They are linked to the way in which the body processes oxytocin, otherwise known as the "cuddle hormone" in the brain.

Naturally released in the body the hormone is believed to be one of the reasons why humans beings develop emotional bonds with one another.

Researchers tested the effects of the genes on 200 university students, who were all shown pictures of faces and asked to infer emotions from the photographs.

Figure 1.1 Genetic association basics. (a) A Mendelian mutation ◇ that is causal for a phenotype. Other "neutral" variants are nearby. (b) Popular news highlighting the discovery of the gene responsible for a phenotype. In many cases, all that is observed is a correlation between a mutation and the phenotype. The causality is assumed based on some knowledge of the function of the protein encoded by the gene. Figure reprinted by permission. © Telegraph Media Group Limited 2011.

each polymorphic location, check for a correlation of the variants with case/control status. In Figure 1.1, we see that the occurrence of the ◇ correlates highly with the case status and conclude that the mutation is causal. Given that the mutation lies in the SCKO gene, we conclude that SCKO is responsible. The popular media is peppered with accounts of discoveries of genes responsible for a phenotype.

The intelligent reader will immediately question this premise because these "discoveries" are often not the final confirmation, but simply an observed correlation between the occurrence of the mutation and the phenotype. First, what is the chance that we are even testing with the causal mutation? Typically, genotypes are determined using the technology of DNA chips. The individual DNA is extracted (often from saliva or serum) and washed over the chip. The chip allows us to sample, in parallel, close to 0.5–1 M polymorphic locations, and determine the allelic values at these locations. This fast and inexpensive test allows us to investigate a large population of cases and controls, and makes genetic association possible. However, we do not test *each* location (there are three billion). It is very possible that the causal mutation is not even sampled, and that we may not find correlations even when they exist. Second, even if we do find a correlation, there is no guarantee that we have found the right one. Surely, a simple correlation at one of 1 M markers could have arisen just by chance. How can that be a clue towards the causal gene?

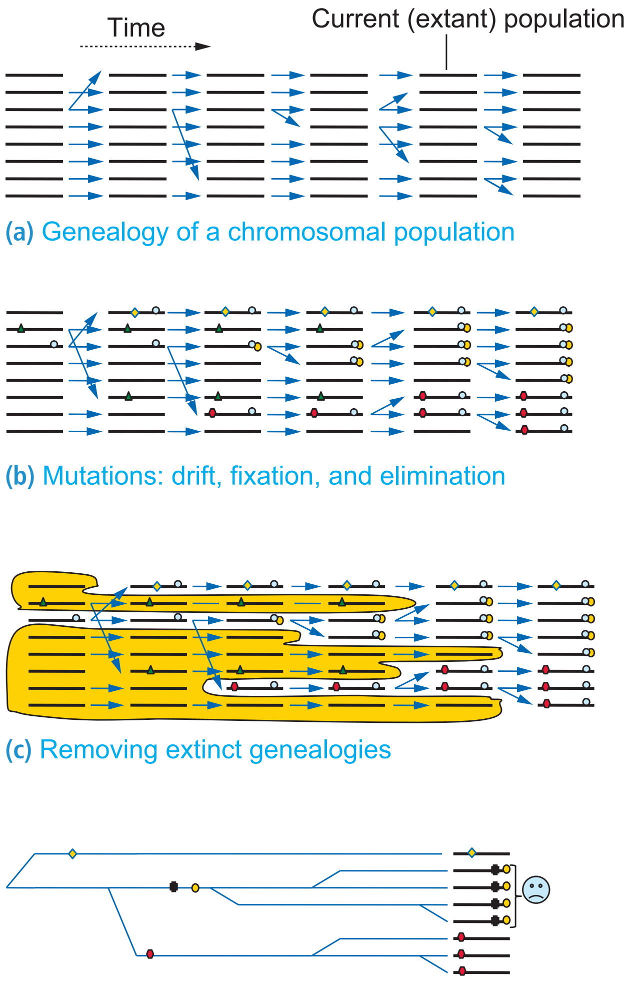
The answer might surprise some. Nature helps us in two ways: first, it establishes a correlation between SNPs that are close to the causal mutation, so any of the SNPs in the region (that contains the relevant gene) are correlated with the mutation. Second, it "destroys" the correlation as the distance from the causal mutation increases. Therefore, a correlation is indeed a strong suggestion that we are in the right location, and any gene in that region is worth a closer look. The next section is devoted to an explanation of the underlying genetic principles, and is followed by a description of the statistical tests used to quantify the extent of the correlation.

