A Graphical Approach to Multi-Locus Match Probability Computation: Revisiting the Product Rule

Yun S. Song[a] and Montgomery Slatkin[b]

[a]Department of Computer Science, University of California, Davis, CA 95616, USA

[b]Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

E-mail addresses: yssong@cs.ucdavis.edu (Y.S. Song) and slatkin@berkeley.edu (M. Slatkin)

Corresponding Author:

Yun S. Song

Department of Computer Science

University of California at Davis

2063 Kemper Hall

One Shields Avenue

Davis, CA 95616

U.S.A.

E-mail: yssong@cs.ucdavis.edu

Phone: +1 530 754 9586

Fax: +1 530 752 4767

To Appear in Theoretical Population Biology

Abstract

The genealogical relationships of individuals in a finite population can create statistical non-independence of alleles at unlinked loci. In this paper, we introduce a flexible graphical method for computing the probabilities that two individuals in a finite, randomly-mating population have the same haplotype or genotype at several loci. This method allows us to generalize the analysis of Laurie and Weir (2003) to cases with more loci and other models of mating. We show that monogamy increases the probabilities of genotypic matches at unlinked loci and that the effect of monogamy increases with the number L of loci. We conjecture a sharp upper bound on the effect of monogamy for a given L.

Keywords: match probability; product rule; unlinked; linkage disequilibrium; monogamy; match graph

1 Introduction

The probability of a complete genotypic match of two unrelated individuals at two or more unlinked loci is of importance to the forensic use of DNA typing. The question that often arises is the extent to which a genotypic match at several unlinked loci between a suspect and a blood or other sample from a crime scene indicates that the suspect is the source of the crime-scene sample (Evett and Weir, 1998). The standard procedure in US criminal courts is to assume that the probability of a genotypic match between two unrelated individuals in the same population can be obtained by assuming statistical independence of the loci. With that assumption, the probability a genotypic match at all loci, called the random match probability (RMP), is obtained by multiplying the probabilities of genotypic matches at each locus, which are obtained from Hardy-Weinberg frequencies (Evett and Weir, 1998). This assumption, which is called the *product rule* in US courts, is the basis for computing such low RMPs that juries are usually convinced that a suspect whose genotype matches that from a crime-scene sample at several loci was indeed at the crime scene.

The product rule is based on the well-established population genetics theory that shows that recombination in an infinite population eliminates statistical dependence between pairs of loci, i.e., linkage disequilibrium (LD). In finite populations, however, genealogical relationships between unrelated individuals can create LD even between unlinked loci. For two loci the effect is very small (Hill and Robertson, 1968; Ohta and Kimura, 1969). Although this result supports the use of the product rule, it does not ensure that consistent deviations from the predictions of the product rule will not emerge when more than two loci are considered together. At present, 13 tetranucleotide microsatellite loci, called the Combined DNA Index System (CODIS) loci, are generally typed in the US and many other populations (the CODIS web-site is http://www.fbi.gov/hq/lab/codis/index1.htm). Because there are 78 pairs of CODIS loci, it is possible that subtle LD between each pair could result in substantial errors in the RMP for all 13 loci. In a detailed study of a very large data set of

genotypes at 9 loci, Weir (2004) found approximate agreement between the numbers of individuals who had the same genotypes at 5 of 9 loci and the predictions of the product rule, provided that a large enough correction (denoted θ) for excess homozygosity was assumed.

Laurie and Weir (2003) presented a way to compute the probability that two unrelated individuals match at two and three loci in a finite randomly mating population. They showed that the product rule works quite well unless the mutation rate to new neutral alleles is unreasonably high. Their results are obtained from a system of coupled linear recurrence equations. The equilibrium match probabilities are found by assuming stationarity.

Although the method of Laurie and Weir (2003) is simple in principle, setting up the systems of recurrence equations becomes increasingly difficult for more than two unlinked loci. For the standard Wright-Fisher model of random mating, Laurie and Weir succeeded in computing the genotypic match probability for two loci and the haplotypic match probability for two and three loci, but they concluded that finding the genotypic match probability for more than two loci or the haplotypic match probability for more than three loci, "would be combinatorially very difficult."

In this paper, we develop a simpler and more flexible framework for computing match probabilities. Using this framework, we can consider more than three loci and other models of mate choice. Our strategy is to represent match probabilities in terms of graphs. By performing a set of prescribed operations on a given graph at generation t, we determine how it is related to a linear combination of graphs at generation $t-1$. The graphical method makes the combinatorial structure of the problem easier to understand. For constructing the required systems of equations, it is possible to implement our method in a fully automated program, thus reducing the chance of human error in finding the recurrence equations for a particular model. We have written such a program in *Mathematica* that can compute genotypic match probabilities for up to three loci and haplotypic match probabilities for up to five loci. It should be possible to analyze more loci by

implementing our algorithm in a faster programming language such as *C*. If mutation rates at all loci are the same, then certain match probabilities become equal; this reduction in the number of independent variables should allow us to handle about twice as many loci.

In addition to the standard Wright-Fisher model of random mating, we consider a mating scheme with perfect monogamy. We show that the effect of monogamy on the L-locus match probability increases as L increases. Furthermore, for a given number of loci, we conjecture sharp upper bounds on the effect of monogamy on the haplotypic and genotypic match probabilities.

This paper is organized as follows. The models considered in this paper are described in Section 2. Our graphical framework is described in detail in Section 3, where we explain the correspondence between match probabilities and graphs, as well as the operations that one needs to perform on the graphs. Simple examples are provided in Section 4 and the main results on match probabilities are discussed in Section 5, where we also describe an approximation method and discuss the aforementioned sharp upper bounds on the effect of monogamy on match probabilities. We conclude with discussion in Section 6.

2 Model Description

Some frequently used symbols are listed in Table 1. Throughout, we assume a neutral infinite-alleles model for a single population containing N diploid individuals where N is assumed to be large. By a *gamete*, we simply mean a collection of loci; different loci may physically reside on different chromosomes. We assume that generations are non-overlapping and that mutations occur at locus i with probability μ_i per gamete per generation, independently of other loci.

We use x_i to denote the allele at locus i in gamete x. When many gametes are considered, a superscript is sometimes used to distinguish different gametes. For example, x_i^k denotes the allele at locus i in gamete x^k. Our convention differs from that of Laurie and Weir (2003), who use

Table 1: Frequently used notation.

Notation	Explanation
$2N$	Number of gametes in each generation.
L	Number of loci.
μ_i	Per gamete per generation mutation rate at locus i.
x_i	Allele at locus i in either a haplotypic or a genotypic sequence (it will be clear which from context).
$\boldsymbol{x}$	A haplotypic or a genotypic sequence $\boldsymbol{x} = x_1 x_2 \ldots x_L$.
$x_i \equiv y_i$	Allele x_i matches allele y_i.
$\boldsymbol{x} \equiv \boldsymbol{y}$	Allele x_i matches allele y_i for all loci $i = 1, \ldots, L$.
$\mathbb{P}_h(x_i \equiv y_i)$	One-locus haplotypic match probability for locus i.
$\mathbb{P}_h(\boldsymbol{x} \equiv \boldsymbol{y})$	L-locus haplotypic match probability.
$\mathbb{P}_g(x_i \equiv y_i)$	One-locus genotypic match probability for locus i.
$\mathbb{P}_g(\boldsymbol{x} \equiv \boldsymbol{y})$	L-locus genotypic match probability.
R_h^U, R_h^M	The ratio $\mathbb{P}_h(\boldsymbol{x} \equiv \boldsymbol{y}) / \prod_{i=1}^{L} \mathbb{P}_h(x_i \equiv y_i)$ under *unconstrained* and *perfect monogamy* mating schemes, respectively.
R_g^U, R_g^M	The ratio $\mathbb{P}_g(\boldsymbol{x} \equiv \boldsymbol{y}) / \prod_{i=1}^{L} \mathbb{P}_g(x_i \equiv y_i)$ under *unconstrained* and *perfect monogamy* mating schemes, respectively.

subscripts to denote gamete labels. In their notation a_i denotes the allele at locus a in gamete i.

2.1 Mating schemes

How gametes in the next generation are produced from those in the current generation depends on the assumed mating scheme. In this paper we consider the following two random mating schemes:

Unconstrained mating: Randomly sample two gametes, each with replacement. The same gamete may be sampled twice under this mating scheme. A new gamete is produced as a mosaic of the two samples (as described below). This is the standard Wright-Fisher model and the work of Laurie and Weir (2003) pertains to this model. With probability μ_i, the offspring gamete has an allele at locus i that has never been seen before.

Perfect monogamy: Before sampling, first randomly partition the $2N$ gametes into a set of N disjoint pairs. To create an offspring gamete, randomly sample a pair from the set of pairs, replacing the pair after sampling. As in unconstrained mating, a new gamete is produced as a mosaic of the two sampled gametes (see below), and with probability μ_i, the offspring gamete has an allele at locus i that has never been seen before. Unlike in unconstrained mating, the two parental gametes are always different gametes, though they may be identical by state.

2.2 Inheritance pattern of the offspring gamete

Two loci: Let x_1x_2 and y_1y_2 denote the two sampled parental gametes. Then, the inheritance pattern of the offspring gamete is x_1x_2, y_1y_2, x_1y_2, or y_1x_2, with probability $\frac{1}{2}(1-r)$, $\frac{1}{2}(1-r)$, $\frac{1}{2}r$, or $\frac{1}{2}r$, respectively. Note that $r = 1/2$ corresponds to the case of unlinked loci.

More than two loci: Let $x_1x_2 \ldots x_L$ and $y_1y_2 \ldots y_L$ denote the two sampled parental gametes with L loci. For ease of discussion, we focus on a set of loci that are pairwise unlinked, as was done previously by other authors (Strobeck and Golding, 1983; Laurie and Weir, 2003). Hence, in the offspring gamete $z_1z_2 \ldots z_L$, the allele z_i at locus i is equally likely to have descended from x_i or y_i. The probability of any particular inheritance pattern is $1/2^L$.

3 Graphical Framework: Overall Idea

In this section, we lay out our strategy, explaining the correspondence between match probabilities and graphs, and that between the events in the assumed reproduction model and certain operations on graphs. In the previous section, we described a forward perspective on genealogy. Here, we adopt a backward point of view and determine how a match probability at generation t is related to a combination of match probabilities at generation $t-1$. Henceforward, L denotes the number of loci.

