2.3 Linkage, recombination, and LD.

Within small linked regions of the genome, the coalescent process generates correlations between the genotypes at different SNPs. This is known as linkage disequilibrium (LD). Meanwhile, at larger distances, recombination breaks down LD by shuffling genotypes. Here we discuss how the opposing forces of linkage and recombination shape genetic variation.

The concepts of linkage, recombination, and LD appear in almost every topic in human genetics, including natural selection, population history, population admixture and introgression, and the genetics of complex traits.

A first look at haplotype structure. The first time anyone sequenced the same locus in multiple individuals was in 1983. In a landmark study, Marty Kreitman, who was a graduate student at the time, sequenced the ADH gene in 11 lines of the fly *Drosophila melanogaster* [174]. The figure below shows a simplified version of the complete data set from that paper:

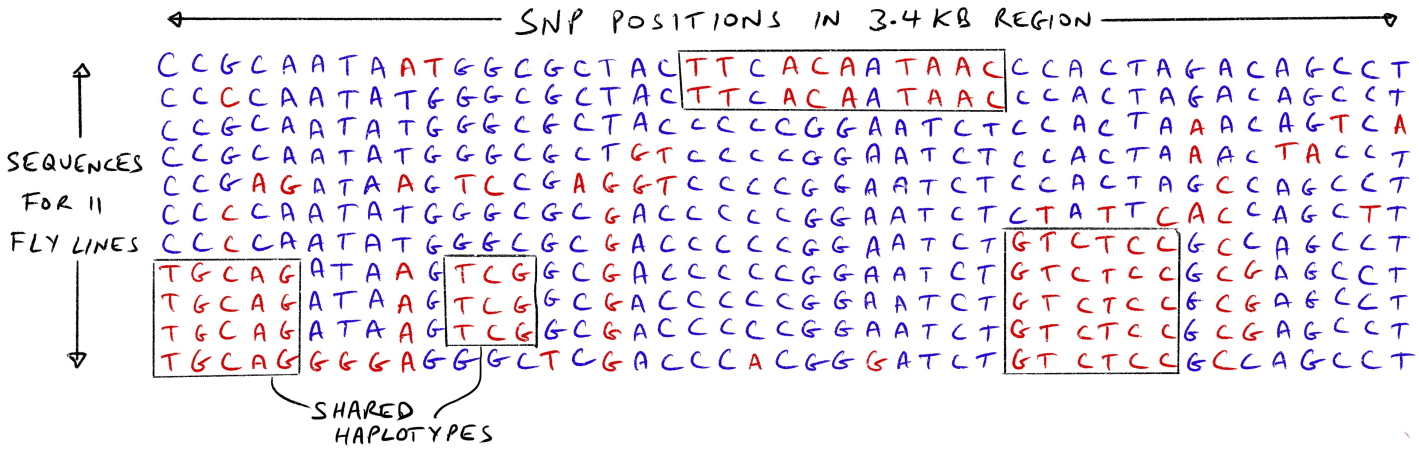


Figure 2.36: Haplotype structure. *Each row shows the genotype for a single fly line, and the columns show genotypes at SNP positions (most sites are not shown as they were identical in all 11 lines). The major allele at each position is shown in blue. Examples of blocks of shared haplotypes are indicated.* *Note that each line was constructed to carry only a single haplotype. For simplicity, a few indels are not shown. Data: Martin Kreitman (1983) [Link].*

Each row shows the sequence of alleles found on a particular chromosome copy in the population. We refer to the set of alleles found at variant positions within a linked region as a **haplotype** [a].

[a] *For another early example, this time from human data, see Figure 1.31.*

Looking at these haplotypes, one feature may jump out at you: **particular combinations of alleles at different SNPs frequently appear together**. For example, on the left, a block of alleles TGCAG is shared among four lines, all of which (and one other) later carry another block: GTCTCC.

This is a very typical feature of genetic data: particular alleles at nearby SNPs often appear together more often than expected by chance. This nonrandom assortment of alleles at different sites is referred to as

linkage disequilibrium (LD). *How do we understand this?*

In the next couple of pages, we'll talk about how **linkage generates LD, while recombination tends to break down LD**. As before, both the backward-in-time models (the coalescent) and forward-in-time models (Wright-Fisher) provide complementary kinds of intuition, and we'll use both approaches.

Linkage generates haplotype structure (or equivalently, LD). Sites that are close together in the genome are usually inherited together. This is called **linkage**.

When we introduced the coalescent in the last chapter, we ignored the possibility of recombination, focusing on sequences that are **completely linked**. In this setting, there is a clear relationship between the branching structure of the tree, and the corresponding haplotypes:

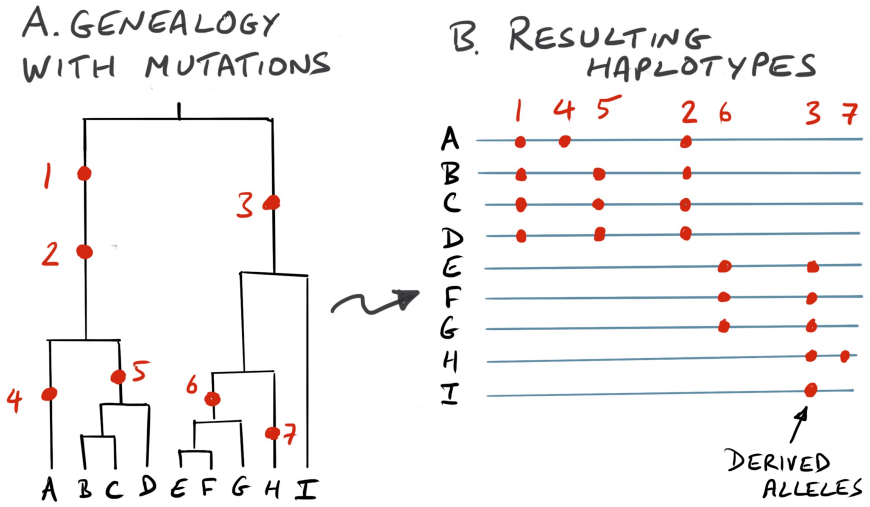


Figure 2.37: **The coalescent process generates haplotype structure. A** *shows a coalescent tree without recombination. The red circles indicate mutations.* **B** *shows the corresponding haplotypes; the red circles indicate derived alleles using the same numbering as in panel A.* *The positions of the mutations within the sequenced region are random.*

For example in the tree above, mutations 1 and 2 occurred on the same branch, and hence those two derived alleles always appear together. Mutations 1 and 5 are on adjacent branches, and so derived alleles 1 and 5 *usually* appear together (although haplotype A has the derived allele at 1 but not at 5).

For a tree without recombination, there are very strong constraints on the possible configurations of the derived alleles across haplotypes. For example, if we focus on two SNPs at a time, you might expect that there could be four possible haplotypes. If we label the ancestral and derived alleles as A/a at the first SNP, and B/b at the second SNP, then in principle the haplotypes could be A-B, A-b, a-B, a-b.

But in the absence of recombination we can only get either two or three of the four possible haplotypes, depending on where the mutations occur. As you can see in the examples below, we can get either 2 or 3 of the possible combinations, but not all 4:

Furthermore, looking across all SNPs together, there are additional con-

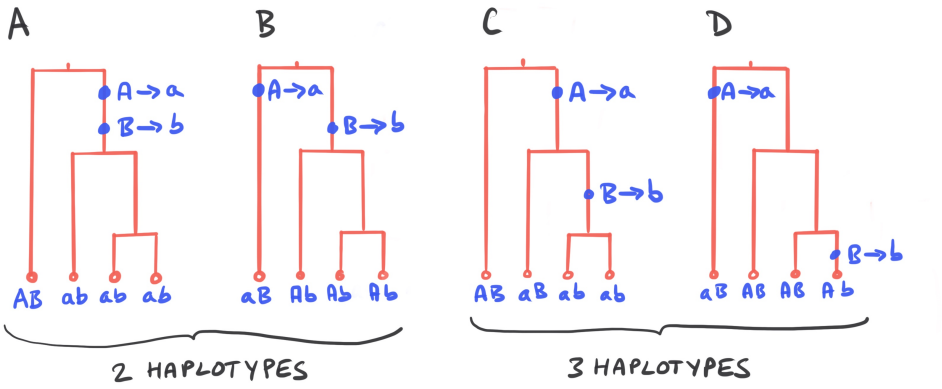


Figure 2.38: **Pairwise LD in the absence of recombination.** *For any pair of SNPs we can observe either 2 or 3 out of the 4 possible haplotypes (depending on where the mutations lie on the tree). While this is illustrated here for 4 samples, it is true regardless of sample size.*

straints: the alleles must be nested in a way that is consistent with existence of a single tree. Haplotypes that are consistent with a single tree are said to form a **perfect phylogeny** [Link]). I suggest that you draw some examples of trees with mutations, to see what configurations are possible.

