Topological linkage disequilibrium calculated from coalescent genealogies

Johannes Wirtz, Martina Rauscher, Thomas Wiehe*

Institut für Genetik, Universität zu Köln

Abstract

We revisit the classical concept of two-locus linkage disequilibrium (LD) and introduce a novel way of looking at haplotypes. In contrast to defining haplotypes as allele combinations at two marker loci, we concentrate on the clustering of sampled chromosomes induced by their coalescent genealogy. The root of a binary coalescent trees defines two clusters of chromosomes. At two different loci this assignment may be different as a result of recombination. We show that the amount of shared chromosomes among clusters at two different loci, measured by the squared correlation, constitutes a natural measure of LD . We call this *topological LD* (tLD) since it is induced by the topology of the coalescent tree. We find that its rate of decay decreases more slowly with distance between loci than that of conventional LD . Furthermore, tLD has a smaller coefficient of variation, which should render it more accurate for any kind of mapping purposes than conventional LD . We conclude with a practical application to the LCT region in human populations.

Keywords: Linkage Disequilibrium, Topological Linkage Disequilibrium, Kingman Coalescent, Ancestral Recombination Graph, Sequential Markov Coalescent

1. Introduction

Maintenance and decay of haplotype structure and allelic association are of prominent interest when studying the evolutionary dynamics of recombining chromosomes. One possibility of quantifying this association is by the well-known concept of *Linkage Disequilibrium* (LD), reviewed in several arti-

*Corresponding author

Email addresses: wirtzj0@uni-koeln.de (Johannes Wirtz), martina.rauscher@uni-koeln.de (Martina Rauscher), twiehe@uni-koeln.de (Thomas Wiehe)

2 Materials and Methods

cles, for instance [1, 2, 3]. LD measures the degree of statistical dependence among alleles at two - or more - loci and is commonly defined via a combination of haplotype frequencies. In the simplest non-trivial case two alleles at each of two loci can be distinguished. In the framework of the polarized infinite sites model [4], two allele classes - derived and ancestral - can be distinguished for each polymorphic site. The derived alleles can be traced back to a single mutation event, marking the origin of the derived class at this locus. This event may lie on any branch of the underlying coalescent tree [5] and divide the leaves in two classes: those belonging to the subtree rooted by the mutation, and those belonging to the complement. In this work, we consider the most recent common ancestor, i.e. the root of the entire coalescent tree, rather than of a single mutation event as such a class marker, which divides leaves into a 'left' (from the root) and a 'right' class. Along a chromosome tree topology may change as a consequence of recombination and thereby re-shuffle leaves among the two classes. This gives rise to an extended definition of linkage disequilibrium: considering two coalescent trees along a recombining chromosome, we call the correlation between left and right class members *topological linkage disequilibrium*, tLD for short. While tLD is a theoretically simple concept, coalescent topology is in practice generally unknown and has to be estimated from polymorphism data. Still, tLD has a number of interesting properties. For instance, tLD decays more slowly with distance than conventional LD , because only a subset of recombination events affects tree topology at the root. This is of practical interest when measuring long-range LD and searching for possible interactions between chromosomally distant loci. To estimate tree topology several SNPs - on the order of about ten or more - are required. However, this jointly considering of SNPs, rather than calculating LD for all SNP-pairs, introduces a smoothing effect compared to conventional LD . In fact, the coefficient of variation (CoV) of tLD is much smaller than that of LD , especially for short distances between loci. For larger distances, the CoV of tLD is almost constant and only slightly larger than 1, while the CoV of conventional LD is around 3, indicating a much higher dispersion relative to tLD (Figure 6).

2. Materials and Methods

2.1. Linkage Disequilibrium

Consider a population of constant size $2N$ chromosomes and let α, β be two loci with alleles a, A and b, B with allele frequencies $p(a), p(A), p(b)$ and $p(B)$. Let the four haplotypes ab, aB, Ab and AB have frequencies $p(ab) = x_1, p(aB) = x_2, p(Ab) = x_3$ and $p(AB) = x_4$. Two-locus linkage

2 Materials and Methods

Figure 1: Alignment of $2N$ haplotypes $i_1, \dots, i_{2N}$, indicated by vertical lines, with two bi-allelic loci. Frequencies given below the parentheses.

disequilibrium is

$$D_{\alpha,\beta} := x_1x_4 - x_2x_3. \quad (1)$$

This can also be written as $D_{\alpha,\beta} = x_1 - (x_1 + x_2)(x_1 + x_3) = p(ab) - p(a) \cdot p(b)$. A configuration of $x_1, \dots, x_4$ such that $D_{\alpha,\beta} = 0$, is called *linkage equilibrium*. In this case, all haplotype frequencies are identical to the product of the involved allele frequencies. In practice, allele and haplotype frequencies in the population must usually be estimated from sample frequencies. In what follows, we do not distinguish between samples and populations and view all frequencies as population frequencies.

Since LD depends on allele frequencies, several standardizations have been introduced. One is Pearson's correlation coefficient of the allelic association

$$r_{\alpha,\beta} := \frac{D_{\alpha,\beta}}{\sqrt{(p(a)p(A)p(b)p(B))}} \quad (2)$$

which can take values between -1 and 1 , where the sign depends on labeling of alleles. Furthermore,

$$D_{\alpha,\beta} = 0 \iff r_{\alpha,\beta} = 0,$$

and

$$r_{\alpha,\beta}^2 = \frac{D_{\alpha,\beta}^2}{(p(a)p(A)p(b)p(B))} \quad (3)$$

is the χ^2 -value of the allelic association. An alternative, frequently used, standardization is

$$D'_{\alpha,\beta} = \frac{|D_{\alpha,\beta}|}{\min\{p(A)(1 - p(B)), p(B)(1 - p(a))\}}, \quad (4)$$

2 Materials and Methods

known as "Lewontin's D " [2]. All three quantities, $r_{\alpha,\beta}$, $r_{\alpha,\beta}^2$ and $D'_{\alpha,\beta}$ have been used to estimate recombination rates, divergence times between species, or to statistically test the neutral evolution hypothesis [6, 7, 8]. Here, we focus on $r_{\alpha,\beta}^2$.

