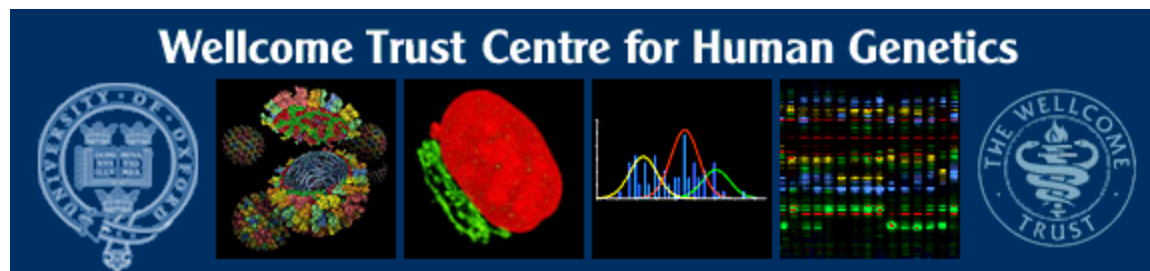# Calculation of IBD probabilities

David Evans

University of Oxford

Wellcome Trust Centre for Human Genetics

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)
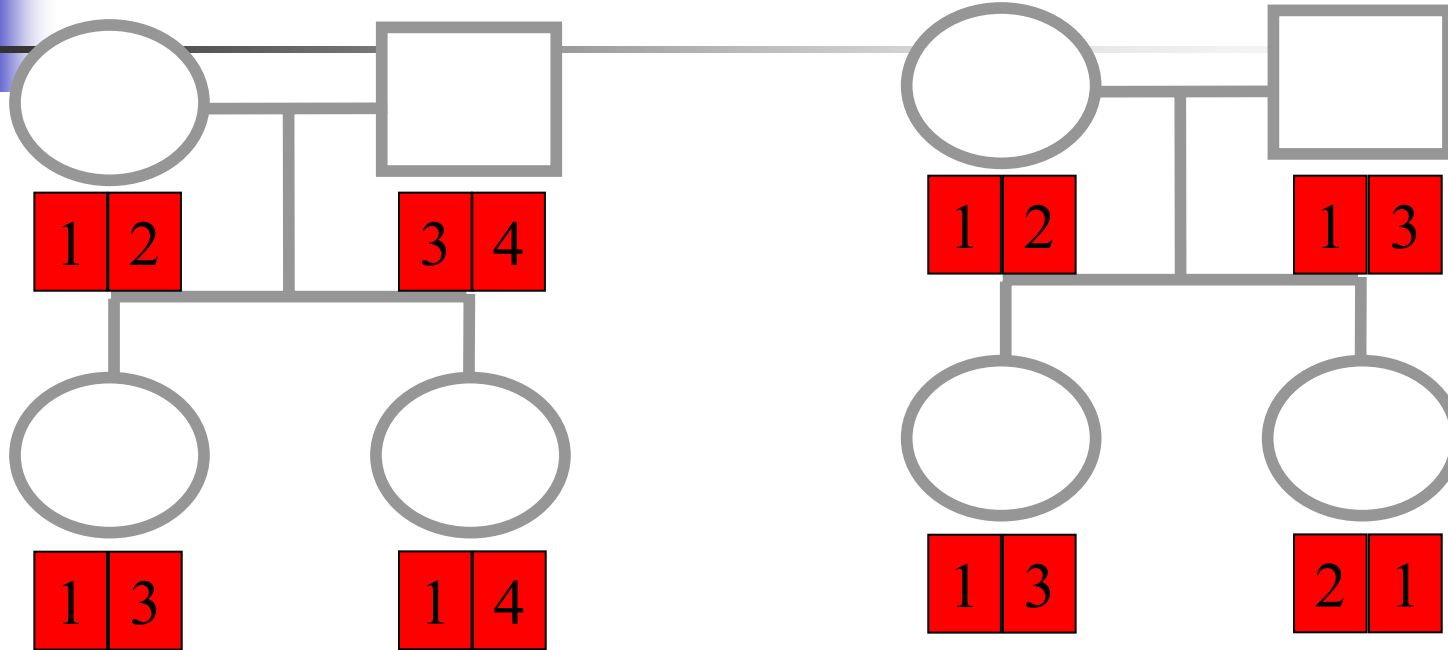
Identical by Descent   Identical by state only

