2 Recombination

In diploid organisms recombination happens during *meiosis* (the production of gametes). Recombination mixes paternal and maternal material before it is transferred to the next generation. Each gamete that is produced by an individual therefore contains material from the maternal and the paternal side. To see what this means, take a look at your two chromosomes number 1, one of which came from your father and one from your mother. The one that stems from your father is in fact a mosaic of pieces from his mother and his father, your two paternal grandparents. In humans these mosaics are such that a chromosome is made of a couple of chunks or recombination blocks. There is generally more than one such block, but rarely more than ten per generation. Chromosomes that do not recombine are not mosaics. The Y-chromosome does not recombine at all, males inherit it completely from their father and paternal grandfather, etc. Mitochondrial DNA also does not normally recombine, both females and males inherit mitochondria from their mother, maternal grandmother, etc. The X-chromosome only recombines when it is in a female.

There are various mechanisms for recombination. The most well-known one is *crossing over*, where matching regions in homologous chromosomes (which pair during meiosis) experience a *double strand break* and subsequently are reconnected to the other chromosome (see Fig. 2.1). There are other recombination mechanisms like *gene conversion*, where a stretch of DNA is copied from one chromosome to the matching region of its homologous partner. Exchange of genetic material can also happen in haploid individuals. In this case two different individuals exchange pieces of their genome.

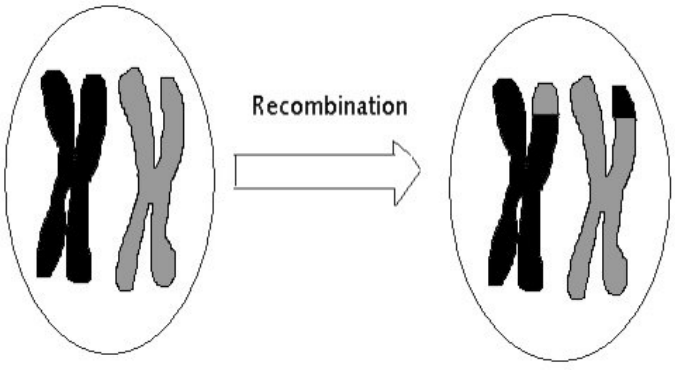

Figure 2.1: Single recombination event by *crossing over* of chromosomes during meiosis. The figure shows a pair of homologous chromosomes after the initial duplication, before and after recombination. Black and gray parts derive from different parents. Subsequently, the duplicated pair will segregate into four gametes, two recombined and two not recombined.

2.1 Linkage and linkage disequilibrium

Linkage

Mendel's second law (of *independent assortment*) states that genes are inherited independently of each other. It means that the probability of inheriting a gene at some locus $\mathcal{A}$ from one grandmother is independent of whether or not a gene at a different locus $\mathcal{B}$ has been inherited from the same grandmother. This "law" is generally only true for gene loci that are located on different chromosomes: they are *unlinked*. On the other hand, if genes are on the same chromosome, they are said to be *physically linked*. Linked genes are not inherited independently of each other. In particular, if gene loci are very close to each other, recombination between them is rare and they are typically inherited together. Mathematically, this is expressed by the *recombination fraction* $r = r_{AB}$ between loci $\mathcal{A}$ and $\mathcal{B}$, which defines the probability that genes inherited from different grandparents at these loci end up on the same parental gamete (sperm, egg, pollen) that contributes to the offspring genotype,

$$\binom{a_1 b_1}{a_2 b_2} \quad \longrightarrow \quad \begin{cases} \left.\begin{matrix} a_1 b_1 \\ a_2 b_2 \end{matrix}\right\} & \text{freq. } \frac{1}{2}(1-r) \text{ each} \\ \left.\begin{matrix} a_1 b_2 \\ a_2 b_1 \end{matrix}\right\} & \text{freq. } \frac{1}{2} \cdot r \text{ each} \end{cases} . \tag{2.1}$$

Here, a_1 and b_1 (res. a_2 and b_2) do not denote an allelic state, but only the origin of the gene either from grandparent 1 or 2.