Let c be the recombination probability per generation per chromosome between loci α and β . The random variable $r_{\alpha,\beta}^2$ changes due to the action of drift and recombination.

We consider first its expectation with respect to c . When $\alpha = \beta$, we have $x_1 = p(a) = p(b)$ and $x_4 = p(A) = p(B)$, and therefore

$$\mathbb{E}(r_{\alpha,\alpha}^2) = \mathbb{E}\left(\frac{((p(a)p(A))^2)}{((p(a)p(A))^2)}\right) = \mathbb{E}(1) = 1.$$

A fortiori, two identical loci do not recombine, i.e. $c = 0$. However, this may still hold also if $\alpha \neq \beta$, for instance in non-recombining chromosomes. In this case LD may take any value D_0 (r_0^2) and, on average, remains at this value, regardless of time and initial allele frequency distribution, i.e.

$$\mathbb{E}(r_{\alpha,\beta}^2(c = 0)) = r_0^2. \quad (5)$$

Let x_i (D) denote haplotype frequency (LD) in the current generation and x'_i (D') denote haplotype frequency (LD) in the next generation. We have

$$\mathbb{E}(x'_i|x) = x_i + \eta_i c D, i = 1, \dots, 4,$$

where $\eta_1 = \eta_4 = -1$ and $\eta_2 = \eta_3 = +1$, and therefore

$$\mathbb{E}(D'|D) = (1 - c)D,$$

expected LD decreases at rate c per generation for any value $c > 0$. Since, on average, allele frequencies do not change across generations, we have

$$\mathbb{E}((r_{\alpha,\beta}^2)'|r_{\alpha,\beta}^2) = (1 - c)^2 r_{\alpha,\beta}^2 \approx (1 - 2c) r_{\alpha,\beta}^2,$$

where the last approximation holds for small c . As c increases $\mathbb{E}(r_{\alpha,\beta}^2)$ decreases. If alleles from both loci are assembled into haplotypes completely independently from the previous generation (i.e., $c = 1$, as the theoretical maximum), the expected value in this limiting case is

$$\mathbb{E}(r_{\alpha,\beta}^2) = \frac{1}{2N - 1}, \quad (6)$$

as originally derived by Haldane [9]. We state

2 Materials and Methods

Lemma 1. Consider a sample of $n \in \mathbb{N}$ chromosomes with two loci α, β and with fixed allele frequencies $f(a) = s, f(A) = n - s$ at α and $f(b) = u, f(B) = n - u$ at β . Assume that chromosomes with allele a and allele b are chosen uniformly and independently, and that the assignment of a -alleles at α is independent from b -alleles at β . Then,

- (a) $\mathbb{E}(r_{\alpha,\beta}^2) = \frac{1}{n-1}$.
- (b) This also holds if the allele frequencies s and u are randomly chosen according to discrete distributions p_α, p_β on $1, \dots, n-1$.
- (c) In the case of fixed allele frequencies (as in (a)) one has

$$\text{Var}(r_{\alpha,\beta}^2) = \frac{\kappa - 1}{(n-1)^2},$$

where κ is the fourth standardized moment (kurtosis) of a hypergeometric random variable.

Proof. In the scenario described in Lemma 1 (a), the number of individuals of type (a, b) is distributed as a hypergeometric $H(n, s, u)$ random variable X . Thus, the expected squared correlation can be expressed by the formula

$$\mathbb{E}(r_{\alpha,\beta}^2) = \sum_{k=0}^s \frac{\binom{u}{k} \binom{n-u}{s-k}}{\binom{n}{u}} \frac{n^2 (k - s \frac{u}{n})^2}{s(n-s)u(n-u)}$$

The denominator of the term $\frac{n^2 (k - s \frac{u}{n})^2}{s(n-s)u(n-u)}$ is independent of k and may be extracted from the summation. The remainder of the summation can then be written as

$$\sum_{k=0}^s \frac{\binom{u}{k} \binom{n-u}{s-k}}{\binom{n}{u}} n^2 \left(k - s \frac{u}{n}\right)^2 = n^2 \text{Var}(X) = n^2 \frac{s(n-s)u(n-u)}{n^2(n-1)},$$

such that the initial equation can be transformed to

$$\mathbb{E}(r_{S,U}^2) = \frac{s(n-s)u(n-u)}{s(n-s)u(n-u)(n-1)} = \frac{1}{n-1}$$

To prove Lemma 1 (b), it suffices to note that the same expectation with respect to the distributions of s and u can be expressed by

$$\mathbb{E}_{p_\alpha, p_\beta}(r_{\alpha,\beta}^2) = \sum_{s=1}^{n-1} \sum_{u=1}^{n-1} p_\alpha(s) p_\beta(u) \left(\sum_{k=0}^s \frac{\binom{u}{k} \binom{n-u}{s-k}}{\binom{n}{u}} \frac{n^2 (k - s \frac{u}{n})^2}{s(n-s)u(n-u)} \right).$$

2 Materials and Methods

Since both distributions integrate to 1, the result remains unchanged. Finally, we obtain Lemma 1 (c) by first writing

$$\text{Var}(r_{\alpha,\beta}^2) = \mathbb{E}(r_{\alpha,\beta}^4) - \mathbb{E}(r_{\alpha,\beta}^2)^2 = \mathbb{E}(r_{\alpha,\beta}^4) - \frac{1}{(n-1)^2}.$$

Then,

$$\mathbb{E}(r_{\alpha,\beta}^4) = \sum_{k=0}^s \frac{\binom{u}{k} \binom{n-u}{s-k}}{\binom{n}{u}} \frac{n^4 (k - s \frac{u}{n})^4}{(s(n-s)u(n-u))^2},$$

which we can identify as

$$\frac{1}{(n-1)^2} \mathbb{E} \left(\frac{X - \mu}{\sigma} \right)^4$$

where X is the hypergeometric random variable mentioned above, $\mu = \frac{su}{n}$ is its expectation and $\sigma^2 = \frac{s(n-s)u(n-u)}{n^2(n-1)}$ its variance. Therefore, the last expression is $\frac{1}{(n-1)^2} \kappa_X$, and κ_X is the kurtosis of X . □

Determining $\mathbb{E}(r_{\alpha,\beta}^2)$ with respect to c for the non-limiting case is non-trivial. Sved [3] proposed an approximate solution by relating $\mathbb{E}(r_{\alpha,\beta}^2)$ to the *conditional probability of linked identity by descent*, denoted by a parameter Q . He obtained the approximation

$$\mathbb{E}(r_{\alpha,\beta}(c)^2) \approx \frac{1}{1 + 4Nc^{\frac{1-c}{2}}} \approx \frac{1}{1 + 4Nc}, \quad (7)$$

where the last approximation holds for $c \ll 1$ (see also [10]).

