Calculation of IBD probabilities

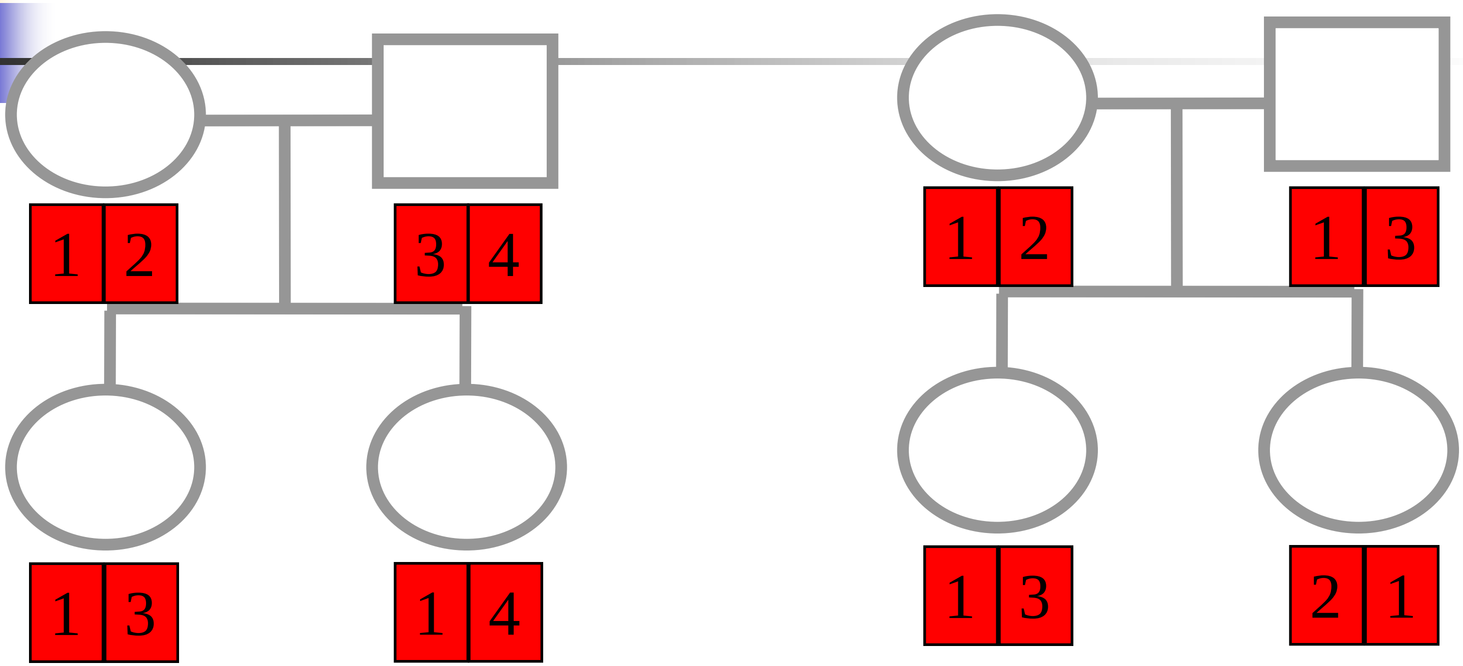
David Evans

University of BRISTOL

This Session …

- Identity by Descent (IBD) vs Identity by state (IBS)
- Why is IBD important?
- Calculating IBD probabilities
    - Lander-Green Algorithm (MERLIN)
        - Single locus probabilities
        - Hidden Markov Model => Multipoint IBD
    - Other ways of calculating IBD status
        - Elston-Stewart Algorithm
        - MCMC approaches
- MERLIN
- Practical Example
    - IBD determination
    - Information content mapping
        - SNPs vs micro-satellite markers?

Identity By Descent (IBD)



Identical by Descent

Identical by state only

Two alleles are IBD if they are descended from the same ancestral allele

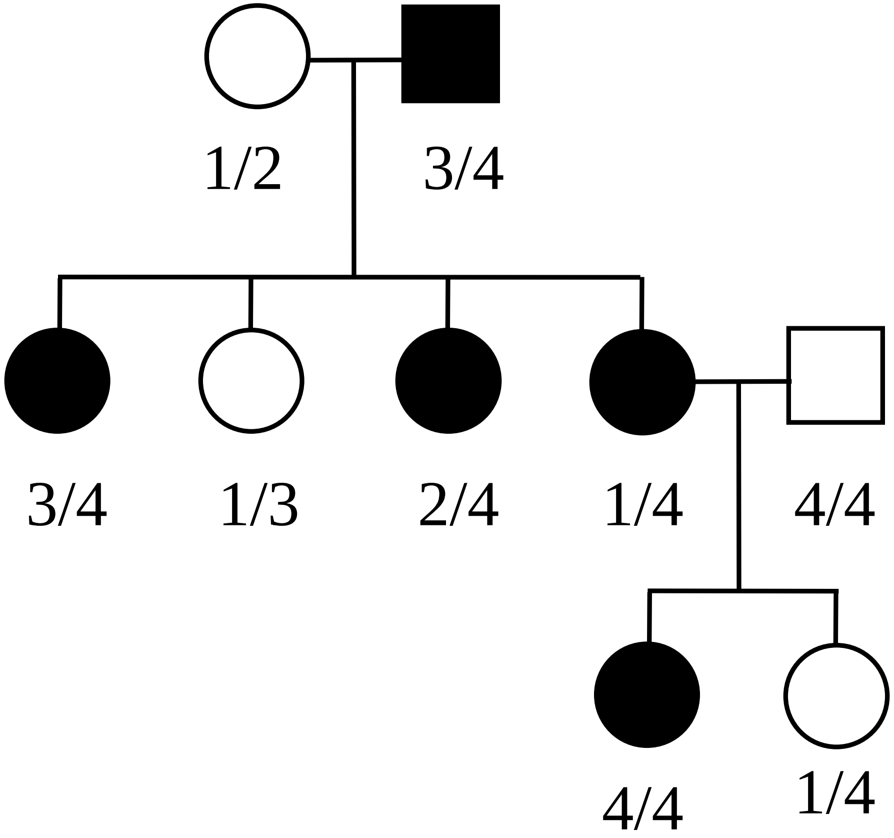
Example: IBD in Siblings

Consider a mating between mother AB x father CD:

		Sib1			
		AC	AD	BC	BD
Sib 2	AC	2	1	1	0
	AD	1	2	0	1
	BC	1	0	2	1
	BD	0	1	1	2

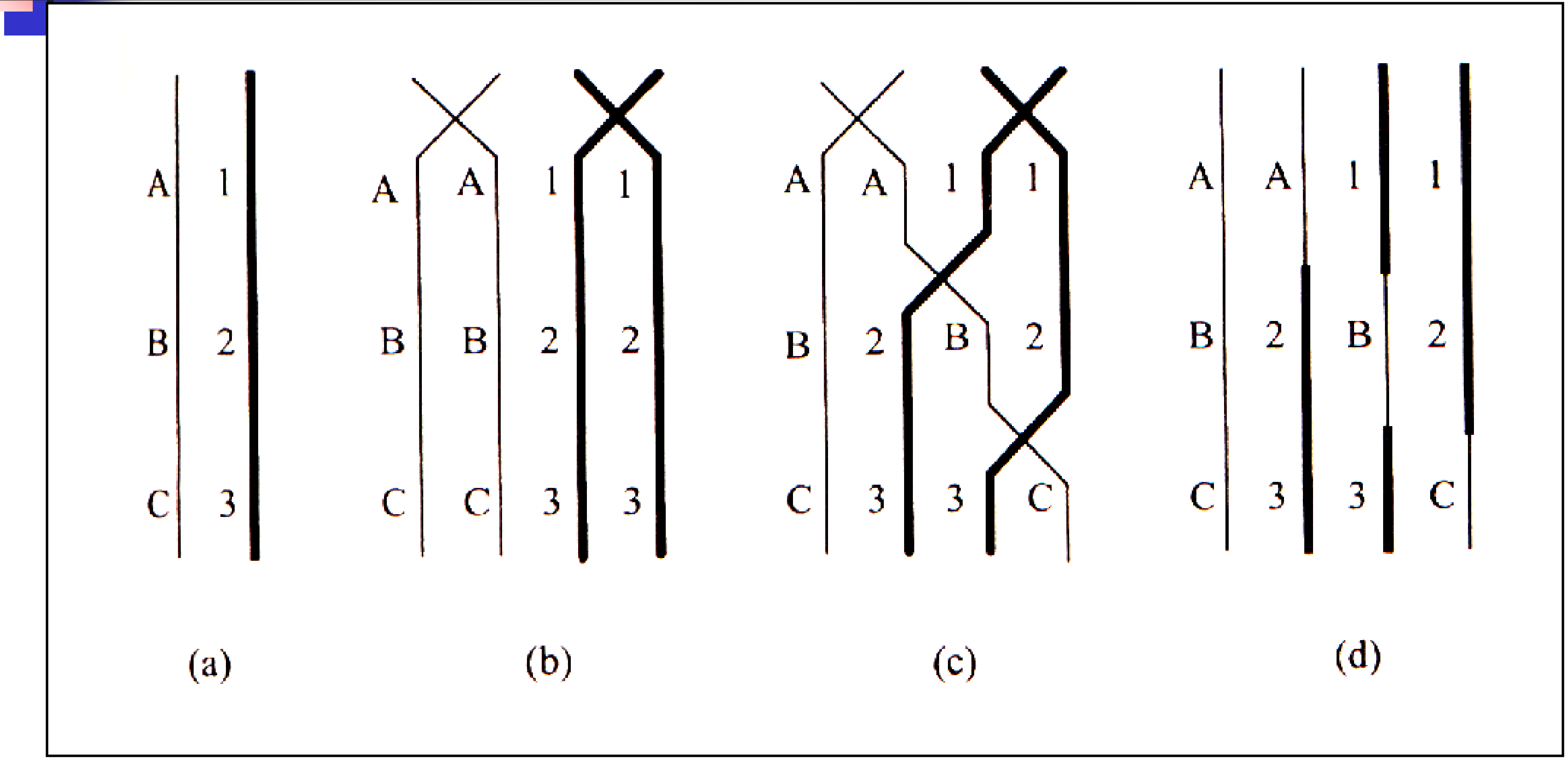
IBD 0 : 1 : 2 = 25% : 50% : 25%

Why is IBD Sharing Important?