Figure 1: Examples of *fully-labeled* graphs. Vertex labels correspond to gamete labels and edge labels denote loci. The graph G_1 represents the match probability $\mathbb{P}(x_1 \equiv y_1, x_2 \equiv y_2, x_3 \equiv z_3)$, whereas G_2 represents $\mathbb{P}(x_1 \equiv y_1, x_2 \equiv y_2, y_3 \equiv z_3)$. Ignoring the vertex labels, these graphs are isomorphic as *edge-labeled* graphs. Under random mating, $\mathbb{P}(x_1 \equiv y_1, x_2 \equiv y_2, x_3 \equiv z_3) = \mathbb{P}(x_1 \equiv y_1, x_2 \equiv y_2, y_3 \equiv z_3)$, and G_1 and G_2 are considered equivalent.

3.1 Graphical representation of match probabilities

We use $x_i \equiv y_i$ to denote that alleles at locus i are identical in gametes x and y. To a particular match probability (e.g., the probability of $(x_i \equiv y_i) \wedge (x_j \equiv z_j) \wedge (y_k \equiv z_k)$), we associate a *fully-labeled* graph as follows:

- **Vertex:** Create a vertex labeled x for gamete x.
- **Edge:** Draw an edge labeled i between vertices x and y if and only if $x_i \equiv y_i$.

For example, shown in Figure 1 are two graphs G_1 and G_2 which correspond to the match probabilities $\mathbb{P}(x_1 \equiv y_1, x_2 \equiv y_2, x_3 \equiv z_3)$ and $\mathbb{P}(x_1 \equiv y_1, x_2 \equiv y_2, y_3 \equiv z_3)$, respectively. Under random mating, note that these two probabilities are equal. More generally, any two match probabilities are equal under random mating if they are related by some permutation of the gamete labels. In terms of our graphical representation, this equality of match probabilities translates to the following equivalence relation: two *fully-labeled* graphs (i.e., all vertices and edges are labeled) are equivalent if they are isomorphic as *edge-labeled* graphs (i.e., ignoring vertex labels). In Figure 1, G_1 and G_2 are equivalent since they are isomorphic as edge-labeled graphs. In terms of this graphical framework, our objective is as follows.

Main Goal: To develop a graphical method of setting up systems of equations that correctly relate *edge-labeled* graphs, in the same way that corresponding match probabilities are related.

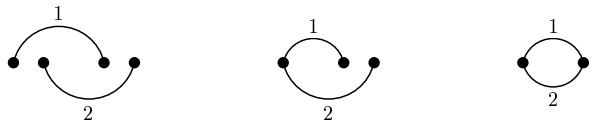

Figure 2: Two-locus match probabilities each proportional to $(1-\mu_1)^2(1-\mu_2)^2$.

3.2 Mutations (Vertex Count)

Let $\{x_i^1, x_i^2, \ldots, x_i^k\}$ denote a set of alleles at locus i in k gametes at time t. Under an infinite-alleles model, the alleles $\{x_i^1, x_i^2, \ldots, x_i^k\}$ all match only if their parental alleles at time $t-1$ all match and no mutation occurs between times $t-1$ and t in the lineages relating $\{x_i^1, x_i^2, \ldots, x_i^k\}$ to their parents. Hence, the probability of any match relation at time t that requires $x_i^1 \equiv x_i^2 \equiv \cdots \equiv x_i^k$ must contain an overall factor of $(1-\mu_i)^k$ when written in terms of match probabilities at time $t-1$. This fact translates to the following statement in our graphical representation:

Given a graph G, let $V(G)$ denote the set of all vertices in G, and, for $v \in V(G)$, define

$$\delta_i(v) := \begin{cases} 1, & \text{if at least one edge labeled } i \text{ is incident with } v, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

That is, $\delta_i(v)$ is an indicator variable that says whether the gamete associated with vertex v is involved in a match relation at locus i. The total number of gametes involved in match relations at locus i is denoted by $\delta_i(G) := \sum_{v \in V(G)} \delta_i(v)$. When relating G to graphs in the previous generation, there will be an overall factor of

$$\prod_{i=1}^{L} (1-\mu_i)^{\delta_i(G)}.$$

For instance, each of the graphs shown in Figure 2 has $\delta_1(G) = \delta_2(G) = 2$, so the corresponding probability of each graph is proportional to $(1-\mu_1)^2(1-\mu_2)^2$.

3.3 Inheritance pattern across loci for each gamete (Vertex Split)

Here, we consider only a single gamete at time t and investigate the inheritance pattern across its loci. When more than one gamete is considered at time t, we also need to consider how they can share parental gametes. That will be discussed in the next subsection.

By "δ-degree" of a vertex v, we mean the sum $\sum_{i=1}^{L} \delta_i(v)$, where $\delta_i(v)$ is defined in (1); it is equal to the total number of *distinctly* labeled edges incident with v. In the graphs corresponding to haplotypic match probabilities, each edge label appears at most once, so the δ-degree of any vertex coincides with its ordinary degree, the total number of edges incident with the vertex.

Two loci: Consider the case of two loci. Let x and y denote the two gametes sampled at time $t-1$, giving rise a child gamete h at time t. With probability r, one of the two loci in h has descended from x and the other from y, while with probability $1-r$, both loci in h have descended from a single parental gamete.

Let R denote a match relation at time t and G the corresponding match graph. If only one of the two loci in a gamete is involved in R (e.g., in $R = (x_1 \equiv y_1) \wedge (y_2 \equiv z_2)$, locus 2 of gamete x is not involved in the match relation. Similarly, locus 1 of gamete z is not involved in the match relation.), then, since we only need to track ancestral loci, we do not need to consider the possibility of the gamete having two parental gametes. Suppose that both loci in gamete h are involved in R, so that the vertex labeled h in G has δ-degree 2. If gamete h has two parental gametes, each contributing one locus to h, then that is represented in our graphical framework by splitting the vertex h into two vertices, distributing the edges that used to be incident with h such that each new vertex has δ-degree 1. An example is shown on the left hand side of Figure 3.

A graph obtained from splitting *zero or more* δ-degree-2 vertices in G is called a *split graph* of G, and G is called a *pivot graph*. The two new vertices that result from a vertex split are called a *split pair*. If G contains at least one δ-degree-2 vertex, then more than one inequivalent split graph

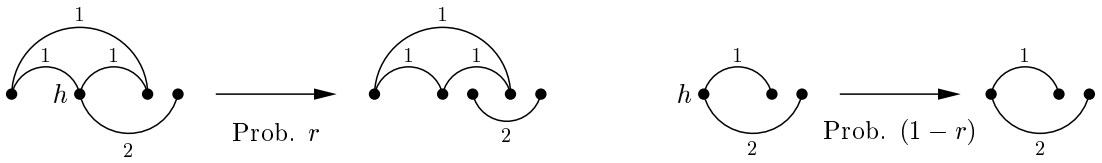

Figure 3: Illustration of vertex split operations on match graphs for two loci. Vertex h has δ-degree 2. On the left hand side, vertex h is split into two vertices, and the edges that used to be incident with h are divided between the two new vertices such that each new vertex has δ-degree 1. On the right hand side, zero vertex split operation is performed.

can be obtained. Note that a split graph is only an intermediate graph that is useful for relating a pivot graph at time t to a set of relevant match graphs at time $t-1$.

More than two loci: Suppose that $L > 2$. For ease of discussion, we focus on a set of loci that are pairwise unlinked. A case with linked loci can easily be accommodated in our framework by introducing more parameters (recombination rates) and putting constraints on vertex split operations.

Let $D = \{1, 2, \ldots, n\}$, where $n \leq L$, denote the set of *distinct* loci in gamete h that are involved in a match relation R. Let $B_1 \sqcup B_2$ denote a bipartition of D into two disjoint subsets, such that the loci in B_1 and those in B_2 come from different parental gametes. (Note that if the bipartition is $\varnothing \sqcup D$, then effectively there is only one parental gamete.) There are 2^{n-1} inequivalent bipartitions of D, and we assume that each bipartition has probability $1/2^{n-1}$. In the graph G corresponding to R, the vertex labeled h has δ-degree n, and the bipartition of D into $\{i_1, \ldots, i_k\} \sqcup \{i_{k+1}, \ldots, i_n\}$ corresponds to splitting h into two vertices v_1 and v_2, such that of all edges that used to be incident with h in G, those that had labels in B_i now becomes incident with v_i, for $i = 1, 2$. An example is shown in Figure 4.

3.4 Sharing of parental gametes (Vertex Merge)

As described above, a vertex split operation is used to capture that a gamete at time t has inherited at least one locus from each of the two sampled gametes at time $t-1$ (c.f., Section 2.1). We now

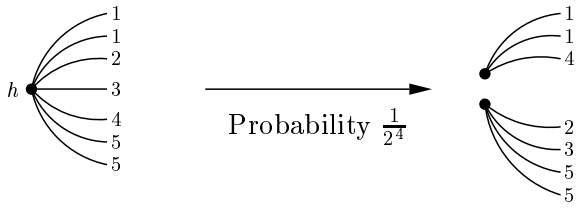

Figure 4: A split of a δ-degree-5 vertex in a model with unlinked loci. This vertex split corresponds to a bipartition of $\{1, \ldots, 5\}$ into $\{1, 4\}$ and $\{2, 3, 5\}$. These are not entire graphs; only the parts relevant for illustrating a vertex split are shown here.

need to consider the possibility of a gamete at time $t-1$ being a common parental gamete of two or more gametes at time t. This sharing of a parental gamete translates to merging relevant vertices in the split graph into a single vertex. The precise pattern of allowed sharing of parental gametes depends on the assumed mating scheme, and so do the allowed set of vertex merge operations and their associated probabilities. In what follows, we adopt the following convention:

Convention 1 *When a set of vertices merge into a single vertex, we remove all edges that used to join any pair of vertices in that set.*

Consider the example shown in Figure 5. The leftmost graph G_P is a pivot graph corresponding to the probability of the match relation $(x_1 \equiv y_1) \wedge (x_2 \equiv y_2)$ at time t. Since there are two vertices in G_P each with δ-degree greater than 1, we can perform zero, one or two vertex splits in G_P. Shown in the middle of Figure 5 is the split graph G_S obtained from two vertex splits in G_P. We have given different labels to the vertices in G_S for ease of discussion, but we are not saying that they necessarily correspond to distinct gametes at time $t-1$. Graph G_{M_1} on the right hand side of Figure 5 does correspond to the case in which all four vertices are associated with distinct gametes. If more than one vertex in G_S in fact corresponds to the same gamete at time $t-1$, then that is represented by merging those vertices into a single vertex.