- r is often also called a *recombination rate*, but it is really a probability in discrete generation models. We generally have $r = 1/2$ as upper limit for unlinked loci on different chromosomes and $0 \leq r < 1/2$ for linked loci.

- We can define a molecular recombination probability ρ as the probability for recombination between neighboring base pairs along a chromosome. Typical values are $\rho \approx 10^{-8}$ per generation. However, ρ generally depends strongly on the genomic position x. The estimation of recombination maps $\rho(x)$ from data is an important task of genomics.

- For a given recombination map, we can define a *recombination distance* d along a chromosome in units of *Morgans* (named after Thomas Morgan). A distance of $d = 1M$ indicates that there is on average one recombination breakpoint per generation within the stretch (e.g., due to crossing over). Typical lengths of chromosome regions measure in *centi-Morgans* (cM).

- The recombination fraction r between loci on the same chromosome is the probability of an odd number of recombination breakpoints between these loci. Ignoring interference of recombination events in neighboring regions, r relates to the recombination

distance d via *Haldane's mapping function*

$$r = \frac{1}{2}\big(1 - \exp[-2d]\big)\,. \tag{2.2}$$

Linkage disequilibrium

Assume now that there are k alleles $\{A_1, \ldots, A_k\}$ at locus $\mathcal{A}$ and l alleles $\{B_1, \ldots, B_l\}$ at locus $\mathcal{B}$. There are then $k \times l$ gametes (or haplotypes) A_iB_j with frequency denoted as $P_{A_iB_j}$. The allele frequencies derive as

$$P_{A_i} = \sum_{j=1}^{l} P_{A_iB_j}\;; \quad P_{B_j} = \sum_{i=1}^{k} P_{A_iB_j}\,. \tag{2.3}$$

As a measure of non-random association of alleles A_i and B_j at different loci on the same gamete (or haplotype), we define the *linkage disequilibrium* (LD)

$$D_{A_iB_j} = P_{A_iB_j} - P_{A_i}P_{B_j}\,. \tag{2.4}$$

If the linkage disequilibrium is zero, $D_{A_iB_j} = 0$, we say that alleles A_i and B_j are in *linkage equilibrium* (LE).

- Mathematically, D is simply the covariance of two indicator random variables that take value 1 if a randomly picked haplotype shows the corresponding allele at locus $\mathcal{A}$ resp. $\mathcal{B}$, and zero otherwise. Linkage disequilibria depend strongly on the allele frequencies and (since $P_{A_iB_j} \leq \max[P_{A_i}, P_{B_j}]$) we see that

  $$D_{A_iB_j} \leq \max[P_{A_i}(1 - P_{B_j}), P_{B_j}(1 - P_{A_i})]\,.$$

  In order to make disequilibria between different pairs of alleles better comparable, one therefore often uses the normalized measure

  $$r^2_{A_iB_j} = \frac{D^2_{A_iB_j}}{P_{A_i}(1 - P_{A_i})P_{B_j}(1 - P_{B_j})}\,, \tag{2.5}$$

  which corresponds to the (squared) correlation coefficient of the indicator variables.

- In addition to two-locus disequilibria, we can also define higher-order linkage disequilibria between alleles at three or more loci (e.g. as higher-order cross-locus cumulants, see chapter 5 of the book by R. Bürger).

- Note that *linkage* and *linkage disequilibrium* are concepts on different levels. While linkage is a property of loci and manifests in each individual, linkage disequilibrium is a population property and related to allele/haplotype frequencies. Unlinked loci can certainly have non-zero linkage disequilibria among their alleles, while alleles at linked loci (even with $r = 0$) can be in linkage equilibrium.