Despite the later attempts (e.g. [11]) to improve this formula, the unsatisfactory discrepancy between eqs (6) and (7) still persists.

2.2. Topological Linkage

A slightly modified concept of linkage disequilibrium can be defined very intuitively in the framework of coalescent theory. Consider a sample of n recombining chromosomes of unit length and with recombination rate $\rho = 2Nc > 0$ between the ends of the chromosome. The genealogical history of such a sample can be modelled by the *ancestral recombination graph* (ARG) A_n [12]. The projection of the ARG on any point $\gamma \in [0, 1]$ yields a Kingman coalescent G_γ . Trees at different positions γ and γ' may be different due to changes of tree topology by recombination. In fact, $(G_\gamma)_{\gamma \in [0,1]}$ can be viewed as a (non-Markovian) stochastic process on the set of Kingman

2 Materials and Methods

coalescent genealogies with state changes caused by recombination events. The genealogy G_γ persists for some nonempty interval $[a \leq \gamma, \gamma \leq b)$ with expected length $(\sum_{k=1}^{n-1} \rho/k)^{-1}$. We call $S = [a, b)$ a *segment* of the chromosome and denote its genealogy by G_S .

As a Markovian approximation of the ARG, the so-called *Sequential Markov Coalescent* (SMC) was introduced by [13]. In this construction A_n as the unified history of coalescence and recombination events of the sample is omitted. Instead, genealogies change along the chromosome by uniformly choosing a branch of the current genealogy, removing the subtree below and replacing it somewhere else in the tree. This *prune-regraft* operation will be described in more detail below.

The SMC-construction is considered a reasonably accurate approximation of the ARG process, with the advantage that computation is much easier. Under the SMC-construction $(G_\gamma)_{\gamma \in [0,1]}$ becomes a continuous-time Markov chain, where the distance between changes in G_γ is the average segment length $(\sum_{k=1}^{n-1} \rho/k)^{-1}$, as in the ARG. For the remainder of this manuscript we assume that $(G_\gamma)_{\gamma \in [0,1]}$ is a realization of the SMC.

Any genealogy G_S extracted from $(G_\gamma)_{\gamma \in [0,1]}$ and valid on some segment S naturally provides a separation of the sample into two disjoint sets: Let S_1 contain all sample members found on the "left-hand" side, and S_2 all those on the "right-hand" side of the root node of G_S . Note that "left" and "right" can also be interpreted as two different alleles, originating by a point mutation along one of the root branches. Moving from segment S to another segment U along the sequence $(G_\gamma)_{\gamma \in [0,1]}$, the genealogical tree G_U may differ from tree G_S as a result of recombination. As a consequence, also the left (U_1) and right (U_2) descendants below the root of G_U may differ from S_1 and S_2 . Correlation can be measured in a way analogous to conventional linkage disequilibrium: Let

$$p(S_i) = |S_i|/n, \quad p(U_i) = |U_i|/n \quad (i = 1, 2)$$

and

$$x_1 = |S_1 \cap U_1|/n, \quad x_2 = |S_1 \cap U_2|/n, \quad x_3 = |S_2 \cap U_1|/n, \quad x_4 = |S_2 \cap U_2|/n.$$

Then we call

$$r_{S,U}^2 = \frac{(x_1x_4 - x_2x_3)^2}{p(S_1)p(S_2)p(U_1)p(U_2)} \quad (8)$$

topological linkage disequilibrium of the segments S and U , in short *tLD*.

The choices of "left" and "right" are arbitrary, as much as the naming of alleles, and does not affect the value $r_{S,U}^2$. Note, that we have $r_{S,U}^2 = 1$, if

80 and only if $S_1 = U_1$ or $S_1 = U_2$, where segments S and U may be different.

3 Theory

Figure 2: Example of the topological association of twelve chromosomes 1 – 12 in two random coalescent trees. The above configuration yields an r^2 -coefficient of 0.04375.

Such a configuration is called *complete linkage*.

tLD combines the concepts of LD and coalescent. Although tLD and LD have similar properties, they differ, for instance, in their rate of decay with distance. While any recombination event and point mutation may affect LD , tLD is affected only by topological changes at the root of a coalescent tree. We explore this in more detail now.

3. Theory

3.1. Limit behaviour of tLD in the SMC

Disregard for a moment branch lengths of individual coalescent trees G_γ of the SMC and consider only their branching pattern. Such trees are also called *coalescent tree topologies* (see [14] and Figure 3). The set $\mathcal{T}_n$ of coalescent tree topologies of n chromosomes has size $|\mathcal{T}_n| = n!(n-1)!2^{1-n}$ [15]. Omitting branch lengths, but introducing integer-labellings of internal tree nodes which respect the time-order of these branchings, Kingman's coalescent induces the uniform distribution on $\mathcal{T}_n$. Note that the sequence of topologies, $(T_\gamma)_{\gamma \in [0,1]}$ with $T_\gamma \in \mathcal{T}_n$ for $\gamma \in [0,1]$ that is induced by the SMC, is a *finite-state* Markov chain.

