Forensic DNA analysis and multi-locus match probability in finite populations:

A fundamental difference between the Moran and Wright-Fisher models

Yun S. Song
Departments of EECS and Statistics
UC Berkeley

DIMACS
April 27, 2009

Outline

1. Introduction
   - Random match probability
   - Cold hit
2. Models of Random Mating
   - Recurrence equations
3. Graphical Framework
   - Match graphs
   - Operations on graphs
   - Topological ordering and graph enumeration
4. Results
   - Accuracy of the product rule
   - Wright-Fisher vs. Moran
   - Excluding siblings
5. Other Works
   - Perfect Monogamy Model
   - Subdivided populations

Random match probability

Given

Two random individuals from a population.

Question

What is the probability that their DNA profiles match?

Art source: René Magritte

Forensic science context

The question that often arises is the extent to which a complete match of DNA profiles between a suspect and a crime-scene sample indicates that the suspect is the source of the sample.

Art source: René Magritte

Match probability depends on many factors, including

- The number of loci in the DNA profile.
- Mutation rates.
- Population history.

Art source: René Magritte

Short Tandem Repeats (a.k.a microsatellites)

Repetitions of words usually $2 \sim 6$ base-pairs in length

Simple Examples of STR:

Word Length	Locus	DNA Repeat Sequence	Copy Number Variation in Population
2 bp	APOA2	ACACACAC$\cdots$AC	$[\mathrm{AC}]_{8\sim22}$
3 bp	Huntingtin	CAGCAGCAG$\cdots$CAG	$[\mathrm{CAG}]_{6\sim35}$ (Normal)
			$[\mathrm{CAG}]_{36\sim120}$ (Pathogenic)
4 bp	TPOX	AATGAATG$\cdots$AATG	$[\mathrm{AATG}]_{5\sim14}$

Allele

Useful genetic STR markers have a typical copy number of $10 \sim 30$. Copy numbers will be called *alleles*.

Short Tandem Repeats (a.k.a microsatellites)

Repetitions of words usually $2 \sim 6$ base-pairs in length

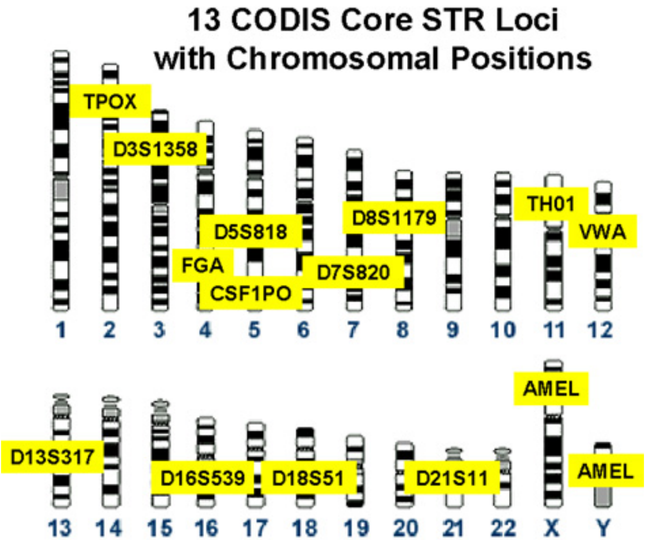
Simple Examples of STR:

Word Length	Locus	DNA Repeat Sequence	Copy Number Variation in Population
2 bp	APOA2	ACACACAC· · · AC	$[AC]_{8\sim22}$
3 bp	Huntingtin	CAGCAGCAG· · · CAG	$[CAG]_{6\sim35}$ (Normal)
			$[CAG]_{36\sim120}$ (Pathogenic)
4 bp	TPOX	AATGAATG· · · AATG	$[AATG]_{5\sim14}$

Allele

Useful genetic STR markers have a typical copy number of $10 \sim 30$. Copy numbers will be called *alleles*.

At present, 11 to 13 unlinked autosomal microsatellite loci are typed for forensic use.



Source: http://www.cstl.nist.gov/div831/strbase/fbicore.htm

FBI's CODIS (COmbined DNA Index System) Short Tandem Repeat loci (tetranucleotide)

AATGAATG· · · AATG

Mostly on different chromosomes

Amelogenin Gene
On X: 106 bp
On Y: 112 bp

Example: an individual's CODIS profile

Chromosome	Locus	Genotype (Unordered Pair)
2	TPOX	7, 8
3	D3S1358	15, 18
4	FGA	19, 24
5	D5S818	11, 13
5	CSF1PO	11, 11
7	D7S820	10, 11
8	D8S1179	12, 13
11	THO1	8, 12
12	VWA	16, 16
13	D13S317	11, 16
16	D16S539	11, 14
18	D18S51	12, 13
21	D21S11	29, 31
	AMEL	106bp, 112bp

The DNA Identification Act of 1994

Authorized the FBI to establish a national DNA index for law enforcement purposes.

Combined DNA Index System (operational since 1998)

Three levels of hierarchy

1. National DNA Index System
   Allows labs between states to exchange DNA profiles
2. State DNA Index System
   Allows labs within states to exchange DNA profiles
3. Local DNA Index System
   DNA profiles are collected at the local level

Number of “offender” profiles

	As of Oct 2007	As of Dec 2008
Nation-wide	5,265,258	6,539,919

Number of “offender” profiles

	As of Oct 2007	As of Dec 2008
Nation-wide	5,265,258	6,539,919
California	893,147	1,073,768
Florida	397,500	533,670
Texas	314,366	395,374
Virginia	260,403	285,851
Illinois	276,339	320,132
Michigan	221,354	255,274
New York	216,083	294,498
Wyoming	197	8,722
Rhode Island	834	3,890

Number of “offender” profiles

	As of Oct 2007	As of Dec 2008
Nation-wide	5,265,258	6,539,919
California	893,147	1,073,768
Florida	397,500	533,670
Texas	314,366	395,374
Virginia	260,403	285,851
Illinois	276,339	320,132
Michigan	221,354	255,274
New York	216,083	294,498
Wyoming	197	8,722
Rhode Island	834	3,890

Usually, but not always, conviction for some type of criminal offense is required to be included in the database.

L-Locus Match Probability (MP)

The probability of a complete match at *L* unlinked loci between two individuals randomly chosen from a population.

L-Locus Match Probability (MP)

The probability of a complete match at L unlinked loci between two individuals randomly chosen from a population.

The Product Rule (currently used in US criminal courts)

- Assume statistical independence across all L loci.
- Multiply the 1-locus MPs at those loci.

L-Locus Match Probability (MP)

The probability of a complete match at *L* unlinked loci between two individuals randomly chosen from a population.

The Product Rule (currently used in US criminal courts)

- Assume statistical independence across all *L* loci.
- Multiply the 1-locus MPs at those loci.

Warning

In a finite population, the genealogical relationships of individuals can create statistical non-independence of alleles at unlinked loci.

L-Locus Match Probability (MP)

The probability of a complete match at *L* unlinked loci between two individuals randomly chosen from a population.

The Product Rule (currently used in US criminal courts)

- Assume statistical independence across all *L* loci.
- Multiply the 1-locus MPs at those loci.

Warning

In a finite population, the genealogical relationships of individuals can create statistical non-independence of alleles at unlinked loci.

Question

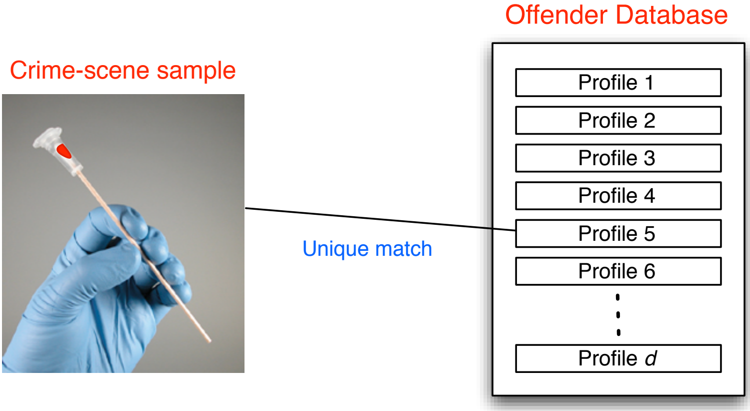
Then, how accurate is the product rule, which assumes independence between loci?

Question on Question

In any case, everyone believes that the true 13-locus MP is a very small number. Then, why are we interested in computing it accurately?

Cold Hit

A crime-scene sample is found to match a known profile in a database, resulting in the identification of a suspect based only on genetic evidence.



Cold hits and erroneous attribution

- Consider a hypothetical series of cold hit cases.

Cold hits and erroneous attribution

- Consider a hypothetical series of cold hit cases.
- The average probability that there exists another person in the population whose profile matches the crime-scene sample but who is not in the database is

$$\frac{1 + n \times AMP - (1 - AMP)^n}{1 + n \times AMP},$$

where *AMP* is the average match probability and *n* is the total number of people *not* in the database.

(Song, Patil, Murphy, Slatkin, *J. Forensic Sciences*, 2009.)

Cold hits and erroneous attribution

- Consider a hypothetical series of cold hit cases.
- The average probability that there exists another person in the population whose profile matches the crime-scene sample but who is not in the database is

$$\frac{1 + n \times AMP - (1 - AMP)^n}{1 + n \times AMP},$$

where *AMP* is the average match probability and n is the total number of people *not* in the database.

(Song, Patil, Murphy, Slatkin, *J. Forensic Sciences*, 2009.)

- This probability is approximately equal to $2n \times AMP$.