Recombination. But in practice, most regions of the genome are subject to recombination. Recombination plays a crucial role in shuffling haplotypes, and producing combinations that would be impossible in the absence of recombination.

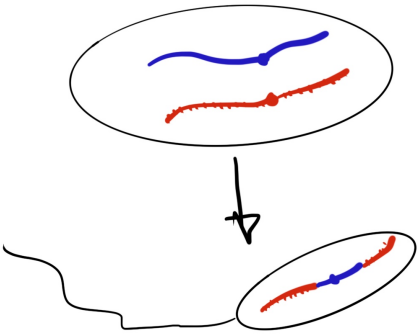


Figure 2.39: **Recombination.** *Sperm and eggs carry recombined mixtures of the parental chromosomes; typically there are around 1-2 switches per chromosome, known as* ***crossovers****, positioned randomly along each chromosome.*

A quick refresher on recombination. Recall that during the production of eggs and sperm, the chromosomes go through meiosis. In humans, this reduces the number of chromosomes from 46 to 23. During this process, the maternal and paternal chromosomes are broken and then joined back together so that chromosomes in the resulting gametes are mixtures of the parental chromosomes. This is called **recombination**, or **crossover** [175]. Crossover events are positioned more-or-less randomly across the genome with an average of 26 crossovers per sperm and 42 per egg.

Genetic distance. It will be helpful to talk about genetic distance, which measures the rate of crossover, between different positions along a chromosome. Genetic distance is measured in terms of **centiMorgans (cM)**.

We define the genetic distance x, between two points on a chromosome to be x cM if the average number of crossovers between those two points is x/100 per meiosis. For example, if two points are 10 cM apart, then we expect 0.1 crossovers per meiosis.

Figure 2.40: **Crossover observed in laboratory whiteboard pens.**

Furthermore, we'll be most interested in short genetic distances, for which **we can also interpret genetic distance in centiMorgans as the percent probability of a crossover** in the specified interval [176]. For example, if the genetic distance between two sites is 1 cM, then there is about 1% probability of a crossover per meiosis in that interval.

Lastly, it's helpful to define relate genetic distances to base pair distance in the DNA sequence. For this purpose we define the **recombination rate**. This is commonly measured in cM/Mb: that's 100 times the expected number of cross overs per megabase. The average recombination rate in the human genome is about 1.2 cM per Mb. In other words, there

is about a 1.2% probability of a crossover event per megabase [b] [177].

So, why do we care about this here? Recombination is central to our story because it shuffles haplotypes:

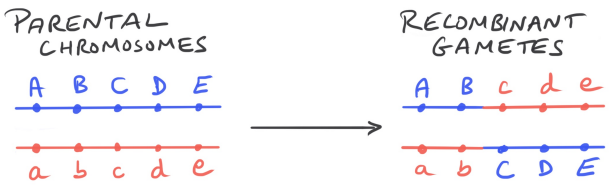


Figure 2.41: **Recombination mixes haplotypes,** *often creating new combinations that did not exist previously.*

In this way, **recombination generates new combinations of alleles that would not be possible with complete linkage**. For example, when there is recombination we *do* expect to see all four possible haplotypes for a pair of SNPs, unlike what I showed you above. However, the rates depend crucially on genetic distance.

As we shall see, for SNPs that are close together in the genome (less than ∼0.01–0.1 cM, or about 10–100 Kb) linkage is a stronger force than recombination and there tends to be strong haplotype structure. At larger distances (more than ∼0.1 cM), recombination is highly effective at shuffling genotypes, and LD is generally weak. In the next sections we'll see why this is.

Measuring LD between pairs of SNPs. To make this discussion more precise, it's helpful to define some measures of LD that we can study in models, and in real data.

Imagine two SNPs. One has alleles A and a, with allele frequencies p_A and p_a; the other SNP has alleles B and b, with frequencies p_B and p_b. Then there are four possible haplotypes: AB, Ab, aB, and ab, with frequencies p_{AB}, P_{Ab}, and so on:

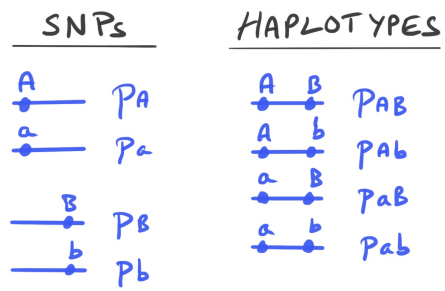


Figure 2.42: **Notation for allele and haplotype frequencies at two SNPs.** *Here p denotes a frequency.*

If I didn't tell you anything about these SNPs in advance, what would you guess for the haplotype frequency p_{AB}? The most natural thing would be to guess that the alleles are independent of each other, in which case $p_{AB} = p_A p_B$.

This intuition is captured by a measure called D, which is the difference between the observed and expected frequency of the AB haplotype:

$$D = p_{AB} - p_A p_B. \tag{2.39}$$

If genotypes at the two SNPs are independent (i.e., the SNPs are in linkage equilibrium), then $D = 0$. [c]

It may seem arbitrary to define D in terms of just the AB haplotype, but a little algebra will show that if we redefined D in terms of a different haplotype (e.g., with respect to Ab), then the only thing that would happen is that D would switch signs to become $-D$. Since the allele labeling is usually arbitrary, in practice we'll only pay attention to the absolute value $|D|$.

The second important measure of LD is known as D'. A weakness of D

[b] *It's useful to remember that the human recombination rate is about 1% per megabase.*

[c] *Note: If the alleles are labeled 0 and 1 at each SNP, then D can be interpreted as the statistical covariance between alleles at the two SNPs.*

as a measure of LD is that its possible range depends on the allele frequencies of the two SNPs, so it doesn't immediately tell us if LD between two SNPs is weak or strong.

To solve this limitation, D' is adjusted to range between -1 and 1 regardless of the allele frequencies:

$$D' = \frac{D}{D_{\max}} = \frac{D}{\min(p_A p_b, p_a p_B)} \quad \text{for } D > 0 \tag{2.40}$$

$$= \frac{D}{\min(p_A p_B, p_a p_b)} \quad \text{for } D < 0 \tag{2.41}$$

As with D, the sign of D' depends on the labeling of the alleles, so most papers use the absolute value, $|D'|$.

The formula for $|D'|$ is a bit messy, but a little algebra reveals a crucial property: **$|D'| < 1$ if and only if all four possible haplotypes are present. In other words, $|D'| < 1$ implies that there must have been recombination between the two sites**.

The third important measure of LD is called r^2, and again builds off D:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}. \tag{2.42}$$

The value of r^2 ranges from 0 to 1, where 0 means that the SNPs are completely independent. **A value of $r^2 = 1$ is referred to as perfect LD** [d], **and occurs if and only if there are just two of the four possible haplotypes: i.e., only AB/ab or only Ab/aB.**

[d] *Note: r^2 can also be interpreted as the squared correlation coefficient: the statistical correlation between genotypes at two SNPs is $r = D/\sqrt{p_A p_a p_B p_b}$.*

As we'll see later in the book, **r^2 is the natural parameter for measuring the contribution of LD to genetic associations in complex trait genetics** [178].

Strong recombination breaks down LD. We can now show a key result for how LD behaves in a model with strong recombination (and no drift).

Suppose we create an artificial population where two SNPs start out in strong LD, with an initial value of $D{=}D_0$ in generation 0. Let c be the probability of crossover, per generation, between these two SNPs. (See the Box below for a precise definition of c and a derivation.)

In the next generation, the LD (denoted D_1) is predicted to be:

$$D_1 = (1 - c)D_0, \tag{2.43}$$

and over successive generations the decay of D simply multiplies, so that in generation t we have:

$$D_t = (1 - c)^t D_0. \tag{2.44}$$

This implies that unless the recombination rate is very small, LD decays very quickly. For two SNPs that are 20 Mb apart, say, we expect that $c \sim 0.2$. After ten generations, $(1 - 0.2)^{10}{=}0.1$, meaning that within ten

generations D decays to just 10% of its starting value. In this time, LD between unlinked parts of the genome ($c = 0.5$) would essentially disappear. Here's a plot showing decay of D over time:

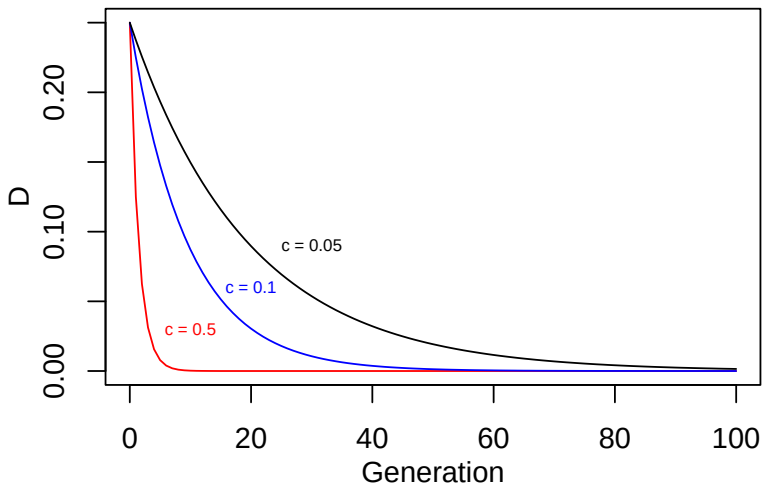


Figure 2.43: **D decays within a few generations for large recombination rates** *(Equation 2.44, assuming $D_0 = 0.25$). Timescales of tens of generations are very short compared to timescales of drift – which takes place over tens of thousands of generations.*

This result also holds for D' since, without drift, the denominator is unaffected by recombination.