The k -th *layer* of a coalescent tree topology T is the tree slice in which it has k branches (Figure 3). In particular, the 1-st layer is the (imaginary) slice with only one branch, which extends from the root back into the past. Treating each segment of a branch from the original Coalescent genealogy extending between two layers $k, k+1$ as a separate branch in the associated Coalescent tree topology, each topology of size n has exactly $\frac{n(n+1)}{2}$ branches. Under the SMC, T_γ switches states at recombination sites γ according to a

3 Theory

Figure 3: *Left*: Example of a Coalescent tree topology (T_1). It is a binary tree equipped with an integer labelling representing the order of coalescent events backward in time, and identifiers (A-D) for the chromosomes at the bottom. *Middle, Right*: A *srARG* of size 4, represented by its associated topologies T_2 and T_3 . A subtree of T_2 is selected for pruning (p) and is regrafted (r) at a layer equal or smaller than where T_2 was pruned. Note that the prune-regraft operation affects the internal labelling of T_3 .

prune-regraft operation: select a branch at layer k , say, to place a pruning site on the tree; cut the underlying subtree; select a branch at a layer less than k to place a re-grafting site; re-attach the cut subtree at the regrafting site. We call such a transition a *single recombination SMC*, for short *srSMC*.

110 Assuming an 'infinite recombination site' model, the complete *SMC* can be viewed as a sequence of *srSMCs*.

Consider a random *srSMC* on n chromosomes and extract the topologies T_1 and T_2 of the two coalescent genealogies G_1 and G_2 . T_1 and T_2 are still uniformly distributed, but they are not independent. Given a random topology T_1 , assume a pruning event in layer k at a randomly chosen branch b_k . Assume the branch b_j for regrafting to be chosen *uniformly* from the $\frac{k(k+1)}{2}$ possible branches. A *srSMC* can be viewed as a triplet (T_1, b_k, b_j) . To obtain the conditional distribution $P(T_2|T_1 = T)$ one has to average over all possible pruning and re-grafting sites. The probability of pruning in layer $k > 1$ is given by (see [14])

$$\Pr(\text{pruning in level } k) = \frac{(k-1)^{-1}}{h_{n-1}}, \quad (9)$$

where $h_{n-1} = \sum_{i=1}^{n-1} (1/i)$ is the $(n-1)$ -st harmonic number, keeping in mind that this probability does depend on the duration of the layer, i.e. on the coalescent branch lengths. The prune-regraft-operation describes how states

3 Theory

115 in the Markov chain $(T_\gamma)_{\gamma \in [0,1]}$ are changed. In particular, the set of all *Aldous moves* on a topology T_γ (see [16]) is a subset of the possible transformations T_γ can undergo. It follows that $(T_\gamma)_{\gamma \in [0,1]}$ is recurrent and aperiodic. We now consider the leftmost and rightmost chromosomal segments $S = [0, \gamma_1]$ and $U = [\gamma_2, 1]$, $\gamma_1 \leq \gamma_2$, along the interval $[0, 1]$. We have the following
120 result on the limiting behaviour of the squared correlation $r_{S,U}^2$.

Theorem 1. *Given segments S and U with topologies T_S and T_U and the topological groupings (S_1, S_2) and (U_1, U_2) . Then*

$$\mathbb{E}(r_{S,U}^2(\rho)) \rightarrow \frac{1}{n-1}$$

for $\rho \rightarrow \infty$.

Proof. $(T_\gamma)_{\gamma \in [0,1]}$ is a Markov chain with stationary distribution $\Pr^*(T) = \frac{2^{n-1}}{n!(n-1)!}$ for all coalescent tree topologies T . Since it is recurrent, there exists an integer $M \in \mathbb{N}$ (*mixing time*), depending only on the sample size n , such that for $\epsilon > 0$

$$\|\Pr(T_U = T|T_S) - \Pr^*(T)\|_{\mathbb{L}^1} = \sum_{T \in \mathcal{T}_n} |\Pr(T|T_S) - \Pr^*(T)| \leq \epsilon$$

given there are M or more state changes in $(T_\gamma)_{\gamma \in [0,1]}$. By choosing ρ sufficiently large, it is possible to ensure that the probability of M or more changes is close to 1. Therefore, $\Pr(T_U = T|T_S)$ may be brought arbitrarily
125 close to $\Pr^*(T)$.

Let $\rho \rightarrow \infty$ and consider the random variable $k_\rho = |S_1 \cap U_1|$ of individuals that are on the left of both trees T_S and T_U under the *SMC* with recombination rate ρ . As $\rho \rightarrow \infty$, k_ρ converges in distribution to a random variable X which, given $|S_1|$ and $|U_1|$, is distributed $X_{|S_1|,|U_1|} \sim \text{Hyp}(n, |S_1|, |U_1|)$. $|S_1|$ and $|U_1|$ themselves are uniformly distributed on $\{1, \dots, n-1\}$.
130

Therefore, Lemma 1 is applicable to the limiting r.v. X . Convergence in distribution of the k_ρ implies convergence of $\mathbb{E}(r_{S,U}^2)$ to $\frac{1}{n-1}$. $\square$

Over large distances, we therefore recover the limiting value predicted by Haldane [9], see eq (6). Note that by item (c) of Lemma 1 we can also obtain
135 an exact expression of the variance of *tLD* in this situation.

3.2. Decline of *tLD* with distance

Regarding the expected decline of *tLD* along the chromosome, we state the following

3 Theory

Lemma 2. *Let T_1, T_2 be the two topologies resulting from a srSMC. The probability of a topological change, i.e. breaking of complete linkage, between T_1 and T_2 is asymptotically $\frac{1}{3} + \mathcal{O}\left(\frac{1}{\log(n)}\right)$.*

Proof. Recombination events that have an effect on tLD can be subdivided into two groups: events which shift a non-root branch above the root or events which move a branch from the left to the right (or vice versa) root-subtree without changing the root. We call the latter *switching events*. The probability of root-changing events is $\mathcal{O}\left(\frac{1}{\log(n)}\right)$ [14]. To calculate the probability $\Pr(\text{switch})$ assume w.l.o.g. that a branch is moved from left to right. Suppose pruning takes place in layer k . The right side has $1 \leq r \leq k-1$ branches, where each number r has probability $1/(k-1)$. The probability of selecting a branch on the left for pruning is $\frac{k-r}{k}$, and some arithmetics leads to the probability of selecting a branch on the right for regrafting, which is $\frac{(k-1)(r+1)}{k(k+1)}$ when averaged over all k -sized coalescent tree topologies. This needs to be multiplied by $\frac{1}{kh_{n-1}}$ (compare equation 9), and then summed over all levels k . Therefore, we obtain

$$\Pr(\text{switch}) = 2 \sum_{k=2}^n \sum_{r=1}^{k-1} \frac{k-r}{k} \frac{(k-1)(r+1)}{k(k+1)} \frac{1}{k-1} \frac{1}{kh_{n-1}}$$

where the factor 2 accounts for the two possibilities, switching from left to right or vice versa. After some simplifications, this can be rewritten as

$$\Pr(\text{switch}) = \frac{1}{3} + \frac{\sum_{k=2}^n \frac{2}{k^2(k+1)} + \frac{2}{3k^2}}{h_{n-1}}$$