2.2 Two-locus model

Only recombination

Consider the two-locus model as described above. Without mutation or selection (or drift), the single-locus allele frequencies in the population stay constant, $P'_{A_i} = P_{A_i}$. However, recombination will change the haplotype frequencies. Assuming HW proportions in the germ cells prior to meiosis (and recombination), we obtain

$$P'_{A_iB_j} = (1-r)P_{A_iB_j} + r \cdot P_{A_i}P_{B_j} = P_{A_iB_j} - r \cdot D_{A_iB_j} \,. \tag{2.6}$$

Indeed, a fraction of $(1-r)$ of all gametes that contribute to the new generation has not undergone any recombination. In this part of the population, haplotype frequencies maintain their value from the previous generation. Conversely, a fraction of r of new gametes are recombination products. In HW equilibrium, the probability for them to result in a A_iB_j haplotype is $P_{A_i}P_{B_j}$. For the change in linkage disequilibrium, we obtain

$$D'_{A_iB_j} = P'_{A_iB_j} - P'_{A_i}P'_{B_j} = (1-r)P_{A_iB_j} + r \cdot P_{A_i}P_{B_j} - P_{A_i}P_{B_j} = (1-r) \cdot D_{A_iB_j} \,. \tag{2.7}$$

- We thus see that for $r > 0$ all linkage disequilibria decay to zero at geometric rate $(1-r)$. The population approaches linkage equilibrium among all alleles, $P_{A_iB_j} = P_{A_i}P_{B_j}$.
- Note that, in contrast to HW equilibrium, linkage equilibrium among alleles at different loci is *not* reached in a single generation, but only asymptotically – even for unlinked loci with $r = 1/2$.

Recombination and selection in discrete time

Consider a model with two loci under selection and focus on the case of two alleles at each locus. We can write the fitness schemes for haploid or diploid individuals as follows

	B	b
A	w_{AB}	w_{Ab}
a	w_{aB}	w_{ab}

;

	BB	Bb	bb
AA	w_{ABAB}	w_{ABAb}	w_{AbAb}
Aa	w_{ABaB}	w_{ABab}	w_{Abab}
aa	w_{aBaB}	w_{aBab}	w_{abab}

The diploid scheme assumes that the fitness of a genotype depends only on the number and type of alleles in the genotype, but not on the association of the allele to a particular haplotype (no *position effect*). I.e., the fitness of the diploid genotype (Ab, aB) is the same as the one of (AB, ab). Assuming HW proportions for diploids in zygote state, marginal fitness values for the 2-locus haplotypes follow in the usual way, $w_{AB} = w_{ABAB}\,P_{AB} + w_{ABAb}\,P_{Ab} + w_{ABaB}\,P_{aB} + w_{ABab}\,P_{ab}$, etc. The mean fitness for both haploids and diploids is

$$\bar{w} = w_{AB}\,P_{AB} + w_{Ab}\,P_{Ab} + w_{aB}\,P_{aB} + w_{ab}\,P_{ab} \,.$$

It is convenient to write the linkage disequilibrium as

$$\begin{aligned} D_{AB} &= P_{AB} - P_A P_B \\ &= P_{AB}(P_{AB} + P_{Ab} + P_{aB} + P_{ab}) - (P_{Ab} + P_{AB})(P_{aB} + P_{AB}) \\ &= P_{AB}P_{ab} - P_{Ab}P_{aB}\,. \end{aligned} \tag{2.8}$$

It is easy to verify that

$$D := D_{AB} = D_{ab} = -D_{Ab} = -D_{aB}\,.$$

Discrete time dynamics

Like for the mutation-selection model, we can construct a recombination-selection model by including both events as separate steps into a life cycle. This is best done on the level of haplotype frequencies. Indeed, with random mating, whole genotype frequencies decompose into haplotype frequencies also in a diploid population. On the other hand, haplotype frequencies do not factor into allele frequencies as long as $D \neq 0$. Starting with zygotes, we first have selection, followed by recombination during reproduction. This results in

$$P'_{AB} = \hat{P}_{AB} - r\hat{D}\,, \tag{2.9a}$$

$$\begin{aligned} D' &= (\hat{P}_{AB} - r\hat{D})(\hat{P}_{ab} - r\hat{D}) - (\hat{P}_{Ab} + r\hat{D})(\hat{P}_{aB} + r\hat{D}) \\ &= \hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{Ab}\hat{P}_{aB} - r\hat{D}\,, \end{aligned} \tag{2.9b}$$

and similar expressions for the other haplotype frequencies. $\hat{P}_{\cdot\cdot}$ and $\hat{D}$ are the values for the frequencies and for LD after selection. We have