Cold hits and erroneous attribution

- Consider a hypothetical series of cold hit cases.
- The average probability that there exists another person in the population whose profile matches the crime-scene sample but who is not in the database is

$$\frac{1 + n \times AMP - (1 - AMP)^n}{1 + n \times AMP},$$

where *AMP* is the average match probability and *n* is the total number of people *not* in the database.

(Song, Patil, Murphy, Slatkin, *J. Forensic Sciences*, 2009.)

- This probability is approximately equal to $2n \times AMP$.
- If the *AMP* is as large as 10^{-9}, there is a considerable risk that someone not in the database has the same profile.

Challenge

Analytically computing true multi-locus match probability has remained a very difficult problem.

Plan of the talk

1. We will introduce a flexible graphical framework to compute multi-locus MPs analytically.
2. We will consider two standard models of random mating, namely the Wright-Fisher and Moran models. (We will reach the magic number 13 for the Moran model.)
3. We will describe a striking fundamental difference between the two models which becomes transparent only when many loci are considered in a finite population.
4. We will discuss the accuracy of the product rule.
5. If time permits, we will discuss the biparental diploid model (Chang, 1999).

Challenge

Analytically computing true multi-locus match probability has remained a very difficult problem.

Plan of the talk

1. We will introduce a flexible graphical framework to compute multi-locus MPs analytically.
2. We will consider two standard models of random mating, namely the Wright-Fisher and Moran models. (We will reach the magic number 13 for the Moran model.)
3. We will describe a striking fundamental difference between the two models which becomes transparent only when many loci are considered in a finite population.
4. We will discuss the accuracy of the product rule.
5. If time permits, we will discuss the biparental diploid model (Chang, 1999).

Challenge

Analytically computing true multi-locus match probability has remained a very difficult problem.

Plan of the talk

1. We will introduce a flexible graphical framework to compute multi-locus MPs analytically.
2. We will consider two standard models of random mating, namely the Wright-Fisher and Moran models. (We will reach the magic number 13 for the Moran model.)
3. We will describe a striking fundamental difference between the two models which becomes transparent only when many loci are considered in a finite population.
4. We will discuss the accuracy of the product rule.
5. If time permits, we will discuss the biparental diploid model (Chang, 1999).

Challenge

Analytically computing true multi-locus match probability has remained a very difficult problem.

Plan of the talk

1. We will introduce a flexible graphical framework to compute multi-locus MPs analytically.
2. We will consider two standard models of random mating, namely the Wright-Fisher and Moran models. (We will reach the magic number 13 for the Moran model.)
3. We will describe a striking fundamental difference between the two models which becomes transparent only when many loci are considered in a finite population.
4. We will discuss the accuracy of the product rule.
5. If time permits, we will discuss the biparental diploid model (Chang, 1999).

Challenge

Analytically computing true multi-locus match probability has remained a very difficult problem.

Plan of the talk

1. We will introduce a flexible graphical framework to compute multi-locus MPs analytically.
2. We will consider two standard models of random mating, namely the Wright-Fisher and Moran models. (We will reach the magic number 13 for the Moran model.)
3. We will describe a striking fundamental difference between the two models which becomes transparent only when many loci are considered in a finite population.
4. We will discuss the accuracy of the product rule.
5. If time permits, we will discuss the biparental diploid model (Chang, 1999).

Challenge

Analytically computing true multi-locus match probability has remained a very difficult problem.

Plan of the talk

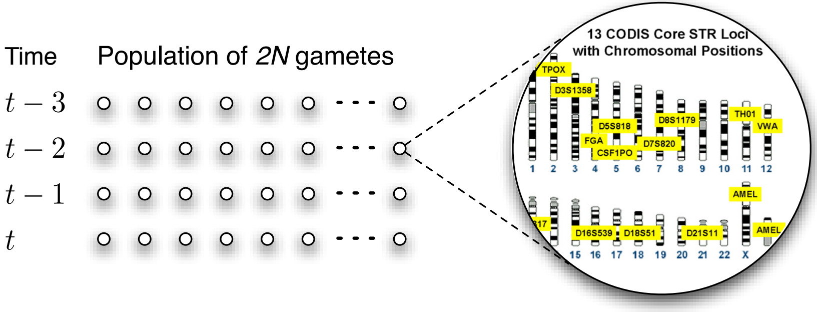
1. We will introduce a flexible graphical framework to compute multi-locus MPs analytically.
2. We will consider two standard models of random mating, namely the Wright-Fisher and Moran models. (We will reach the magic number 13 for the Moran model.)
3. We will describe a striking fundamental difference between the two models which becomes transparent only when many loci are considered in a finite population.
4. We will discuss the accuracy of the product rule.
5. If time permits, we will discuss the biparental diploid model (Chang, 1999).

Outline

1. Introduction
   - Random match probability
   - Cold hit
2. Models of Random Mating
   - Recurrence equations
3. Graphical Framework
   - Match graphs
   - Operations on graphs
   - Topological ordering and graph enumeration
4. Results
   - Accuracy of the product rule
   - Wright-Fisher vs. Moran
   - Excluding siblings
5. Other Works
   - Perfect Monogamy Model
   - Subdivided populations

Assumptions

- Constant population size.
- Random mating.
- Infinite alleles model of mutation.



Time Population of *2N* gametes

$t-3$ ○ ○ ○ ○ ○ ○ $\cdots$ ○

$t-2$ ○ ○ ○ ○ ○ ○ $\cdots$ ○

$t-1$ ○ ○ ○ ○ ○ ○ $\cdots$ ○

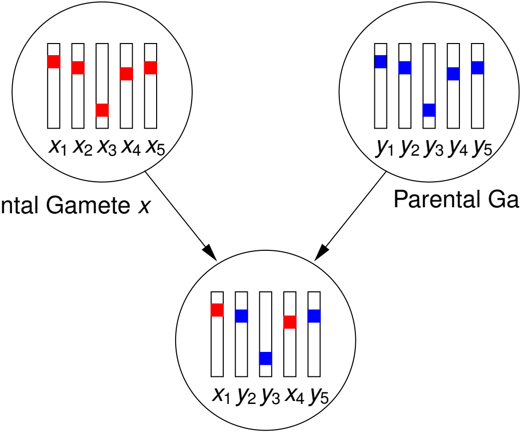
t ○ ○ ○ ○ ○ ○ $\cdots$ ○

A *gamete* refers to a collection of alleles at 13 unlinked loci.

Generating a newborn

Randomly sample two gametes, each with replacement, and create a new gamete as an assortment of the two samples.

Generation t


x_1 x_2 x_3 x_4 x_5
y_1 y_2 y_3 y_4 y_5
x_1 y_2 y_3 x_4 y_5

Parental Gamete x

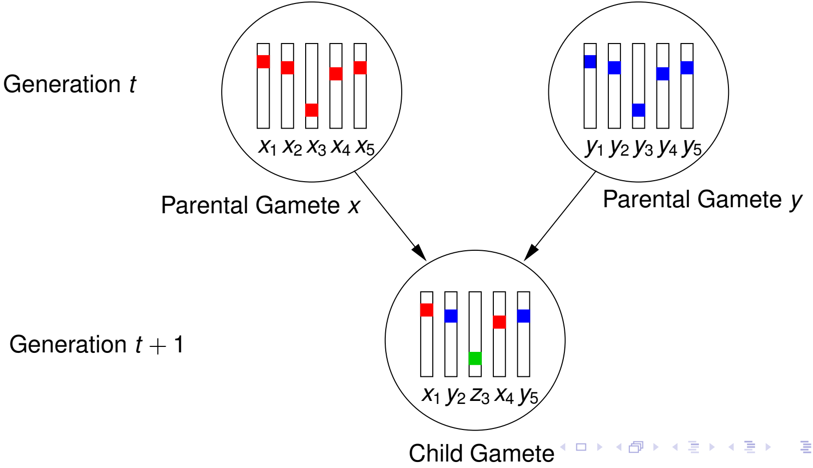
Parental Gamete y

Generation $t + 1$

Child Gamete

Infinite-alleles model of mutation

With probability μ_i, the child gamete has an allele (copy number) at locus *i* that has never been seen before.


Generation t
x_1 x_2 x_3 x_4 x_5
y_1 y_2 y_3 y_4 y_5
Parental Gamete x
Parental Gamete y
Generation $t+1$
x_1 y_2 z_3 x_4 y_5
Child Gamete

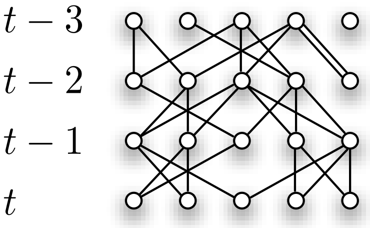
Wright-Fisher model

- $2N_{\mathrm{WF}}$ gametes.
- Non-overlapping generations. (The entire population gets replaced every generation.)

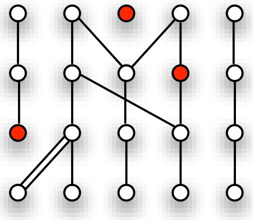
Moran model

- $2N_{\mathrm{M}}$ gametes.
- Overlapping generations. (Exactly one individual gets replaced every generation. All other individuals survive to the next generation.)

Wright-Fisher Model



Moran Model



Wright-Fisher model

- $2N_{\mathrm{WF}}$ gametes.
- Non-overlapping generations. (The entire population gets replaced every generation.)

Moran model

- $2N_{\mathrm{M}}$ gametes.
- Overlapping generations. (Exactly one individual gets replaced every generation. All other individuals survive to the next generation.)