The series in the numerator converges to $(4\pi^2 - 33)/9 < 1$ and $h_n \sim \log(n)$, which conclude the proof. $\square$

Knowing the proportion of recombination events contributing to the decay of $r_{S,U}^2$, we calculate now the expected proportion of chromosomes affected by a switching event between two segments. Denote by $L_{S,U}$ the number of chromosomes whose assignment to either the left or right class has *not* been affected by switching. Note that the quantity $L_{S,U}$ is similar to Q in [3]. However, we are here interested in a *joint* probability of identity by descent, and not in a probability conditioned on identity at one of the loci. $L_{S,U} = n$ means perfect linkage and $L_{S,U} < n$ means some switching took place (Figure 4). Lemma 1 is not readily applicable here, because its conditions require that $L_{S,U}$ vanishes.

Lemma 3. $\mathbb{E}(L_{S,U})$ declines geometrically with rate approximately $\frac{1}{3 \log(n)}$ per recombination event.

3 Theory

Figure 4: A single recombination event moves a branch of the left root-subtree to the right side. In the resulting tree, chromosomes marked by red dots remain in the same left/right-grouping as before recombination took place. Their number is $L_{S,U}$.

Proof. According to Lemma 2 about $1/3$ of all recombination events move some individuals from one side of the tree to the other. To determine how many are moved, it suffices to note that recombination events are distributed uniformly over the branches of a given genealogy T in the *SMC*, such that the size of the subtree T_r below a recombination event is distributed according to the neutral frequency spectrum [17]. Thus,

$$\Pr(|T_r| = k) \sim \frac{1}{k \cdot h_{n-1}}. \quad (10)$$

The expected proportion of chromosomes affected by a recombination event is the expectation of the above distribution divided by n , i.e. $h_{n-1}^{-1} \approx (\gamma + \log(n))^{-1}$, where γ is Euler's constant. We can combine those two results in a recursion formula for $\mathbb{E}(L_{S,U})$. Let U_1, U_2 be two neighboring segments, then we have

$$\begin{aligned} \mathbb{E}(L_{U_1, U_2}) &= 2 \cdot \frac{1}{3} \cdot \left(\mathbb{E}(L_{U_1, U_1}) + \frac{1}{2} \mathbb{E}(L_{U_1, U_1}) \left(1 - \frac{1}{\gamma + \log(n)} \right) \right) \\ &= \mathbb{E}(L_{U_1, U_1}) \left(1 - \frac{1}{3(\gamma + \log(n))} \right) \end{aligned}$$

Iterating this formula with initial value

$$\mathbb{E}(L_{S,S}) = n$$

shows that

$$\mathbb{E}(L_{S,U}) = n \left(1 - \frac{1}{3(\gamma + \log(n))} \right)^{c_{S,U}}$$

with $c_{S,U}$ representing the number of recombination events encountered moving along the chromosome from S to U , which depends only on ρ . $\square$

3 Theory

Figure 5: At segment S , the sample members are grouped by tree topology into left (*white*) and right (*black*). Moving from S to U , a number of $L_{S,U} = l + m$ members remains unaffected by recombination, while the rest of the members might switch sides.

3.3. Numerical approximation of $\mathbb{E}(r_{S,U}^2)$

The above results about the decline of the parameter $L_{S,U}$ suggest an approximation scheme for the expectation of tLD with respect to the number of recombination events separating two segments. $L_{S,U}$ can be written as $L_{S,U} = l + m$, where l (m , respectively) is the number of $L_{S,U}$ -chromosomes on the left (right) side of both trees. There are $p = |S_1| - l$ additional individuals on the left side of T_S and $q = |U_1| - l$ on the left side of T_U . See Figure 5 for an example.

To calculate $r_{S,U}^2$, we need to determine how many of those additional p chromosomes on the left side of T_S are also on the left side of T_U by chance. We assume that this number is given by a hypergeometric distribution. This is an approximation, because in both *ARG* and *SMC* the number of possible left-right configurations is smaller than the number of all hypergeometrically possible configurations. However, this restriction becomes negligible with increasing distance between S and U , i.e. with increasing number of recombination events. Let k denote the number of individuals which end up on the left side of T_S and T_U by this hypergeometric assignment, such that there are $k + l$ chromosomes in total that are on the left side of both trees.

Under these assumptions, the expected tLD is

$$\mathbb{E}(r_{S,U}^2 | p, q, l, m) = \sum_{k=0}^q \frac{\binom{p}{k} \binom{n-p-L_{S,U}}{q-k}}{\binom{n-L_{S,U}}{q}} \cdot \frac{\left(\frac{k+l}{n} - \frac{(p+l)(q+l)}{n^2} \right)^2}{\frac{|S_1|}{n} \cdot \left(1 - \frac{|S_1|}{n} \right) \frac{|U_1|}{n} \cdot \left(1 - \frac{|U_1|}{n} \right)}.$$