$$\hat{P}_{AB} = \frac{w_{AB}}{\bar{w}}\,P_{AB}\,.$$

For $\hat{D}$, we need to distinguish the haploid and diploid dynamics. For haploids that recombine after random union of gametes after the selective phase, we obtain

$$\hat{D} = \hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{Ab}\hat{P}_{aB} = \frac{w_{AB}w_{ab}}{\bar{w}^2}P_{AB}P_{ab} - \frac{w_{Ab}w_{aB}}{\bar{w}^2}P_{Ab}P_{aB}$$

and thus

$$D' = (1-r)\left(\frac{w_{AB}w_{ab}}{\bar{w}^2}P_{AB}P_{ab} - \frac{w_{Ab}w_{aB}}{\bar{w}^2}P_{Ab}P_{aB}\right)\,. \tag{2.10}$$

For diploids, recombination occurs in the diploid phase after selection and we get

$$\hat{D} = \widehat{P_{AB}P_{ab}} - \widehat{P_{Ab}P_{aB}} = \frac{w_{ABab}}{\bar{w}}\big(P_{AB}P_{ab} - P_{Ab}P_{aB}\big) = \frac{w_{ABab}}{\bar{w}}\,D$$

resulting in

$$D' = \frac{w_{AB}w_{ab}}{\bar{w}^2}P_{AB}P_{ab} - \frac{w_{aB}w_{Ab}}{\bar{w}^2}P_{Ab}P_{aB} - r\frac{w_{ABab}}{\bar{w}}\,D \tag{2.11}$$

Assume that, initially, $D = P_{AB}P_{ab} - P_{Ab}P_{aB} = 0$. Eqs. (2.10) and (2.11) show that selection will create positive or negative LD, depending on the fitness values for haplotypes and on the so-called level of *epistasis*. In both cases

$$w_{AB}w_{ab} - w_{Ab}w_{aB} \begin{cases} > 0 & \text{positive epistasis, creates positive LD} \quad D' > 0 \\ = 0 & \text{no epistasis, maintains LE} \quad D' = D = 0 \\ < 0 & \text{negative epistasis, creates negative LD} \quad D' < 0\,. \end{cases} \tag{2.12}$$

For haploids, we can normalize the fitness of the *wildtype* (ab) to 1 and set

$$w_{ab} = 1\,; \quad w_{Ab} = v_A\,; \quad w_{aB} = v_B\,; \quad w_{AB} = v_A v_B + \varepsilon$$

where v_A and v_B are the single-mutant fitness values and the *epistasis parameter* measures the deviation of the double mutant fitness from the multiplicative effects of the single mutants. Obviously, $\varepsilon > 0$ ($\varepsilon < 0$) implies positive (negative) epistasis and leads to positive (negative) LD if evolution starts in LE. For the diploid case, the haplotype fitnesses are marginal fitnesses and depend on the haplotype frequencies. We can still verify that epistasis vanishes for all frequencies if the genotype fitnesses are multiplicative across loci ($w_{ABAb} = v_{AA}v_{Bb}$, etc). In contrast to the haploid case, there is more than one epistasis parameter needed to parametrize deviations from multiplicative fitnesses in the full fitness scheme.

In the absence of epistasis, we see that the LE manifold is invariant in both the haploid and the diploid case. We can thus consider the dynamics restricted to this manifold and search for equilibria. However, in general the LE manifold need not be attracting and there may be additional equilibria with $D' \neq 0$ (see the books by Bürger, chapter 2 and by Nagylaki, chapter 8 for solutions in special cases).