Of course, while the basic premise is correct, and simply stated, it is (like everything else in biology) simplistic. In the following sections, we look at issues that can confound the statistical tests for association, and how they are resolved. The resolution of these problems requires a mix of ideas from genetics, statistics, and algorithms.

2 Genetic variation: mutation, recombination, and coalescence

Dobzhansky famously said that "nothing in biology makes sense except in the light of evolution," and that is where we will start. You might recall from your high-school biology that each of us has two copies of each chromosome, each inherited from one parent.[1] Having two parents makes it tricky to study the ancestral history (the genealogy) of an individual. Therefore, we work with a population of chromosomes, where every individual does have a single parent. In this abstraction, the individual is simply "packaging" for the chromosomes, two at a time. We also make the assumption (absurd, but useful) that all individuals reproduce at the same time. Finally, we assume that the population size does not change from generation to generation. Figure 1.2a shows the basic process. Time is measured in reproductive generations. In each generation, an individual chromosome is created by "choosing" a single parent from the previous generation. To see how this helps, go back in time, starting with the extant population. Every time two chromosomes choose the same parent (coalesce), the number of ancestral chromosomes reduces by 1, and never increases again. Once this ancestry reduces to a single chromosome (the most recent common ancestor, or MRCA), we can stop because the history prior to that event has been lost forever. As each individual has a single parent, the entire history from the MRCA to the extant generation is

[1] Not quite, but we will consider recombinations in a bit.



(a) Genealogy of a chromosomal population

(b) Mutations: drift, fixation, and elimination

(c) Removing extinct genealogies

(d) Causal and correlated mutations

Figure 1.2 An evolving population of chromosomes. (a) The Wright Fisher model is an idealized model of an evolving population where the number of individuals stays fixed from generation to generation, and each child chooses a single parent uniformly from the previous generation. (b) Mutations are inherited by all descendants, and drift until they are fixed or eliminated. (c) We only consider the history that connects the existing population to its most recent common ancestor. (d) The underlying data are presented as a SNP matrix (with a hidden genealogy). The genealogy leads to correlations between SNPs.

described by a tree (*the coalescent tree*). Other genealogical events that occurred after MRCA but are not part of the coalescent tree are useless because the lineages died out before reaching the current generation (Figure 1.2c). The only historical events that will concern us are ones in the underlying coalescent tree.

Now, let us consider mutations. Each chromosome is identical to its parent, except when a mutation modifies a specific location. Given the short time frame of evolution of the human population relative to the number of mutating positions, most locations are modified at most once in history. To simplify things, we assume that this is true for all variants (the *infinite sites assumption*): once a location mutates to a new allelic value, it maintains that allele, and all descendants of the chromosome inherit the mutation. As individuals choose their parents and inherit mutations, the frequency of mutations changes (*drifts*) from generation to generation. This principle is illustrated in Figure 1.2b. The mutation denoted by the blue ○ arises before the MRCA, and is therefore *fixed* in the current population. On the other hand, △ arises in a lineage that was *eliminated* and is not observed. Other mutations, such as the ⬡, arose sometime after the MRCA, and present as polymorphisms when sampled in the existing population. This is illustrated in Figure 1.2d. Here, we have removed the generation information, and represent time simply by the branch-lengths. When we sample a population with DNA microchips, we create a matrix of polymorphisms; rows correspond to individuals, columns represent polymorphic locations, and the entries represent allelic values representing the consequence of historical mutations on the coalescent tree. The tree itself is invisible, although likely trees can be reconstructed using phylogenetic techniques.