Unconstrained Mating: Under unconstrained mating, recall that the same gamete may be sampled twice, and each of the sampled gametes may transmit genetic material to its offspring. Hence,

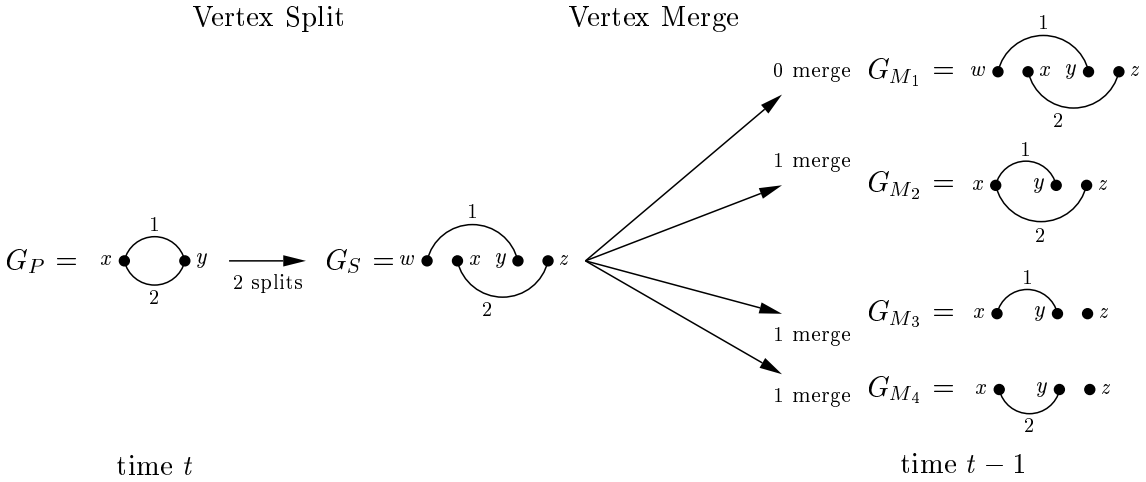

Figure 5: Examples of vertex split and merge operations under *unconstrained mating* scheme. There are other possible vertex merge operations not shown here. Further, there are other split graphs, obtained from either zero or one vertex split.

going backwards in time, an offspring gamete splits into two parental gametes as a consequence of "recombination" and then the latter two gametes may immediately find a common ancestor in the previous generation. Analogously, two vertices in G_S that are a split pair (e.g., vertices w and x or y and z in G_S in Figure 5), may merge into the same vertex. More generally, following a similar line of reasoning, we see that any set of vertices in G_S may merge into a single vertex under unconstrained mating. This fact simplifies things considerably since we do not need to keep track of which vertices are a split pair.

Under unconstrained mating, determining the probability associated with a given merge operation on a given split graph is straightforward. Suppose that a split graph G_S contains n vertices labeled by $[n] = \{1, 2, \ldots, n\}$. Then, under unconstrained mating, there exists a one-to-one correspondence between the set of all vertex merge operations on G_S and the set of all partitions of $[n]$ into non-empty subsets; each subset of $[n]$ corresponds to those vertices that merge. A partition of $[n]$ into k non-empty subsets defines a particular case of assigning n labeled gametes to k distinct unlabeled parental gametes, with each of those k parents having at least one child. It is easy to see

Figure 6: Examples of i-equivalent graphs. Two graphs are said to be i-equivalent, denoted by $\overset{i}{\sim}$, if they become isomorphic as edge-labeled graphs after dropping *isolated* vertices.

that the probability of such a choice under unconstrained mating is given by

$$f(n,k) := \frac{(2N)_{(k)}}{(2N)^n}, \tag{2}$$

where $z_{(k)}$ denotes the falling factorial $z(z-1)\cdots(z-k+1)$. Hence, the probability of a particular set of vertex merges in G_S such that k vertices remain, is given by $f(n,k)$. It is important to note that different sets of merges can produce graphs that are equivalent. For example, consider G_{M_2} on the right hand side of Figure 5. There are four different merge operations on G_S—namely, merge w with x, w with z, y with x, or y with z—that produce match graphs equivalent to G_{M_2} as edge-labeled graphs. Hence, the probability of obtaining G_{M_2} from G_S through merge operations is $4 \times f(4,3)$. In contrast, there exists a unique merge operation that produces G_{M_3} from G_S, and therefore the probability of obtaining G_{M_3} from G_S is $f(4,3)$. The same goes true for G_{M_4}.

Note that graphs G_{M_3} and G_{M_4} each contain an isolated vertex (a vertex with no incident edges). Such a vertex is not involved in any match relation and therefore can be ignored. We say that two graphs are i-equivalent, denoted by $\overset{i}{\sim}$, if they become isomorphic as edge-labeled graphs after dropping *isolated* vertices. (See Figure 6 for examples.) Two i-equivalent graphs correspond to the same match probability. If a graph only contains isolated vertices, then it defines no match relation, and the associated probability is defined to be 1.

Perfect Monogamy: In the case of perfect monogamy, vertex merge operations need to be constrained and merge probabilities modified. One needs to keep track of which vertices in each split graph are a split pair, to determine allowed merge operations. So, in drawing a split graph,

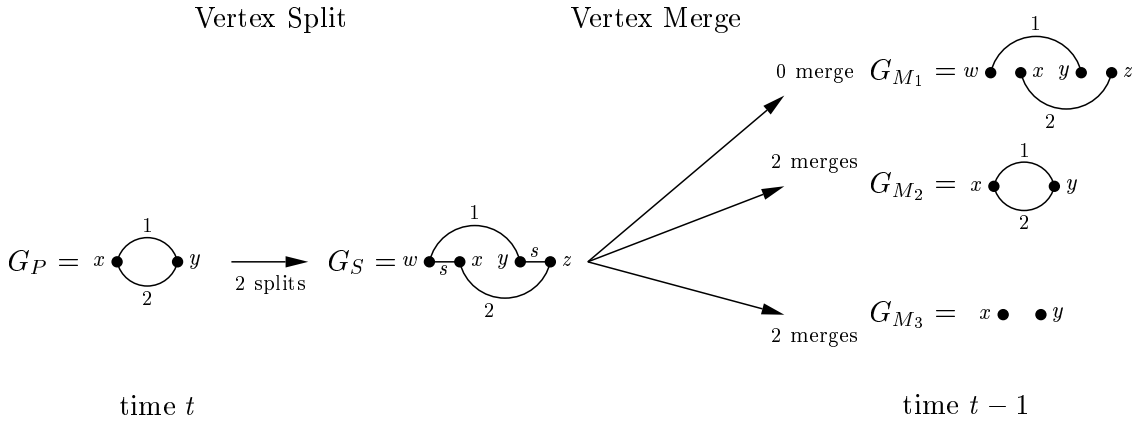

Figure 7: Examples of vertex split and merge operations under *perfect monogamy*. In G_S, an edge labeled "s" joins two vertices that are a split pair. No other vertex merge operations are possible for the given G_S. There still are other split graphs, obtained from either zero or one vertex split.

we add a new edge labeled "s" between the two vertices in each split pair. The perfect monogamy condition imposes the following two constraints on vertex merges:

1. Two vertices joined by an edge labeled "s" may not merge. (Two gametes sampled under perfect monogamy, as described in Section 2.1, are necessarily different gametes, so if the offspring gamete is obtained via "recombination", it must have two different parental gametes.)

2. Vertex merges may not produce a non-cyclic length-2 path with both edges labeled "s". (If two gametes at time t each have two parental gametes at time $t-1$, then their sets of parental gametes are either disjoint or the same, i.e., there can be no half-sibs.)

In addition to Convention 1, we remove all edges labeled "s" after vertex merge operations are complete. The above constraints imply that, under perfect monogamy, G_{M_2}, G_{M_3}, and G_{M_4} in Figure 5 cannot be obtained from G_S; i.e., the corresponding merge operations have probability zero under perfect monogamy. The graphs that can be obtained from allowed merge operations on G_S are shown in Figure 7.

For a given split graph G_S of a pivot graph G_P, label the vertices in the split graph with $[n]$. Let $\mathcal{P} = \{X_1, \ldots, X_k\}$ denote a partition of $[n]$ into k non-empty subsets $X_1, \ldots, X_k$. The partition $\mathcal{P}$

defines a set of merges in G_S, collapsing all vertices in X_i into a single vertex, for each $i = 1, \ldots, k$. Let G_M denote the graph resulting from those merge operations, and define

$$\begin{aligned} S &:= \{i \in [n] \mid i \text{ arose from splitting a vertex in } G_P\}, \\ T &:= \{X \in \mathcal{P} \mid S \cap X \neq \varnothing\}, \\ U &:= \{X \in \mathcal{P} \mid S \cap X = \varnothing\}. \end{aligned}$$

Note that $|T| + |U| = k$ is the number of vertices in G_M, before dropping any isolated vertices. The set T corresponds to the vertices in G_M that the vertices in S will map to under the merge operation defined by $\mathcal{P}$, whereas the set U corresponds to the remaining vertices in G_M. Then, as described in Appendix A, the probability of the set of vertex merges corresponding to $\mathcal{P}$ is given by

$$\frac{1}{(2N)^{n-(|S|+|T|)/2}} \left(\prod_{i=1}^{\frac{|T|}{2}-1} \frac{N-i}{N} \right) \prod_{j=1}^{|U|} (2N - |T| - j + 1), \tag{3}$$

provided that the merges are consistent with the aforementioned two constraints for perfect monogamy. Otherwise, the probability is defined to be zero. For a split graph obtained from zero split operation, $S = \varnothing, T = \varnothing$, and $|U| = k$; and therefore (3) reduces to (2). (We use the convention that a product of form $\prod_{i=1}^{l} g(i)$ is defined to be 1 if $l \leq 0$.)