Facts

1. For the two models to have the same effective population size N_e, we need to set $N_{\mathrm{M}} = 2N_{\mathrm{WF}}$.
2. The two models converge to the same diffusion limit.

Genotypic Match Probability

Randomly choose two pairs of gametes without replacement. At stationarity, what is the probability that the two pairs have a complete genotypic match at L unlinked loci?

Haplotypic Match Probability

Randomly choose two gametes without replacement. At stationarity, what is the probability that the two gametes have a complete copy number match at L unlinked loci?

Pair 1

Locus	Genotype
1	7,8
2	15,16
3	19,20
4	11,11
5	29,31

Pair 2

Locus	Genotype
1	7,8
2	15,16
3	19,20
4	11,11
5	29,31

Genotypic Match Probability

Randomly choose two pairs of gametes without replacement. At stationarity, what is the probability that the two pairs have a complete genotypic match at L unlinked loci?

Haplotypic Match Probability

Randomly choose two gametes without replacement. At stationarity, what is the probability that the two gametes have a complete copy number match at L unlinked loci?

Gamete x

Locus	Copy Number
1	7
2	15
3	19
4	11
5	29

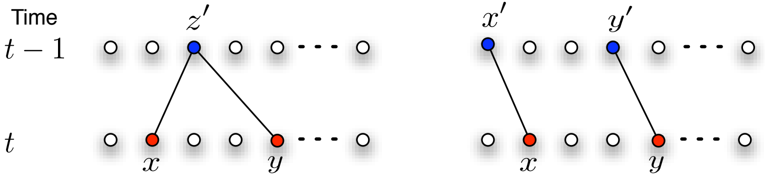
Gamete y

Locus	Copy Number
1	7
2	15
3	19
4	11
5	29

Recurrence equations

Consider two gametes $x = (x_1, \ldots, x_L)$ and $y = (y_1, \ldots, y_L)$.

Two possible ancestries for locus i under the WF model



Probability: $\frac{1}{2N_{WF}}$ $\frac{2N_{WF}-1}{2N_{WF}}$

Recurrence equation

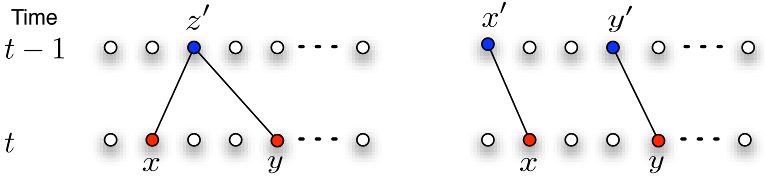
◂ Graphs

$$\mathbb{P}(x_i = y_i) = (1-\mu_i)^2 \left[\frac{1}{2N_{\mathrm{WF}}} + \frac{2N_{\mathrm{WF}}-1}{2N_{\mathrm{WF}}} \mathbb{P}(x_i' = y_i') \right]$$

At stationarity, $\mathbb{P}(x_i = y_i) = \mathbb{P}(x_i' = y_i')$, so we can solve for the stationary probability $\mathbb{P}(x_i = y_i)$.

Consider two gametes $x = (x_1, \ldots, x_L)$ and $y = (y_1, \ldots, y_L)$.

Two possible ancestries for locus i under the WF model



Probability: $\dfrac{1}{2N_{WF}}$ $\dfrac{2N_{WF}-1}{2N_{WF}}$

Recurrence equation

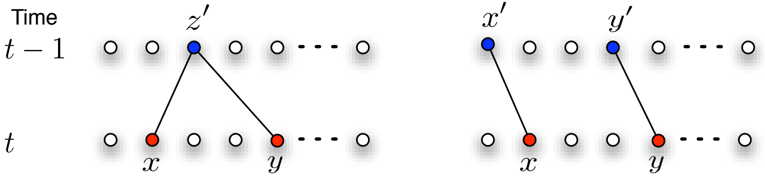
◂ Graphs

$$\mathbb{P}(x_i = y_i) \;=\; (1-\mu_i)^2 \left[\frac{1}{2N_{\mathrm{WF}}} + \frac{2N_{\mathrm{WF}}-1}{2N_{\mathrm{WF}}} \mathbb{P}(x_i' = y_i') \right]$$

At stationarity, $\mathbb{P}(x_i = y_i) = \mathbb{P}(x_i' = y_i')$, so we can solve for the stationary probability $\mathbb{P}(x_i = y_i)$.

Consider two gametes $x = (x_1, \dots, x_L)$ and $y = (y_1, \dots, y_L)$.

Two possible ancestries for locus i under the WF model



Probability: $\frac{1}{2N_{WF}}$ $\frac{2N_{WF}-1}{2N_{WF}}$

Recurrence equation

◂ Graphs

$$\mathbb{P}(x_i = y_i) \;=\; (1-\mu_i)^2 \left[\frac{1}{2N_{\mathrm{WF}}} + \frac{2N_{\mathrm{WF}}-1}{2N_{\mathrm{WF}}}\mathbb{P}(x_i' = y_i')\right]$$

At stationarity, $\mathbb{P}(x_i = y_i) = \mathbb{P}(x_i' = y_i')$, so we can solve for the stationary probability $\mathbb{P}(x_i = y_i)$.

The ultimate goal

Want to compute $\mathbb{P}[(x_1, \ldots, x_L) = (y_1, \ldots, y_L)]$.

General strategy

Given a match relation R, use

$$\mathbb{P}(R) = \sum_{\text{Ancestry}} \mathbb{P}(R \mid \text{Ancestry})\, \mathbb{P}(\text{Ancestry})$$

to generate a recurrence equation of form $\mathbb{P}(R) = \sum_k c_k \mathbb{P}(R'_k)$,

where c_k are coefficients which depend on N and $\mu_1, \ldots, \mu_L$.
Laurie and Weir (2003) adopted the same strategy.

Problem

For large L, there are many ancestries and many match relations to consider.

The ultimate goal

Want to compute $\mathbb{P}[(x_1, \ldots, x_L) = (y_1, \ldots, y_L)]$.

General strategy

Given a match relation R, use

$$\mathbb{P}(R) = \sum_{\text{Ancestry}} \mathbb{P}(R \mid \text{Ancestry})\, \mathbb{P}(\text{Ancestry})$$

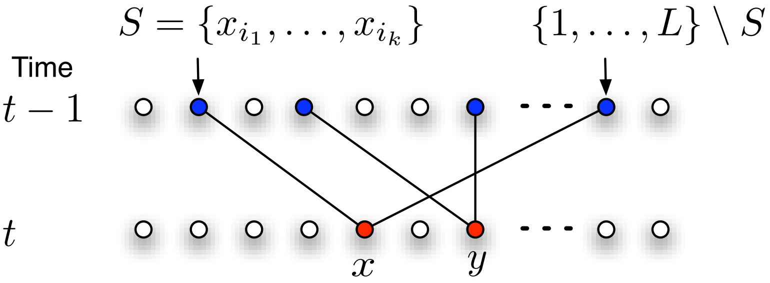
to generate a recurrence equation of form $\mathbb{P}(R) = \sum_k c_k \mathbb{P}(R'_k)$, where c_k are coefficients which depend on N and $\mu_1, \ldots, \mu_L$. Laurie and Weir (2003) adopted the same strategy.

Problem

For large L, there are many ancestries and many match relations to consider.

The ultimate goal

Want to compute $\mathbb{P}[(x_1, \ldots, x_L) = (y_1, \ldots, y_L)]$.


$S = \{x_{i_1}, \ldots, x_{i_k}\}$
$\{1, \ldots, L\} \setminus S$
Time
$t-1$
t
x
y

Problem

For large L, there are many ancestries and many match relations to consider.

Question

How many inequivalent match relations do we need to consider for the 13-locus haplotypic match probability computation?

Question

How many inequivalent match relations do we need to consider for the 13-locus haplotypic match probability computation?

General case

For arbitrary mutation rates $\mu_1, \ldots, \mu_{13}$, we need to consider 2021616201559793 inequivalent match relations.

Question

How many inequivalent match relations do we need to consider for the 13-locus haplotypic match probability computation?

General case

For arbitrary mutation rates $\mu_1, \ldots, \mu_{13}$, we need to consider 2021616201559793 inequivalent match relations.

A special case

For $\mu_1 = \mu_2 = \cdots = \mu_{13}$, we need to consider 3112753 inequivalent match relations.

Question

How many inequivalent match relations do we need to consider for the 13-locus haplotypic match probability computation?

General case

For arbitrary mutation rates $\mu_1, \ldots, \mu_{13}$, we need to consider 2021616201559793 inequivalent match relations.

A special case

For $\mu_1 = \mu_2 = \cdots = \mu_{13}$, we need to consider 3112753 inequivalent match relations.

Question

How do we generate the recurrence relations satisfied by those match relations?

Outline

1. Introduction
   - Random match probability
   - Cold hit
2. Models of Random Mating
   - Recurrence equations
3. Graphical Framework
   - Match graphs
   - Operations on graphs
   - Topological ordering and graph enumeration
4. Results
   - Accuracy of the product rule
   - Wright-Fisher vs. Moran
   - Excluding siblings
5. Other Works
   - Perfect Monogamy Model
   - Subdivided populations

Match graphs

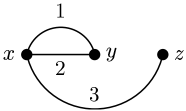
We have developed a simple and flexible graphical framework for computing match probabilities. (Song and Slatkin, 2007)

From match probabilities to match graphs

- Match graph:
  - **Vertex:** Create a vertex labeled x for gamete x.
  - **Edge:** Draw an undirected edge labeled i between vertices x and y if and only if $x_i = y_i$.
- 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).