What is the point of all this? It is simply that the underlying tree imposes a correlation between mutations. Let the black circle ● in Figure 1.2d represent a causal mutation. Individuals display a phenotype if and only if they carry this mutation. However, every mutation in this matrix is correlated to some extent. For example, the presence of the yellow mutation (which is on the same branch) is equally predictive of the phenotype, and the red ⬡ (which occurs on a different lineage) implies that the individual does not carry the phenotype. We call this the principle of *linkage*: mutations that are part of an evolutionary lineage are correlated. Thus, it is not necessary to sample all mutations to identify the gene of interest. However, this is not enough. If all SNPs on the chromosome are correlated (albeit to varying degree), they cannot help to narrow the search for the causal locus. We are helped again by the natural phenomenon of *recombination*. In meiosis (production of gametes), a crossing over of the two parental chromosomes might occur. The child therefore gets a mix of the two parental chromosomes, as shown schematically in Figure 1.3a,b. Now consider a population. Recombination events between two locations change the underlying coalescent tree. With increasing distance between loci, the number of historical recombination events increases and destroys the correlations. In Figure 1.3c, the yellow and black ○ are proximal and remain correlated. However, recombination events destroy the correlations (the linkage) between the red ⬡ and causal (black) ●. This establishes a second principle: *correlation between mutations is destroyed with increasing distance between loci due to the accumulation of recombination events.*

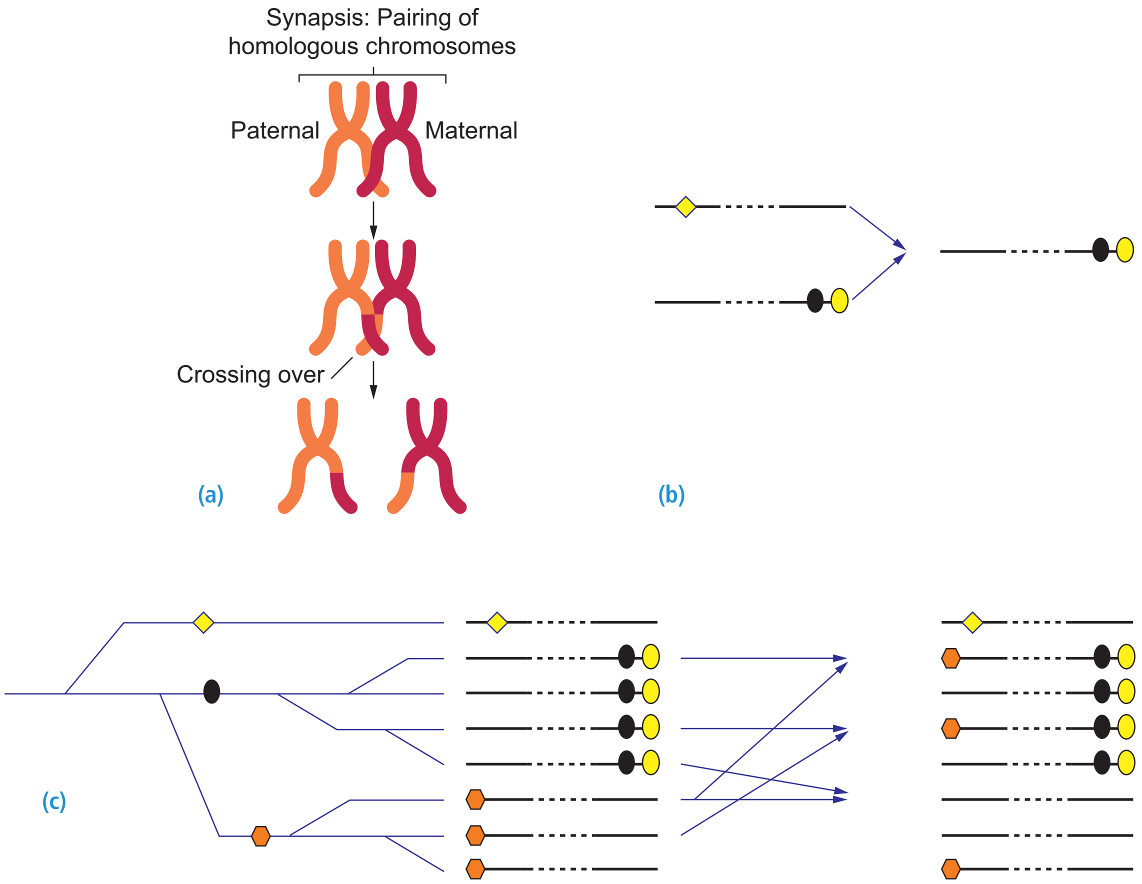


Figure 1.3 Recombination events change genealogical relationships, and destroy correlation between SNPs. (a) Crossover during meiosis. (b) Schematic of a crossover and its effect of linkage between mutations. (c) Multiple recombination events destroy linkage between SNPs.