3 Theory

It is useful to note that the term

$$\sum_{k=0}^q \frac{\binom{p}{k} \binom{n-p-L_{S,U}}{q-k}}{\binom{n-L_{S,U}}{q}} \cdot k^2$$

is the second moment of a hypergeometric variable with parameters $n - L_{S,U}, p, q$ and with expectation $\frac{qp}{n-L_{S,U}}$. We calculate therefore

$$\frac{(p+l)(q+l)}{n^2} - \frac{qp}{n \cdot (n-L_{S,U})} = \frac{\frac{npq}{L_{S,U}-n} + (l+p)(l+q)}{n^2}$$

and use this to rewrite the original expression as

$$\begin{aligned} \mathbb{E}(r_{S,U}^2 | p, q, l, m) &= \sum_{k=0}^q \frac{\binom{p}{k} \binom{n-p-L_{S,U}}{q-k}}{\binom{n-L_{S,U}}{q}} \cdot \frac{\left(\frac{k}{n} - \frac{qp}{n \cdot (n-L_{S,U})} - \frac{\frac{npq}{L_{S,U}-n} + (l+p)(l+q)}{n^2} + \frac{l}{n} \right)^2}{\frac{|S_1|}{n} \cdot \left(1 - \frac{|S_1|}{n}\right) \frac{|U_1|}{n} \cdot \left(1 - \frac{|U_1|}{n}\right)} \\ &= \sum_{k=0}^q \frac{\binom{p}{k} \binom{n-p-L_{S,U}}{q-k}}{\binom{n-L_{S,U}}{q}} \cdot \left(\frac{\left(\frac{k}{n} - \frac{qp}{n \cdot (n-L_{S,U})} \right)^2}{\frac{|S_1|}{n} \cdot \left(1 - \frac{|S_1|}{n}\right) \frac{|U_1|}{n} \cdot \left(1 - \frac{|U_1|}{n}\right)} \right. \\ &\quad \left. - 2 \frac{\left(\frac{k}{n} - \frac{qp}{n \cdot (n-L_{S,U})} \right) \cdot \left(\frac{\frac{npq}{L_{S,U}-n} + (l+p)(l+q)}{n^2} - \frac{l}{n} \right)}{\frac{|S_1|}{n} \cdot \left(1 - \frac{|S_1|}{n}\right) \frac{|U_1|}{n} \cdot \left(1 - \frac{|U_1|}{n}\right)} \right. \\ &\quad \left. + \frac{\left(\frac{\frac{npq}{L_{S,U}-n} + (l+p)(l+q)}{n^2} - \frac{l}{n} \right)^2}{\frac{|S_1|}{n} \cdot \left(1 - \frac{|S_1|}{n}\right) \frac{|U_1|}{n} \cdot \left(1 - \frac{|U_1|}{n}\right)} \right). \end{aligned} \quad (\#)$$

This expression can be simplified. The middle term of the summation (line # above) vanishes because of symmetry; the first summand contains the variance of a $Hyp(n-l-m, p, q)$ random variable divided by some constant terms. Thus, we arrive at

$$\mathbb{E}(r_{S,U}^2 | p, q, l, m) = \frac{\frac{pq}{n-L_{S,U}} \cdot \left(1 - \frac{p}{n-L_{S,U}}\right) \frac{n-L_{S,U}-q}{n-L_{S,U}-1} \cdot \frac{1}{n^2}}{\frac{|S_1|}{n} \cdot \left(1 - \frac{|S_1|}{n}\right) \frac{|U_1|}{n} \cdot \left(1 - \frac{|U_1|}{n}\right)} \quad (\text{HYP})$$

$$+ \frac{\left(\frac{\frac{npq}{L_{S,U}-n} + (l+p)(l+q)}{n^2} - \frac{l}{n} \right)^2}{\frac{|S_1|}{n} \cdot \left(1 - \frac{|S_1|}{n}\right) \frac{|U_1|}{n} \cdot \left(1 - \frac{|U_1|}{n}\right)}. \quad (\text{PSP})$$

(11)

In this form, the contribution that arises from the hypergeometric random assignment, labelled HYP, and the remaining parameter-specific (PSP) terms

3 Theory

are separated. This decomposition is useful in two ways. First, under the above assumptions, an upper bound can be obtained for $\mathbb{E}(r_{S,U}^2)$ at least if $L_{S,U}$ is small in relation to n (See Lemma 4). Second, by averaging over all configurations it is possible to approximate $\mathbb{E}(r_{S,U}^2)$ regardless of the tree topologies at segments S and U . Since the size $|S_1|$ of the left side of T_S is uniform on $\{1, \dots, n-1\}$, we start by choosing $|S_1|$ randomly according to the uniform distribution. The $L_{S,U}$ -sized portion of individuals not having undergone recombination when going from S to U is then subdivided into l individuals which are on the left side both in T_S and T_U , and m individuals which are on the right side in both trees by choosing hypergeometrically from the assignment at S , which implicitly determines the parameters p, l and m . The number q of additional individuals on the left side of T_U is determined by drawing uniformly from $\{1, \dots, n - L_{S,U}\}$. These calculations are easily performed by computational algebra. Note that the explicit calculation of SNP-based LD is much harder because the sizes of the classes are not uniform. The resulting approximation of the expected tLD with respect to $L_{S,U}$ has to be scaled with respect to the expected decay of $L_{S,U}$ itself. Assuming that recombination events are uniformly distributed across a chromosome, the approximation of $\mathbb{E}(r_{S,U}^2)$ can be expressed in terms of physical distance in base pairs, given a constant recombination rate per bp.

3.4. An upper bound to $\mathbb{E}(r_{S,U}^2)$

Lemma 4. Assume that the parameter $L_{S,U}$ is small in relation to the sample size n . Then, we have:

- (a) The expectation of the parameter-specific contribution ("PSP") in equation 11 is of order $\mathcal{O}\left(\left(\frac{L_{S,U} \log(n)}{n}\right)^2\right)$.
- (b) The proposed approximation of $\mathbb{E}(r_{S,U}^2)$ is bounded from above by $\frac{1}{n-1} + c_n \left(1 - \frac{1}{3 \log(n)}\right)^i$, where i denotes the number of recombination events separating S and U and c_n is some constant.

Proof. Under the assumption of $L_{S,U} < n$, it is possible to write $\frac{n}{n-L_{S,U}} = \sum_{i=0}^{\infty} \left(\frac{L_{S,U}}{n}\right)^i$. Furthermore, $l \cdot (n - p - q - l) \in [-l \cdot n, l \cdot n]$. This allows us to

3 Theory

rewrite the numerator of the parameter-specific term in the following way:

$$\begin{aligned} & \left(\frac{(l+p)(l+q) - \frac{npq}{n-L_{S,U}}}{n^2} - \frac{l}{n} \right)^2 \\ &= \left(\frac{pq \sum_{i=1}^{\infty} \left(\frac{L_{S,U}}{n} \right)^i + l(n-p-q-l)}{n^2} \right)^2 \\ &\leq \left(\sum_{i=1}^{\infty} \left(\frac{L_{S,U}}{n} \right)^i + \frac{l}{n} \right)^2 \\ &\in \mathcal{O} \left(\left(\frac{L_{S,U}}{n} \right)^2 \right) \end{aligned}$$

The last statement is true because $l/n \leq L_{S,U}/n$.