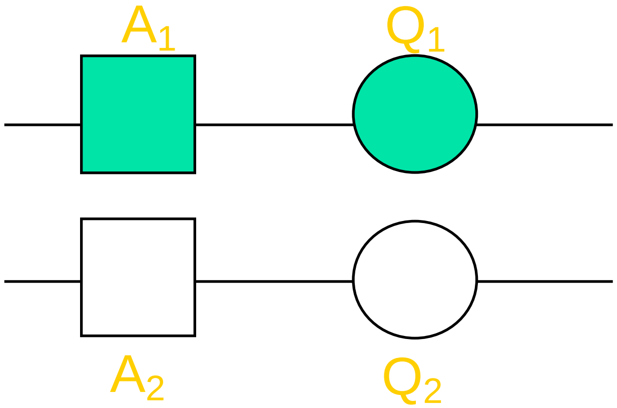
- Affected relatives not only share disease alleles IBD, but also tend to share marker alleles close to the disease locus IBD more often than chance
- IBD sharing forms the basis of non-parametric linkage statistics

Crossing over between homologous chromosomes

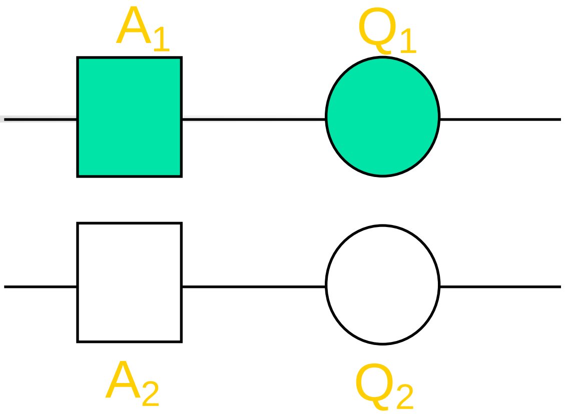


Cosegregation => Linkage

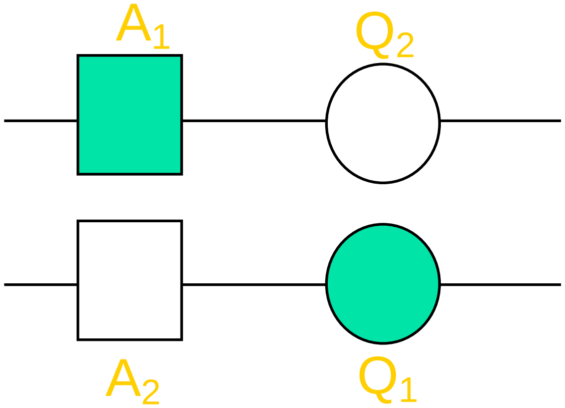
Parental genotype



Non-recombinant
Parental
genotypes
(many, $1 - \theta$)

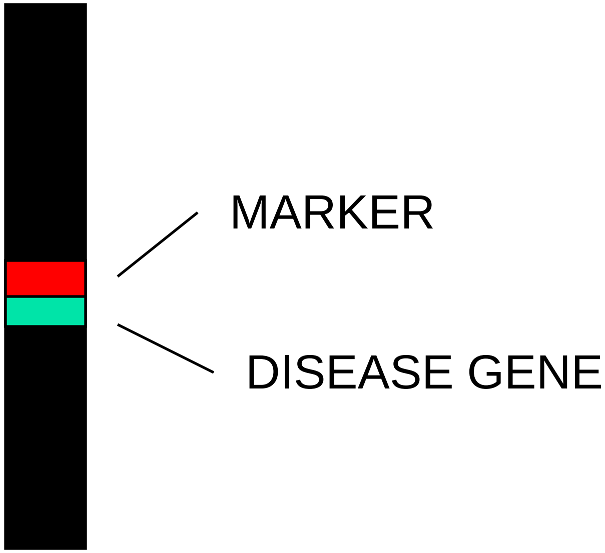


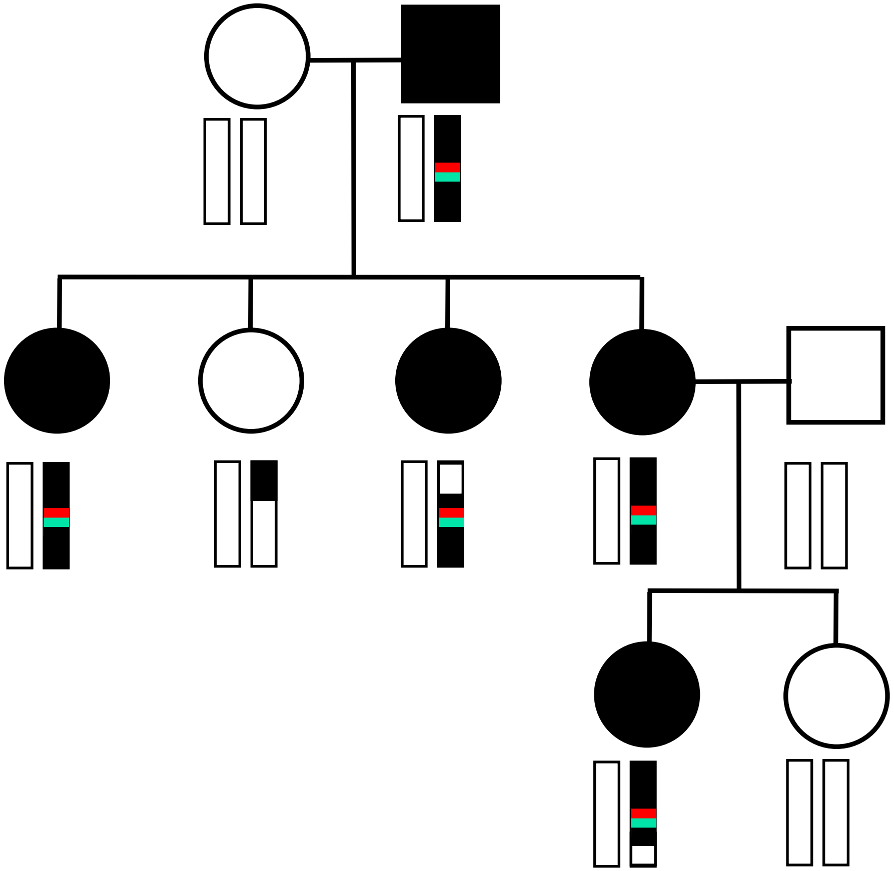
Recombinant
genotypes
(few, θ)



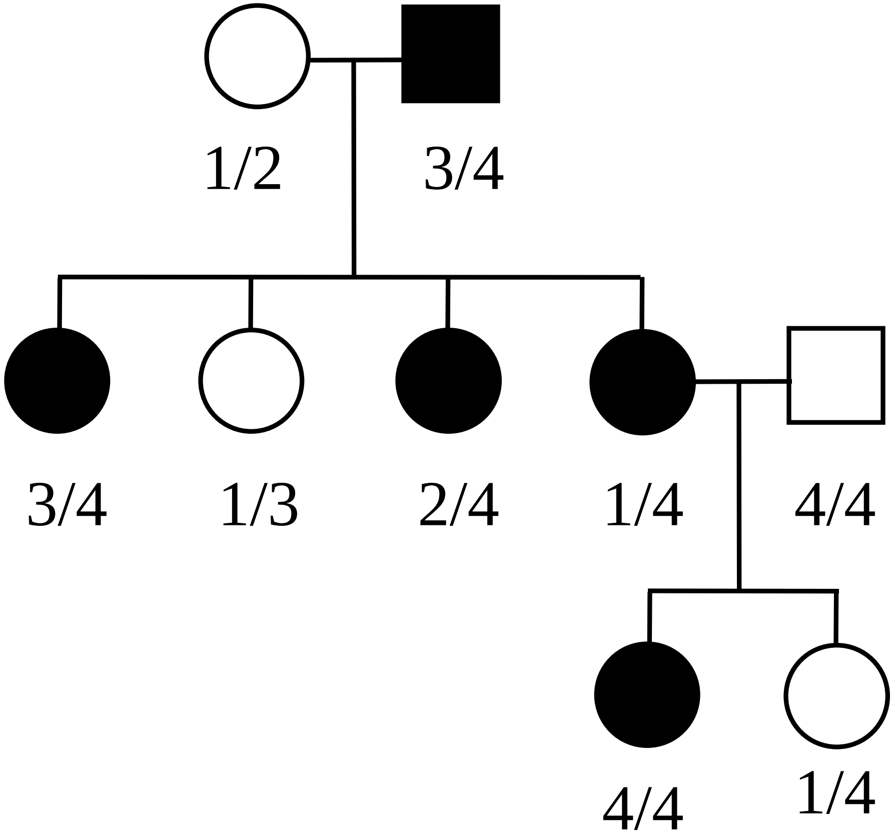
Alleles close together on the same chromosome tend to stay together in meiosis; therefore they tend be co-transmitted.