3 Statistical tests

Let us digress and consider a simple experiment to statistically test for correlation between two events: thunder and lightning. It is intuitively clear that the two are correlated, but we will formalize this. Let $x_i = 1$ indicate the event that we saw lightning on the ith day. Respectively, let $y_i = 1$ indicate the event that we heard thunder on the ith day. Let P_x (respectively, P_y) denote $\Pr(x_i = 1)$ (respectively, $\Pr(y_i = 1)$) for a randomly chosen day. Assume that we see lightning 35 days in a year, so that $P_x = 35/365 \simeq 0.1$. Likewise, let $P_y \simeq 0.1$. What is the chance of seeing both on the same day? Formally, denote the chance of joint occurrence by $P_{xy} = \Pr(x_i = 1 \text{ and } y_i = i)$. If the two were not correlated, we would not observe both very often. In other words,

$P_{xy} = P_x P_y \simeq 0.01$, and so only 3–4 days a year are expected to present both events. If we observe 30 days of thunder and lightning, then we can conclude that they are correlated. What if we observe 10 days of thunder and lightning? This is the question we will consider.

Denote two loci as x, y, and let x_i denote the allelic value for the ith chromosome. If we make the assumption of infinite sites, x_i will take one of two possible allelic values. Without loss of generality, let $x_i \in \{0, 1\}$. The generalization to multi-allelic loci will be considered in Section 4.2. Let P_x denote $\Pr(x_i = 1)$ for a randomly sampled chromosome i at locus x. Correspondingly, $P_{\bar{x}} = 1 - P_x$ represents the probability that $x_i = 0$. Denote the joint probabilities as

$$P_{xy} = \Pr(x_i = 1, y_i = 1) = P_x \Pr(y_i = 1 | x_i = 1)$$

$$P_{\bar{x}y} = \Pr(x_i = 0, y_i = 1) = P_{\bar{x}} \Pr(y_i = 1 | x_i = 0)$$

and so on. If x, y are proximal then $\Pr(y_i = 1 | x_i = 1)$ is very different from P_y. See, for example, the black and yellow ○ in Figure 1.3c. By contrast, if x, y are very far apart so that recombination events have destroyed any correlation, then

$$P_{xy} \simeq P_x P_y$$

$$P_{\bar{x}y} \simeq P_{\bar{x}} P_y.$$

As the recombination events destroy correlation over time, we use the term *Linkage Equilibrium* to denote the lack of correlation. The converse of this, often termed *Linkage Disequilibrium* (LD), or *association*, describes the correlation between the proximal loci. A straightforward statistic to measure $\mathrm{LD}(x, y)$ is given by

$$D = P_{xy} - P_x P_y. \tag{1.1}$$

Note that the choice of allele does not matter. The interested reader can verify that

$$\begin{aligned} |D| &= \left| P_{xy} - P_x P_y \right| \\ &= \left| P_{\bar{x}y} - P_{\bar{x}} P_y \right| \\ &= \left| P_{x\bar{y}} - P_x P_{\bar{y}} \right| \\ &= \left| P_{\bar{x}\bar{y}} - P_{\bar{x}} P_{\bar{y}} \right|. \end{aligned}$$

The larger the value of $|D|$, the greater the correlation. Apart from its historical significance, the D-statistic is used more as a relative, rather than an absolute measure. Instead, a scaled statistic D' is defined as

$$D' = \frac{D}{D_{\max}} = \begin{cases} \frac{D}{\min\{P_{\bar{x}} P_y, P_x P_{\bar{y}}\}} & D \geq 0 \\ \frac{D}{-\min\{P_x P_y, P_{\bar{x}} P_{\bar{y}}\}} & D < 0 \end{cases}. \tag{1.2}$$