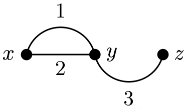
$$\mathbb{P}(x_1 = y_1, x_2 = y_2, x_3 = z_3)$$

$G_1 =$



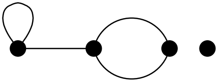
$$\mathbb{P}(x_1 = y_1, x_2 = y_2, y_3 = z_3)$$

$G_2 =$

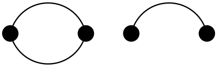


Observation

There is a 1-to-1 correspondence between the set of L-locus match graphs and the set of loopless multigraphs with L edges and non-isolated vertices.



Looped multigraph



Loopless multigraph

General case

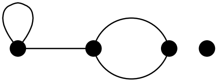
For arbitrary mutation rates $\mu_1, \ldots, \mu_{13}$, we need to consider loopless multigraphs with k labeled edges, for $k = 1, \ldots 13$.

A special case

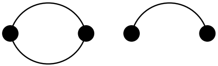
For $\mu_1 = \mu_2 = \cdots = \mu_{13}$, we need to consider consider loopless multigraphs with k unlabeled edges, for $k = 1, \ldots 13$.

Observation

There is a 1-to-1 correspondence between the set of L-locus match graphs and the set of loopless multigraphs with L edges and non-isolated vertices.



Looped multigraph



Loopless multigraph

General case

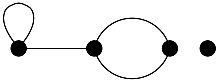
For arbitrary mutation rates $\mu_1, \ldots, \mu_{13}$, we need to consider loopless multigraphs with k labeled edges, for $k = 1, \ldots 13$.

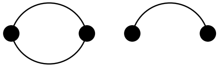
A special case

For $\mu_1 = \mu_2 = \cdots = \mu_{13}$, we need to consider consider loopless multigraphs with k unlabeled edges, for $k = 1, \ldots 13$.

Observation

There is a 1-to-1 correspondence between the set of L-locus match graphs and the set of loopless multigraphs with L edges and non-isolated vertices.


Looped multigraph


Loopless multigraph

General case

For arbitrary mutation rates $\mu_1, \ldots, \mu_{13}$, we need to consider loopless multigraphs with k labeled edges, for $k = 1, \ldots 13$.

A special case

For $\mu_1 = \mu_2 = \cdots = \mu_{13}$, we need to consider consider loopless multigraphs with k unlabeled edges, for $k = 1, \ldots 13$.

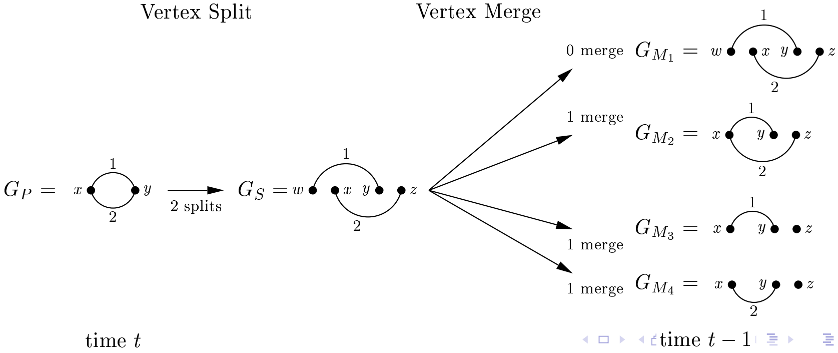
	Number of loopless multigraphs with *L* edges	
L	Edge labeled	Edge unlabeled
1	1	1
2	3	3
3	16	8
4	139	23
5	1 750	66
6	29 388	212
7	624 889	686
8	16 255 738	2 389
9	504 717 929	8 682
10	18 353 177 160	33 160
11	769 917 601 384	132 277
12	36 803 030 137 203	550 835
13	1 984 024 379 014 193	2 384 411
Total	2 021 616 201 559 793	3 112 753

Labelle (2000), Harary and Palmer (1973)

Finding recurrence equations

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 $t-1$.

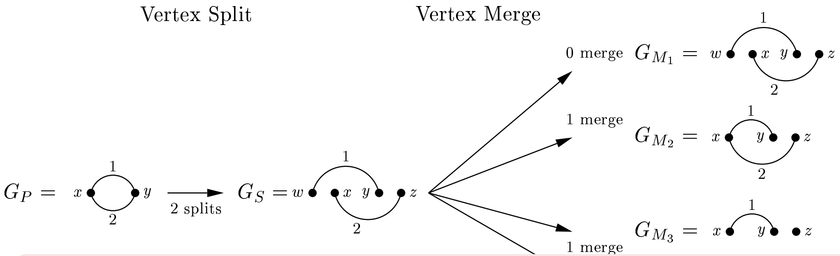
1. Vertex Split (inheritance pattern across loci for each gamete)
2. Vertex Merge (sharing of parents by two or more gametes)



Finding recurrence equations

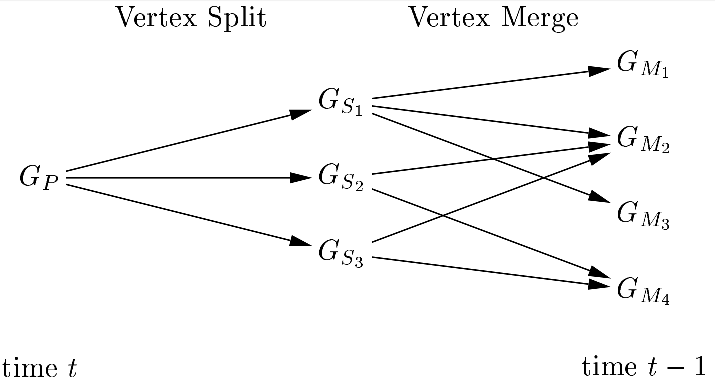
By performing a set of prescribed operations on a given graph at generation *t*, we determine how it is related to a linear combination of graphs at generation *t* − 1.

1. Vertex Split (inheritance pattern across loci for each gamete)
2. Vertex Merge (sharing of parents by two or more gametes)

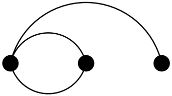

Vertex Split
Vertex Merge
$G_P =$ x y 1 2
2 splits
$G_S =$ w x y z 1 2
0 merge $G_{M_1} =$ w x y z 1 2
1 merge $G_{M_2} =$ x y z 1 2
1 merge $G_{M_3} =$ x y z 1

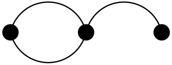
Split-merge operations have associated probabilities which appear as coefficients in recurrence equations.

Summary

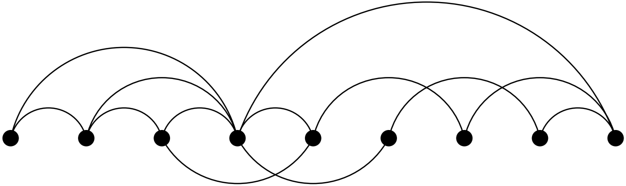

Vertex Split
Vertex Merge
G_P
G_{S_1}
G_{S_2}
G_{S_3}
G_{M_1}
G_{M_2}
G_{M_3}
G_{M_4}
time t
time $t-1$

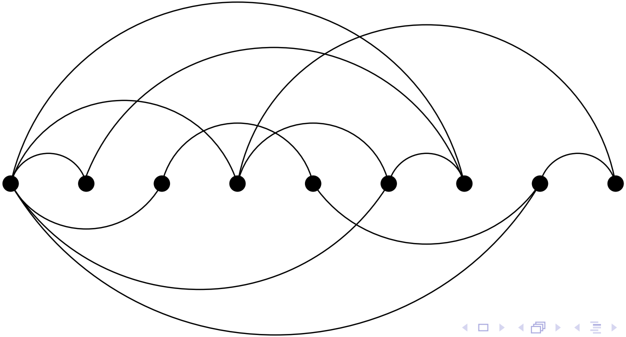
Clearly, these graphs are isomorphic.



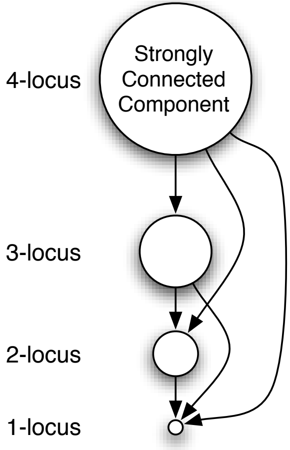


How about these?