Under the assumptions from section 3.3, and if $L_{S,U}$ is small in comparison to n , then the sizes $|S_1|$ and $|U_1|$ of the left sides of the genealogies are approximately independent and uniformly distributed on $\{1, \dots, n-1\}$. Thus, the expectation of the denominator in

$$\mathbb{E} \left(\frac{1}{\frac{|S_1|}{n} \cdot \left(1 - \frac{|S_1|}{n}\right) \frac{|U_1|}{n} \cdot \left(1 - \frac{|U_1|}{n}\right)} \right),$$

converges to

$$\frac{n^4}{(n-1)^2} \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} \left(\frac{1}{k(n-k)l(n-l)} \right) = 4 \frac{n^2}{(n-1)^2} h_{n-1}^2$$

as $L_{S,U}/n$ becomes small. This term is of order $\mathcal{O}(\log(n)^2)$, allowing us to conclude that

$$\mathbb{E}(\text{PSP}) \in \mathcal{O} \left(\left(\frac{L_{S,U}}{n} \right)^2 \cdot \log(n)^2 \right),$$

establishing claim (a).

To show (b), we recall Hölder's inequality

$$\mathbb{E}(X^2) \leq \mathbb{E}(X) \cdot \max X,$$

for a nonnegative random variable X . Let $X = \frac{L_{S,U}}{n}$, assuming exactly i recombination events between S and U . The maximal value of this random variable is 1 (no decline at all). The expectation of X given the number of

4 Simulations and Application

recombination events i between S and U has been calculated in section 3.2 as $\left(1 - \frac{1}{3\log(n)}\right)^i$. Thus, $\mathbb{E}\left(\left(\frac{L_{S,U}}{n}\right)^2 | i\right) \leq 1 \cdot \left(1 - \frac{1}{3\log(n)}\right)^i$, and therefore

$$\mathbb{E}(\text{PSP}|i) \leq c_n \cdot \left(1 - \frac{1}{3\log(n)}\right)^i,$$

with $c_n \in \mathbb{R}$ depending on n and of maximally (squared) logarithmic growth.

190 The expectation of the hypergeometric contribution is 0 for $L_{S,U} = n$ and converges to $\frac{1}{n-1}$ from below for $L_{S,U}/n \rightarrow 0$, finishing the proof. $\square$

The above calculations hold in the setting assumed in 3.3 and are an approximation of the true *SMC*. Furthermore, since the term $4h_{n-1}^2$ is not bounded from above, this upper bound is only of relevance for large n .

195 4. Simulations and Application

We use the program `ms` [18] to generate samples of n chromosomes with recombination rate r and mutation rate θ . Computing *tLD* requires two steps. First, select two disjoint loci (sub-intervals of $[0, 1]$). Second, determine the top-most left and right root sub-tree clusters for both loci. We consider two possibilities: (a) use the true tree structure provided by the coalescent simulations (option "-T" selected) or (b) estimate the clusters from SNP data. In the latter case, this is achieved by a two-means clustering approach: We first determine the two most diverged haplotypes ('antipodes') and then assign the remaining haplotypes to either of the antipodes based on minimal Hamming distance. The estimated clusters agree well with the true (simulated) clusters if sufficient SNPs are available for estimation. A minimum number of 10 SNPs gives good results (M. Rauscher, unpubl. data; see Figure 7). This is confirmed by the excellent agreement of the summary statistics (average and variance) of *tLD* determined from actual and estimated clusters (see Figure 6) and by the very good agreement of individual *LD* estimates (see Figure heatmap). Hence, true and estimated values of *tLD* agree in general. Importantly, this observation does not depend much on distance. However, the correlations depend heavily on the relation between mutation and recombination rates.

215 Comparing *tLD* with conventional *LD* in simulated data, we find that *tLD* declines comparably slowly and smoothly, although the variance is large. In contrast, conventional *LD* vanishes very quickly over similar distances (Figure 6). Although average *LD* is small, its variance is high, in particular for short distances. In this regime the variance of *tLD* is much smaller relative to the average than that of conventional *LD* (coefficient of variation

4 Simulations and Application

Figure 6: *tLD* vs. SNP-*LD*. Average (top left), variance (top right), coefficient of variation (bottom left) and index of dispersion (bottom right) of r^2 . Data from a single simulation run performed with the program `ms` [18]. Parameter settings: `ms 200 1 -t 100 -r 100 1000 -T`. For a population size of $N = 10^4$ and a recombination rate of $1\text{cM}/\text{Mb}$ the simulated region corresponds to 0.25cM or 250kb physical distance. Red: *tLD* calculated from the actual coalescent trees (i.e., using the trees obtained by setting the parameter `-T`). Orange: *tLD* calculated from estimated tree topology (see text). Blue: Conventional *LD* calculated from SNP pairs. Coefficient of variation: σ/μ ; index of dispersion σ^2/μ .

σ/μ shown in Figure 6). These observations are theoretically supported by Lemma 2.

In practice tree topology is unknown. To estimate tree topology at the tree root, we apply the two-means clustering approach described above. We use the human $1k$ genomes data [19] to determine *tLD* across the LCT region on chromosome 2 in the CEU (Central Europe) and YRI (Yoruba) populations. We estimate *tLD* using a window size of 5kb per locus. Most of these windows contain 10 or more SNPs. We find a strongly elevated level of *tLD* in the CEU population compared to YRI. As can be seen from the heat maps, there is a much higher, and a longer-ranging level of correlation to be observed for *tLD* than for conventional *LD* (Figure 8).

4 Simulations and Application

Figure 7: Heatmaps of true tLD calculated from tree topologies and tLD calculated from estimated tree topologies, performed on the same simulated dataset used in Figure 6

4 Simulations and Application

Figure 8: The heatmaps of conventional (left) and topological LD generated for the LCT and neighbouring regions of the CEU (top) and YRI subsamples of the human 1k genomes project [19].