Segregating Chromosomes



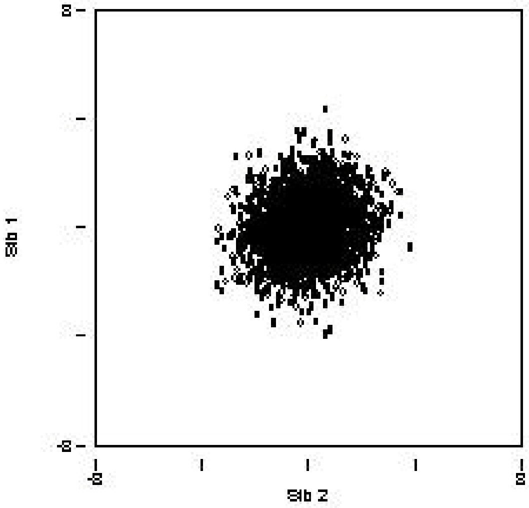


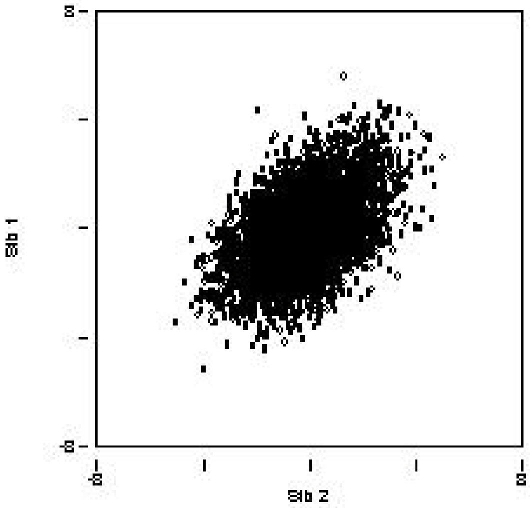
Marker Shared Among Affecteds

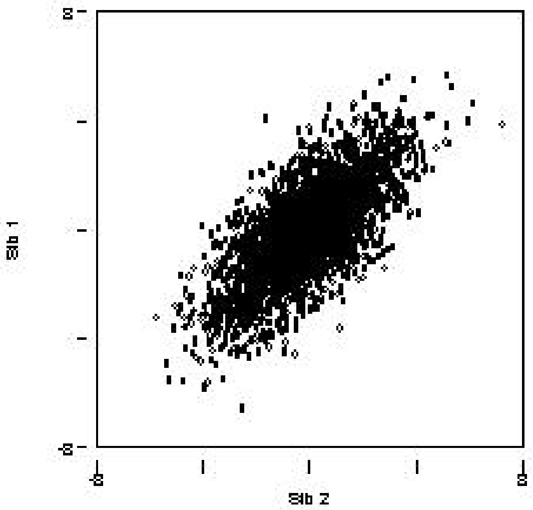


Genotypes for a marker with alleles {1,2,3,4}

Linkage between QTL and marker

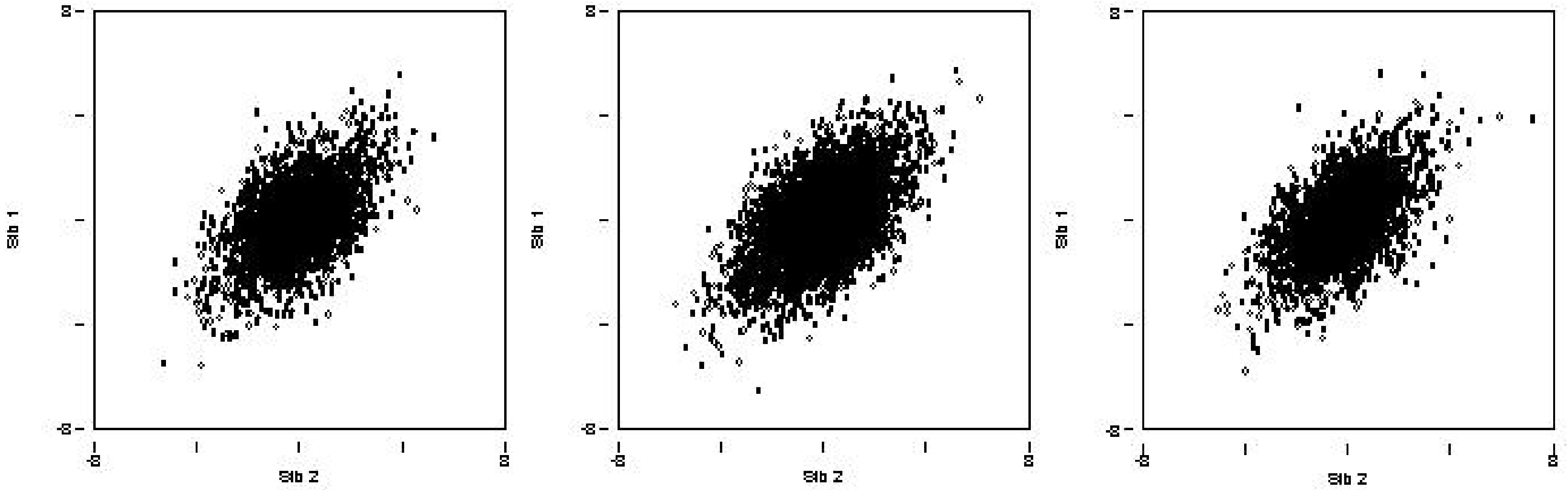






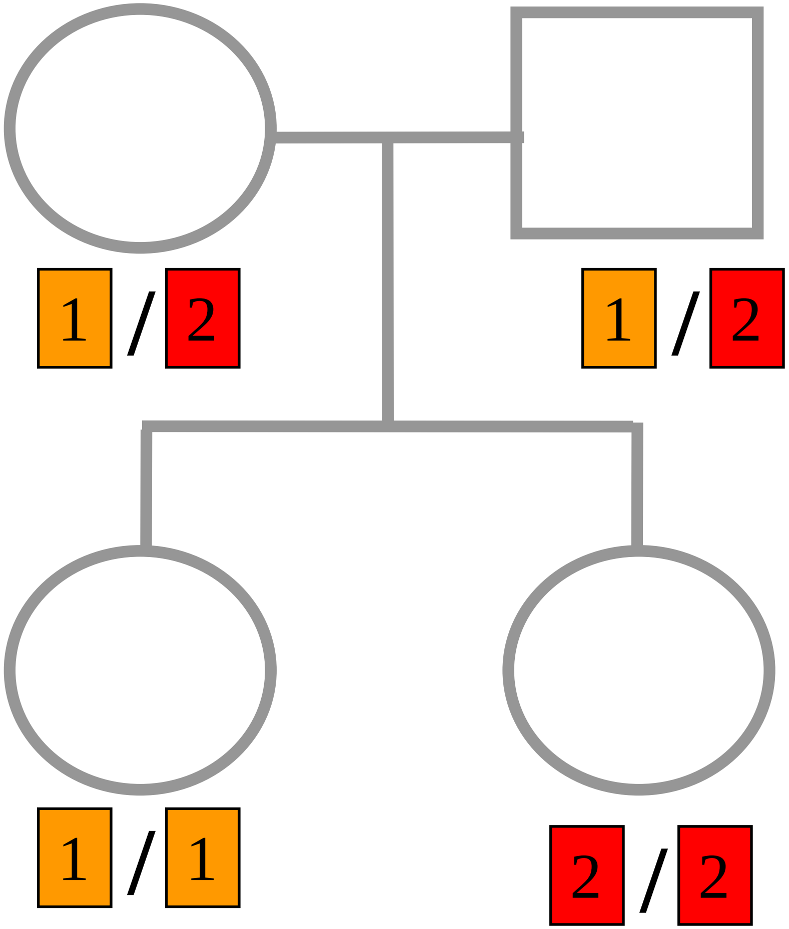
QTL	IBD 0	IBD 1	IBD 2
Marker	**IBD 0**	**IBD 1**	**IBD 2**