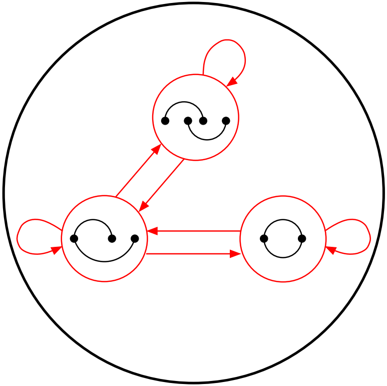




Topological Ordering of the System

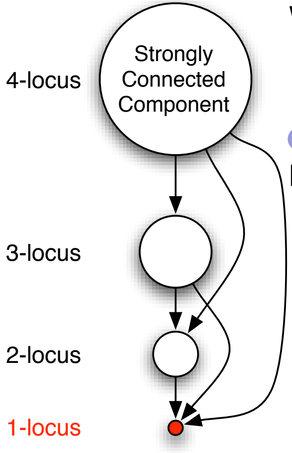


A closer look at the 2-locus SCC for the Moran model



1-locus case: 1 equation

Topological Ordering of the System



Wright-Fisher model:

$$\overset{\frown}{\bullet\;\;\bullet} = (1-\mu)^2\left[\frac{2N_{\mathrm{WF}}-1}{2N_{\mathrm{WF}}}\,\overset{\frown}{\bullet\;\;\bullet}+\frac{1}{2N_{\mathrm{WF}}}\right]$$

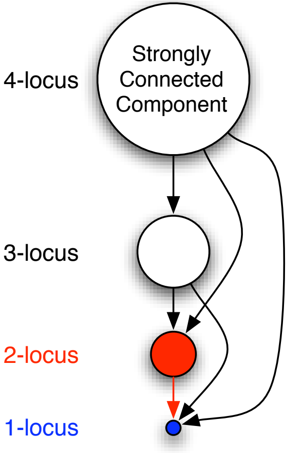
▸ Ancestry

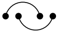
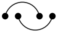
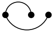
Moran model:

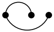
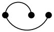
$$\overset{\frown}{\bullet\;\;\bullet} = \left[\frac{2N_{\mathrm{M}}-2}{2N_{\mathrm{M}}}+\frac{2N_{\mathrm{M}}-1}{(2N_{\mathrm{M}})^2}2(1-\mu)\right]\overset{\frown}{\bullet\;\;\bullet}+\frac{2(1-\mu)}{(2N_{\mathrm{M}})^2}$$

2-locus case: 3 coupled equations

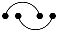
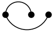
Topological Ordering of the System



 $= \Big[\frac{2N_{\rm M}-4}{2N_{\rm M}} + \frac{2N_{\rm M}-3}{(2N_{\rm M})^2}\cdot 4(1-\mu)\Big]$  $+\frac{2(1-\mu)}{(2N_{\rm M})^2}\Big(4$  $+2$ $\Big)$

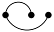
 $= \Big\{\frac{2N_{\rm M}-3}{2N_{\rm M}} + \frac{2N_{\rm M}-2}{(2N_{\rm M})^2}[2(1-\mu)+(1-r)(1-\mu)^2]\Big\}$ 

$+\frac{1}{(2N_{\rm M})^2}\Big\{2(1-\mu)$ $+2[(1-r)(1-\mu)^2+(1-\mu)]$ $\Big\}$

$+\frac{(1-\mu)^2}{(2N_{\rm M})^3}\cdot r\Big\{(2N_{\rm M}-2)(2N_{\rm M}-3)$  $+3(2N_{\rm M}-2)$ 

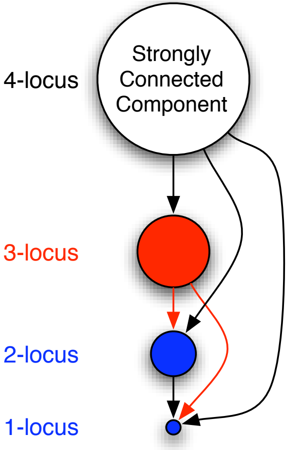
$+$ $+2(2N_{\rm M}-1)$ $+1\Big\}$

$= \Big[\frac{2N_{\rm M}-2}{2N_{\rm M}} + \frac{2N_{\rm M}-1}{(2N_{\rm M})^2}\cdot 2(1-\mu)^2(1-r)\Big]$ $+\frac{1}{(2N_{\rm M})^2}2(1-\mu)^2(1-r)$

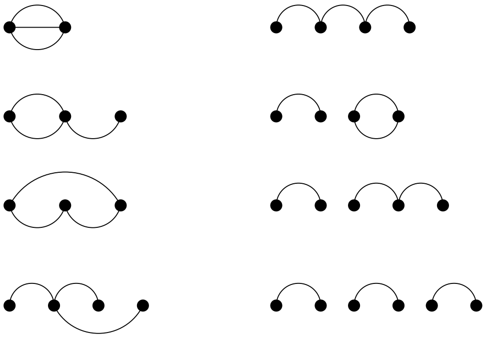
$+\frac{(1-\mu)^2}{(2N_{\rm M})^3}\cdot 2r\Big\{(2N_{\rm M}-1)(2N_{\rm M}-2)$  $+(2N_{\rm M}-1)\Big[2$ $+$ $\Big]+1\Big\}$

1-locus match graph appears as a known constant.

Topological Ordering of the System

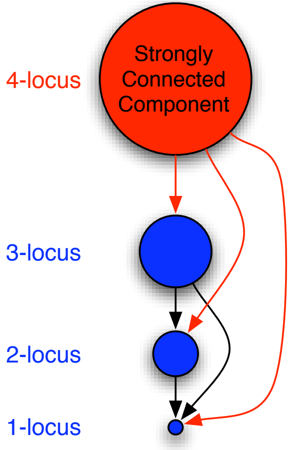


3-locus case: 8 coupled equations



1-locus and 2-locus match graphs are treated as known constants.

Topological Ordering of the System

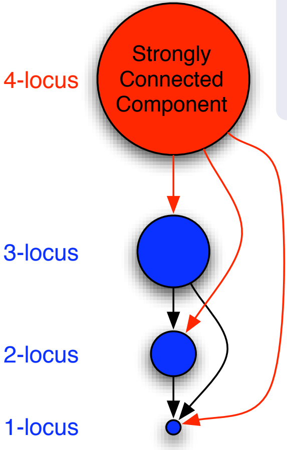


4-locus case: 23 coupled equations

So and so forth.

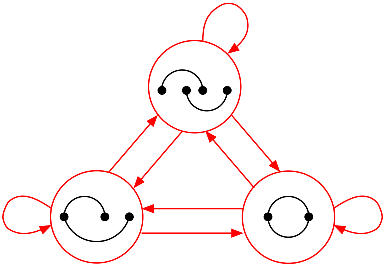
Topological ordering and graph enumeration

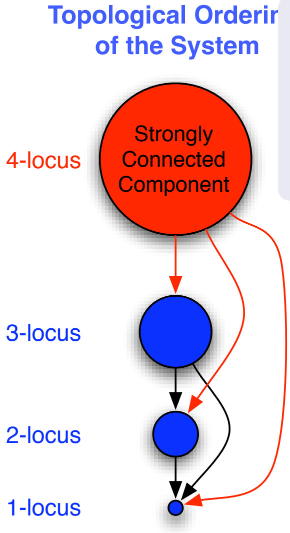
Topological Orderir of the System



- WF and Moran models have exactly the same set of match graphs.
- But, the WF model has significantly more directed edges in each strongly connected component.

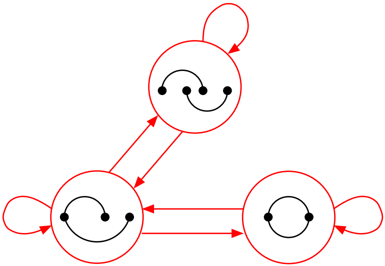
2-locus SCC for the WF model





- WF and Moran models have exactly the same set of match graphs.
- But, the WF model has significantly more directed edges in each strongly connected component.

2-locus SCC for the Moran model



- Our graphical approach makes the combinatorial structure of the problem easier to understand.
- We implemented our method in a fully automated program, thus reducing the chance of human error.

Related Problems

1. Graph isomorphism testing. (We used the *nauty* package.)
2. Canonical encoding of graphs.
3. Equivalence of split-merge operations. Two different vertex split-merge operations on a graph with symmetries may produce isomorphic match graphs.
4. Solving a large linear system of equations. (We used the iterative Successive Over-Relaxation method.)

Outline

1. Introduction
   - Random match probability
   - Cold hit
2. Models of Random Mating
   - Recurrence equations
3. Graphical Framework
   - Match graphs
   - Operations on graphs
   - Topological ordering and graph enumeration
4. Results
   - Accuracy of the product rule
   - Wright-Fisher vs. Moran
   - Excluding siblings
5. Other Works
   - Perfect Monogamy Model
   - Subdivided populations

Moran model MPs for $N_e = 10,000$ and $\mu_i = \mu$ for all loci i:

L	Prod. Rule	True $MP(L)$	Prod. Rule	True $MP(L)$	Prod. Rule	True $MP(L)$
	$\mu = 1 \times 10^{-4}$		$\mu = 2 \times 10^{-4}$		$\mu = 3 \times 10^{-4}$	
1	2.00×10^{-1}	2.00×10^{-1}	1.11×10^{-1}	1.11×10^{-1}	7.69×10^{-2}	7.69×10^{-2}
2	4.00×10^{-2}	4.00×10^{-2}	1.23×10^{-2}	1.24×10^{-2}	5.91×10^{-3}	5.94×10^{-3}
3	8.00×10^{-3}	8.01×10^{-3}	1.37×10^{-3}	1.38×10^{-3}	4.55×10^{-4}	4.66×10^{-4}
4	1.60×10^{-3}	1.61×10^{-3}	1.52×10^{-4}	1.59×10^{-4}	3.50×10^{-5}	4.03×10^{-5}
5	3.20×10^{-4}	3.25×10^{-4}	1.69×10^{-5}	2.01×10^{-5}	2.69×10^{-6}	5.29×10^{-6}
6	6.40×10^{-5}	6.68×10^{-5}	1.88×10^{-6}	3.51×10^{-6}	2.07×10^{-7}	1.52×10^{-6}
7	1.28×10^{-5}	1.44×10^{-5}	2.09×10^{-7}	1.08×10^{-6}	1.59×10^{-8}	7.00×10^{-7}
8	2.56×10^{-6}	3.48×10^{-6}	2.32×10^{-8}	4.94×10^{-7}	1.22×10^{-9}	3.63×10^{-7}
9	5.11×10^{-7}	1.05×10^{-6}	2.57×10^{-9}	2.60×10^{-7}	9.39×10^{-11}	1.93×10^{-7}
10	1.02×10^{-7}	4.16×10^{-7}	2.86×10^{-10}	1.42×10^{-7}	7.22×10^{-12}	1.03×10^{-7}
11	2.05×10^{-8}	2.06×10^{-7}	3.18×10^{-11}	7.84×10^{-8}	5.55×10^{-13}	5.54×10^{-8}
12	4.09×10^{-9}	1.15×10^{-7}	3.53×10^{-12}	4.35×10^{-8}	4.27×10^{-14}	2.98×10^{-8}
13	8.18×10^{-10}	6.69×10^{-8}	3.92×10^{-13}	2.41×10^{-8}	3.28×10^{-15}	1.60×10^{-8}