Optional: Decay of LD due to recombination. Here we sketch out the argument for how D decays over time (Equation 2.44).

First, my definition of c above was a bit sloppy. To be more precise, c will be the probability that the two alleles passed into a gamete both came from the same parent (i.e., both from the mother, or both from the father). At short genetic distances, c is closely approximated by the probability of at least one cross-over between the two SNPs, but this definition implies that the maximum value of c is 0.5, which corresponds to random assortment of alleles on different chromosomes [179].

Our derivation assumes that recombination is happening much faster than drift. Specifically we assume the allele frequencies p_A and p_B stay constant, while the haplotype frequencies, p_{AB}, etc, change due to recombination. (To be more precise, this approximation will be accurate when the change in D due to recombination, cD, is much larger than the rate of drift, which is $\mathcal{O}(1/N)$.)

Let p_{AB} be the frequency of the AB haplotype in the current generation. What do we expect for p^*_{AB}, the frequency in the next generation? Here I use the notation * to indicate the value of a parameter in the next generation.

The haplotype frequency p^*_{AB} depends on two effects. First, recombination breaks apart AB haplotypes at a rate $c \times p_{AB}$. Second, recombination creates AB haplotypes at a rate $c \times p_A p_B$: this is the probability of randomly assembling an AB haplotype as a result of recombination. So we have

$$p^*_{AB} = p_{AB} - cp_{AB} + cp_A p_B \tag{2.45}$$
$$= (1 - c)p_{AB} + cp_A p_B \tag{2.46}$$

Subtracting $p_A p_B$ from both sides we write this in terms of D:

$$p^*_{AB} - p_A p_B = (1-c)p_{AB} + c p_A p_B - p_A p_B \tag{2.47}$$
$$= (1-c)p_{AB} - (1-c)p_A p_B \tag{2.48}$$
$$D^* = (1-c)D. \tag{2.49}$$

Then, using the same logic over multiple generations, the decay of LD follows

$$D_t = (1-c)^t D_0. \tag{2.50}$$

To summarize, so far we have seen that:
• if there is no recombination, the basic properties of the coalescent genealogy tell us to expect strong LD;
• at large distances (for SNPs more than a few centiMorgans apart, say, and certainly for SNPs on different chromosomes), recombination rapidly eliminates LD.

We now need to explore what happens at intermediate distance scales, between ∼1 kb to 100 kb, where recombination and coalescence compete against each other.

The coalescent with recombination: the ARG. To understand these models, we can incorporate recombination into the coalescent. This produces a more complex structure called an **ancestral recombination graph (ARG)** [180].

To begin, we'll look at two positions in a sequence, labeled L (left) and R (right). Going backward in time, we now have *two* kinds of events: both coalescence and recombination. **As before, coalescence joins lineages, but now recombination can split sequences apart so that each side of a breakpoint becomes a separate lineage.** This is visualized here:

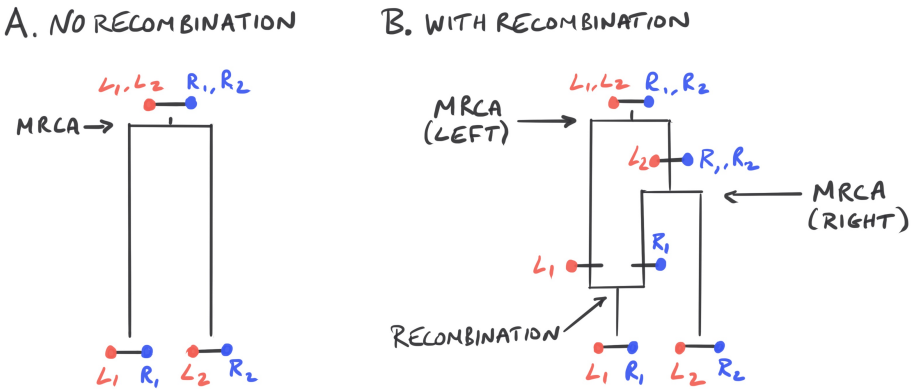


Figure 2.44: **An ARG for two samples.** *L and R indicate the left and right-hand ends of haplotypes 1 and 2.* **A.** *Coalescence without recombination.* **B.** *Going backwards in time Lineage 1 splits due to recombination. L_1 takes the left-hand path, while R_1 takes the right. The two ends of the locus have different MRCAs as indicated.*

In the figure above, you can see in panel B that, going backwards in time, Lineage 1 is split apart by a recombination event so that we have a different coalescent time for the left-hand side of the region (red) versus the right-hand side (blue).

We can extend this idea to consider more samples and more recombina-

tion events. Instead of considering just two sites, we consider a sequence of length c (in units of genetic distance), and recombination can occur anywhere within this region. We indicate the position of a recombination event by a fraction: for example, recombination at 0.3 occurs 30% of the way along the sequence.

In this figure, the full ARG is shown at the left. This contains a series of four so-called "marginal" trees at different positions across the sequence, shown in red, each with a different branching pattern. As you can see, the ARGs can become very complicated:

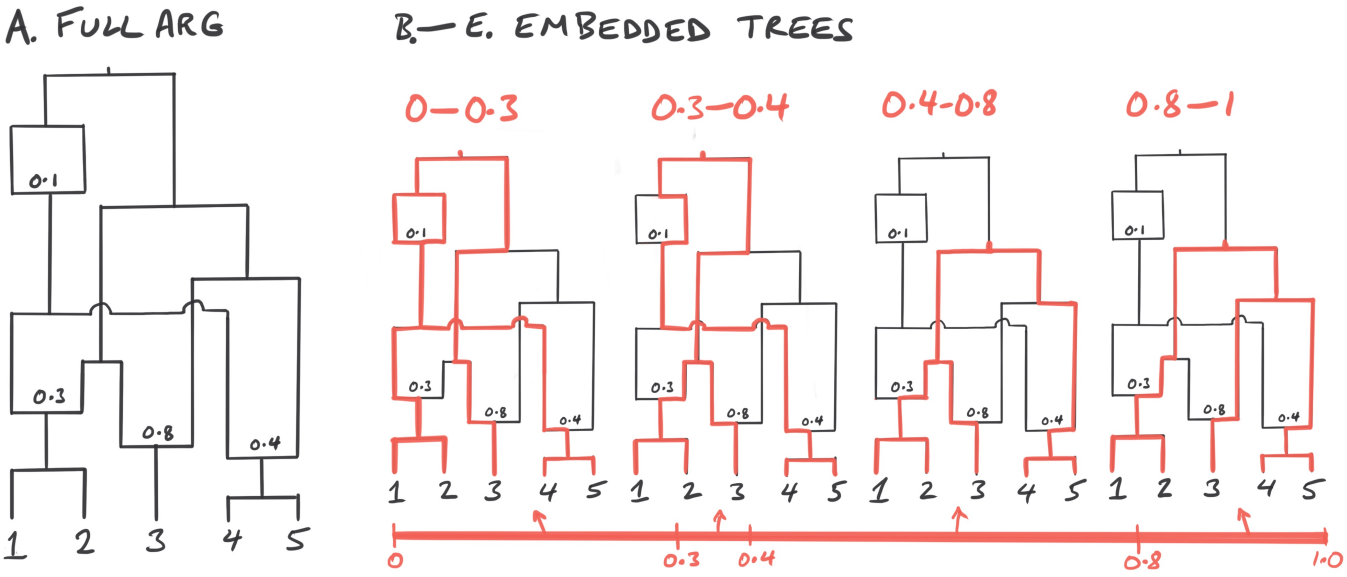


Figure 2.45: ARG and embedded "marginal" trees. *Split points represent recombination events; we use the convention that the left-hand side of the sequence follows the left-hand branch, and the right-hand side follows the right branch. The red numbers at top show which part of the sequence is relevant to each marginal tree: e.g., the first tree covers positions 0–0.3. Note that the recombination at 0.1 does not affect the marginal trees.*

The graph above contains 4 recombination events. These split the region into four distinct blocks, each with a different coalescent tree. However, the trees don't change entirely: haplotypes 1 and 2, as well as 4 and 5, are closely related across the entire region.