NO Linkage between QTL and marker



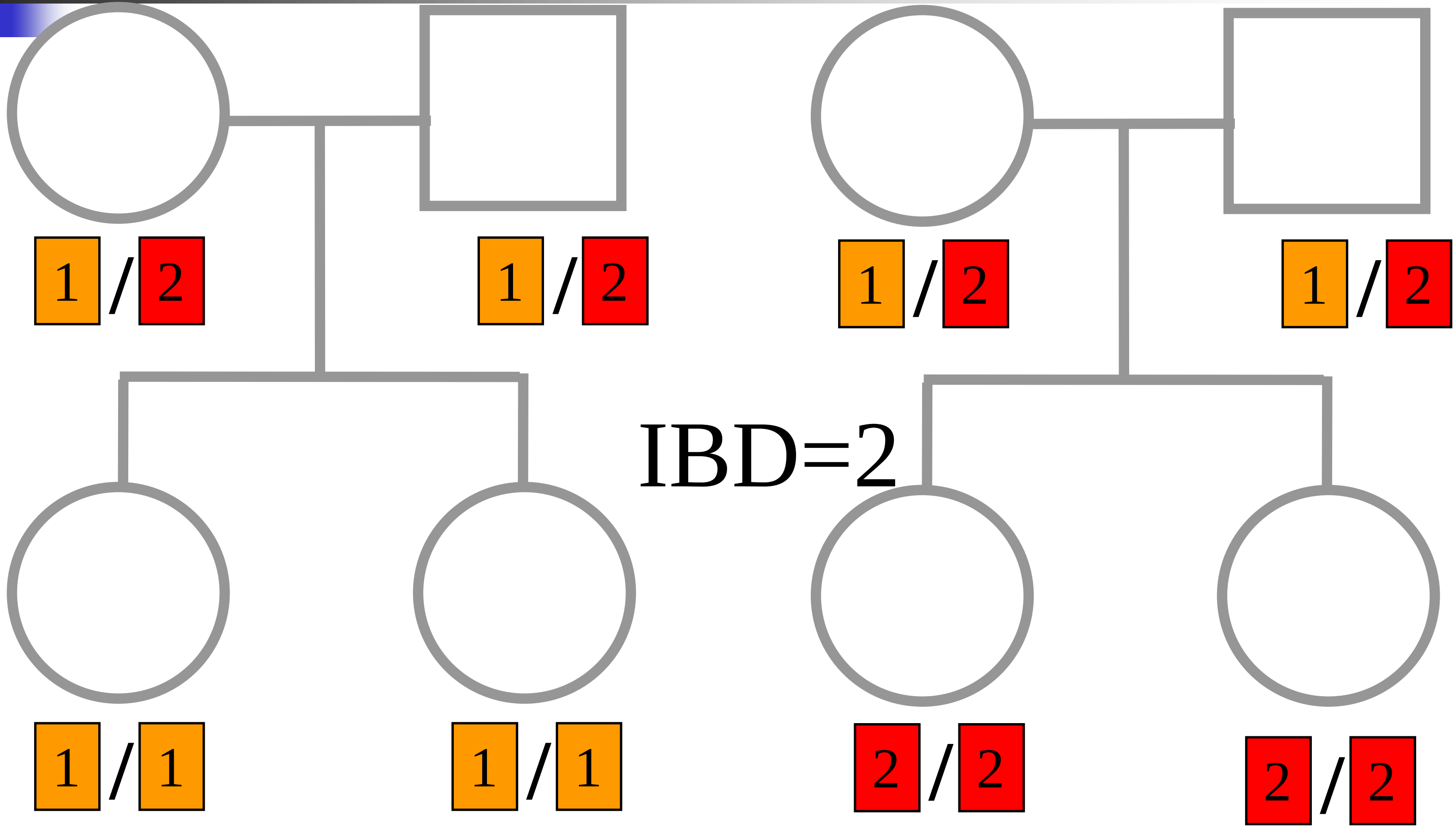
Marker IBD 0 IBD 1 IBD 2

IBD can be trivial…



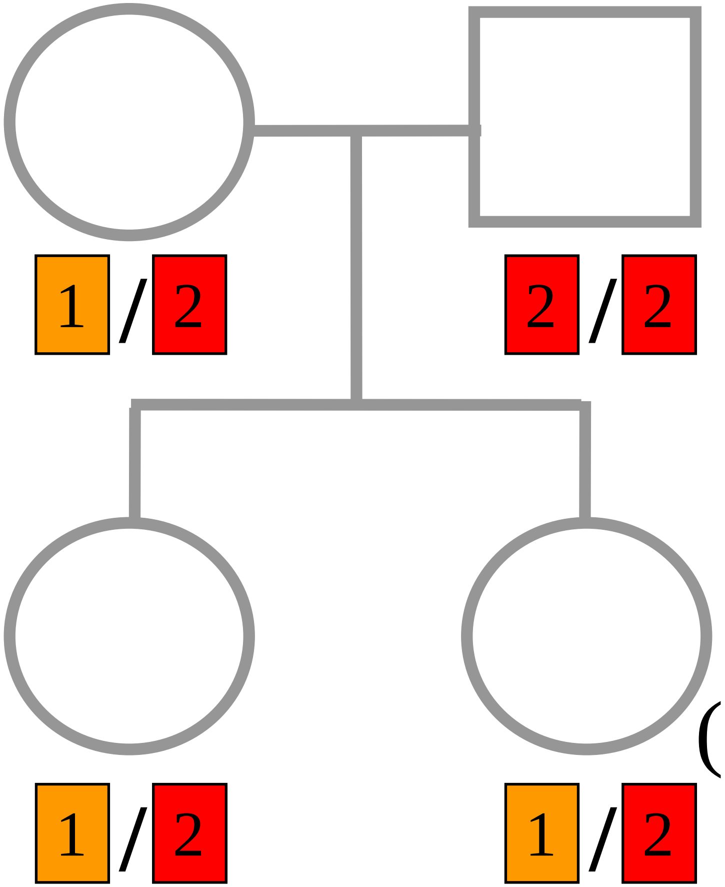
IBD=0

Two Other Simple Cases…



IBD=2

A little more complicated...



1/2

2/2

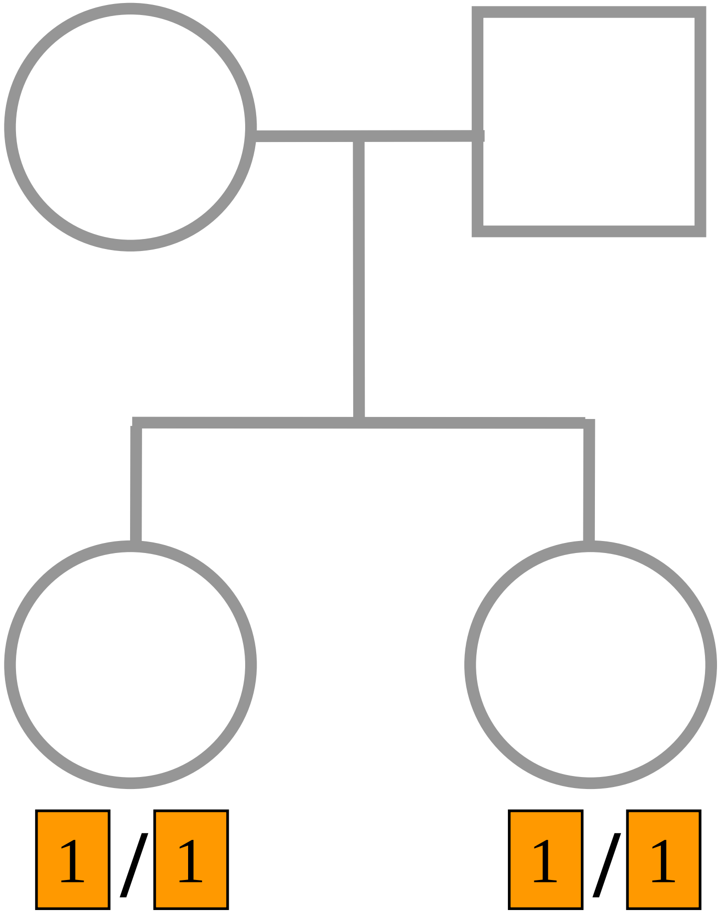
IBD=1
(50% chance)

1/2

1/2

IBD=2
(50% chance)

And even more complicated...



IBD=?

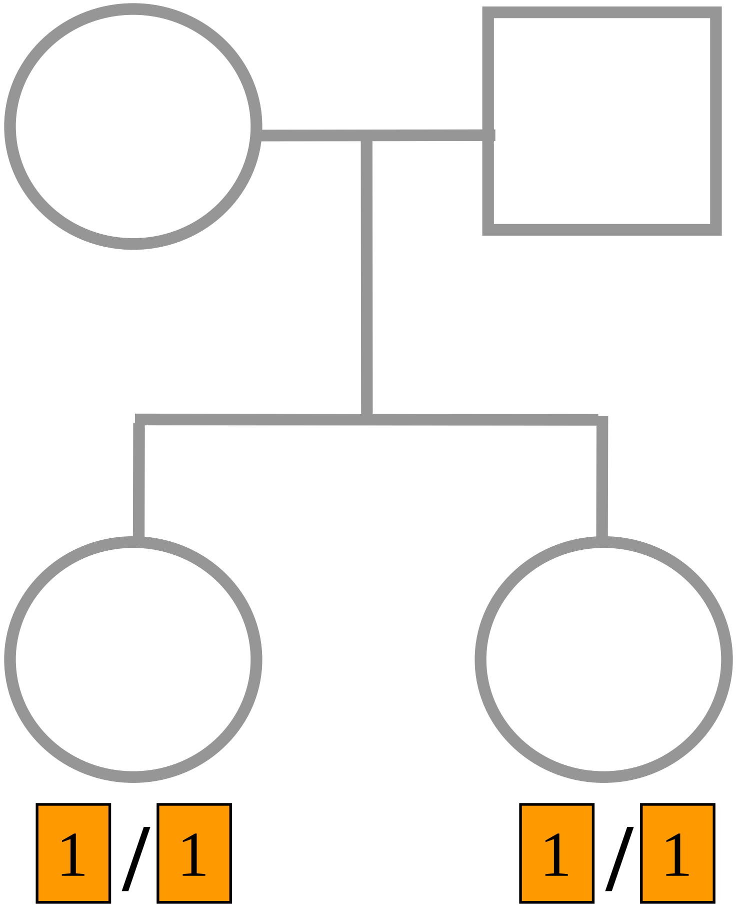
Bayes Theorem for IBD Probabilities

$$
\begin{aligned}
P(IBD = i \mid G) &= \frac{\mathrm{P}(\mathrm{IBD} = i, G)}{P(G)} \\
&= \frac{P(IBD = i)P(G \mid IBD = i)}{P(G)} \\
&= \frac{P(IBD = i)P(G \mid IBD = i)}{\sum_{j=0,1,2} P(IBD = j)P(G \mid IBD = j)}
\end{aligned}
$$

P(Genotype | IBD State)

Sib 1	Sib 2	P(observing genotypes | *k* alleles IBD)		
		k=0	*k*=1	*k*=2
A_1A_1	A_1A_1	p_1^4	p_1^3	p_1^2
A_1A_1	A_1A_2	$2p_1^3p_2$	$p_1^2p_2$	0
A_1A_1	A_2A_2	$p_1^2p_2^2$	0	0
A_1A_2	A_1A_1	$2p_1^3p_2$	$p_1^2p_2$	0
A_1A_2	A_1A_2	$4p_1^2p_2^2$	p_1p_2	$2p_1p_2$
A_1A_2	A_2A_2	$2p_1p_2^3$	$p_1p_2^2$	0
A_2A_2	A_1A_1	$p_1^2p_2^2$	0	0
A_2A_2	A_1A_2	$2p_1p_2^3$	$p_1p_2^2$	0
A_2A_2	A_2A_2	p_2^4	p_2^3	p_2^2

Worked Example



$p_1 = 0.5$

$P(G \mid IBD=0) =$

$P(G \mid IBD=1) =$

$P(G \mid IBD=2) =$

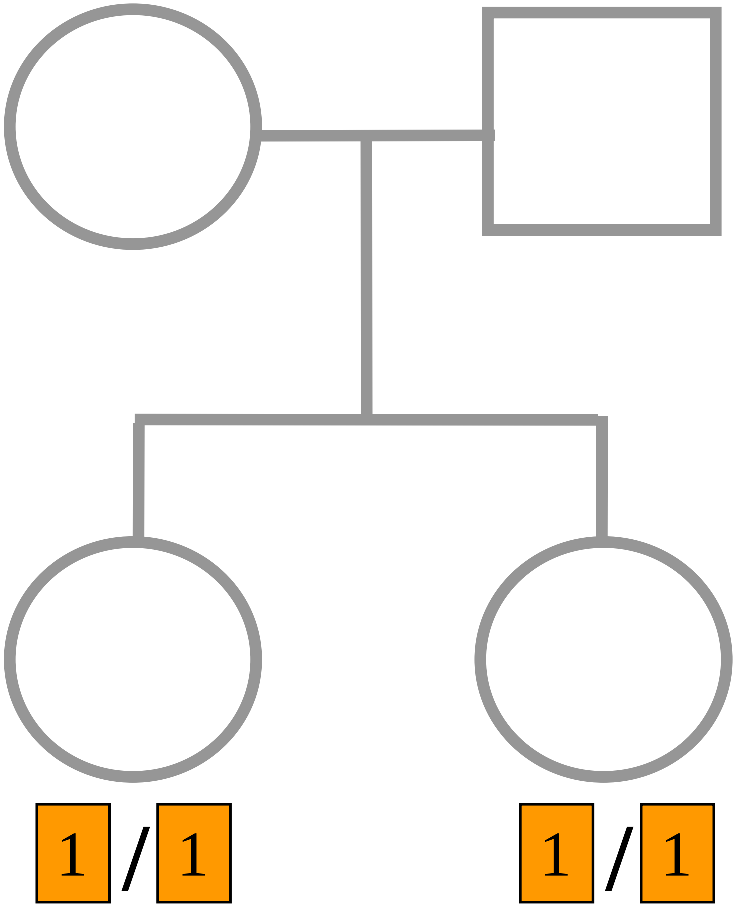
$P(G) =$

$P(IBD=0 \mid G) =$

$P(IBD=1 \mid G) =$

$P(IBD=2 \mid G) =$

Worked Example



$$p_1 = 0.5$$

$$P(G \mid IBD = 0) = p_1^4 = \tfrac{1}{16}$$

$$P(G \mid IBD = 1) = p_1^3 = \tfrac{1}{8}$$

$$P(G \mid IBD = 2) = p_1^2 = \tfrac{1}{4}$$

$$P(G) = \tfrac{1}{4} p_1^4 + \tfrac{1}{2} p_1^3 + \tfrac{1}{4} p_1^2 = \tfrac{9}{64}$$