Recently, we succeeded in computing haplotypic MPs for up to 10 loci in the WF model, and up to 13 loci in the Moran model.
(Bhaskar and Song, *ISMB 2009, in press*)

Moran model MPs for $N_e = 10,000$ and $\mu_i = \mu$ for all loci i:

L	Prod. Rule	True $MP(L)$	Prod. Rule	True $MP(L)$	Prod. Rule	True $MP(L)$
	$\mu = 1 \times 10^{-4}$		$\mu = 2 \times 10^{-4}$		$\mu = 3 \times 10^{-4}$	
1	2.00×10^{-1}	2.00×10^{-1}	1.11×10^{-1}	1.11×10^{-1}	7.69×10^{-2}	7.69×10^{-2}
2	4.00×10^{-2}	4.00×10^{-2}	1.23×10^{-2}	1.24×10^{-2}	5.91×10^{-3}	5.94×10^{-3}
3	8.00×10^{-3}	8.01×10^{-3}	1.37×10^{-3}	1.38×10^{-3}	4.55×10^{-4}	4.66×10^{-4}
4	1.60×10^{-3}	1.61×10^{-3}	1.52×10^{-4}	1.59×10^{-4}	3.50×10^{-5}	4.03×10^{-5}
5	3.20×10^{-4}	3.25×10^{-4}	1.69×10^{-5}	2.01×10^{-5}	2.69×10^{-6}	5.29×10^{-6}
6	6.40×10^{-5}	6.68×10^{-5}	1.88×10^{-6}	3.51×10^{-6}	2.07×10^{-7}	1.52×10^{-6}
7	1.28×10^{-5}	1.44×10^{-5}	2.09×10^{-7}	1.08×10^{-6}	1.59×10^{-8}	7.00×10^{-7}
8	2.56×10^{-6}	3.48×10^{-6}	2.32×10^{-8}	4.94×10^{-7}	1.22×10^{-9}	3.63×10^{-7}
9	5.11×10^{-7}	1.05×10^{-6}	2.57×10^{-9}	2.60×10^{-7}	9.39×10^{-11}	1.93×10^{-7}
10	1.02×10^{-7}	4.16×10^{-7}	2.86×10^{-10}	1.42×10^{-7}	7.22×10^{-12}	1.03×10^{-7}
11	2.05×10^{-8}	2.06×10^{-7}	3.18×10^{-11}	7.84×10^{-8}	5.55×10^{-13}	5.54×10^{-8}
12	4.09×10^{-9}	1.15×10^{-7}	3.53×10^{-12}	4.35×10^{-8}	4.27×10^{-14}	2.98×10^{-8}
13	8.18×10^{-10}	6.69×10^{-8}	3.92×10^{-13}	2.41×10^{-8}	3.28×10^{-15}	1.60×10^{-8}

- For a give mutation rate μ, the product rule becomes less accurate as the number of loci increases.
- Furthermore, for a large number L of loci, a slight change in μ causes the product rule MP to decrease by a large amount.

Moran model MPs for $N_e = 10,000$ and $\mu_i = \mu$ for all loci i:

L	Prod. Rule	True $MP(L)$	Prod. Rule	True $MP(L)$	Prod. Rule	True $MP(L)$
	$\mu = 1 \times 10^{-4}$		$\mu = 2 \times 10^{-4}$		$\mu = 3 \times 10^{-4}$	
1	2.00×10^{-1}	2.00×10^{-1}	1.11×10^{-1}	1.11×10^{-1}	7.69×10^{-2}	7.69×10^{-2}
2	4.00×10^{-2}	4.00×10^{-2}	1.23×10^{-2}	1.24×10^{-2}	5.91×10^{-3}	5.94×10^{-3}
3	8.00×10^{-3}	8.01×10^{-3}	1.37×10^{-3}	1.38×10^{-3}	4.55×10^{-4}	4.66×10^{-4}
4	1.60×10^{-3}	1.61×10^{-3}	1.52×10^{-4}	1.59×10^{-4}	3.50×10^{-5}	4.03×10^{-5}
5	3.20×10^{-4}	3.25×10^{-4}	1.69×10^{-5}	2.01×10^{-5}	2.69×10^{-6}	5.29×10^{-6}
6	6.40×10^{-5}	6.68×10^{-5}	1.88×10^{-6}	3.51×10^{-6}	2.07×10^{-7}	1.52×10^{-6}
7	1.28×10^{-5}	1.44×10^{-5}	2.09×10^{-7}	1.08×10^{-6}	1.59×10^{-8}	7.00×10^{-7}
8	2.56×10^{-6}	3.48×10^{-6}	2.32×10^{-8}	4.94×10^{-7}	1.22×10^{-9}	3.63×10^{-7}
9	5.11×10^{-7}	1.05×10^{-6}	2.57×10^{-9}	2.60×10^{-7}	9.39×10^{-11}	1.93×10^{-7}
10	1.02×10^{-7}	4.16×10^{-7}	2.86×10^{-10}	1.42×10^{-7}	7.22×10^{-12}	1.03×10^{-7}
11	2.05×10^{-8}	2.06×10^{-7}	3.18×10^{-11}	7.84×10^{-8}	5.55×10^{-13}	5.54×10^{-8}
12	4.09×10^{-9}	1.15×10^{-7}	3.53×10^{-12}	4.35×10^{-8}	4.27×10^{-14}	2.98×10^{-8}
13	8.18×10^{-10}	6.69×10^{-8}	3.92×10^{-13}	2.41×10^{-8}	3.28×10^{-15}	1.60×10^{-8}

- The observed homozygosity at the CODIS microsatellite loci typically ranges between 0.1 and 0.3, with the average over all 13 loci being about 0.2 (Budowle *et. al*, 2001).
- Under the infinite alleles model with $N_e = 10,000$, homozygosity $= 0.2$ corresponds to $\mu = 10^{-4}$.

Moran model MPs for $N_e = 10,000$ and $\mu_i = \mu$ for all loci i:

L	Prod. Rule	True $MP(L)$	Prod. Rule	True $MP(L)$	Prod. Rule	True $MP(L)$
	$\mu = 1 \times 10^{-4}$		$\mu = 2 \times 10^{-4}$		$\mu = 3 \times 10^{-4}$	
1	2.00×10^{-1}	2.00×10^{-1}	1.11×10^{-1}	1.11×10^{-1}	7.69×10^{-2}	7.69×10^{-2}
2	4.00×10^{-2}	4.00×10^{-2}	1.23×10^{-2}	1.24×10^{-2}	5.91×10^{-3}	5.94×10^{-3}
3	8.00×10^{-3}	8.01×10^{-3}	1.37×10^{-3}	1.38×10^{-3}	4.55×10^{-4}	4.66×10^{-4}
4	1.60×10^{-3}	1.61×10^{-3}	1.52×10^{-4}	1.59×10^{-4}	3.50×10^{-5}	4.03×10^{-5}
5	3.20×10^{-4}	3.25×10^{-4}	1.69×10^{-5}	2.01×10^{-5}	2.69×10^{-6}	5.29×10^{-6}
6	6.40×10^{-5}	6.68×10^{-5}	1.88×10^{-6}	3.51×10^{-6}	2.07×10^{-7}	1.52×10^{-6}
7	1.28×10^{-5}	1.44×10^{-5}	2.09×10^{-7}	1.08×10^{-6}	1.59×10^{-8}	7.00×10^{-7}
8	2.56×10^{-6}	3.48×10^{-6}	2.32×10^{-8}	4.94×10^{-7}	1.22×10^{-9}	3.63×10^{-7}
9	5.11×10^{-7}	1.05×10^{-6}	2.57×10^{-9}	2.60×10^{-7}	9.39×10^{-11}	1.93×10^{-7}
10	1.02×10^{-7}	4.16×10^{-7}	2.86×10^{-10}	1.42×10^{-7}	7.22×10^{-12}	1.03×10^{-7}
11	2.05×10^{-8}	2.06×10^{-7}	3.18×10^{-11}	7.84×10^{-8}	5.55×10^{-13}	5.54×10^{-8}
12	4.09×10^{-9}	1.15×10^{-7}	3.53×10^{-12}	4.35×10^{-8}	4.27×10^{-14}	2.98×10^{-8}
13	8.18×10^{-10}	6.69×10^{-8}	3.92×10^{-13}	2.41×10^{-8}	3.28×10^{-15}	1.60×10^{-8}

- For this value of μ, the product rule is reasonably accurate, especially for $L \leq 10$.
- But, for $\mu = 2 \times 10^{-4}$, which corresponds to homozygosity = 0.11, the product rule produces considerably less accurate MPs.