I've described all this in terms of the full ARG process, but it's worth noting that the sequence data only depend on the marginal trees at each position (the red trees), and it can be easier to think about the process just in terms of these trees, and the fact that they change as you move along the sequence [e].

Breakdown of LD within the ARG. How does recombination affect haplotype variation? Crucially, **recombination can create mixtures of haplotypes. This is illustrated in the example below, where addition of a single recombination event produces all four possible haplotypes** – remember that this would not possible in the absence of recombination:

[e] *We won't cover inference methods in detail, but in practice the modern inference methods focus on estimating marginal trees rather than the full graph including all recombination events.*

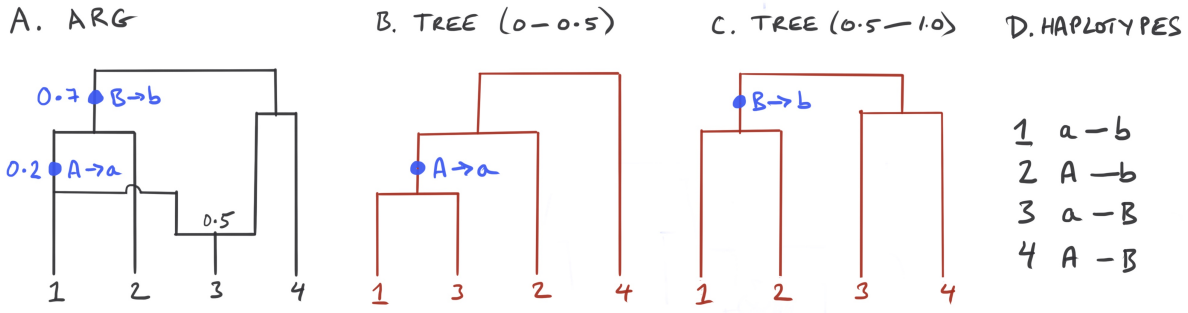


Figure 2.46: Recombination can generate all four haplotypes for two SNPs. *The mutations are at positions 0.2 and 0.7 along the sequence, as indicated; the recombination is at 0.5.*

The tug-of-war between coalescence and recombination. It's difficult to get really deep intuition for properties of the ARG. But I think it's helpful to think about it as a competition between two key processes: the tree-structure of the coalescent creates haplotype structure, while recombination tends to break it apart.

The outcome of this competition is determined by a compound parameter, $4Nc$: the ratio of the rate of recombination (c) to the rate of coalescence ($1/2N$) (and an extra factor of 2, see below). A large value of $4Nc$ basically means there is a high rate of recombination per unit rate of coalescence, so recombination tends to be the winner (and conversely for small $4Nc$).

The next box explains why $4Nc$ is a natural parameter.

Optional: Timescales in the ARG. Let's start with two samples, and a region of length c. What's the probability that these two samples coalesce without recombination?

Going backwards in time, as usual, coalescence occurs at a rate $1/2N$ per generation. Meanwhile, recombination occurs at a rate c per generation (in either lineage), so $2c$ in total. So the probability of at least one recombination before coalescence is

$$\frac{2c}{2c+1/2N} = \frac{4Nc}{4Nc+1}. \tag{2.51}$$

(This result uses a method for "competing exponentials"; don't worry if it's not familiar.)

More generally, take a slice through the ARG at any time. Suppose that we have k lineages. What are the waiting times to the next event (backwards in time, as usual)? Coalescence decreases the number of lineages from k to $k-1$; this occurs at a rate $k(k-1)/4N$ per generation as before. Meanwhile, recombination increases the number of lineages from k to $k+1$; this occurs at a rate ck, where c is the total recombination rate across the segment of interest [181].

So the probability that the next event is a recombination event is

$$\frac{kc}{kc+k(k-1)/4N} = \frac{4Nc}{4Nc+k-1}. \tag{2.52}$$

This formula suggests the following:

- $4Nc$ is the natural parameter to describe the role of recombination in an ARG.

• In large samples, coalescence predominates at recent timescales (when k is large), while recombination is more effective at scrambling the lineages further back in time (when k is small).

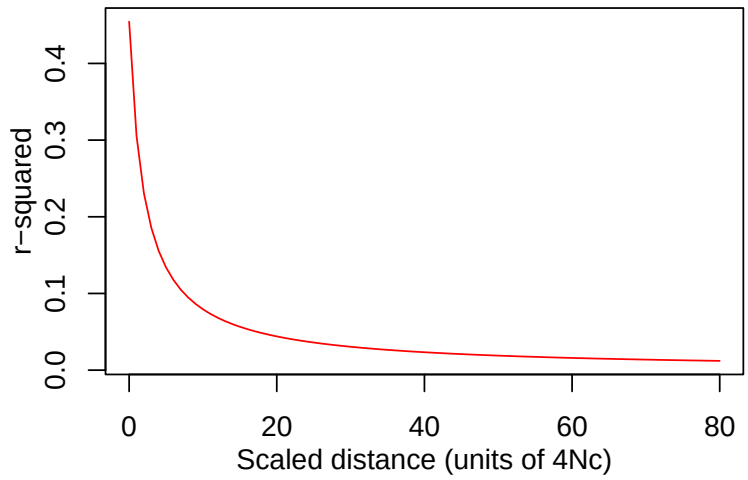
To summarize, consideration of the ARG highlights a few key points:

- sites that are close together tend to share the same genealogy, hence SNPs are in high LD;
- genealogies become less and less correlated with increasing genetic distance, thus reducing LD;
- the scale of LD depends on the product of N and c (usually written as $4Nc$);
- in large samples, the most recent coalescent events occur faster than recombination, so closely related haplotypes can be shared over large recombination distances, even at distances where overall LD is low.

Decay of r^2 with distance. So we have a qualitative prediction that LD should decay with genetic distance. Can we predict this more precisely?

It turns out that the expected r^2 can be approximated as a ratio of covariances in coalescence times among sequences at different distances [f]. I won't present the math for this (it's a bit fiddly) but you can read about it here [Link] [182].

And here's how average r^2 decays as a function of distance:



To give you a sense of scale, on average 100 kb in humans is around $4Nc = 80$; this model predicts that LD should decay to be low within around $10 - 100$kb, which is fairly typical in practice.

Recombination and LD in human data. Most of this basic theory was already understood by the 1980s and 1990s. But for a long time we didn't have the tools to measure this in real data [183].

[f] *Note: these results focus on the typical levels of LD at very short distances with recombination and coalescence; as such it differs from the earlier results predicting rapid decay of D starting from an initial condition of unnaturally high LD.*

Figure 2.47: **Predicted decay of mean r^2** *between pairs of SNPs, as a function of distance. To interpret the x-axis, note that* ***for humans 4Nc=80 corresponds to c=0.1 cM or ∼100 kb*** *at the genome-average recombination rate.* *The function plotted here is $(10+4Nc)/(22+13(4Nc)+(4Nc)^2)$, which approximates the mean of r^2 between common SNPs. See [Link].*

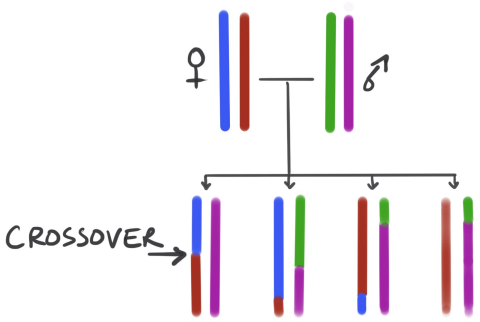


Figure 2.48: **Pedigree studies of recombination.** *Traditional genetic mapping studies used a scaffold of genetic markers to count crossover events within pedigrees – shown here for a single chromosome in parents and four kids.*

This started to change in the 1990s, alongside the Human Genome Project. At that time, one goal was to create genetic and physical maps of the genome. One main approach was to genotype a genome-wide scaffold of genetic markers (STRs) in families, and count recombination events directly. With these data it was possible to estimate recombination rates along each chromosome, as shown here in this recombination map of Chromosome 1, made in 2002:

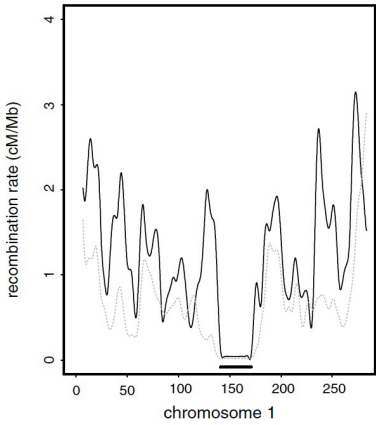


Figure 2.49: **Pedigree-based recombination map for Chromosome 1**. *These data, from a 2002 pedigree study, show estimated* ***recombination rates at megabase scale*** *(females, solid line; males, dotted line).* *The region without recombination at around 150 Mb marks the centromere. Credit: From Figure 2 of Augustine Kong et al (2002) [Link] Used with permission.*

As you can see here, at this scale recombination rates vary from about 1–3 cM/Mb (except for the centromere), and are higher near the telomeres, which is pretty typical of the genome. Female rates (solid line) are generally higher than male rates (dotted), consistent with the fact that genome-wide rates in females are 1.6× the male rates.