$$P(IBD = 0 \mid G) = \frac{\tfrac{1}{4} p_1^4}{P(G)} = \tfrac{1}{9}$$

$$P(IBD = 1 \mid G) = \frac{\tfrac{1}{2} p_1^3}{P(G)} = \tfrac{4}{9}$$

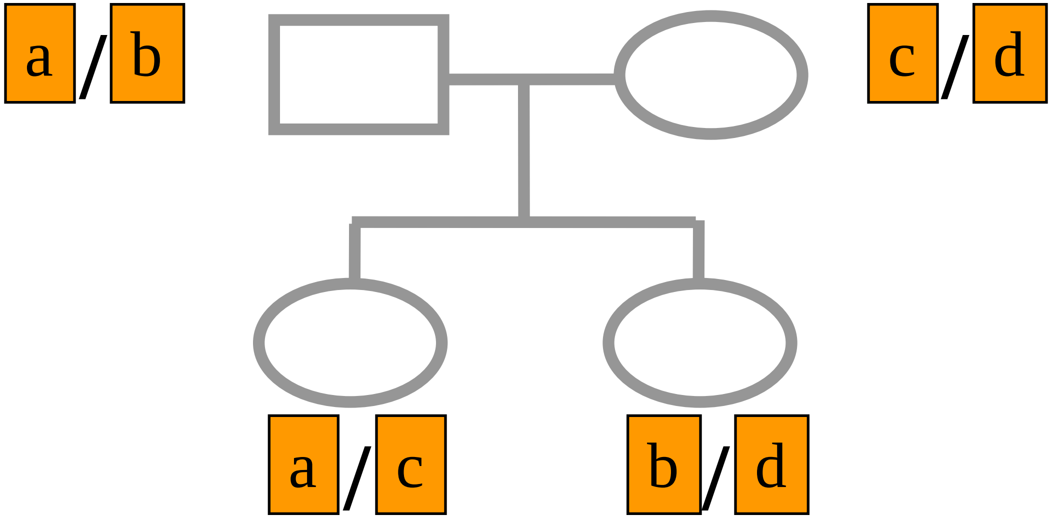
$$P(IBD = 2 \mid G) = \frac{\tfrac{1}{4} p_1^2}{P(G)} = \tfrac{4}{9}$$

For ANY PEDIGREE the inheritance pattern at any point in the genome can be completely described by a binary inheritance vector of length *2n*:

$$v(x) = (p_1, m_1, p_2, m_2, \ldots, p_n, m_n)$$

whose coordinates describe the outcome of the paternal and maternal meioses giving rise to the *n* non-founders in the pedigree

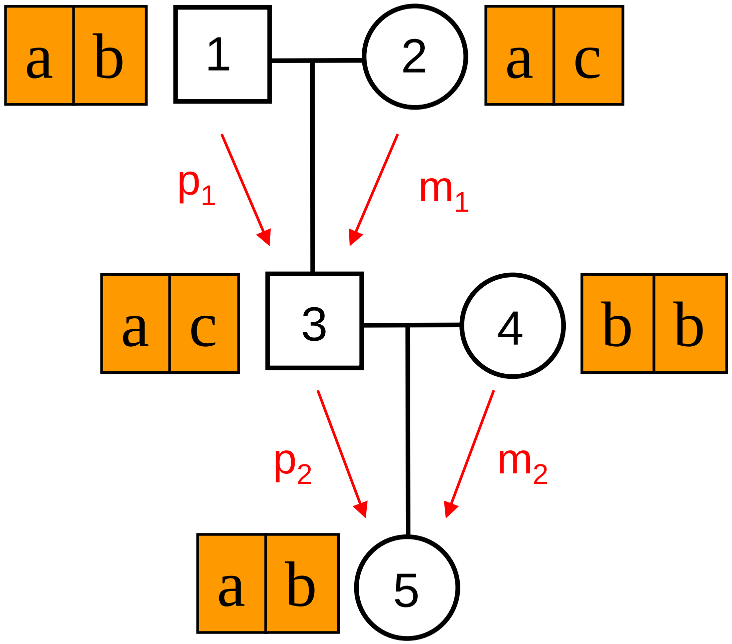
p_i (m_i) is 0 if the grandpaternal allele transmitted
p_i (m_i) is 1 if the grandmaternal allele is transmitted



v(x) = [0,0,1,1]

Inheritance Vector

In practice, it is not possible to determine the true inheritance vector at every point in the genome, rather we represent partial information as a probability distribution of the possible inheritance vectors

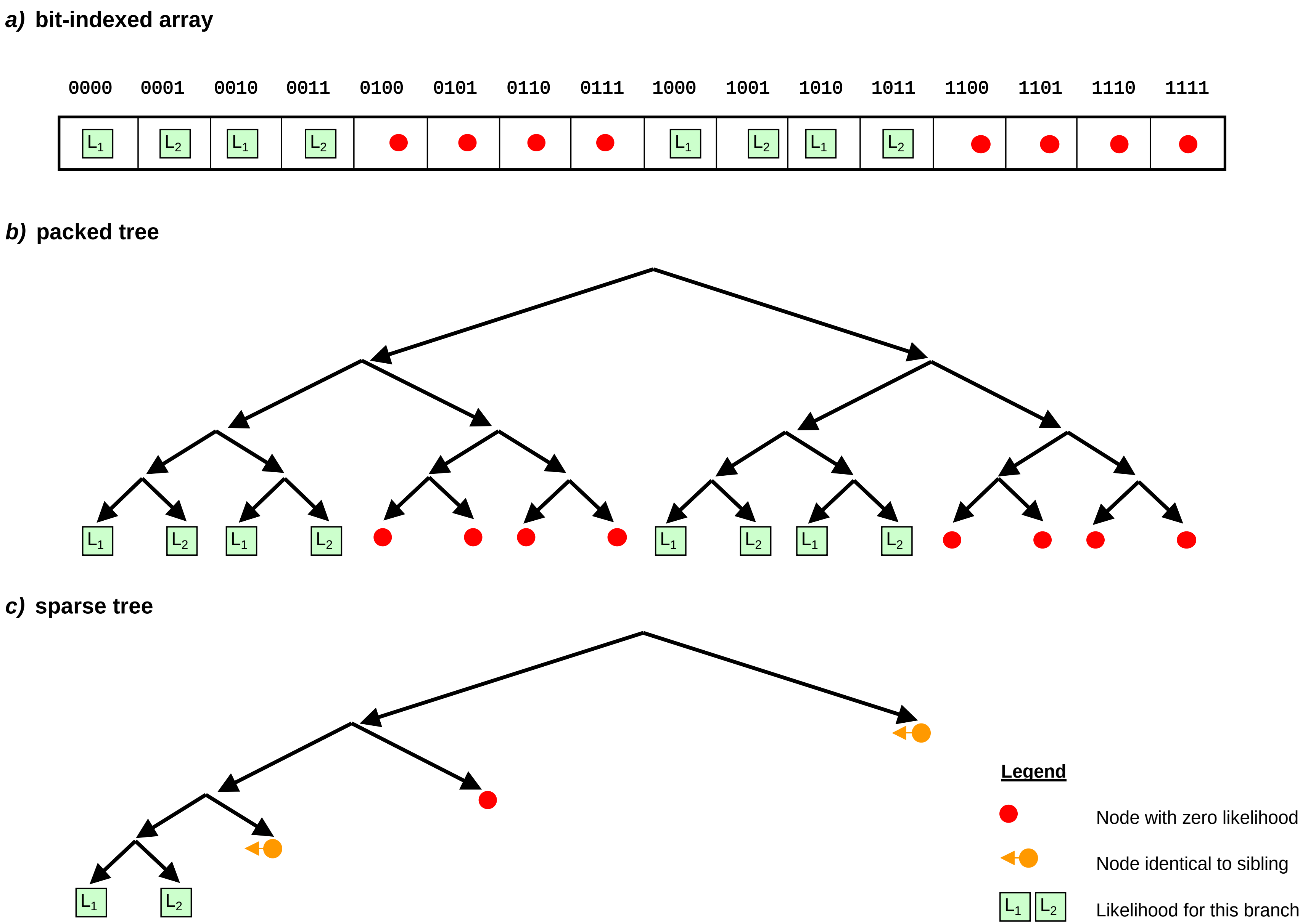


Inheritance vector	Prior	Posterior
0000	1/16	1/8
0001	1/16	1/8
0010	1/16	0
0011	1/16	0
0100	1/16	1/8
0101	1/16	1/8
0110	1/16	0
0111	1/16	0
1000	1/16	1/8
1001	1/16	1/8
1010	1/16	0
1011	1/16	0
1100	1/16	1/8
1101	1/16	1/8
1110	1/16	0
1111	1/16	0

Computer Representation

- At each marker location ℓ
- Define inheritance vector $\mathbf{v}_\ell$
    - Meiotic outcomes specified in index bit
    - Likelihood for each gene flow pattern
        - Conditional on observed genotypes at location ℓ
    - 2^{2n} elements !!!