Example: Consider the split graph G_S shown in Figure 7. To distinguish edge labels from vertex labels, we have labeled the four vertices in G_S with $\Psi = \{w, x, y, z\}$ instead of $[4] = \{1, 2, 3, 4\}$. Since w, x and y, z are both split pairs in G_S, we obtain $S = \Psi, T = \mathcal{P}$, and $U = \varnothing$ for all partitions $\mathcal{P}$ of Ψ. The partition $\mathcal{P} = \{\{w, x\}, \{y\}, \{z\}\}$ is not compatible with perfect monogamy (since w, x are a split pair, they are not allowed to merge). The partition $\mathcal{P} = \{\{w\}, \{x\}, \{y\}, \{z\}\}$ is compatible with perfect monogamy and the corresponding merge operation produces G_{M_1}. Using

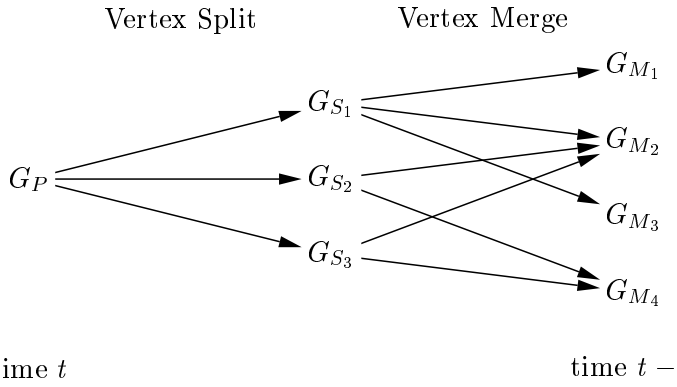

Figure 8: Schematic summary of our graphical approach. For each pivot graph G_P, all allowed vertex split and merge operations are considered, keeping track of the corresponding probabilities. The pivot G_P can be written as a linear combination of the resulting G_{M_i}.

$n = 4, |S| = 4, |T| = 4, |U| = 0$ in (3), we obtain $(N-1)/N$ for the probability of that merge operation. The partition $\mathcal{P} = \{\{w, z\}, \{x, y\}\}$ produces G_{M_2} and using $n = 4, |S| = 4, |T| = 2, |U| = 0$ in (3) produces $1/(2N)$. The partition $\mathcal{P} = \{\{w, y\}, \{x, z\}\}$ produces G_{M_3} and, again, using $n = 4, |S| = 4, |T| = 2, |U| = 0$ in (3) produces $1/(2N)$. More examples can be found in Section 4.3.

3.5 Summary

Schematically illustrated in Figure 8 is our method of generating the equation that relates a match probability at time t to appropriate match probabilities at times $t-1$. Our strategy is to express a pivot graph G_P at time t in terms of G_{M_i} at time $t-1$, by considering all allowed vertex split and merge operations. In this framework, it is easy to keep track of the combinatorial factors and the probabilities associated with inheritance patterns and sharing of parental gametes.

Here is how our graphical framework can be used in practice: Suppose the match probability associated with a particular graph H is not known. To compute it, we need to find a closed system $\mathscr{E}$ of equations that has H as one of its unknown variables. Let $\mathcal{K}$ denote the set of all graphs whose associated match probability values have already been determined. In what follows, $\mathcal{G}$ denotes the

set of graphs on which vertex split and merge operations need to be performed; $\mathcal{N}$ the set of new unknown graphs reached from $\mathcal{G}$ via vertex split and merge operations; $\mathcal{V}$ the set of all variables in $\mathscr{E}$. With $\mathcal{G} = \{H\}, \mathcal{N} = \varnothing$, and $\mathcal{V} = \varnothing$ as initialization, our algorithm for constructing $\mathscr{E}$ goes as follows:

1. For each pivot graph $G_P \in \mathcal{G}$, consider all possible vertex split operations, producing a set $\mathcal{S}_{G_P}$ of split graphs. Record the probability of obtaining each split graph.

2. For all $G_P \in \mathcal{G}$, in any order, carry out the following steps:

   (a) For each graph in $\mathcal{S}_{G_P}$, consider all allowed vertex merge operations, again keeping track of the associated probabilities. Let $\mathcal{M}_{G_P}$ denote the set of all graphs obtained after considering the entire $\mathcal{S}_{G_P}$. Now, G_P can be written in terms of the graphs in $\mathcal{M}_{G_P}$, with appropriate coefficients determined by split, merge and mutation probabilities.

   (b) Update $\mathcal{N}$ by setting $\mathcal{N} \leftarrow \mathcal{N} \cup \mathcal{M}_{G_P} \setminus (\mathcal{G} \cup \mathcal{K})$

3. Set $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{G}$.

4. If $\mathcal{N} \neq \varnothing$, set $\mathcal{G} \leftarrow \mathcal{N}$ and $\mathcal{N} \leftarrow \varnothing$. Then, go back to step 1. If $\mathcal{N} = \varnothing$, then a closed system of equations has been obtained for the graphs in $\mathcal{V}$ and it can be solved.

Some explicit examples are provided in the following section.

4 Examples of Closed Systems of Equations

In this section, we consider some simple examples to elucidate the graphical framework described in the previous section. We adopt the following notational convention when discussing two-locus examples:

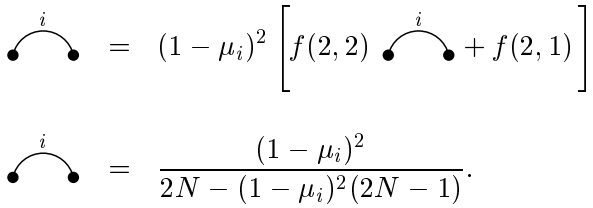

Figure 9: The equilibrium equation satisfied by the one-locus match probability $\mathbb{P}_h(x_i \equiv y_i)$. Here, $f(n,k)$ is defined as in (2) and the factor $(1-\mu_i)^2$ arises as explained in Section 3.2. In deriving the recurrence equation, one needs to recall that a graph consisting of a single isolated vertex has probability 1.

Convention 2 *For two loci, there are only two edge types. So, to simplify notation, we adopt the convention of drawing edges for locus* 1 *(respectively, locus* 2*) as arcs above (respectively, below) vertices.*

4.1 Simplest example

Most mating schemes have the same expression for the probability of $x_i \equiv y_i$, a one-locus match relation involving two gametes. As illustrated in Figure 9, the recurrence equation for $\mathbb{P}(x_i \equiv y_i)$ and its solution at stationarity can easily be obtained using the graphical approach described above.

4.2 Unconstrained mating example

We consider two-locus examples in the remainder of this section. Assuming stationarity and unconstrained mating, it is straightforward to obtain the system of coupled linear equations shown in Figure 10. Let G_1, G_2, and G_3 denote the graphs on the left hand sides of those three equations, respectively, from top to bottom. Note that G_1 does not contain any vertex with δ-degree greater than 1, so no vertex split is possible. Modulo $(1-\mu_1)^2(1-\mu_2)^2$, the expression on the right hand side of the equation for G_1 is obtained from considering all possible merge operations on G_1. The same combination of terms, denoted Ω_1, also appear in the equation for G_2, since G_1 can be obtained from a vertex split operation on G_2 and there are no constraints on vertex merges. The

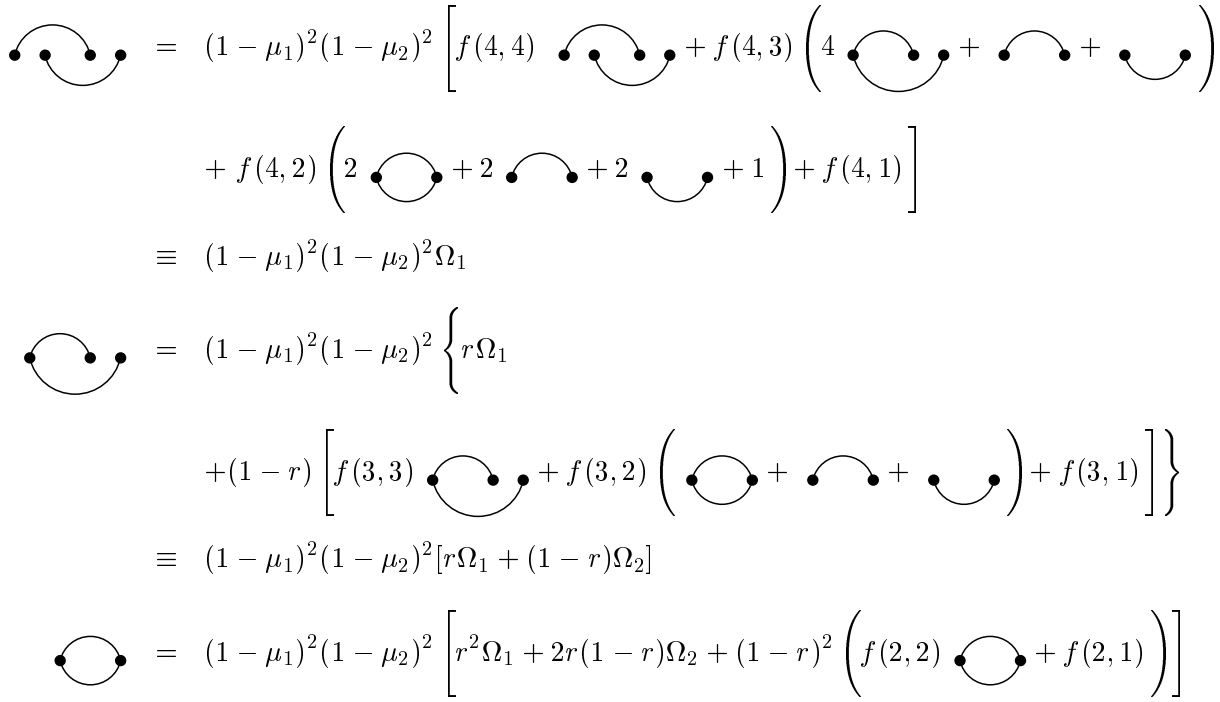

Figure 10: A closed system of coupled equations under unconstrained mating. We use G_1, G_2 and G_3 to refer to the graphs on the left hand side of the first, the second, and the third equation, respectively. These equations should be compared with the equations for perfect monogamy in Figure 11.

remaining terms, denoted Ω_2, arise from performing all possible vertex merges in G_2 without any vertex split. Note that Ω_1 and Ω_2 appear in the equation for G_3, corresponding to performing two and one vertex splits, respectively, in G_3, followed by all possible vertex merges. Notice the factor of 2 in $2r(1-r)\Omega_2$; it comes from the fact that the two possible ways of applying a single vertex split in G_3 produces equivalent split graphs.