But the resolution of this type of map was limited by the number of available STRs (about 2 per Mb), which meant that they could not study fine-scale variation in rates [184].

So when the first high-resolution SNP data came along, it was a big surprise to find that the LD data revealed something much more striking!

But before we get to this, I need to explain a little about how to visualize LD. Let's suppose that we genotype a bunch of SNPs across a region. One thing we could do is to show colored haplotypes in the style of Figure 1.31, but it's hard to get a quantitative sense of the data from this. Instead, a commonly-used approach computes a matrix of r^2 or D' between all pairs of SNPs, and displays it with a color scale, like this:

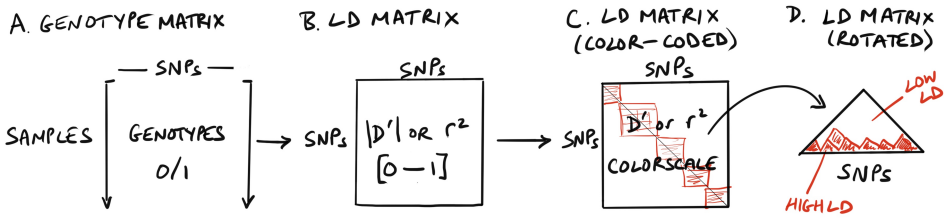


Figure 2.50: **Visualization of LD patterns.** *Panel* **A** *displays the full haplotype data but is difficult to interpret quantitatively.* **B** *Instead it's common to display the data as a matrix of pairwise LD; and often color-coded* **C**. **D** *Finally, the matrix is rotated, and only the top half is shown.*

In panel D above, the SNP pairs that are close together are plotted near the base of the triangle, and SNPs that are far apart are higher up. Thus,

we expect that LD will usually be high (red) near the base, and decrease going up.

By the early 2000s, as it became possible to collect SNP data at higher density, some very interesting patterns started to emerge [185]. Our model above suggests that LD might be expected to decay smoothly with distance, but this is not the case at all. Instead, LD structure forms striking blocks of high LD (so-called **haplotype blocks**), separated with lower LD between blocks. Here's a typical example from a 500 Kb region of the genome:

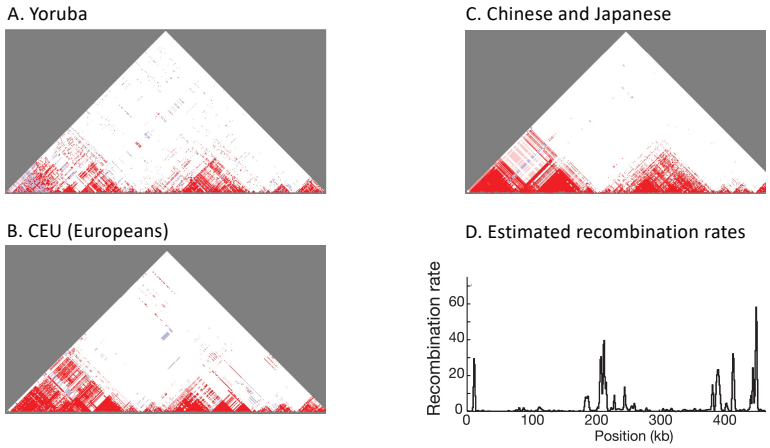


Figure 2.51: **Fine-scale patterns of LD and recombination** *for a 500 Kb region on Chromosome 2, in three population samples (***A–C***). Red entries indicate pairs of sites with* $|D'| = 1$*; white sites indicate* $|D'| < 1$*. Notice that the overall structure is largely shared across populations, but the extent of LD is lowest in the African population (Yoruba) and highest in east Asian population.* **D.** *Estimated recombination rates in cM/Mb across the same region. The peak recombination estimates are much higher than in the pedigree map.* *Credit: Modified from Figure 8 of HapMap (2005) [Link] Used with permission.*

In the plot above, red indicates $|D'| = 1$. Remember that $|D'| < 1$ *(white) indicates that all 4 possible haplotypes are present, and that past recombination must have shuffled genotypes between the two sites*. The blocky structure of the D′ matrices suggests that most recombination is taking place at the boundary points between adjacent blocks.

It's possible to use the LD structure to estimate a fine-scale recombination map (panel D) [186] – this supports the visual impression that most recombination is concentrated into narrow regions with extremely high recombination rates. These locations are referred to as **recombination hotspots**.

These early results have proved to be typical of the genome overall: **the structure of LD tends to form blocks, with generally high LD inside blocks, and lower LD between blocks. This reflects the structure of recombination, which is mainly concentrated into narrow hotspots.**

Genome-wide, more than 30,000 hotspots have been identified, with additional recombination spread among weaker hotspots [187]. **This helps to set the scale of LD, which typically extends around 10–100 Kb, depending on the genomic region** [g].

PRDM9 and the hotspot paradox. The discovery of tens of thousands of recombination hotspots immediately suggested a new question: What controls the locations of hotspots? Work on this question led to a fascinating saga spanning molecular genetics, human genetics, and evolutionary biology.

[g] *The figure above also illustrates another typical pattern, namely that* ***LD is lowest in African populations due to their larger long-term effective population size****.*

The first major progress came in a 2005 paper by Simon Myers and colleagues, which reported that a certain 7-nucleotide sequence motif is highly enriched within hotspots [188]. The presence of this short DNA motif at many hotspots suggested that the locations of hotspots are, at least in part, directed by local DNA sequences [h]. This situation is reminiscent of binding sequences for transcription factors, and suggested that recombination events might be directed by an unknown DNA-binding protein that recognizes this motif.

[h] *Terminology: DNA sequences that control local activity are said to act **in cis**, while external factors such as a DNA-binding protein that recognizes those sequences are said to act **in trans**.*

But this exciting observation immediately raised a theoretical problem known as the **hotspot paradox** [189]. The hotspot paradox argues that, due to the molecular details of recombination, evolution should tend to remove cis-acting hotspot motifs.

To explain this, I need to say a bit about what happens during recombination. During meiosis, the homologous chromosomes pair up. Crossovers are initiated by one of the two homologs (the blue one, in the example below). The initiating chromosome undergoes a double-strand break, and part of the chromosome is chewed back in both directions around the break. Eventually, this damaged region is repaired using the other (red) chromosome as a template, in a process known as **gene conversion**:

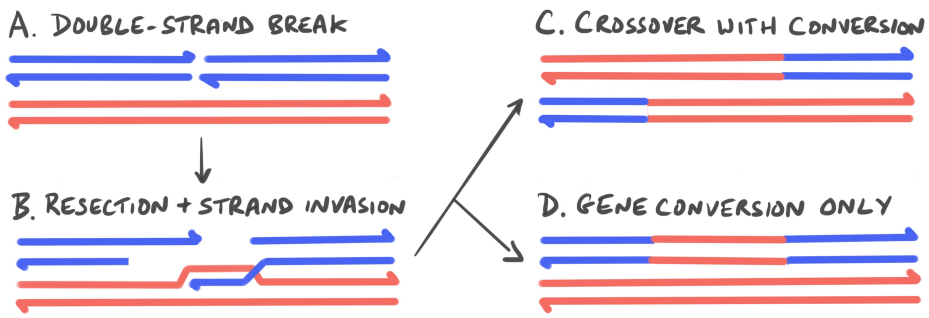


Figure 2.52: **Simplified model of recombination with gene conversion**. **A.** *Recombination is initiated with a double-strand break in the blue chromosome.* **B.** *Several hundred bp around the DSB are resected (chewed away) on the blue chromosome; then one strand invades the red chromosome. This is resolved in one of two endpoints:* **C.** *crossover with gene conversion to repair the damaged section using the red chromosome as template, or* **D.** *gene conversion without crossover. Note: this is a simplified account of a complex process. Figure modified from [Link].*

The key point here is that, within the gene conversion region, it is the initiating chromosome (blue) that is copied over by its partner (red). Both resulting chromosomes end up with the red sequence inside the converted region.