0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L



a) bit-indexed array

b) packed tree

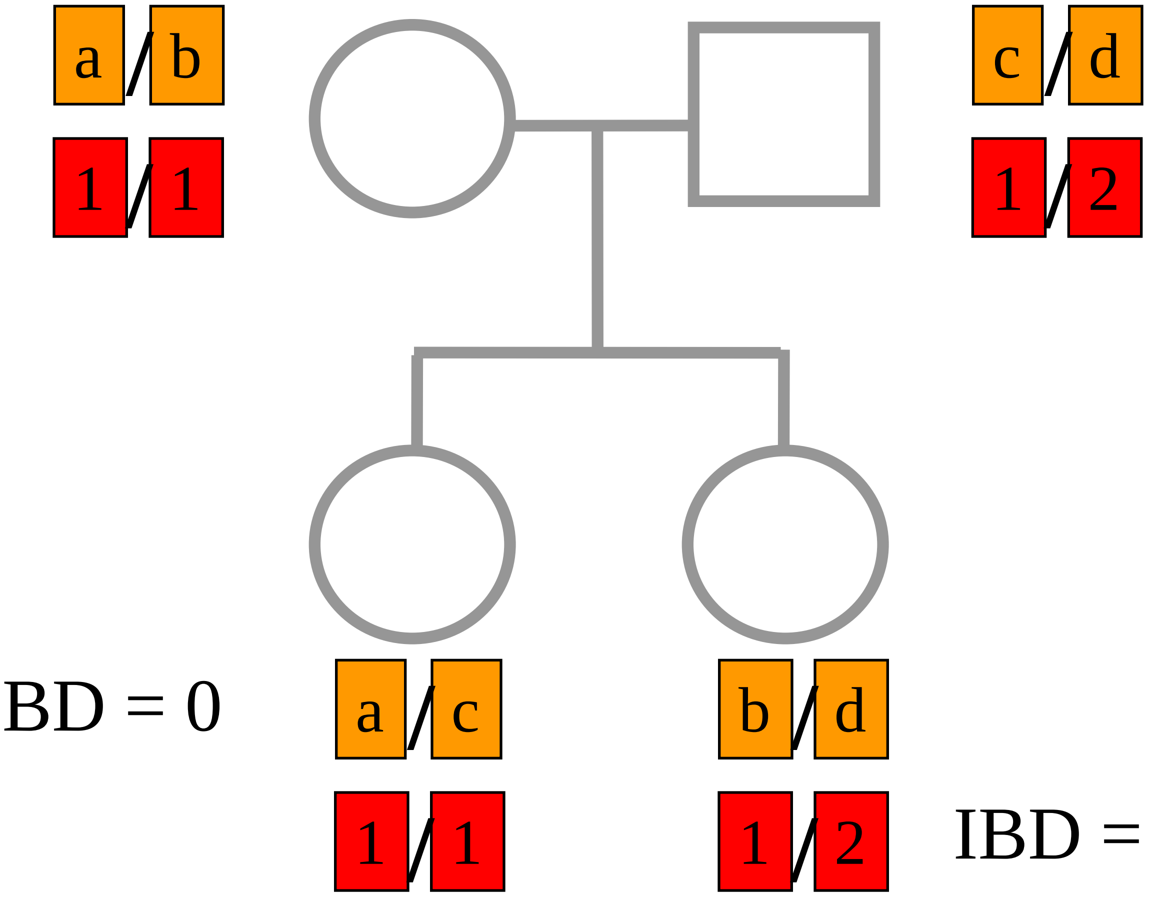
c) sparse tree

Abecasis et al (2002) *Nat Genet* **30:**97-101

Multipoint IBD

- IBD status may not be able to be ascertained with certainty because e.g. the mating is not informative, parental information is not available
- IBD information at uninformative loci can be made more precise by examining nearby linked loci

Multipoint IBD



IBD = 0

IBD = 0 or IBD =1?

Complexity of the Problem in Larger Pedigrees

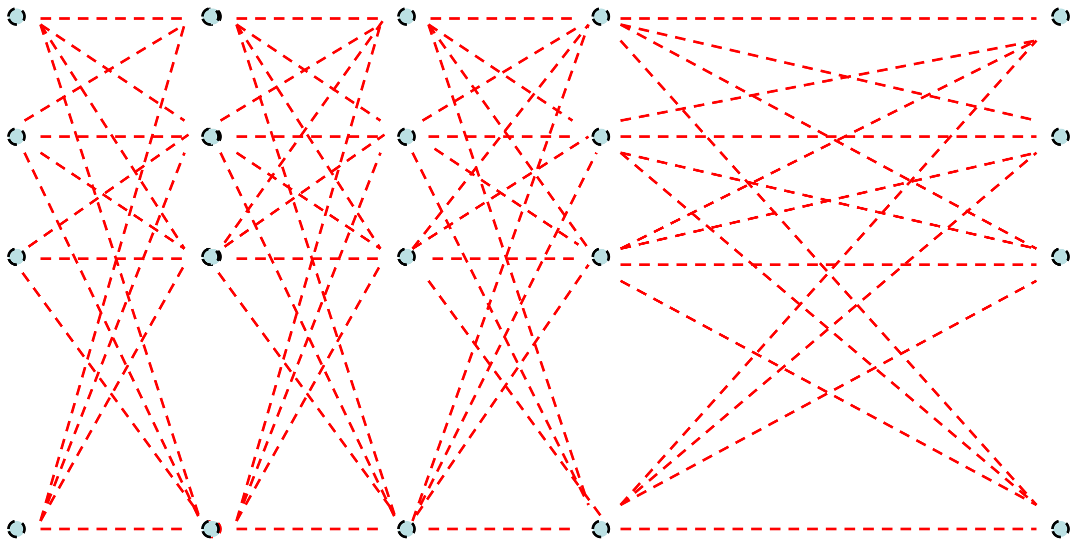
- For each person
  - 2*n* meioses in pedigree with *n* non-founders
  - Each meiosis has 2 possible outcomes
  - Therefore 2^{2n} possibilities for each locus
- For each genetic locus
  - One location for each of *m* genetic markers
  - Distinct, non-independent meiotic outcomes
- Up to 4^{nm} distinct outcomes!!!

Example: Sib-pair Genotyped at 10 Markers

Inheritance vector

P(G | 0000)

$(1 - \theta)^4$



0000

0001

0010

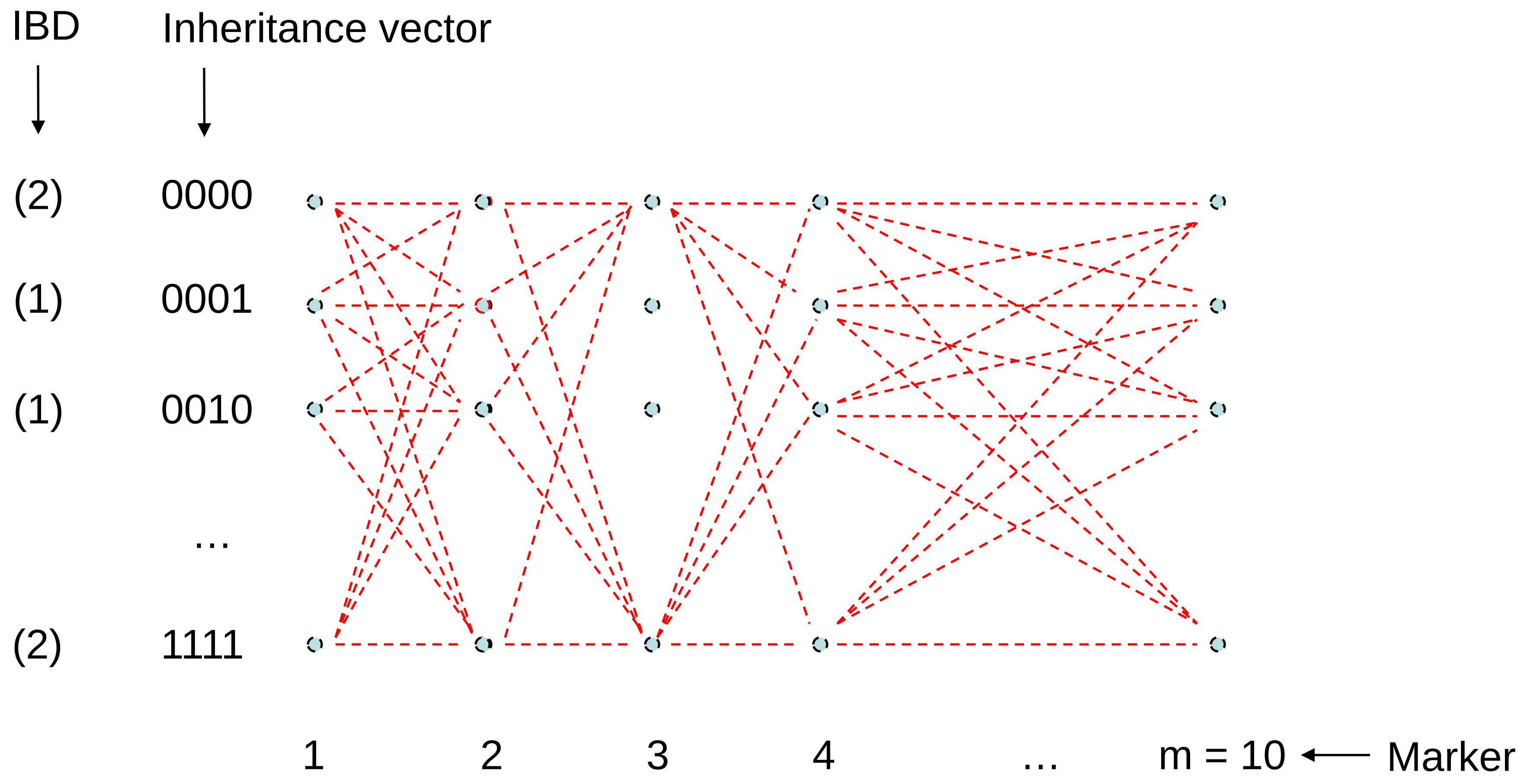
…

1111

1 2 3 4 … m = 10 ← Marker

$(2^{2xn})^m = (2^{2 \text{ x } 2})^{10}$ =~ 10^{12} possible paths !!!