For $\mu_1 = \mu_2 = 0$, all match probabilities are equal to 1, and indeed the right hand side of each equation in Figure 10 sums to 1 in that case. Such consistency conditions are useful for checking that coefficients in recurrence equations have been determined correctly. Since the one-locus match probability $\mathbb{P}_h(x_i \equiv y_i)$ can be determined as shown in Figure 9, the equations in Figure 10 form a closed system of coupled equations that can be solved for G_1, G_2, and G_3.

4.3 Perfect monogamy example

We now consider the same three graphs G_1, G_2, G_3 under the perfect monogamy model. For each graph, we need to consider the same set of vertex split operations as in the unconstrained mating scheme. However, vertex merges are constrained under perfect monogamy, and the allowed merges carry probabilities different from the corresponding merges under unconstrained mating. Using the allowed vertex merges described in Section 3.4 for perfect monogamy and the merge probability given in (3), at stationarity we obtain the set of equations shown in Figure 11. For $\mu_1 = \mu_2 = 0$, the right hand side of each equation correctly sums to 1 when all match probabilities are set to 1. As in the unconstrained mating case, these equations form a closed system of coupled equations, and we can solve it for G_1, G_2, and G_3.

5 Match Probabilities

Given two gametes $\boldsymbol{h} = h_1 h_2 \dots h_L$ and $\boldsymbol{h}' = h'_1 h'_2 \dots h'_L$ randomly sampled *without* replacement, we define $\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h}')$ as the L-locus haplotypic match probability. The product rule probability is given by $\prod_{i=1}^{L} \mathbb{P}_h(h_i \equiv h'_i)$, where $\mathbb{P}_h(h_i \equiv h'_i)$ is the one-locus match probability for locus i. We are interested in studying the following ratio:

$$R_h(L) = \frac{\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h}')}{\prod_{i=1}^{L} \mathbb{P}_h(h_i \equiv h'_i)}.$$

To study genotypic match probabilities, we consider two pairs of gametes sampled *without* replacement. Each pair of gametes defines an individual's genotypic sequence. Let $\boldsymbol{g} = g_1 g_2 \dots g_L$ and $\boldsymbol{g}' = g'_1 g'_2 \dots g'_L$ denote the two genotypic sequences so obtained. We are interested in the ratio

$$R_g(L) = \frac{\mathbb{P}_g(\boldsymbol{g} \equiv \boldsymbol{g}')}{\prod_{i=1}^{L} \mathbb{P}_g(g_i \equiv g'_i)},$$

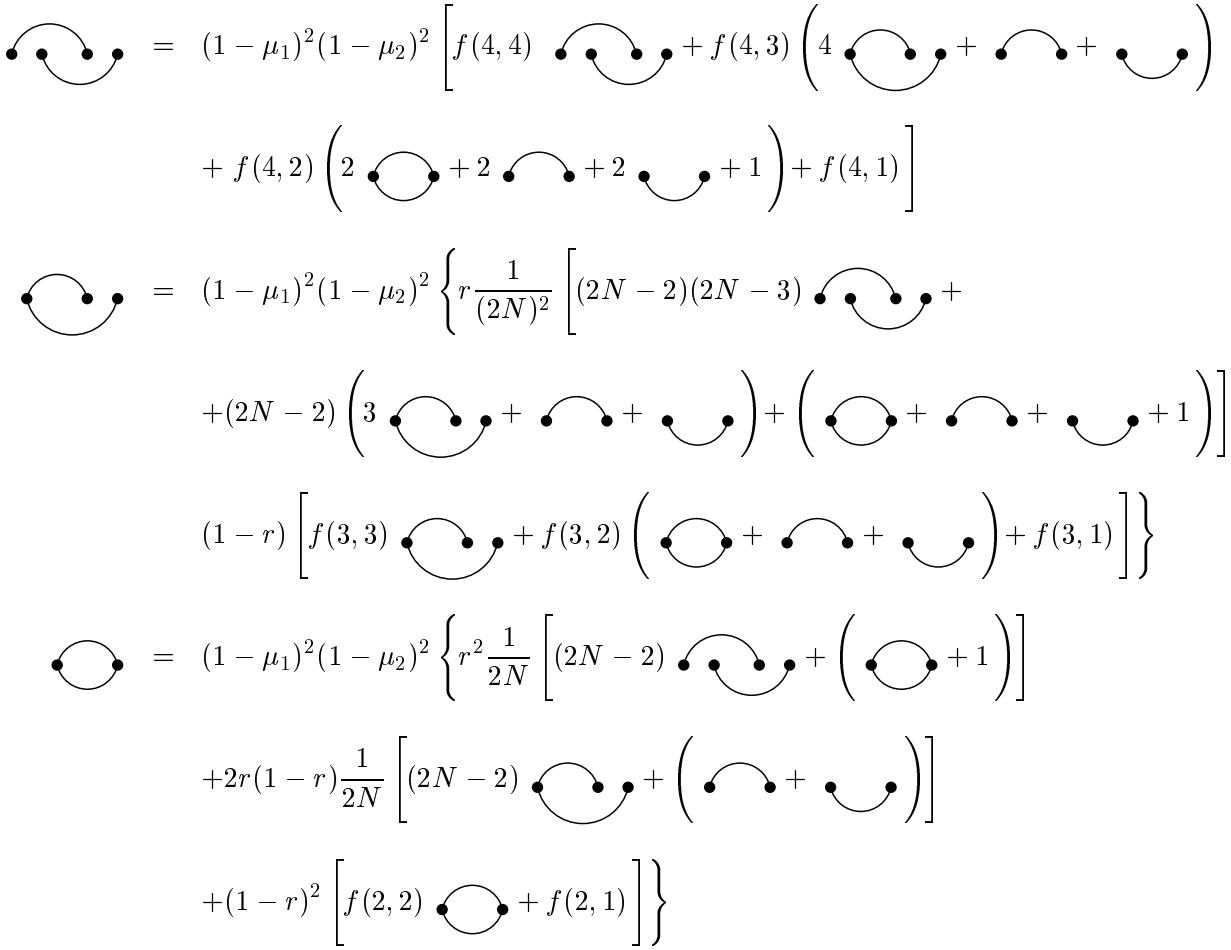

Figure 11: A closed system of coupled equations under perfect monogamy. We use G_1, G_2 and G_3 to refer to the graphs on the left hand side of the first, the second, and the third equation, respectively. These equations should be compared with the equations for unconstrained mating in Figure 10.

with $\mathbb{P}_g(\boldsymbol{g} \equiv \boldsymbol{g}')$ being the L-locus genotypic match probability and $\mathbb{P}_g(g_i \equiv g_i')$ the one-locus genotypic match probability for locus i.

In what follows, the superscript "U" is used to refer to the unconstrained mating scheme, whereas "M" is used to refer to the perfect monogamy model. The one-locus haplotypic match probability $\mathbb{P}_h(h_i \equiv h_i')$ for unconstrained mating is equal to that for perfect monogamy. Similarly, the one-locus genotypic match probability $\mathbb{P}_g(g_i \equiv g_i')$ for unconstrained mating is equal to that for perfect

monogamy. Hence, it follows that

$$\begin{aligned}
\frac{R_h^M(L)}{R_h^U(L)} &= \frac{\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h'}) \text{ for perfect monogamy}}{\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h'}) \text{ for unconstrained mating}}, \\
\frac{R_g^M(L)}{R_g^U(L)} &= \frac{\mathbb{P}_g(\boldsymbol{g} \equiv \boldsymbol{g'}) \text{ for perfect monogamy}}{\mathbb{P}_g(\boldsymbol{g} \equiv \boldsymbol{g'}) \text{ for unconstrained mating}},
\end{aligned}$$

and these ratios capture the effect of monogamy on the L-locus match probability. At the end of this section, we conjecture sharp upper bounds on these ratios.

5.1 Two-locus haplotypic match probability

As a warm-up exercise, we first consider the two-locus *haplotypic* match probability. Given a random pair of gametes $\boldsymbol{h} = h_1 h_2$ and $\boldsymbol{h'} = h_1' h_2'$, we are interested in comparing the two locus haplotypic match probability $\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h'})$ with the product $\mathbb{P}_h(h_1 \equiv h_1')\,\mathbb{P}_h(h_2 \equiv h_2')$. In our graphical framework, $\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h'})$ is as shown in Figure 12. Hence, we can compute $\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h'})$ for unconstrained mating and for perfect monogamy using the systems of coupled equations shown in Figures 10 and 11, respectively. Recall that $\mathbb{P}_h(h_1 \equiv h_1')$ and $\mathbb{P}_h(h_2 \equiv h_2')$ are as shown in Figure 9. Hence, the ratios R_h^U and R_h^M can easily be computed. With $\mu_1 = \mu_2 = u$, some numerical values of R_h^U and R_h^M are shown on the left hand side of Table 2 for $N = 10,000$ and $r = 1/2$. The shown values of R_h^U agree exactly with that of Laurie and Weir (see Table 2 of their paper), thus confirming the correctness of our graphical framework. Note that both ratios R_h^M and R_h^U can be substantially larger than 1, and that $R_h^M \geq R_h^U$ for all u. For two loci, mutation rates need to be rather high for the effect of monogamy to be noticeable. As we discuss later in Section 5.4, the effect of monogamy increases with the number of loci.

$$\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h}') = $$ 

Figure 12: The match graph corresponding to the two-locus haplotypic match probability, using Convention 2.

Table 2: Ratios of the two-locus match probability to the product of one-locus match probabilities for $N = 10,000$, $r = 1/2$, and $\mu_1 = \mu_2 = u$.