Now, let's suppose that one chromosome carries a hotspot motif but the other does not (for example there could be a SNP for which one allele breaks the motif). Then, the chromosome with the motif can initiate the crossover. But that sequence would then be replaced by gene conversion from the other non-motif chromosome. This is known as **biased gene conversion**:

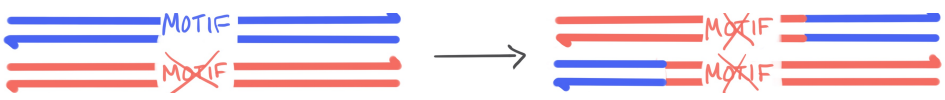


Figure 2.53: **Biased gene conversion**. *An allele that encodes the hotspot motif (blue) will tend to be replaced by an alternative allele that breaks the motif (red).*

In other words, *biased gene conversion tends to remove hotpots!* We haven't

covered selection yet, but *this is mathematically equivalent to a form of selection in favor of alleles that remove hotspots* [190]!

Based on this logic, the hotspot paradox argues that any time a SNP arises inside a hotspot motif, it will tend to spread through the population as if it were positively selected. *Over time, this should tend to eliminate all hotspots. So why are there any hotspots left?*

Around the same time as discovery of the hotspot motif, another intriguing observation was made by comparing LD in humans and chimpanzees.

Recall that chimpanzees are our closest living relatives, and that our genome sequences are extremely similar, differing at only about 1.4% of sites. Given this, you might naively expect most hotspots to be shared if they are controlled by cis-acting motifs. But, remarkably, studies of LD in chimpanzees found no meaningful overlap of hotspot locations between humans and chimps beyond random expectations [191]:

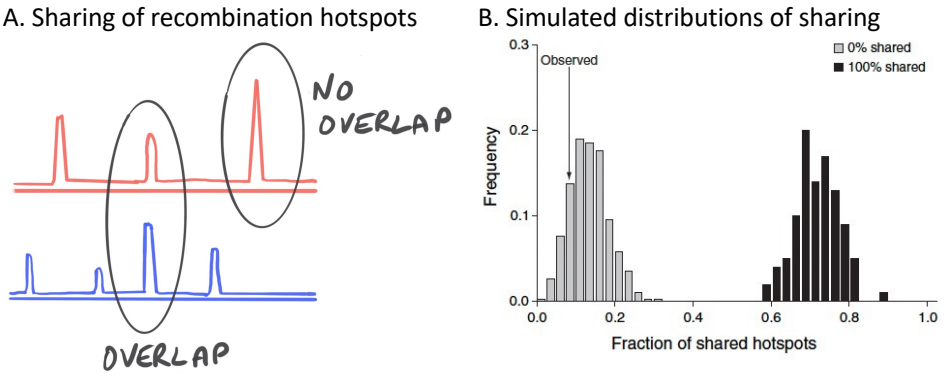


Figure 2.54: **No sharing of hotspots between humans and chimpanzees**. **A.** *One study estimated that just 8% of hotspots overlap between humans and chimpanzees.* **B.** *Simulations showed that this is consistent with no true sharing of hotspots (since hotspot regions are estimated imprecisely, some overlap is expected even by chance).* *Credit: Panel B is Figure 2 from Susan Ptak et al (2005), [Link] Used with permission.*

And an independent study of recombination events in pedigrees in a European-American population showed that, even within humans, not everyone uses the same hotspots at the same rates [192]. All this was very intriguing. If hotspot locations are controlled by local sequences, then shouldn't most hotspots be shared?

Many of these questions started to be resolved by a set of papers in 2010 that identified a gene called PRDM9 as the missing, central, player in this entire saga [193]. PRDM9 encodes a protein with a so-called "zinc finger" domain that is responsible for DNA binding.

The zinc finger domain has a specific affinity to – you guessed it – the previously-discovered hotspot motif. The plot below shows DNA binding predictions from a 2010 paper, based on the protein sequence of the most common European PRDM9 allele. This substantially matches the DNA sequence enriched within hotspots:

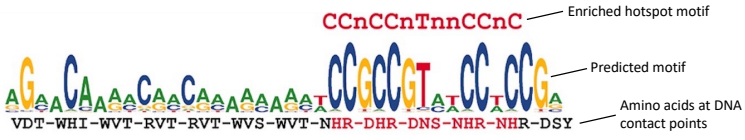


Figure 2.55: **Predicted binding preferences of the PRM9 'A' allele.** *The sequence motif in red at top represents a consensus of the motif enriched at hotspots ('n' indicates no clear consensus). The DNA 'logo' plot shows predicted binding preferences of the PRDM9 protein based on the corresponding zinc fingers; the sizes of the letters reflect the predicted strength of preference for each nucleotide.* *Modified from Figure 2 of Baudat et al (2010) [Link] Used with permission.*

Once PRDM9 binds to the DNA, it recruits additional machinery to

initiate double-strand breaks (which can result in crossovers).

Of particular interest, the zinc finger DNA-binding domain is encoded within a minisatellite repeat section of the gene [i]. Each repeat (or "finger") consists of 28 amino acids (84 bp), of which 4 amino acids touch the DNA and provide binding specificity. Most of the other 24 amino acids are identical across repeats.

Recall that the copy number in such regions is often highly variable due to mispairing during DNA replication. In fact, this is the case at PRDM9, where dozens of alleles have been found in humans. Furthermore, the alleles frequently differ specifically in the amino acids that contact the DNA. The image below shows differences between two of the most common human alleles, A and C, as well as allelic variation across nine different human alleles:

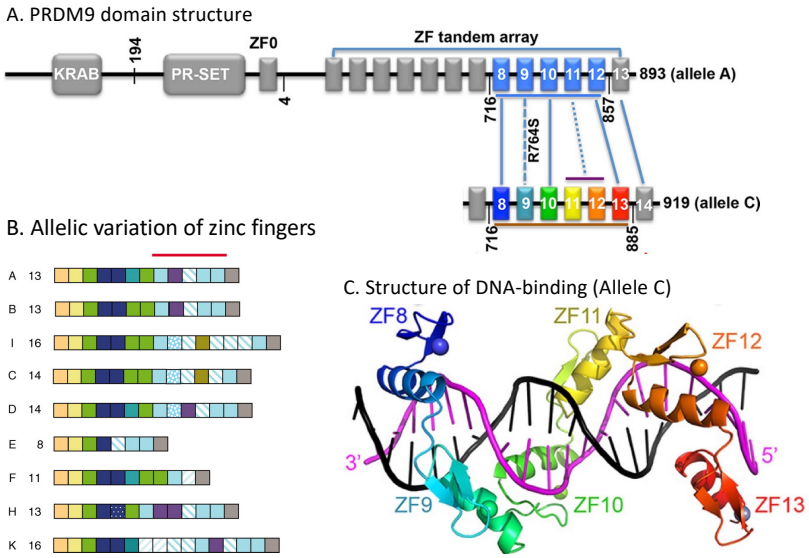


Figure 2.56: **Structure of PRDM9**. **A.** *Domains of the PRDM9 gene. Zinc fingers 8–13 are responsible for DNA binding and differ between Alleles A and C.* **B.** *Diversity of zinc finger structure across nine human PRDM9 alleles. Each box represents a single zinc finger, and colors indicate distinct sequences. Notice that alleles differ both in the numbers of fingers as well as their sequences, especially within the main DNA binding region (red bar).* **C.** *DNA binding structure of the C allele. Fingers 8-13 are responsible for DNA sequence recognition.* *Panel A and C, modified Figure 1 of Patel et al (2017) [Link], CC BY 4; Panel B part of Figure 2 of Baudat et al (2010) [Link] Used with permission.*

[i] *Minisatellites are similar to STRs, but with longer repeat units. They tend to be highly variable due to replication slippage, and are also often referred to as VNTRs: Chapter 1.3; Figure 1.34.*

Importantly, in many cases, the different alleles have different DNA binding preferences. For example, allele C, which is common (36%) in west Africa but rare outside Africa, uses completely different hotspots than allele A (plot at right) [194].

Meanwhile, chimpanzees also have completely different PRDM9 alleles from humans – thus neatly explaining the complete lack of overlap between the human and chimpanzee hotspot maps.

Lastly, the rapid evolution of PRDM9 neatly resolves the hotspot paradox. There has indeed been systematic loss of human hotspots during recent human evolution [195], but this is counteracted by regular jumps in PRDM9 binding preferences due to the evolution of new alleles [196].

As a consequence of all this, the selective pressure imposed by hotspot evolution has made PRDM9 one of the most rapidly evolving vertebrate genes – and a fascinating story involving molecular biology, population genetics, and evolutionary biology.

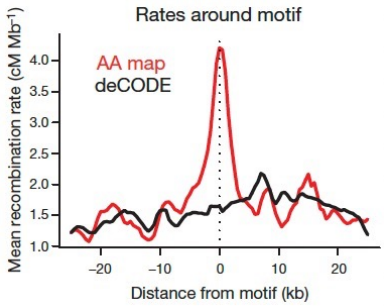


Figure 2.57: **Population differences in hotspot usage at C allele binding motifs.** *Recombination rates averaged across all instances of the C allele binding motif in African Americans (red) and Europeans (black). (Very few other genes exhibit strong functional differences across human populations; in this case it reflects the unique evolutionary pressures acting on PRDM9.)* *From Figure 3 of Anjali Hinch et al (2011) [Link] Used with permission.*

In the last part of this chapter we return to models of LD, but now with a different flavor. The new model is slightly heuristic, but much easier to work with in data analysis – and hopefully more intuitive.