Wright-Fisher vs Moran (for $N_e = 10,000$)

L	WF	Moran	WF	Moran	WF	Moran
	$\mu = 1 \times 10^{-4}$		$\mu = 2 \times 10^{-4}$		$\mu = 3 \times 10^{-4}$	
1	2.00×10^{-1}	2.00×10^{-1}	1.11×10^{-1}	1.11×10^{-1}	7.69×10^{-2}	7.69×10^{-2}
2	4.00×10^{-2}	4.00×10^{-2}	1.24×10^{-2}	1.24×10^{-2}	5.93×10^{-3}	5.94×10^{-3}
3	8.01×10^{-3}	8.01×10^{-3}	1.38×10^{-3}	1.38×10^{-3}	4.60×10^{-4}	4.66×10^{-4}
4	1.60×10^{-3}	1.61×10^{-3}	1.55×10^{-4}	1.59×10^{-4}	3.68×10^{-5}	4.03×10^{-5}
5	3.22×10^{-4}	3.25×10^{-4}	1.78×10^{-5}	2.01×10^{-5}	3.26×10^{-6}	5.29×10^{-6}
6	6.48×10^{-5}	6.68×10^{-5}	2.16×10^{-6}	3.51×10^{-6}	3.80×10^{-7}	1.52×10^{-6}
7	1.31×10^{-5}	1.44×10^{-5}	3.02×10^{-7}	1.08×10^{-6}	6.86×10^{-8}	7.00×10^{-7}
8	2.69×10^{-6}	3.48×10^{-6}	5.41×10^{-8}	4.94×10^{-7}	1.74×10^{-8}	3.63×10^{-7}
9	5.65×10^{-7}	1.05×10^{-6}	1.28×10^{-8}	2.60×10^{-7}	5.08×10^{-9}	1.93×10^{-7}
10	1.24×10^{-7}	4.16×10^{-7}	3.72×10^{-9}	1.42×10^{-7}	1.55×10^{-9}	1.03×10^{-7}

- The two models agree very well in the single locus case.
- However, for large values of L, MPs in the Moran model can be orders of magnitude higher than that in the WF model.
- This difference grows with the number of loci and mutation rates.

The same diffusion limit

Send $\mu \to 0$ and $N_e \to \infty$ while keeping $\theta = 4N_e\mu$ fixed. Then,

$$L\text{-locus MP} \to \left(\frac{1}{1+\theta}\right)^L .$$

in both the WF and Moran models.

The same diffusion limit

Send $\mu \to 0$ and $N_e \to \infty$ while keeping $\theta = 4N_e\mu$ fixed. Then,

$$L\text{-locus MP} \to \left(\frac{1}{1+\theta}\right)^L .$$

in both the WF and Moran models.

Match probabilities for $N_e = 10^4$ and $\mu = 10^{-3}$.

L	$1/(1+\theta)^L$	WF	Moran
1	2.44×10^{-2}	2.44×10^{-2}	2.44×10^{-2}
2	5.95×10^{-4}	6.09×10^{-4}	6.17×10^{-4}
3	1.45×10^{-5}	1.87×10^{-5}	2.39×10^{-5}
4	3.54×10^{-7}	1.42×10^{-6}	4.41×10^{-6}
5	8.63×10^{-9}	2.88×10^{-7}	1.92×10^{-6}
6	2.11×10^{-10}	7.45×10^{-8}	9.38×10^{-7}
7	5.13×10^{-12}	1.99×10^{-8}	4.70×10^{-7}
8	1.25×10^{-13}	5.36×10^{-9}	2.39×10^{-7}
9	3.05×10^{-15}	1.45×10^{-9}	1.21×10^{-7}

The same diffusion limit

Send $\mu \to 0$ and $N_e \to \infty$ while keeping $\theta = 4N_e\mu$ fixed. Then,

$$L\text{-locus MP} \to \left(\frac{1}{1+\theta}\right)^L .$$

in both the WF and Moran models.

Match probabilities for $N_e = 10^9$ and $\mu = 10^{-8}$.

L	$1/(1+\theta)^L$	WF	Moran
1	2.44×10^{-2}	2.44×10^{-2}	2.44×10^{-2}
2	5.95×10^{-4}	5.95×10^{-4}	5.95×10^{-4}
3	1.45×10^{-5}	1.45×10^{-5}	1.45×10^{-5}
4	3.54×10^{-7}	3.54×10^{-7}	3.54×10^{-7}
5	8.63×10^{-9}	8.63×10^{-9}	8.65×10^{-9}
6	2.11×10^{-10}	2.11×10^{-10}	2.20×10^{-10}
7	5.13×10^{-12}	5.34×10^{-12}	9.86×10^{-12}
8	1.25×10^{-13}	1.79×10^{-13}	2.52×10^{-12}
9	3.05×10^{-15}	1.75×10^{-14}	1.22×10^{-12}

MPs conditioned on the event that the two individuals being compared are neither full-sibs nor half-sibs.

- This computation can be carried out by restricting vertex-merge operations.
- The product rule becomes much more accurate if we are provided with the additional information that the individuals being compared are not close relatives.

L	Prod. Rule	WF	Prod. Rule	WF	Prod. Rule	WF
	$\mu = 1 \times 10^{-4}$		$\mu = 5 \times 10^{-4}$		$\mu = 1 \times 10^{-3}$	
1	2.00×10^{-1}	2.00×10^{-1}	4.75×10^{-2}	4.75×10^{-2}	2.43×10^{-2}	2.43×10^{-2}
2	4.00×10^{-2}	4.00×10^{-2}	2.26×10^{-3}	2.26×10^{-3}	5.91×10^{-4}	5.95×10^{-4}
3	7.99×10^{-3}	7.99×10^{-3}	1.07×10^{-4}	1.08×10^{-4}	1.44×10^{-5}	1.48×10^{-5}
4	1.60×10^{-3}	1.60×10^{-3}	5.11×10^{-6}	5.20×10^{-6}	3.49×10^{-7}	3.93×10^{-7}
5	3.19×10^{-4}	3.20×10^{-4}	2.43×10^{-7}	2.54×10^{-7}	8.48×10^{-9}	1.22×10^{-8}
6	6.39×10^{-5}	6.39×10^{-5}	1.15×10^{-8}	1.28×10^{-8}	2.06×10^{-10}	5.19×10^{-10}
7	1.28×10^{-5}	1.28×10^{-5}	5.48×10^{-10}	6.81×10^{-10}	5.01×10^{-12}	3.15×10^{-11}
8	2.55×10^{-6}	2.56×10^{-6}	2.61×10^{-11}	4.02×10^{-11}	1.22×10^{-13}	2.39×10^{-12}
9	5.10×10^{-7}	5.12×10^{-7}	1.24×10^{-12}	2.76×10^{-12}	2.96×10^{-15}	2.00×10^{-13}
10	1.02×10^{-7}	1.03×10^{-7}	5.89×10^{-14}	2.23×10^{-13}	7.19×10^{-17}	1.74×10^{-14}

Excluding siblings

No analogous results for the Moran model.

L	Prod. Rule	WF	Prod. Rule	WF	Prod. Rule	WF
	$\mu = 1 \times 10^{-4}$		$\mu = 5 \times 10^{-4}$		$\mu = 1 \times 10^{-3}$	
1	2.00×10^{-1}	2.00×10^{-1}	4.75×10^{-2}	4.75×10^{-2}	2.43×10^{-2}	2.43×10^{-2}
2	4.00×10^{-2}	4.00×10^{-2}	2.26×10^{-3}	2.26×10^{-3}	5.91×10^{-4}	5.95×10^{-4}
3	7.99×10^{-3}	7.99×10^{-3}	1.07×10^{-4}	1.08×10^{-4}	1.44×10^{-5}	1.48×10^{-5}
4	1.60×10^{-3}	1.60×10^{-3}	5.11×10^{-6}	5.20×10^{-6}	3.49×10^{-7}	3.93×10^{-7}
5	3.19×10^{-4}	3.20×10^{-4}	2.43×10^{-7}	2.54×10^{-7}	8.48×10^{-9}	1.22×10^{-8}
6	6.39×10^{-5}	6.39×10^{-5}	1.15×10^{-8}	1.28×10^{-8}	2.06×10^{-10}	5.19×10^{-10}
7	1.28×10^{-5}	1.28×10^{-5}	5.48×10^{-10}	6.81×10^{-10}	5.01×10^{-12}	3.15×10^{-11}
8	2.55×10^{-6}	2.56×10^{-6}	2.61×10^{-11}	4.02×10^{-11}	1.22×10^{-13}	2.39×10^{-12}
9	5.10×10^{-7}	5.12×10^{-7}	1.24×10^{-12}	2.76×10^{-12}	2.96×10^{-15}	2.00×10^{-13}
10	1.02×10^{-7}	1.03×10^{-7}	5.89×10^{-14}	2.23×10^{-13}	7.19×10^{-17}	1.74×10^{-14}

Summary

1. For a finite population, the accuracy of multi-locus MPs predicted by the product rule is highly sensitive to mutation rates in the range of interest, while the true MPs are not.

Summary

1. For a finite population, the accuracy of multi-locus MPs predicted by the product rule is highly sensitive to mutation rates in the range of interest, while the true MPs are not.
2. We assumed an infinite alleles model, in which identity in allelic state implies identity by descent. Our work studies the effect of shared genealogies in a finite population on the joint probability of identity by descent.