	Haplotypic		Genotypic	
u	R_h^U	R_h^M	R_g^U	R_g^M
1×10^{-1}	2.1691×10^2	4.3279×10^2	2.3535×10^4	9.3698×10^4
2.5×10^{-2}	1.6747×10^1	3.2492×10^1	1.4097×10^2	5.2933×10^2
1×10^{-2}	3.6058	6.2113	7.0176	1.9858×10^1
5×10^{-3}	1.6590	2.3179	1.8782	3.1949
1×10^{-3}	1.0266	1.0532	1.0270	1.0547
1×10^{-4}	1.0003	1.0005	1.0003	1.0005
1×10^{-5}	1.0000	1.0000	1.0000	1.0000

5.2 Two-locus genotypic match probability

Let $\boldsymbol{w} = w_1 w_2$ and $\boldsymbol{x} = x_1 x_2$ denote two gametes forming a genotypic sequence $\boldsymbol{g} = g_1 g_2$, and let $\boldsymbol{y} = y_1 y_2$ and $\boldsymbol{z} = z_1 z_2$ denote two other gametes forming another genotypic sequence $\boldsymbol{g}' = g_1' g_2'$. There are four possible ways, illustrated in Figure 13, that the genotypic match $\boldsymbol{g} \equiv \boldsymbol{g}'$ can happen. These possibilities are not mutually exclusive, and to compute the probability of any one of them being true — that is, the probability of $\boldsymbol{g} \equiv \boldsymbol{g}'$ — we invoke the inclusion-exclusion principle. First, we need to introduce a new definition. Given a set of fully-labeled graphs $H_1, H_2, \ldots, H_k$ with the same labeled vertex sets, we define $H_1 \oplus \cdots \oplus H_k$ as the graph obtained by the following two steps:

1. Let $\mathcal{H}$ denote the match graph obtained by taking a union of the edges in H_a, $a = 1, \ldots, k$.

2. In $\mathcal{H}$, if $x_i \equiv y_i$ is implied by transitivity of match relations but there is no edge labeled i between vertices x and y, then add such an edge. (By transitivity of match relations, we mean that $x_i \equiv z_i$ and $z_i \equiv y_i$ together imply $x_i \equiv y_i$.)

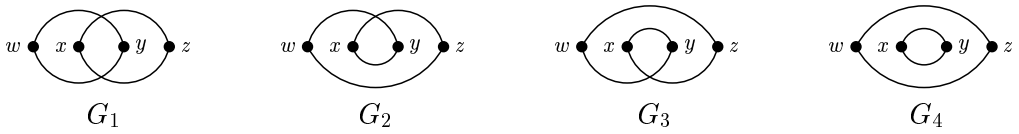

Figure 13: Four possible ways of having two-locus genotypic match. Convention 2 is used here. Gametes w and x form one genotype, and y and z form another. Note that $G_1 \sim G_4$ and $G_2 \sim G_3$, where $\sim$ denotes equivalence as edge-labeled graphs. However, the $\oplus$ operation is defined on G_i as fully labeled graphs.

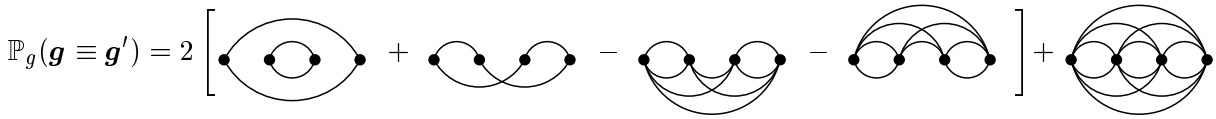

Figure 14: Two-locus genotypic match probability, adopting Convention 2.

Then, by the principle of inclusion-exclusion, we obtain

$$\begin{aligned}\mathbb{P}_g(\boldsymbol{g} \equiv \boldsymbol{g}') &= \sum_{i=1}^{4} G_i - (G_1 \oplus G_2 + G_1 \oplus G_3 + G_1 \oplus G_4 + G_2 \oplus G_3 + G_2 \oplus G_4 + G_3 \oplus G_4) \\ &\quad + (G_1 \oplus G_2 \oplus G_3 + G_1 \oplus G_2 \oplus G_4 + G_1 \oplus G_3 \oplus G_4 + G_2 \oplus G_3 \oplus G_4) - G_1 \oplus G_2 \oplus G_3 \oplus G_4.\end{aligned}$$

Under random mating, this expression simplifies to the graphical representation shown in Figure 14, where we have dropped vertex labels and used the equivalence described in Section 3.1. In a similar vein, it is straightforward to show that the one-locus genotypic match probability $\mathbb{P}_g(g_i \equiv g_i')$ for locus i is as illustrated in Figure 15. The only difference between $\mathbb{P}_g(g_1 \equiv g_1')$ and $\mathbb{P}_g(g_2 \equiv g_2')$ is in their corresponding mutation rates μ_1 and μ_2.

For $\mu_1 = \mu_2 = u$, numerical values of the genotypic ratios R_g^U and R_g^M are shown on the right hand side of Table 2. As mentioned before, our computation of the haplotypic ratio R_h^U agrees exactly with that of Laurie and Weir (2003). However, for $u < 2.5 \times 10^{-2}$, there is a slight

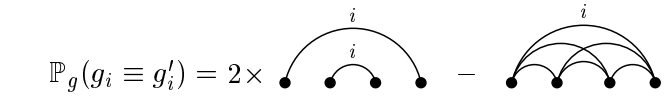

Figure 15: One-locus genotypic match probability $\mathbb{P}_g(g_i \equiv g_i')$. Every edge shown here should be labeled i.

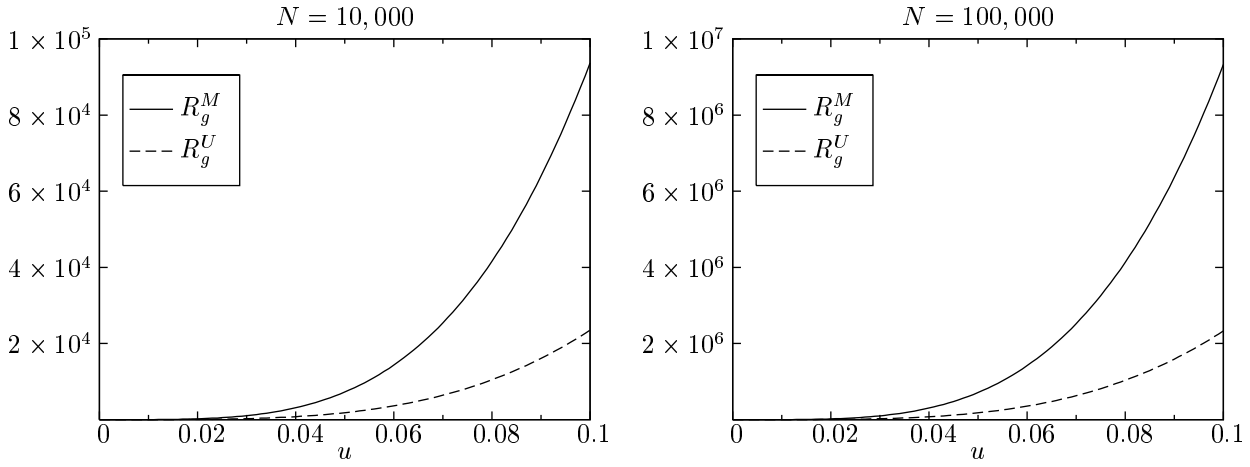

Figure 16: Ratios of two-locus genotypic match probabilities to the product of one-locus match probabilities, assuming $\mu_1 = \mu_2 = u$. As these plots show, the ratio R_g^M for perfect monogamy can be much higher than the ratio R_g^U for unconstrained mating. Both R_g^U and R_g^M significantly increase as N increases.

difference between our computation of the genotypic ratio R_g^U and that reported by Laurie and Weir (see Table 1 of their paper). We found that the difference could be attributed to a minor error in the *Maple* code used to obtain their results. After correcting that error, we verified that their program produces exactly the same results as ours.

Note that $R_g^M \geq R_g^U$ for all u. Illustrated in Figure 16 are plots of R_g^U and R_g^M for $N = 10,000$ and $N = 100,000$. (The human effective population size before expansion into Europe has been estimated to be between $10,000$ and $100,000$. See Harding et al. 1997; Harpending et al. 1998; Takahata 1993; Ayala 1995. Note that Laurie and Weir (2003) also used $N = 10,000$ and $N = 100,000$ in reporting numerical results.) Although both R_g^U and R_g^M significantly increase as N increases, Figure 17 shows that the ratio R_g^M/R_g^U does not depend as much on N, especially for large mutation rates. For low mutation rates, as u increases, R_g^M/R_g^U increases at a faster rate for larger N. Figure 17 suggests that the ratio R_g^M/R_g^U is bounded from above by a finite number. We return to this topic in Section 5.6.

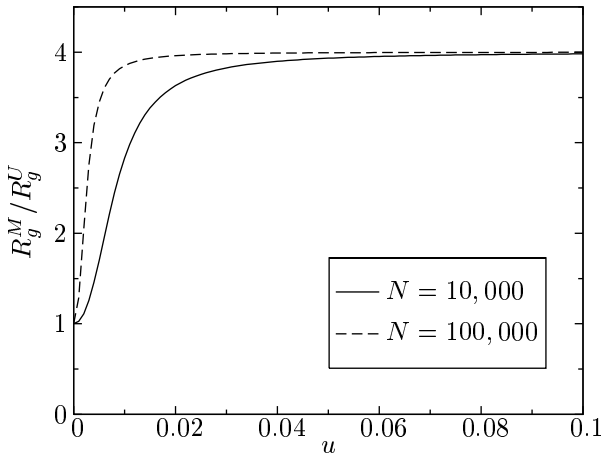

Figure 17: Ratio of the two-locus genotypic match probability R_g^M for perfect monogamy to the probability R_g^U for unconstrained mating, with $\mu_1 = \mu_2 = u$. The ratio R_g^M/R_g^U seems to approach an integer (namely, 4) as u approaches 1 from below. See Section 5.6 for further discussion.