5 Discussion

Figure 9: *Left:* $\mathbb{E}(r_{S,U}^2)$ in a sample of size $n = 100$ with respect to $L_{S,U}$ declining at rate $1/(3 \log(100))$ with respect to the number of recombination events encountered in the sample. *Right, Dots:* Average tLD between pairs of segments with respect to their distance in units of 2500bp (segment size: 5000bp), estimated on the region expanding over approx. $5.51 \cdot 10^6 - 5.52 \cdot 10^6$ on chromosome 5 of the human 1k genomes data (CEU). *Blue Line:* Least-squares fit of expected tLD , estimating $c = \frac{\rho}{4N}$ at $\approx 1.116cM/Mb$.

5. Discussion

We have introduced the concept of topological linkage disequilibrium (tLD). Like conventional LD it is a means to analyse and quantify chromosomal linkage. In contrast to the latter tLD can in principle be calculated for arbitrary pairs of loci, even if not polymorphic, because the concept hinges on coalescent tree topology, or subtree clusters, not on allele pairs. In practice, however, tree topology is usually not directly observable and needs to be estimated from polymorphism data. A moderate number of SNPs, on the order of ten, turns out to be sufficient to determine the two root-subtree clusters (assuming a binary coalescent tree) quite reliably (Figure 7). Given SNP data, we use a 2-means clustering approach, described in more detail above. Even with a single SNP, the correct root-subtree clusters may be recovered. In fact, this happens with probability roughly equal to $(1 + 1/2)/h_{n-1}$, i.e. with the probability that a SNP lands on either one of the two root branches. This is about $1/2$ for $n = 12$ and about $1/4$ for $n = 227$. Alternatively, one may also use other kinds of markers, for instance microsatellites [20], or structural variants, if available, to estimate the root-subtree clusters.

As a function of distance between loci (or markers) average tLD and conventional LD behave qualitatively similarly. Both decay toward the same

5 Discussion

limit, as expressed in Haldane’s equation (see eq 6). However, tLD decays more slowly, and is less dispersed, than conventional LD (Figure 6). The coefficient of variation of tLD is only slightly larger than one, almost independently of distance between loci, and only about a third of that of conventional LD . Conversely, the inverse of the coefficient of variation, in signal processing also called ‘signal to noise ratio’, is about three times as high for tLD than for conventional LD .

The concept of tLD is embedded in coalescent theory. The first moment of tLD and its limiting behavior can be analytically approximated by a simple function using arguments derived from (Kingman-)coalescent properties. A currently open problem is to integrate the concept of tLD into ancestral recombination graph (ARG) theory. Here, we have resorted to its Markovian approximation, represented by the SMC to derive the presented results. We hypothesize that similar results must hold for the ARG in terms of limiting behaviour. In the ARG setting, the decay of tLD is possibly somewhat slower than in the SMC setting, because in the ARG genealogies can be reverted back with a higher probability than in the SMC .

Given that the various multi-genome resequencing projects furnish now whole-genome sequences with sample sizes of several hundred or more, the concept of tLD should prove useful for any analysis of linkage disequilibrium and gene interactions in such experimental studies.

Acknowledgments

This work was supported by the German Research Foundation (DFG-SPP1590).

- [1] M. Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, 9(6):477–485, 2008. doi:10.1038/nrg2361.
- [2] R. C. Lewontin. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, 49(1):49–67, 1964.
- [3] J. A. Sved. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, 2(2):125–141, 1971.
- [4] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, 1969.

5 Discussion

- [5] J. F. C. Kingman. On the genealogy of large populations. *Journal of applied probability*, 19:27–43, 1982. doi:10.2307/3213548.
- [6] J. A. Sved, A. F. McRae, and P. M. Visscher. Divergence between human populations estimated from linkage disequilibrium. *Am. J. Hum. Genet.*, 83(6):737–743, 2008. doi:10.1038/nrg2361.
- [7] S. Ennis. Linkage disequilibrium as a tool for detecting signatures of natural selection. *Methods Mol. Biol.*, 376:59–70, 2007. doi:10.1007/978-1-59745-389-9_5.
- [8] B. P. McEvoy, et al. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res.*, 21(6):821–829, 2011. doi:10.1101/gr.119636.110.
- [9] J. B. S. Haldane. The mean and variance of χ^2 when used as a test of homogeneity, when expectations are small. *Biometrika*, 31(3):3–4, 1940. doi:10.2307/2332614.
- [10] S. Boitard and P. Loisel. Probability distribution of haplotype frequencies under the two-locus wright-fisher model by diffusion approximation. *Theoretical Population Biology*, 71(3):380–391, 2007.
- [11] U. Ober, et al. The expected linkage disequilibrium in finite populations revisited. *ArXiv e-prints*, 2013.
- [12] R. C. Griffiths and P. Marjoram. *An ancestral recombination graph*, pages 257 – 270. Springer, 1997. ISBN 0-387-94944-5. doi:10.1007/978-1-4757-2609-1_16.
- [13] G. A. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1387–1393, 2005. doi:10.1098/rstb.2005.1673.
- [14] L. Ferretti, F. Disanto, and T. Wiehe. The effect of single recombination events on coalescent tree height and shape. *PLoS ONE*, 8(4), 2013. doi:10.1371/journal.pone.0060123.
- [15] F. Murtagh. Counting dendrograms: A survey. *Discrete Applied Mathematics*, 7(2):191 – 199, 1984. ISSN 0166-218X. doi:https://doi.org/10.1016/0166-218X(84)90066-0.
- [16] D. J. Aldous. Mixing time for a Markov chain on cladograms. *Combinatorics, Probability and Computing*, 9(3):191–204, 2000.

5 Discussion

- [17] J. Wakeley. *Coalescent Theory, an Introduction*. Roberts and Company, Greenwood Village, CO, 2005. ISBN 978-0974707754.
- [18] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 2:337–338, 2002.
- [19] A. Auton, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi:10.1038/nature15393.
- [20] H. Li and T. Wiehe. Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. *PLoS Comput Biol*, 9(5):e1003060, 2013. doi:10.1371/journal.pcbi.1003060.