Summary

1. For a finite population, the accuracy of multi-locus MPs predicted by the product rule is highly sensitive to mutation rates in the range of interest, while the true MPs are not.
2. We assumed an infinite alleles model, in which identity in allelic state implies identity by descent. Our work studies the effect of shared genealogies in a finite population on the joint probability of identity by descent.
3. We have revealed a striking difference between the Wright-Fisher and Moran models.

Summary

1. For a finite population, the accuracy of multi-locus MPs predicted by the product rule is highly sensitive to mutation rates in the range of interest, while the true MPs are not.
2. We assumed an infinite alleles model, in which identity in allelic state implies identity by descent. Our work studies the effect of shared genealogies in a finite population on the joint probability of identity by descent.
3. We have revealed a striking difference between the Wright-Fisher and Moran models.
4. Genealogical interpretation? We speculate that the times to the most recent common ancestors at unlinked loci are more correlated in the Moran model than in the WF model.

Summary

1. For a finite population, the accuracy of multi-locus MPs predicted by the product rule is highly sensitive to mutation rates in the range of interest, while the true MPs are not.
2. We assumed an infinite alleles model, in which identity in allelic state implies identity by descent. Our work studies the effect of shared genealogies in a finite population on the joint probability of identity by descent.
3. We have revealed a striking difference between the Wright-Fisher and Moran models.
4. Genealogical interpretation? We speculate that the times to the most recent common ancestors at unlinked loci are more correlated in the Moran model than in the WF model.
5. It is tempting to suspect that other quantities of interest to population genomics may be fundamentally different in the two models, especially when many loci are considered.

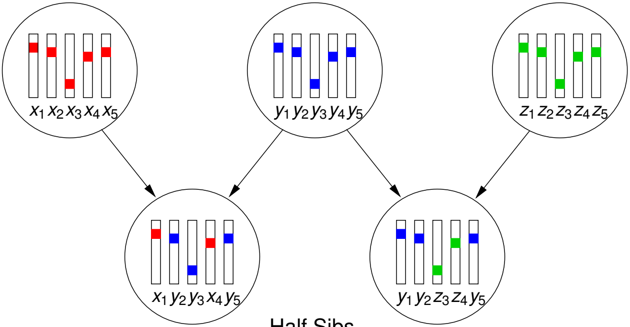
Outline

1. Introduction
   - Random match probability
   - Cold hit
2. Models of Random Mating
   - Recurrence equations
3. Graphical Framework
   - Match graphs
   - Operations on graphs
   - Topological ordering and graph enumeration
4. Results
   - Accuracy of the product rule
   - Wright-Fisher vs. Moran
   - Excluding siblings
5. Other Works
   - Perfect Monogamy Model
   - Subdivided populations

Using our graphical framework, we can consider other models of mating scheme.

Perfect Monogamy

Two gametes cannot be half sibs.

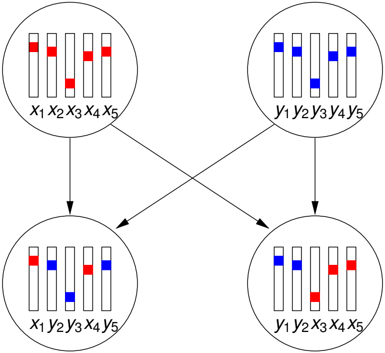

$x_1\, x_2\, x_3\, x_4\, x_5$
$y_1\, y_2\, y_3\, y_4\, y_5$
$z_1\, z_2\, z_3\, z_4\, z_5$
$x_1\, y_2\, y_3\, x_4\, y_5$
$y_1\, y_2\, z_3\, z_4\, y_5$

Half Sibs

Using our graphical framework, we can consider other models of mating scheme.

Perfect Monogamy

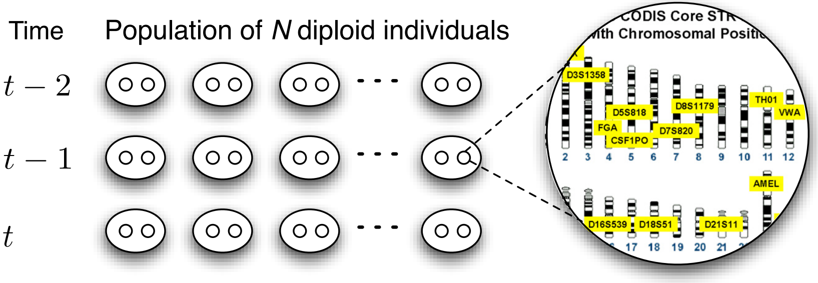
Two gametes cannot be half sibs.


$x_1\,x_2\,x_3\,x_4\,x_5$
$y_1\,y_2\,y_3\,y_4\,y_5$
$x_1\,y_2\,y_3\,x_4\,y_5$
$y_1\,y_2\,x_3\,x_4\,x_5$

Full Sibs

Biparental diploid model

The perfect monogamy haploid model just described is equivalent to a biparental diploid model.



Time

Population of *N* diploid individuals

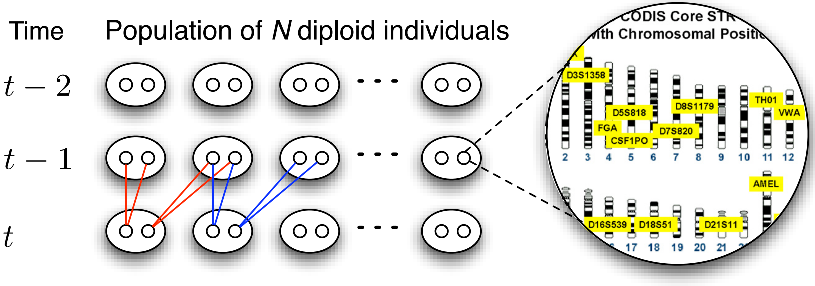
$t-2$

$t-1$

t

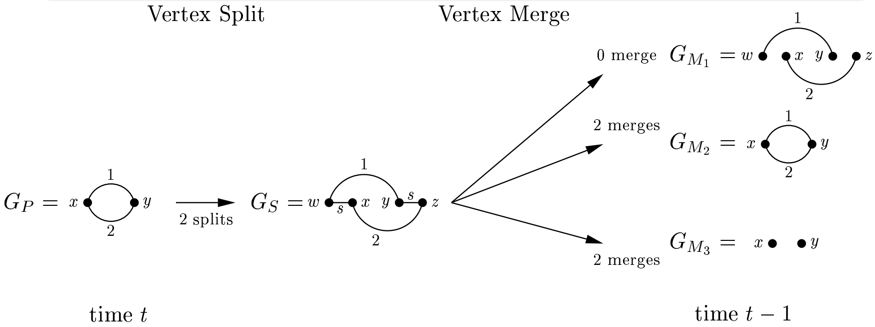
Biparental diploid model

The perfect monogamy haploid model just described is equivalent to a biparental diploid model.



Constraints on vertex merge under Perfect Monogamy

1. Two vertices joined by an edge labeled “*s*” may not merge.
2. Vertex merges may not produce a non-cyclic length-2 path $(\bullet\overset{s}{-\!-\!-\!-}\bullet\overset{s}{-\!-\!-\!-}\bullet)$ with both edges labeled “*s*”.



In a split graph G_S, add a new edge labeled “*s*” between the pair of vertices that arose from splitting a single vertex in G_P.

$$\frac{\text{Perfect monogamy MP}}{\text{Promiscuous mating MP}}$$

	μ					
L	1×10^{-4}	2×10^{-4}	3×10^{-4}	1×10^{-3}	1×10^{-2}	1×10^{-1}
2	1.000	1.001	1.002	1.026	1.723	1.995
3	1.001	1.008	1.024	1.556	3.914	3.992
4	1.006	1.049	1.188	5.184	7.828	7.977
5	1.019	1.259	2.240	12.248	15.573	15.929
6	1.062	2.246	6.994	24.018	30.930	31.768
7	1.192	6.122	19.341	45.882	61.286	63.210
8	1.580	17.218	40.575	87.134	120.899	125.190
9	2.699	39.413	74.664	164.510	236.485	245.708

Summary of results

- The effect of monogamy increases with the number of loci.
- For a given number of loci, the effect of monogamy increases with the mutation rate.

Upper bounds on the effect of monogamy for L loci

Consider the Wright-Fisher model with $\mu_i = \mu$ for all loci i.

Proposition

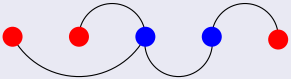
$$\lim_{\mu \uparrow 1} \frac{L\text{-locus MP under perfect monogamy}}{L\text{-locus MP under promiscuous mating}} = 2^{L-1} + O\left(\frac{1}{N_{\mathrm{WF}}}\right).$$

Subdivided populations

It is possible to incorporate population structure in the graphical framework.

Key idea

Use vertex-colored graphs. Different colors for different subpopulations.



(Joint work with Anna Malaspinas and Monty Slatkin.)

Recent California policy on familial search

- California recently implemented a policy for using partial DNA matches to identify potential close relatives of the individual who left a crime-scene sample.
- In addition to the 13-locus CODIS profiles, the policy also calls for using Y-linked markers to provide further evidence of relatedness.
- We just submitted a paper on the population genetics consequences of the policy. Specifically, we have an estimate on the number and ethnic distribution of false leads.
  (Joint work with Erin Murphy and Monty Slatkin.)

Thank you for your attention.

Acknowledgments

UC Berkeley

Monty Slatkin (Integrative Biology)
Erin Murphy (School of Law)
Anand Bhaskar (Computer Science)
Anna Malaspinas (Integrative Biology)

Research supported in part by NIH, NSF, Alfred P. Sloan Research Fellowship, and Packard Fellowship for Science and Engineering