5.3 $1/N$ Expansion

In the L-locus case, a graph that arises in the haplotypic match probability computation can contain up to $2L$ vertices, while a graph in the genotypic case can contain up to $4L$ vertices. Let n denote the number of vertices in a split graph. For $n \geq 12$, the total number of partitions of the set $[n] = \{1, \ldots, n\}$—that is, the Bell number $B(n)$—can be very large (e.g., $B(12) = 4,213,597, B(13) = 27,644,437$, and $B(14) = 190,899,322$). (Recall that a set partition of $[n]$ defines a particular vertex merge operation on a split graph with vertices labeled by $[n]$.) Hence, to handle many loci, we propose an approximation scheme that truncates the equations at certain order in $1/N$, where N is assumed to be substantially large.

Consider the vertex merge operation corresponding to a partition of $[n]$ into k non-empty subsets, merging all vertices within each subset into a single vertex (k corresponds to the number of vertices after merges). Under unconstrained mating, the probability of such a merge operation is of order $1/N^{n-k}$, as can be seen in (2). Hence, in generating the required systems of equations, if we want to keep only those terms with coefficients of order $1/N^m$ where $m \leq 2$—call this *order-2 truncation*—then we only need to consider those partitions of $[n]$ with $k \geq n-2$ non-empty subsets. So, the total

Table 3: *Approximate* two-locus match probability ratios for $N = 10,000$, $r = 1/2$, and $\mu_1 = \mu_2 = u$.

	Haplotypic		Genotypic	
u	R_h^U	R_h^M	R_g^U	R_g^M
1×10^{-1}	2.1691×10^2	4.3279×10^2	2.3529×10^4	9.3691×10^4
2.5×10^{-2}	1.6747×10^1	3.2492×10^1	1.4093×10^2	5.2928×10^2
1×10^{-2}	3.6058	6.2113	7.0162	1.9856×10^1
5×10^{-3}	1.6590	2.3179	1.8780	3.1947
1×10^{-3}	1.0266	1.0532	1.0270	1.0547
1×10^{-4}	1.0003	1.0005	1.0003	1.0005

These results were obtained using truncated systems of equations, ignoring terms with coefficients of $O(1/N^3)$. Comparing this table with Table 2 shows that the proposed approximation method produces very accurate answers.

number of merge operations we need to consider will be $T(n) := S(n,n) + S(n,n-1) + S(n,n-2)$, with $S(n,k)$ being the Stirling number of the second kind. Note that $T(n)$ is substantially smaller than the Bell number $B(n)$ for $n \geq 10$. For example, $T(12) = 1772, T(13) = 2510$, and $T(14) = 3459$. Compare these numbers with the corresponding $B(n)$ shown above.

Truncation in the perfect monogamy model is a bit more subtle. In that case, some partitions with $k = n-3$ or $k = n-4$ have probabilities proportional to $1/N^2$. Therefore, to obtain those terms with coefficients of order $1/N^m$ where $m \leq 2$, we need to consider the partitions of $[n]$ with $k \geq n-4$ non-empty subsets that are consistent with the conditions of the perfect monogamy model (described in Section 3.4).

Shown in Table 3 are two-locus match ratios computed using order-2 truncation. Comparing that table with Table 2, we conclude that the proposed approximation scheme produces very accurate answers. The haplotypic ratios R_h^U and R_h^M in Table 3 are identical to that in Table 2, and we have noticed that even for more loci, R_h^U and R_h^M obtained from order-2 truncation are very close to the exact values. Regarding genotypic match ratios R_g^U and R_g^M, comparing Table 3 with Table 2 shows that the accuracy of order-2 truncation decreases with increasing mutation rate, but still is quite high (about 99.99%).

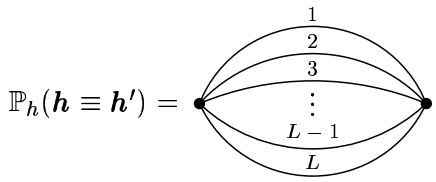

Figure 18: The L-locus haplotypic match probability $\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h}')$.

5.4 Multi-locus haplotypic match probabilities

To compute the L-locus haplotypic match probability $\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h}')$, we need to solve for the graph shown in Figure 18. Taking that graph as a pivot graph, we need to perform all possible vertex split and merge operations, and then iterate the procedure on newly arising graphs, until we obtain a closed system of equations which we can solve. (See Section 3.5 for details. We remark that no two edges have the same label in any haplotypic match graph.) Under unconstrained mating, the same split graph G_S may arise from different pivot graphs. We found that using dynamic programming, which allows one to avoid performing the same vertex merge operations on G_S more than once, can considerably speed up the computation. Further, for both unconstrained mating and perfect monogamy, k-locus graphs, for $k = 2, 3, \ldots, L-1$, will appear in the L-locus computation, so one may again employ dynamic programming and carry out the computation sequentially in increasing number of loci.

The one-locus haplotypic match probability $\mathbb{P}_h(h_i \equiv h_i)$ for locus i is shown in Figure 9. For $L \leq 5$, R_h^U and R_h^M are shown in Table 4. For two and three loci, the R_h^U values shown in that table agree with the corresponding results in Table 2 of Laurie and Weir (2003). To speed up the computation, we used order-2 truncation (described in Section 5.3) for the 5-locus case. Several conclusions can be drawn from this study. First, for a given mutation rate u, both R_h^U and R_h^M increase with the number of loci; the higher the mutation rate, the faster the increase. Second, the effect of monogamy increases with the number of loci, i.e., the ratio R_h^M/R_h^U increases with the number of loci. Third, for a given number of loci, the effect of monogamy increases with the

Table 4: L-locus haplotypic match ratios for $N = 10,000$ and $\mu_i = u$ for all $i = 1, \ldots, L$.

	2-locus			3-locus		
u	R_h^U	R_h^M	R_h^M/R_h^U	R_h^U	R_h^M	R_h^M/R_h^U
1×10^{-1}	2.1691×10^2	4.3279×10^2	1.995	1.7799×10^5	7.1055×10^5	3.992
2.5×10^{-2}	1.6747×10^1	3.2492×10^1	1.940	3.2277×10^3	1.2812×10^4	3.969
1×10^{-2}	3.6058	6.2113	1.723	2.1811×10^2	8.5372×10^2	3.914
5×10^{-3}	1.6590	2.3179	1.397	2.9387×10^1	1.1058×10^2	3.763
1×10^{-3}	1.0266	1.0532	1.026	1.2927	2.0111	1.556
1×10^{-4}	1.0003	1.0005	1.0003	1.0010	1.0025	1.0014

	4-locus			5-locus		
u	R_h^U	R_h^M	R_h^M/R_h^U	R_h^U	R_h^M	R_h^M/R_h^U
1×10^{-1}	1.6479×10^8	1.3145×10^9	7.977	1.5604×10^{11}	2.4855×10^{12}	15.93
2.5×10^{-2}	7.6574×10^5	6.0701×10^6	7.927	1.8809×10^8	2.9735×10^9	15.81
1×10^{-2}	2.0755×10^4	1.6247×10^5	7.828	2.0627×10^6	3.2122×10^7	15.57
5×10^{-3}	1.3677×10^3	1.0481×10^4	7.663	6.8626×10^4	1.0426×10^6	15.19
1×10^{-3}	4.0398	2.0942×10^1	5.184	3.3603×10^1	4.1157×10^2	12.25
1×10^{-4}	1.0027	1.0082	1.0056	1.0060	1.0252	1.0191

All loci are assumed to be pairwise unlinked. For ease of reference, we repeat here the results for two loci. We used order-2 truncation for five loci and the exact computation for all other cases.

mutation rate.

5.5 Three-locus genotypic match probability

We now consider the three-locus genotypic match probability. Let $\boldsymbol{w} = w_1 w_2 w_3$ and $\boldsymbol{x} = x_1 x_2 x_3$ denote two gametes forming a genotypic sequence $\boldsymbol{g} = g_1 g_2 g_3$, and let $\boldsymbol{y} = y_1 y_2 y_3$ and $\boldsymbol{z} = z_1 z_2 z_3$ denote two other gametes forming another genotypic sequence $\boldsymbol{g}' = g_1' g_2' g_3'$. There are eight possible ways that the genotypic match $\boldsymbol{g} \equiv \boldsymbol{g}'$ can happen, as illustrated in Figure 19. As in the case of two loci, these possibilities are not mutually exclusive and we need to use the inclusion-exclusion principle to compute the probability of any one of them being true. More precisely,

$$\mathbb{P}_g(\boldsymbol{g} \equiv \boldsymbol{g}') = \sum_{X \subset \{1,2,\ldots,8\}} (-1)^{|X|+1} \left(\bigoplus_{i \in X} G_i \right),$$

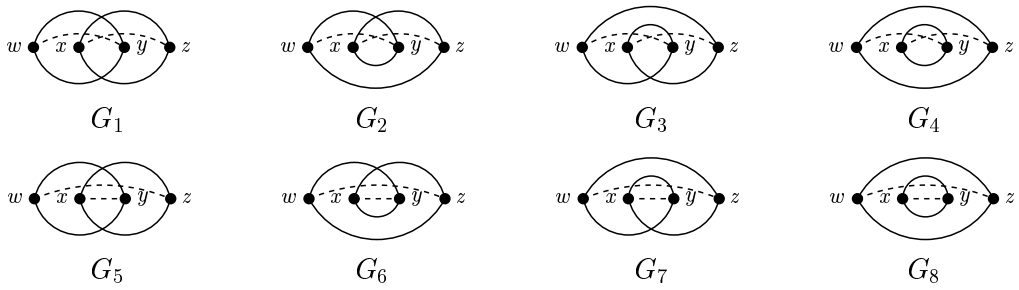

Figure 19: Eight possible ways of having three-locus genotypic match. Gametes w and x form one genotype, and y and z form another. Edge labels are omitted here to avoid clutter; solid arcs above vertices are for locus 1, dotted lines are for locus 2, and solid arcs below vertices are for locus 3. Note that $G_1 \sim G_8, G_2 \sim G_7, G_3 \sim G_6$, and $G_4 \sim G_5$, where $\sim$ denotes equivalence as edge-labeled graphs. Recall that the $\oplus$ operation is defined on G_i as fully labeled graphs.

where X denotes a non-empty subset of $\{1, 2, \ldots, 8\}$ and the $\oplus$ operation is defined as in Section 5.2. This expression simplifies to an expression involving fourteen inequivalent edge-labeled graphs, not shown here. As in the two-locus case, the one-locus genotypic match probability $\mathbb{P}_g(g_i \equiv g_i')$ for locus i is as shown in Figure 15.