Haplotype copying models. While the ARG can be considered an "exact" representation of chromosome ancestry [197], its complexity makes it extremely difficult to use in statistical analysis [198].

But in a landmark 2003 paper, Na Li and Matthew Stephens introduced an alternative framework known as a **haplotype copying model** [199] that approximates key elements of the ARG process, while being much simpler and far more computationally tractable [200]. This model has inspired many methods for a variety of important problems [201].

The central concept of the copying model is to define a *conditional sampling probability* for the "next" haplotype in a sample. Suppose that you have already observed K haplotypes in some region of the genome (by either sequencing or genotyping). You then sequence one more haplotype: before you look at the data, what would you expect this next haplotype to look like?

Intuitively, within a small region of DNA sequence *we should expect the next haplotype to be similar to (i.e., "copy") one we have already seen, but might not be identical due to occasional mutations*.

Secondly, over a larger region, we might expect that *the next haplotype will first copy one haplotype, and then switch to copy a different one, reflecting past recombination events*.

A third key point is that if we have a very large reference panel (large K), then it's more likely that the next haplotype will be similar to something we have already seen, compared to if we were using a small reference panel. So *the rate of both switches and mutations should decrease with K*.

One possible outcome from this process is illustrated here:

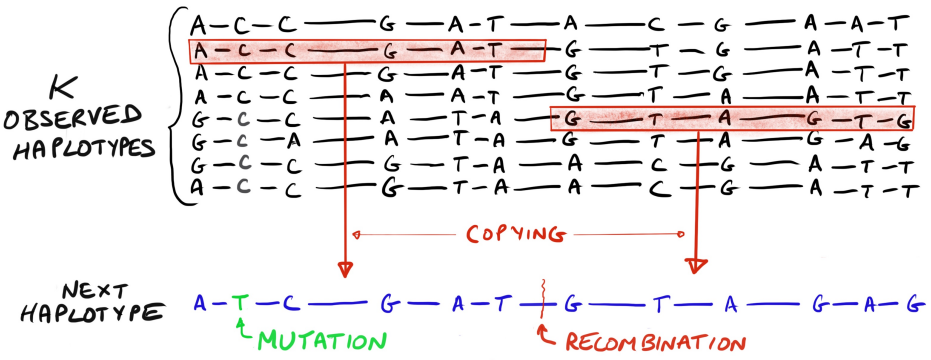


Figure 2.58: **The Haplotype Copying Model** *defines a probability distribution for the next haplotype, modeling it as a mosaic of haplotypes that have already been observed. It allows for occasional differences due to mutation or errors, as well as switches due to past recombination events.*

More formally, these ideas suggest that we could define a **conditional sampling probability**. This allows us to compute the probability of observing any specific sequence of variants as the next haplotype. Under this model, haplotypes that can be generated as simple mosaics of the previous haplotypes are more likely [202].

This model can easily handle recombination hotspots, simply by allowing a higher switch rate any time the copying process passes over a hotspot. Conversely, the model can be used to detect hotspots, as locations where haplotypes often switch parents.

The upcoming box provides some technical detail on the copying process:

Optional: The Conditional Sampling Probability for Haplotype Copying. We define the conditional sampling probability for the next haplotype as follows, assuming a reference panel of $K \geq 1$ haplotypes observed so far. For motivation and details see Li and Stephens (2003).

We focus on genotype data at a set of S SNPs, where the SNP number s ranges from 1 and S. This process defines what is known as a Markov process for the $k+1$ haplotype, conditional on the K haplotypes observed so far:

Initial parent. At the first variant position, s=1, pick a reference panel haplotype k between 1 and K, at random. (We will start copying from this haplotype.)

Next, repeat the following until $s = S$:

- **Determining the allele value.** When copying from haplotype k, we usually copy the allele in haplotype k but we allow for a low probability of single nucleotide mismatches due to mutation (or other events such as genotyping errors or gene conversion). Specifically, at site s, with probability $1-[\theta/(K+\theta)]$ the allele in the new haplotype is set to equal the allele at site s in haplotype k; otherwise we set it to an alternate allele. Here θ reflects the rate of mutations or mismatches in the data [203].
- **Recombination.** When we move from SNP s to SNP $s+1$, we decide whether to switch to copying a different haplotype [204]. Let c_s be the expected number of crossovers between these two SNPs, per generation. With probability $e^{-4Nc_s/K}$ we continue copying from the current haplotype. Otherwise, with probability $1-e^{-4Nc_s/K}$ we introduce a recombination event: in that case we select a new random haplotype parent k' between 1 and K.
- **Increment SNP position.** Set s to $s+1$.

The expression for the switch rate is motivated by noting that the average coalescence time for a new haplotype into an existing panel of K samples is $\sim 2N/K$, so the expected number of recombination events along either branch between the two SNPs is $\sim 4Nc/K$, and the probability of zero recombinations is $e^{-4Nc/K}$.

Notice that here c is measured in units of genetic distance, so it naturally allows for a higher jump rate across hotspots.

One huge advantage of copying models is that they are highly tractable for computational analysis. For example, unlike the ARG, they are amenable to efficient tools for data analysis called Hidden Markov Models (HMMs). The details of HMMs are outside our scope [205], but I'll briefly outline one major application of copying models:

Phasing and imputation. Recall that one of the main ways of collecting genome data on individuals is by genotyping. In genotyping, we measure the genotype of an individual at a pre-specified set of SNPs (com-

monly ~1 million SNPs). Until recently, genotype data has been much cheaper than genome sequencing [j]. However, these data are incomplete in two key ways:

- We do not know the genotype at SNPs that were not on the array;
- We do not know **haplotype phase**: i.e., for heterozygous SNPs, we do not know which allele goes on which homolog.

However, *we can use the concepts of LD and haplotype structure to fill in the missing data*. This is referred to as **phasing** – inferring which heterozygous alleles are from the same homologs; and **imputation** – for inferring genotypes at SNPs that were not on the array. Imputed data are valuable for many purposes as they allow us to approximate whole genome sequencing data at a fraction of the cost. Phased data are needed any time we want to analyze haplotypes and, in any event, most imputation algorithms work by phasing the data simultaneously, as I'll discuss below.

Most applications of phasing/imputation build off a panel of known haplotypes, such as data from the 1000 Genomes Project [206] to enable phasing and imputation in a new sample, as shown here:

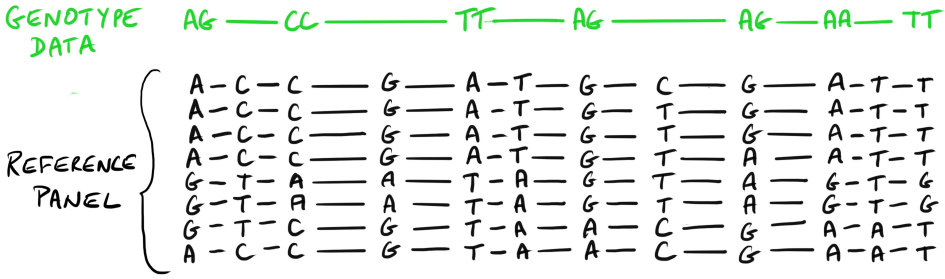


Figure 2.59: **Imputation and Phasing.** *It is common to collect genotype data on a subset of SNPs (green). With the help of a reference panel of known haplotypes (black) we wish to infer haplotype phase and impute the missing genotypes at unmeasured SNPs.*

Under the copying model, we can view the data in a diploid individual as coming from two unknown paths threading through the reference panel. The HMM machinery allows us to identify likely paths and, from this, to infer phase and missing genotypes [207]:

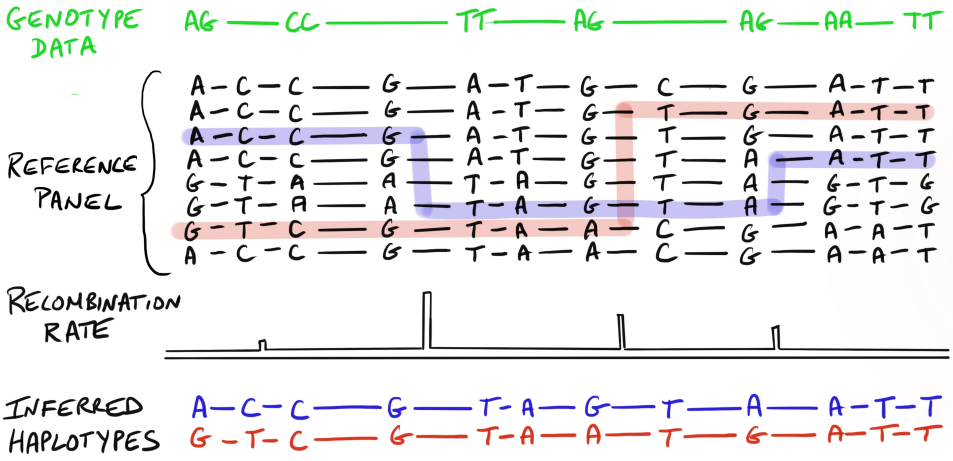


Figure 2.60: **Likely haplotypes inferred from genotype data.** *For a diploid individual, the genotype data result from two independent copying paths through the reference panel. The algorithm finds likely pairs of paths (red and blue) consistent with the genotype data; switch rates are higher at recombination hotspots. There may be multiple likely paths. Once likely paths have been identified we can infer phase and impute variants at ungenotyped SNPs (bottom).*

This type of approach provides the basic structure for how we can phase and impute data from genotypes. While more advanced techniques in-

[j] *These issues also come up for sequencing: in particular, traditional short-read sequencing does not determine haplotype phase.*

clude various bells and whistles and speedups, this type of idea has been used to analyze data from tens of millions of people.