P(IBD) = 2 at Marker Three



IBD

Inheritance vector

(2) 0000

(1) 0001

(1) 0010

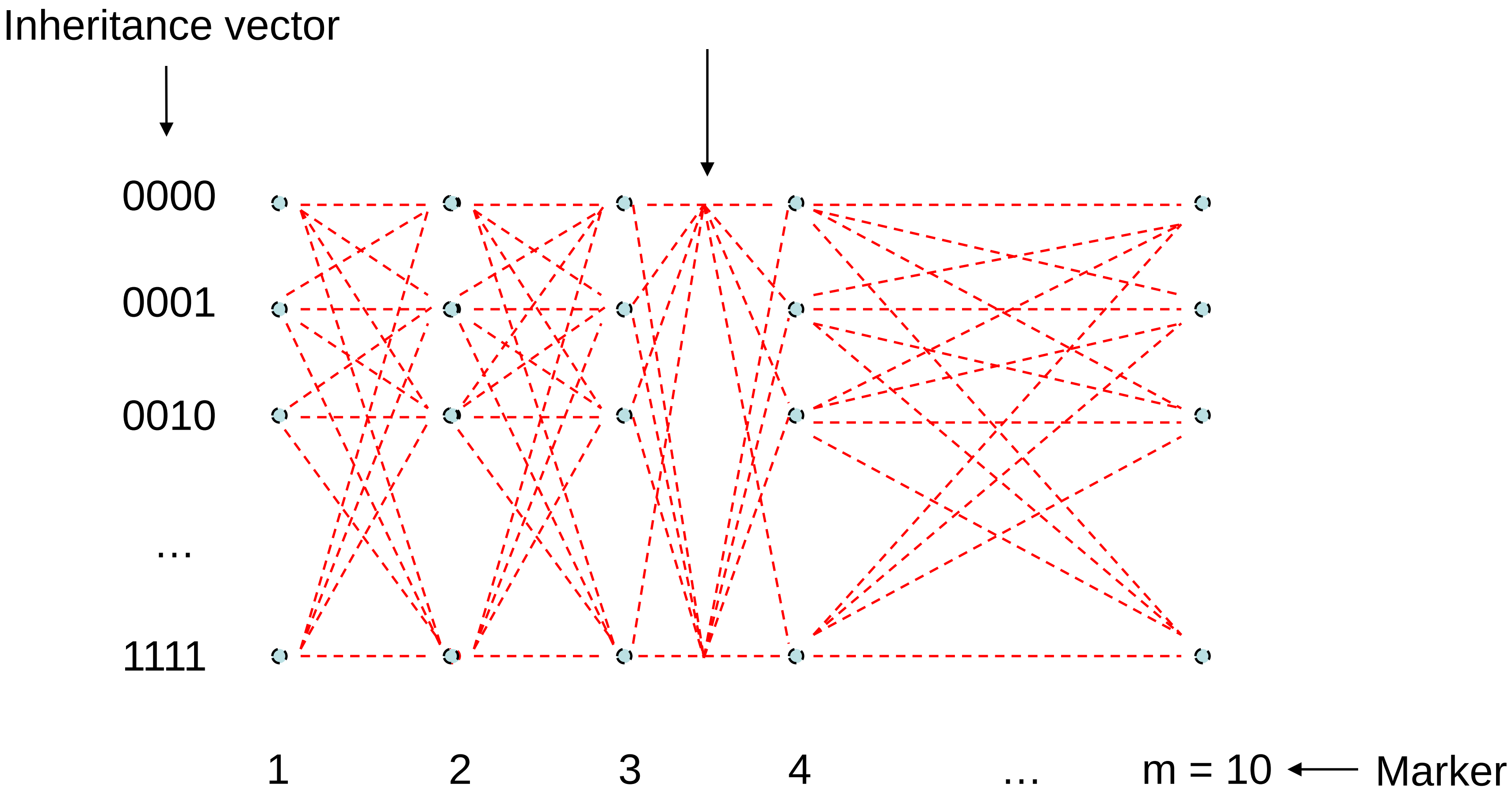
…

(2) 1111

1 2 3 4 … m = 10 ← Marker

(L[0000] + L[0101] + L[1010] + L[1111]) / L[ALL]

P(IBD) = 2 at arbitrary position on the chromosome

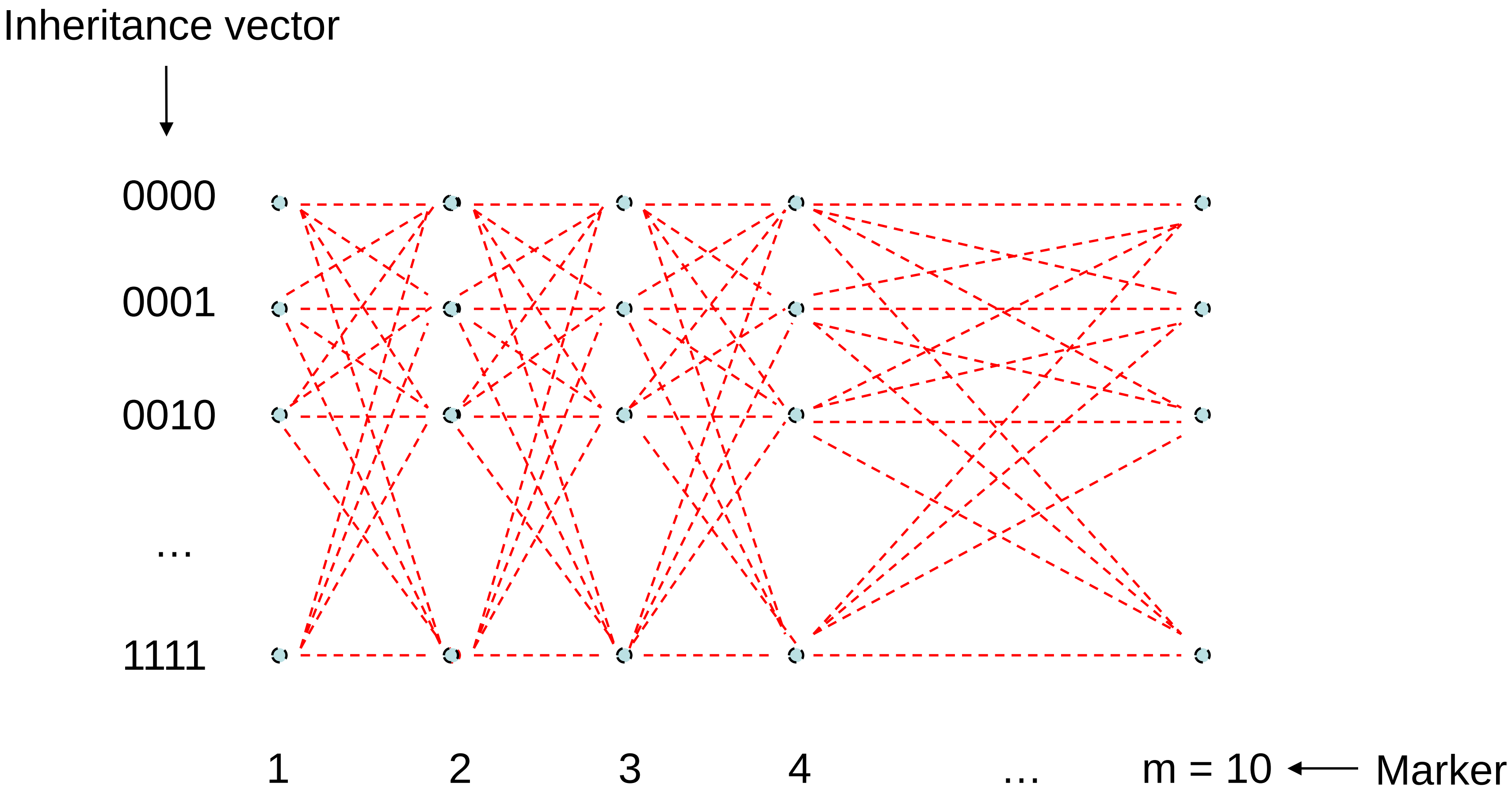


(L[0000] + L[0101] + L[1010] + L[1111]) / L[ALL]

Lander-Green Algorithm

- The inheritance vector at a locus is conditionally independent of the inheritance vectors at all preceding loci given the inheritance vector at the immediately preceding locus (“Hidden Markov chain”)

Lander-Green Algorithm



Inheritance vector

0000

0001

0010

…

1111

1 2 3 4 … m = 10 ← Marker

$M(2^{2n})^2 = 10 \times 16^2 = 2560$ calculations

Lander-Green Algorithm Summary

- Factorize likelihood by marker
  - Complexity $\propto m \cdot e^n$
- Strengths
  - Large number of markers
  - Relatively small pedigrees

Elston-Stewart Algorithm

- Factorize likelihood by individual
  - Complexity $\propto n \cdot e^m$
- Small number of markers
- Large pedigrees
  - With little inbreeding
- VITESSE, FASTLINK etc

Other methods

- Number of MCMC methods proposed
  - ~Linear on # markers
  - ~Linear on # people
- Hard to guarantee convergence on very large datasets
  - Many widely separated local minima
- E.g. SIMWALK

MERLIN-- Multipoint Engine for Rapid Likelihood Inference

Merlin—rapid analysis of dense genetic maps using sparse gene flow trees

Gonçalo R. Abecasis[1,2], Stacey S. Cherny[1], William O. Cookson[1] & Lon R. Cardon[1]

Published online: 3 December 2001, DOI: 10.1038/ng786

Efforts to find disease genes using high-density single-nucleotide polymorphism (SNP) maps will produce data sets that exceed the limitations of current computational tools. Here we describe a new, efficient method for the analysis of dense genetic maps in pedigree data that provides extremely fast solutions to common problems such as allele-sharing analyses and haplotyping. We show that sparse binary trees represent patterns of gene flow in general pedigrees in a parsimonious manner, and derive a family of related algorithms for pedigree traversal. With these trees, exact likelihood calculations can be carried out efficiently for single markers or for multiple linked markers. Using an approximate multipoint calculation that ignores the unlikely possibility of a large number of recombinants further improves speed and provides accurate solutions in dense maps with thousands of markers. Our multipoint engine for rapid likelihood inference (Merlin) is a computer program that uses sparse inheritance trees for pedigree analysis; it performs rapid haplotyping, genotype error detection and affected pair linkage analyses and can handle more markers than other pedigree analysis packages.

Linkage and association studies routinely involve analyzing many markers in related individuals to determine phased haplotypes, test for cosegregation of disease and marker loci or identify problems in genotyping. The shift to dense SNP maps[1,2] poses new problems to pedigree analysis packages[3–9]. Packages based on the Elston-Stewart algorithm[10] can only handle a small number of markers and are not well suited to SNP maps. On the other hand, memory requirements for the Lander-Green algorithm[11] make analyzing hundreds or thousands of markers a severe challenge in all but the smallest pedigrees. Although Markov-Chain Monte-Carlo (MCMC) sampling methods[7,8,12] complement some of the deficiencies in these two approaches, as the number of tightly linked markers increases, it is difficult to guarantee their adequate convergence. Another unresolved issue is undetected genotyping error, which seriously hinders linkage and association studies[13,14]. As most SNP genotyping errors do not lead to mendelian inconsistencies[15], SNPs require specialized quality-control strategies.

The Lander-Green algorithm[11] considers each alternative gene flow pattern in a pedigree separately. Allele-sharing statistics for each set of observed phenotypes and likelihoods conditional on observed marker data are calculated and stored in memory[4]. Because the pattern of gene flow through a pedigree is fully specified by noting whether the grand-maternal or grand-paternal allele is transmitted in each meiosis, the results of these calculations are typically stored in a bit-indexed array (Fig. 1*a*), where each index bit indicates the outcome of one meiosis[4,5,16]. Binary trees provide another natural organization for results that depend on gene flow patterns. Each level in the tree represents one meiosis, and each branch corresponds to transmission of the grand-maternal or grand-paternal allele (Fig. 1*b*). Often, many alternative patterns of gene flow have the same outcome, and we reasoned that sparse binary trees might provide an efficient framework for pedigree analysis and extend the scope of the Lander-Green algorithm to very large data sets. These sparse trees are a reduced representation of the full binary tree, where gene flow patterns with identical outcomes are combined into symmetric and premature leaf nodes (Fig. 1*c*).