Shown in Table 5 are the ratios R_g^U and R_g^M for $N = 10,000$, with $\mu_i = u$ for all $i = 1, \ldots, L$. Two-locus results are repeated there for ease of comparison. Comparing these genotypic results with the haplotypic results in Table 4, we see that for two loci, $R_g^U \geq R_h^U$ and $R_g^M \geq R_h^M$ for any given mutation rate. For three loci, however, these inequalities are violated for low mutation rates (say, $\mu \lesssim 1.2 \times 10^{-3}$). As in the haplotypic case, $R_g^M \geq R_g^U$ for any given mutation rate. The results in Table 5 show that, as in the haplotypic case, the effect of monogamy grows with the number of loci; i.e., the ratio R_g^M / R_g^U increases with the number of loci.

5.6 Sharp upper bounds on the effect of monogamy

Tables 4 and 5 suggest that the L-locus ratios $R_h^M(L)/R_h^U(L)$ and $R_g^M(L)/R_g^U(L)$ stay bounded by a finite number (dependent on L) as the common mutation rate u increases. We have checked numerically that this property still holds for mutation rates higher than 1×10^{-1}. Based on this empirical observation, we make the following two conjectures regarding sharp upper bounds on the

Table 5: Genotypic match ratios for $N = 10,000$ and $\mu_i = u$ for all $i = 1, \ldots, L$, with all loci assumed to be pairwise unlinked.

	2-locus			3-locus		
u	R_g^U	R_g^M	R_g^M/R_g^U	R_g^U	R_g^M	R_g^M/R_g^U
1×10^{-1}	2.35×10^4	9.37×10^4	3.98	7.92×10^9	1.26×10^{11}	16.0
2.5×10^{-2}	1.41×10^2	5.29×10^2	3.76	2.61×10^6	4.12×10^7	15.8
1×10^{-2}	7.016	1.986×10^1	2.840	1.20×10^4	1.84×10^5	15.3
5×10^{-3}	1.878	3.195	1.701	2.21×10^2	3.10×10^3	14.1
1×10^{-3}	1.027	1.055	1.027	1.210	1.861	1.538
1×10^{-4}	1.0003	1.0005	1.0003	1.0009	1.0020	1.0011

effect of monogamy:

Conjecture 1 *Let* $\boldsymbol{h} = h_1 h_2 \ldots h_L$ *and* $\boldsymbol{h}' = h'_1 h'_2 \ldots h'_L$ *denote L-locus haplotypic sequences, and recall that* $R_h^M(L)/R_h^U(L)$ *is equal to the ratio of the L-locus haplotypic match probability* $\mathbb{P}_h(\boldsymbol{h} \equiv \boldsymbol{h}')$ *under perfect monogamy to that under unconstrained mating. Suppose that* $\mu_i = u$ *for all* $i = 1, \ldots, L$*. Then,*

$$\lim_{u \uparrow 1} \frac{R_h^M(L)}{R_h^U(L)} = 2^{L-1},$$

and $R_h^M(L)/R_h^U(L) \leq 2^{L-1}$ *for all* u*.*

Conjecture 2 *Let* $\boldsymbol{g} = g_1 g_2 \ldots g_L$ *and* $\boldsymbol{g}' = g'_1 g'_2 \ldots g'_L$ *denote L-locus genotypic sequences, and recall that* $R_g^M(L)/R_g^U(L)$ *is equal to the ratio of the L-locus genotypic match probability* $\mathbb{P}_g(\boldsymbol{g} \equiv \boldsymbol{g}')$ *under perfect monogamy to that under unconstrained mating. Suppose that* $\mu_i = u$ *for all* $i = 1, \ldots, L$*. Then,*

$$\lim_{u \uparrow 1} \frac{R_g^M(L)}{R_g^U(L)} = 2^{2L-2},$$

and $R_g^M(L)/R_g^U(L) \leq 2^{2L-2}$ *for all* u*.*

The above conjectures are independent of N. However, the larger the N, the faster the rate at which $R_h^M(L)/R_h^M(L)$ and $R_g^M(L)/R_g^M(L)$ approach their respective upper bounds as u increases.

This property is illustrated in Figure 17 for the two-locus genotypic case. Since $R_h^M(L)$ and $R_g^M(L)$ are for perfect monogamy (i.e., the most extreme level of monogamy), the upper bounds shown in the above conjectures are also upper bounds for all intermediate levels of monogamy.

We believe that there may exist a simple combinatorial explanation for the upper bounds 2^{L-1} and 2^{2L-2} appearing in Conjectures 1 and 2, respectively. It would be interesting to study the asymptotic behavior analytically. Further, it would be worthwhile to study the dependence of $R_h^M(L)/R_h^U(L)$ and $R_g^M(L)/R_g^U(L)$ on the mutation rate u, especially for small u. As Figure 17 indicates, it seems that interesting dynamics can happen within a small window of u.

6 Discussion and Conclusions

The goal of this paper is to provide a framework within which multi-locus probabilities that two unrelated individuals have the same genotype at several loci can be analyzed in a relatively simple manner. Although the analysis of models involving two or more loci is necessarily complicated because of the many ways in which identity and nonidentity propagate from one generation to the next, the graphical method introduced here makes the combinatorial structure of the problem clear and the analysis as simple as possible, and it leads to a method for automatic generation of the appropriate recurrence equations that minimizes the problem of human error. The graphical method takes advantage of the underlying symmetry of the inheritance of unlinked loci and can be adapted to the analysis of similar models.

We have shown that the qualitative conclusion of Laurie and Weir (2003) is correct under a wider range of conditions than they were able to consider with their method. In a randomly mating population, the product rule provides a very close approximation to the probability that two unrelated individuals have the same genotype provided that mutation rates are not too large. If the population size is 10,000, then $u = 0.0001$ corresponds to a heterozygosity of 80%, which is

typical of CODIS loci (Budowle et al., 2001). For that value of u, the ratio R is very close to 1 even for the haplotypic match probability at 5 loci and even if there is complete monogamy (see Table 4).

One limitation of our study, as well as that of Laurie and Weir (2003), is that we assume an infinite alleles model of mutation. Consequently, identity in allelic state implies identity by descent. We do not allow for independent origins of the same allele, as can happen with microsatellite loci. Our results show, however, that there is no substantial increase in the joint probability of identity by descent because of shared genealogies in a finite population. That conclusion is true for other mutation models as well.

Acknowledgments

This research is supported in part by NSF grants CCF-0515278 and IIS-0513910 (YSS) and by NIH grant R01-GM40282 (MS). We thank C. Laurie for helpful comments on a preliminary version of this paper and for kindly providing a copy of the *Maple* program used to generate the results in Laurie and Weir (2003).

Appendix A. Derivation of (3)

We briefly describe here how the probability shown in (3) is obtained. The same notation introduced at the end of Section 3.4 is used here. A set partition $\mathcal{P} = \{X_1, \ldots, X_k\}$ of $[n]$ defines a particular case of assigning n labeled gametes to k distinct unlabeled parental gametes, with each of those k parents having at least one child. The elements in T and U correspond to the parents. Suppose that the merge operation defined by $\mathcal{P}$ is consistent with perfect monogamy. Then, $|T|$ is even, since two vertices in each split pair in the split graph map to two distinct subsets X_i, X_j, and two different split pairs map to either the same pair of subsets or two disjoint pairs of subsets. In the perfect monogamy model, recall that there are N pairs of parental gametes. Each split pair can choose a particular pair of parents with probability $1/N$. Two split pairs w, x and y, z can choose the same pair of parents in two ways: either w collides with y and x collides z, or w collides with z and x collides with y. Each possibility has probability $1/(2N)$. Putting all these things together, we conclude that the probability of surjectively assigning $|S|/2$ split pairs to $|T|/2$ disjoint pairs of parents is

$$\frac{1}{(2N)^{\frac{|S|}{2}-\frac{|T|}{2}}} \prod_{i=0}^{\frac{|T|}{2}-1} \frac{N-i}{N}.$$

The remaining $n - |S|$ vertices in the split graph choose parents such that each parent in U has at least one child, and the associated probability is

$$\frac{1}{(2N)^{n-|S|}} \prod_{j=1}^{|U|} (2N - |T| - j + 1).$$

Equation (3) now follows from putting the above two probabilities together.

References

Ayala, F. J. 1995. The myth of Eve: molecular biology and human origins. *Science*, **270,** 1930–1936.

Budowle, B., Shea, B., Niezgoda, S., and Chakraborty, R. 2001. CODIS STR loci data from 41 sample populations. *J. Forensic. Sci.*, **46,** 453–489.

Evett, I. W. and Weir, B. S. *Interpreting DNA Evidence.* Sinauer Associates, Sunderland, Mass., 1998.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., Cox, M. J., Schneider, J. A., Moulin, D. S., and Clegg, J. B. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.*, **60,** 772–789.

Harpending, H. C., Batze, M. A., Gurven, M., Jorde, L. B., Rogers, A. R., and Sherry, S. T. 1998. Genetic traces of ancient demography. *Proc. Nat. Acad. Sci.*, **95,** 1961–1967.

Hill, W. G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, **38,** 226–231.

Laurie, C. and Weir, B. S. 2003. Dependency effects in multi-locus match probabilities. *Theor. Popul. Biol.*, **63,** 207–219.

Ohta, T. and Kimura, M. 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, **63,** 229–238.

Strobeck, C. and Golding, G. B. 1983. The variance of linkage disequilibrium between three loci in a finite population. *Can. J. Genet. Cytol.*, **25,** 139–45.

Takahata, N. 1993. Allelic genealogy and human evolution. *Mol. Biol. Evol.*, **10,** 2–22.

Weir, B. S. 2004. Matching and partially-matching DNA profiles. *J. Forensic. Sci.*, **49,** 1009–1014.