Continuous time dynamics

As in the case of mutation and selection we assume that selection and recombination occur in parallel and independently of each other in continuous time. This is a good approximation, in particular, if selection and recombination are both weak. For simplicity, we focus on the haploid case. We assign Malthusian fitness values to the four hapotypes, m_{ab}, m_{Ab}, m_{aB}, and m_{AB}. The dynamical equations for the haplotype frequencies read

$$\dot{P}_{AB} = P_{AB}(m_{AB} - \bar{m}) - rD \tag{2.13a}$$

$$\dot{P}_{Ab} = P_{Ab}(m_{Ab} - \bar{m}) + rD \tag{2.13b}$$

$$\dot{P}_{aB} = P_{aB}(m_{aB} - \bar{m}) + rD \tag{2.13c}$$

$$\dot{P}_{ab} = P_{ab}(m_{ab} - \bar{m}) - rD \tag{2.13d}$$

where $\bar{m}$ is the mean Malthusian fitness. We focus on the case of no epistasis. On the logarithmic scale of Malthusian fitnesses, this corresponds to additive contributions across loci. Normalizing the wildtype fitness to zero, we have

$$m_{ab} = 0\,; \quad m_{Ab} = m_A\,; \quad m_{aB} = m_B\,; \quad m_{AB} = m_A + m_B\,.$$

The dynamics of the mean fitness then becomes independent of the recombination rate,

$$\begin{aligned}\dot{\bar{m}} &= \dot{P}_{Ab}m_A + \dot{P}_{aB}m_B + \dot{P}_{AB}(m_A + m_B)\\ &= P_{Ab}m_A(m_A - \bar{m}) + P_{aB}m_B(m_B - \bar{m}) + P_{AB}(m_A + m_B)(m_A + m_B - \bar{m})\\ &= P_{Ab}(m_A - \bar{m})^2 + P_{aB}(m_B - \bar{m})^2 + P_{AB}(m_A + m_B - \bar{m})^2 + P_{ab}(0 - \bar{m})^2\,. \qquad (2.14)\end{aligned}$$

We see that mean fitness is non-decreasing (a Lyapunov function), with $\dot{\bar{m}} = 0$ if and only if the allele frequencies are at an equilibrium point. We conclude that $P_{AB}(m_{AB} - \bar{m}) = 0$ and thus with Eq. (2.13a) also $D = 0$ at each equilibrium. The dynamics of the disequilibrium is

$$\begin{aligned}\dot{D} &= \dot{P}_{AB}P_{ab} + P_{AB}\dot{P}_{ab} - \dot{P}_{Ab}P_{aB} - P_{Ab}\dot{P}_{aB}\\ &= P_{AB}P_{ab}(m_{AB} + m_{ab} - 2\bar{m}) - P_{Ab}P_{aB}(m_{Ab} + m_{aB} - 2\bar{m}) - rD\\ &= (m_A + m_B - 2\bar{m} - r)D\,. \qquad (2.15)\end{aligned}$$

Like in discrete time, the dynamics with non-epistatic fitness thus maintains LE $D = 0$. For the search of equilibrium points, we can thus restrict the dynamics to the LE manifold, where we obtain

$$\begin{aligned}\dot{p}_A = \dot{P}_{AB} + \dot{P}_{Ab} &= P_{AB}\big(m_A(1 - p_A) + m_B(1 - p_B)\big) + P_{Ab}\big(m_A(1 - p_A) - m_B p_B\big)\\ &= p_A(1 - p_A)m_A + (P_{AB} - p_A p_B)m_B\\ &= p_A(1 - p_A)m_A\,, \qquad (2.16)\end{aligned}$$

and equivalently for p_B. We see that the dynamics on the LE manifold simply reduces to the single locus dynamics.

- The result shows under which conditions the use of simple single locus models is meaningful in complex biological scenarios: If fitness epistasis is absent, all linkage disequilibria vanish in the continuous-time formulation – and thus, approximately, also in discrete time. Furthermore on the LE manifold, the multi-locus dynamics reduces to the single-locus dynamics. The long-term dynamics can thus be fully described by the single-locus formalism. Absence of epistasis is biologically plausible if both loci affect unrelated traits, where the state of one trait does not affect the fitness effect of the other trait. one locus thus does

- Unless epistasis is very strong and/or linkage very tight, the mutilocus dynamics usually converge to a parameter range very close to the LE manifold. One can then solve the problem under the assumption of LE first and treat linkage disequilibria as a perturbation. This is the idea of the *quasi linkage equilibrium* approximation, see e.g. the book by Bürger.