In this chapter we've talked about how linkage, recombination and drift shape patterns of genetic variation in the genome, including LD. These processes are fundamental to understanding other aspects of human variation including natural selection and disease genetics.

Notes and References.

[174]For a short but fascinating history of Kreitman's seminal paper, see Casey Bergman's blogpost here: [Link]. The paper itself is:

Kreitman M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature. 1983;304(5925):412-7

[175]The terms recombination and crossover are often used interchangeably in the human genetics literature; however many recombination events result in local exchange of material (known as gene conversion) without crossing over. The non-crossover events are difficult to detect from genetic variation data.

[176]Genetic distances (cM) are defined in terms of the expected number of crossovers. This is a sensible way to define the distances so that they add together in a sensible way. However in a lot of practical contexts we actually want the probability of ≥ 1 crossovers. Luckily for short distances – up to about 10 cM, say – these are almost exactly the same (since double crossovers are unlikely) and we can ignore the distinction.

[177]Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 2019;363(6425):eaau1043

[178]Measures of LD and significance of r^2 for tag SNPs:

Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics. 2001;69(1):1-14;

LD scores and LD score regression:

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, of the Psychiatric Genomics Consortium SWG, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. 2015;47(3):291-5.

[179]We define c as the probability that the two alleles passed into a gamete both came from the same parent (i.e., both from the mother, or both from the father). This has the result that the maximum of c is 0.5 (and not 1 as might seem intuitive). Suppose that two SNPs are on different chromosomes, then they are transmitted independently, as predicted from Mendel's laws. In these cases the pairing of alleles is like a coin toss, so c reaches its maximum, $c = 0.5$. This is also true for SNPs on opposite ends of the same chromosome, though it is less obvious as it depends on the mechanics of chromatid pairing in meiosis.

[180]The ARG was first developed (but not really described as such) by Richard Hudson

Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 1983;23(2):183-201

A short but clear description of the ARG is presented by Nordborg 2001 [Link].

[181]Thus the number of lineages, k, forms a Markov chain over time. Since the rate of increases in linear in k, and the rate of decreases is quadratic in k, this will eventually converge to a single ancestor, known at the Ultimate Ancestor (UA). Since the UA likely predates the marginal MRCAs everywhere in the sequence, this is of mathematical but not practical interest.

[182]McVean GA. A genealogical interpretation of linkage disequilibrium. Genetics. 2002;162(2):987-91

[183]For a review of the state of the art in 2001 see Pritchard and Przeworski 2001, cited above.

[184]Pedigree studies are also greatly limited by the number of families analyzed. In this case, the authors measured recombination in 1257 meioses, or in other words, an average of 12 recombination events per cM. This means that they could get adequate estimates at Mb scale, but even with more markers they would not have been able to get a higher resolution map. In general, LD-based maps have higher resolution because they average over many more meioses (i.e., past meioses in the history of population) compared to pedigree-based maps.

[185]I'm slightly oversimplifying the historical narrative here. A few early papers suggested the presence of specific recombination hotspots based on LD data, starting as early as 1984:

Chakravarti A, Buetow K, Antonarakis S, Waber P, Boehm C, Kazazian H. Nonuniform recombination within the human beta-globin gene cluster. American Journal of Human Genetics. 1984;36(6):1239. Meanwhile, Alec Jeffreys (most famous for inventing DNA fingerprinting) and colleagues provided compelling experimental evidence for a small number of hotspots in a series of papers around 2000:

Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics. 2001;29(2):217-22

But the fact that LD patterns are mostly dictated by hotspot locations was not fully evident until a series of papers in 2001-2005.

[186]Later in the chapter I'll give some intuition for one method to estimate this, based on the Li and Stephens model.

These plots used a different approach based on McVean 2002 (cited above)

[187] McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581-4

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005;310(5746):321-4.

[188] Myers et al (2005), cited above. The originally-reported motif was CCTCCCT, although this is modified in later papers. Myers 2006.

[189] This paradox was first pointed out by Rosie Redfield and colleagues in a 1997 paper, motivated by observations from yeast.

Boulton A, Myers RS, Redfield RJ. The hotspot conversion paradox and the evolution of meiotic recombination. Proceedings of the National Academy of Sciences. 1997;94(15):8058-63

[190] Hotspot selection reference

[191] Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, et al. Fine-scale recombination patterns differ between chimpanzees and humans. Nature Genetics. 2005;37(4):429-34

Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, et al. Comparison of fine-scale recombination rates in humans and chimpanzees. Science. 2005;308(5718):107-11

[192] Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. science. 2008;319(5868):1395-8

Note: to be fair to these earlier papers, several of them invoked the possibility of an unknown trans-acting factor that might be variable within or between species, thereby explaining both varied hotspot use and a solution to the hotspot paradox. For example, Coop et al noted that "A single change in the recombination machinery could create many new hotspots in the genome, counteracting the removal of individual hotspots from the population by biased gene conversion".

[193] Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 2010;327(5967):836-40,

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science. 2010;327(5967):876-9,

Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. Science. 2010;327(5967):835-5,

Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, May CA, et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature Genetics. 2010;42(10):859-63

[194] Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, et al. The landscape of recombination in African Americans. Nature. 2011;476(7359):170-5

[195] Myers et al (2010).

[196] Recent work suggests that PRDM9 has to bind the same hotspots on both homologs for efficient crossover. For this reason, it's particularly bad to lose the *hottest* hotspots, as these are the ones most likely to have double binding. Moreover, these sites are precisely the ones that are lost most rapidly through biased gene conversion. For more on this model see

Baker Z, Przeworski M, Sella G. Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence. bioRxiv. 2022:2022-09.

[197] The ARG is "exact" in the sense that if we make a bunch of assumptions – a version of WF dynamics, a mutation, and recombination model – then it's possible to derive the ARG. But of course, any mathematical model of the world is an approximation of a more-complex reality, so you can think of the ARG as corresponding exactly to our best (but approximate) model of population genetics.

[198] There are infinitely many ARGs that can produce any given data set, and it's very difficult to compute, or even approximate, basic statistical quantities such as the likelihood.

[199] Elsewhere in the literature this model is also referred to as *Li and Stephens* or, following the original paper, the *PAC-likelihood* (for "product of approximate conditionals").

[200] Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213-33

[201] Perspective piece by Yun Song:

Song YS. Na Li and Matthew Stephens on modeling linkage disequilibrium. Genetics. 2016;203(3):1005-6.

[202]The Copying model can be thought of as a **generative** model: i.e., a specific model for the evolutionary process that generates the data. In this way it is analogous to the ARG, which is also a generative model but far more complicated.

[203]The modeling for θ is a bit complicated. The notation is motivated by the tradition definition of θ in population genetics $4N_e\mu$. But here, the expression is intended as a slightly heuristic model of the mismatch probability, and may depend on the nature of the data. For example, if we are looking at ascertained SNPs, we do know that there should be at least 1 mutation per site, somewhere within the observed genealogy, and Li and Stephens suggest scaling θ by the expected genealogy length. Furthermore, θ here is implicitly doing some extra work: it should also be able to incorporate sequencing errors, gene conversions, and other types of deviations from the copying model. You can read more about this in Li and Stephens (2003).

[204]We do this only if $s < S$

[205]HMMs are beyond the scope of this book but some googling will lead you to plenty of tutorials of different flavors, eg [Link].

[206]1000 Genomes Project: [Link];
Haplotype Reference Consortium Consortium" THR. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics. 2016;48(10):1279-83

[207]For already-phased haplotypes, the run-time is proportional to the size of the reference panel K. If we need to perform phasing at the same time, then each individual traces two paths through the reference panel, and the run-time is proportional to K^2. In practice this gets rather slow for large panels. Consequently, there has been a great deal of methods development that uses these (or similar) ideas to develop much faster algorithms.