We first evaluated the performance of gene flow trees in single marker analyses using simulated replicates of pedigree D (Fig. 2), which includes 40 meioses. Usually the maternal or paternal origin of founder alleles cannot be discerned, and only 2^{32} representative outcomes must be considered[4]. If outcomes were enumerated in an array, this analysis would exceed the storage capacity of most modern workstations. In comparison, trees describing gene flow pattern likelihoods for SNP markers with equifrequent alleles and 20% missing data have a median size of less than 900 nodes, and are even smaller for more informative markers or smaller amounts of missing data (Table 1). This saves significant amounts of both storage and computing time, and similar savings result when allele-sharing statistics are calculated for most pedigrees.

Table 1 • Complexity of inheritance tree for pedigree D[a]

Missing genotypes	Info[b]	Total nodes			Leaf nodes[c]
		Mean	Median	95% C.I.	
four-allele marker with equifrequent alleles					
–	0.72	154.7	72	64–603	5.2
5%	0.68	245.2	122	64–1,166	9.9
10%	0.64	446.3	171	65–2,429	24.1
20%	0.55	1,747.4	405	69–15,943	107.3
50%	0.28	19,880.6	2,882	154–140,215	2,574.5
two-allele marker with equifrequent alleles					
–	0.42	706.0	151	57–5,447	66.9
5%	0.39	1,299.8	225	57–8,443	159.6
10%	0.36	2,157.7	329	61–15,361	148.9
20%	0.31	8,595.9	872	64–42,592	1,293.9
50%	0.14	55,639.1	4,477	135–383,407	9,173.5

[a] Pedigree D is represented in Fig. 2. [b] Average marker informativeness[4]. [c] The average number of leaf nodes, which correspond to full likelihood evaluations. These statistics summarize 1,000 replicates.

[1] *The Wellcome Trust Center for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.* [2] *Present address: Center for Statistical Genetics, Department for Biostatistics, School of Public Health, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, USA. Correspondence should be addressed to G.R.A. (e-mail: goncalo@umich.edu).*

Capabilities

- Linkage Analysis
  - NPL and K&C LOD
  - Variance Components
- Haplotypes
  - Most likely
  - Sampling
  - All
- IBD and info content
- Error Detection
  - Most SNP typing errors are Mendelian consistent
- Recombination
  - No. of recombinants per family per interval can be controlled
- Simulation

MERLIN Website

www.sph.umich.edu/csg/abecasis/Merlin

- Reference
- FAQ
- Source
- Binaries

- Tutorial
  - Linkage
  - Haplotyping
  - Simulation
  - Error detection
  - IBD calculation

Input Files

- Pedigree File
  - Relationships
  - Genotype data
  - Phenotype data
- Data File
  - Describes contents of pedigree file
- Map File
  - Records location of genetic markers

Example Pedigree File

```
<contents of example.ped>
1   1   0  0  1   1      x   3 3   x x
1   2   0  0  2   1      x   4 4   x x
1   3   0  0  1   1      x   1 2   x x
1   4   1  2  2   1      x   4 3   x x
1   5   3  4  2   2  1.234   1 3   2 2
1   6   3  4  1   2  4.321   2 4   2 2
<end of example.ped>
```

Encodes family relationships, marker and phenotype information

Data File Field Codes

Code	**Description**
M	Marker Genotype
A	Affection Status.
T	Quantitative Trait.
C	Covariate.
Z	Zygosity.
S[n]	Skip n columns.

Example Data File

```
<contents of example.dat>
T   some_trait_of_interest
M   some_marker
M   another_marker
<end of example.dat>
```

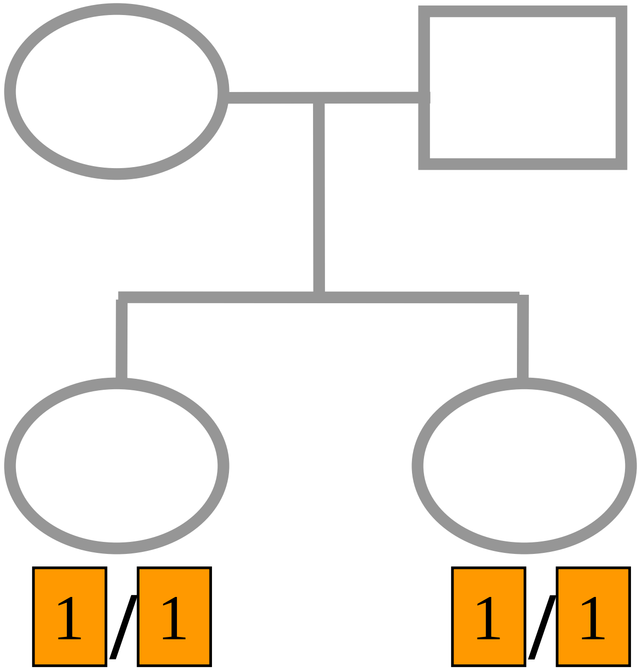
Provides information necessary to decode pedigree file

Example Map File

```
<contents of example.map>
CHROMOSOME    MARKER    POSITION
2             D2S160    160.0
2             D2S308    165.0
…
<end of example.map>
```

Indicates location of individual markers, necessary to derive recombination fractions between them

Worked Example



$p_1 = 0.5$

$P(IBD=0|G) = \frac{1}{9}$

$P(IBD=1|G) = \frac{4}{9}$

$P(IBD=2|G) = \frac{4}{9}$

```
merlin –d example.dat –p example.ped –m example.map --ibd
```

Application: Information Content Mapping

- Information content: Provides a measure of how well a marker set approaches the goal of completely determining the inheritance outcome

- Based on concept of entropy
  - $E = -\Sigma P_i \log_2 P_i$ where P_i is probability of the *i*th outcome

- $I_E(x) = 1 – E(x)/E_0$
  - Always lies between 0 and 1
  - Does not depend on test for linkage
  - Scales linearly with power

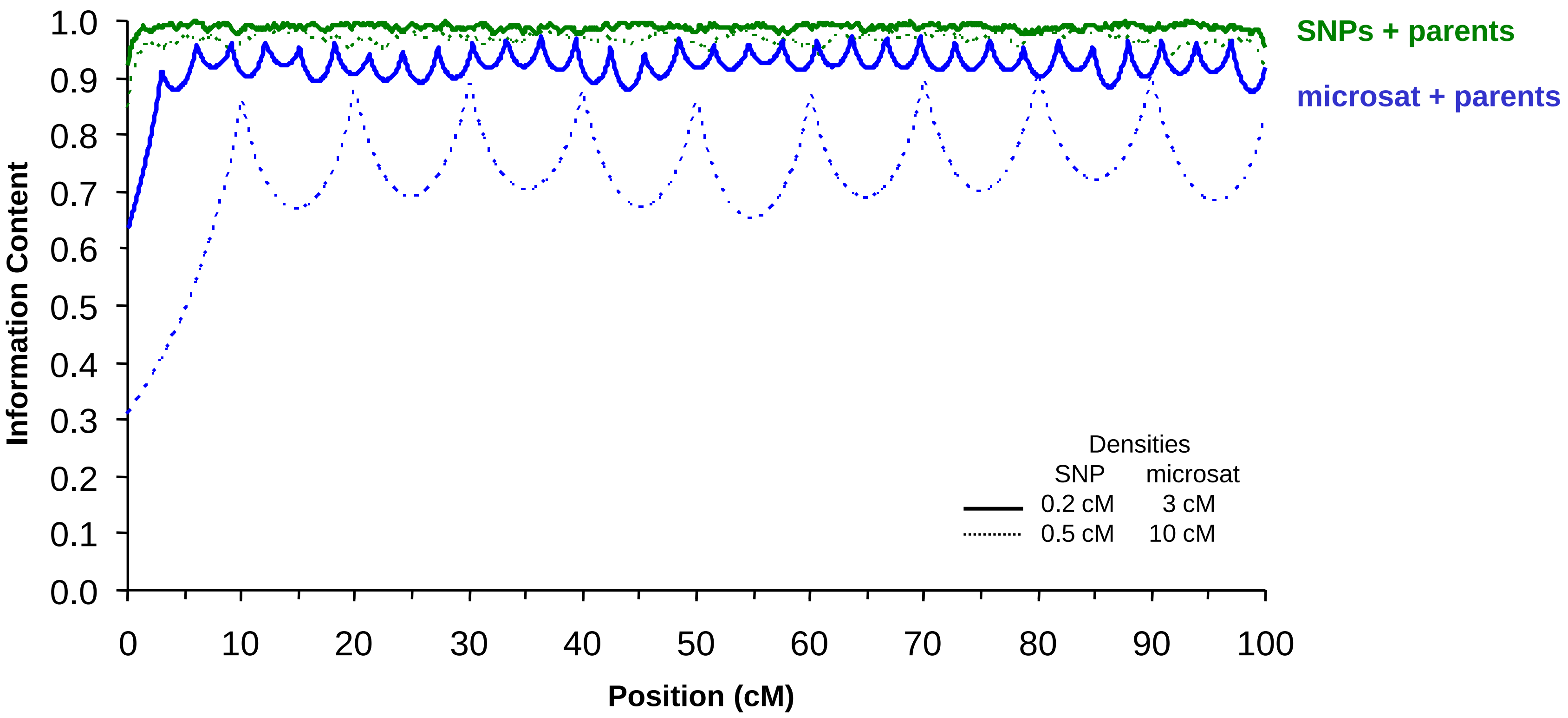
Application: Information Content Mapping

- Simulations
  - ABI (1 micro-satellite per 10cM)
  - deCODE (1 microsatellite per 3cM)
  - Illumina (1 SNP per 0.5cM)
  - Affymetrix (1 SNP per 0.2 cM)

- Which panel performs best in terms of extracting marker information?

```
merlin –d file.dat –p file.ped –m file.map --information
```

SNPs vs Microsatellites


SNPs + parents
microsat + parents
Information Content
Position (cM)
Densities
SNP
microsat
0.2 cM
3 cM
0.5 cM
10 cM