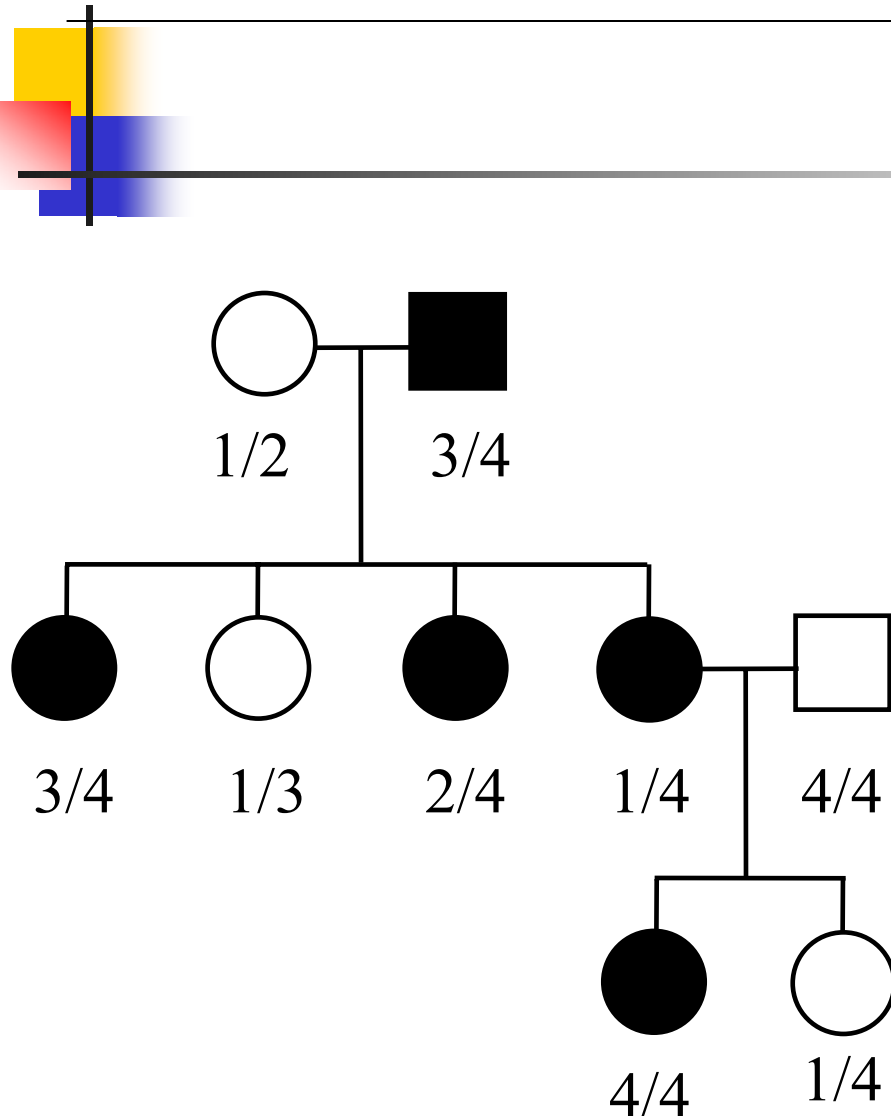Two alleles are IBD if they are descended from the same ancestral allele

Example: IBD in Siblings

Consider a mating between mother AB x father CD:

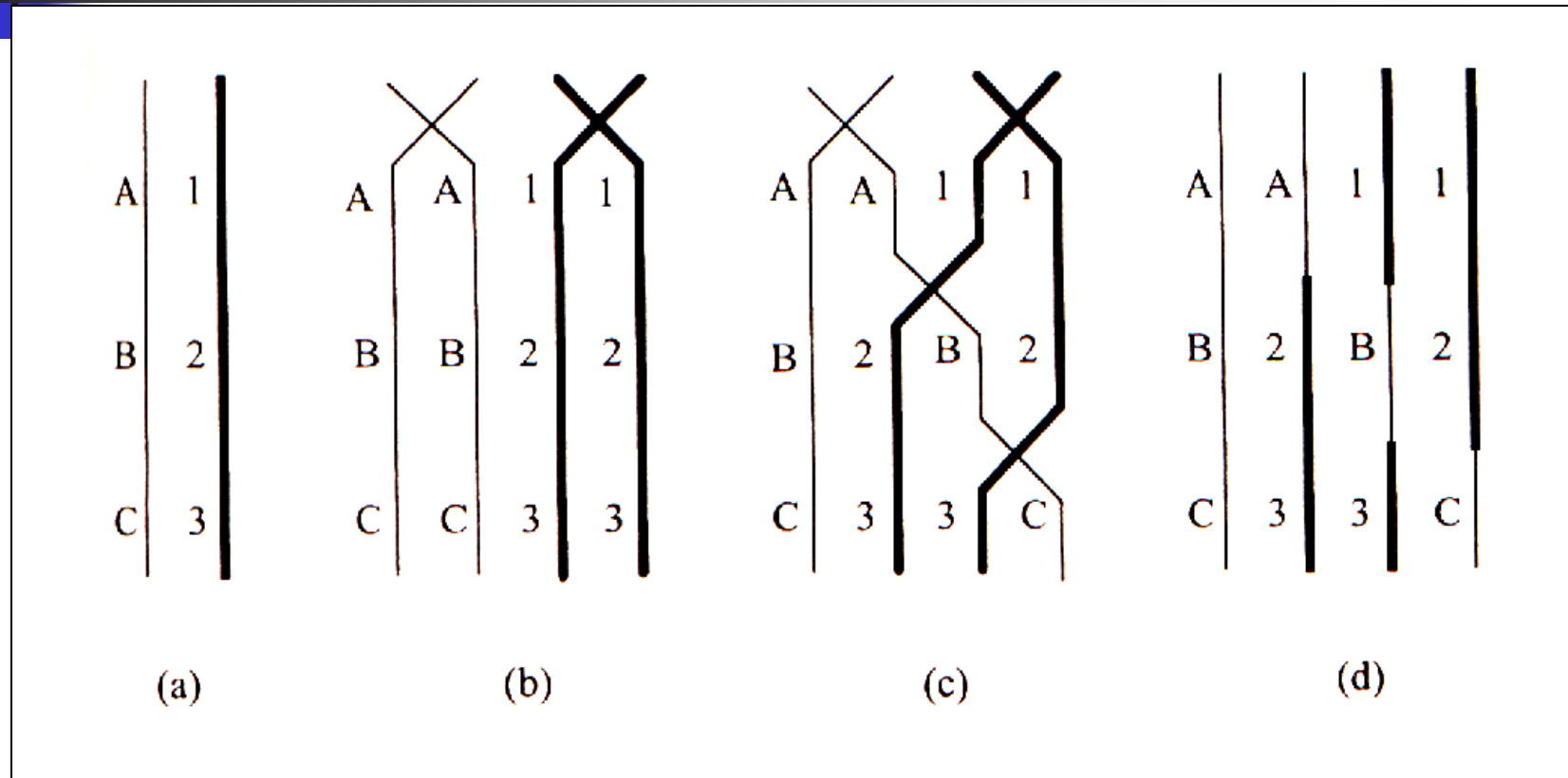|		Sib1			
		AC	AD	BC	BD
Sib 2	AC	2	1	1	0
	AD	1	2	0	1
	BC	1	0	2	1
	BD	0	1	1	2

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

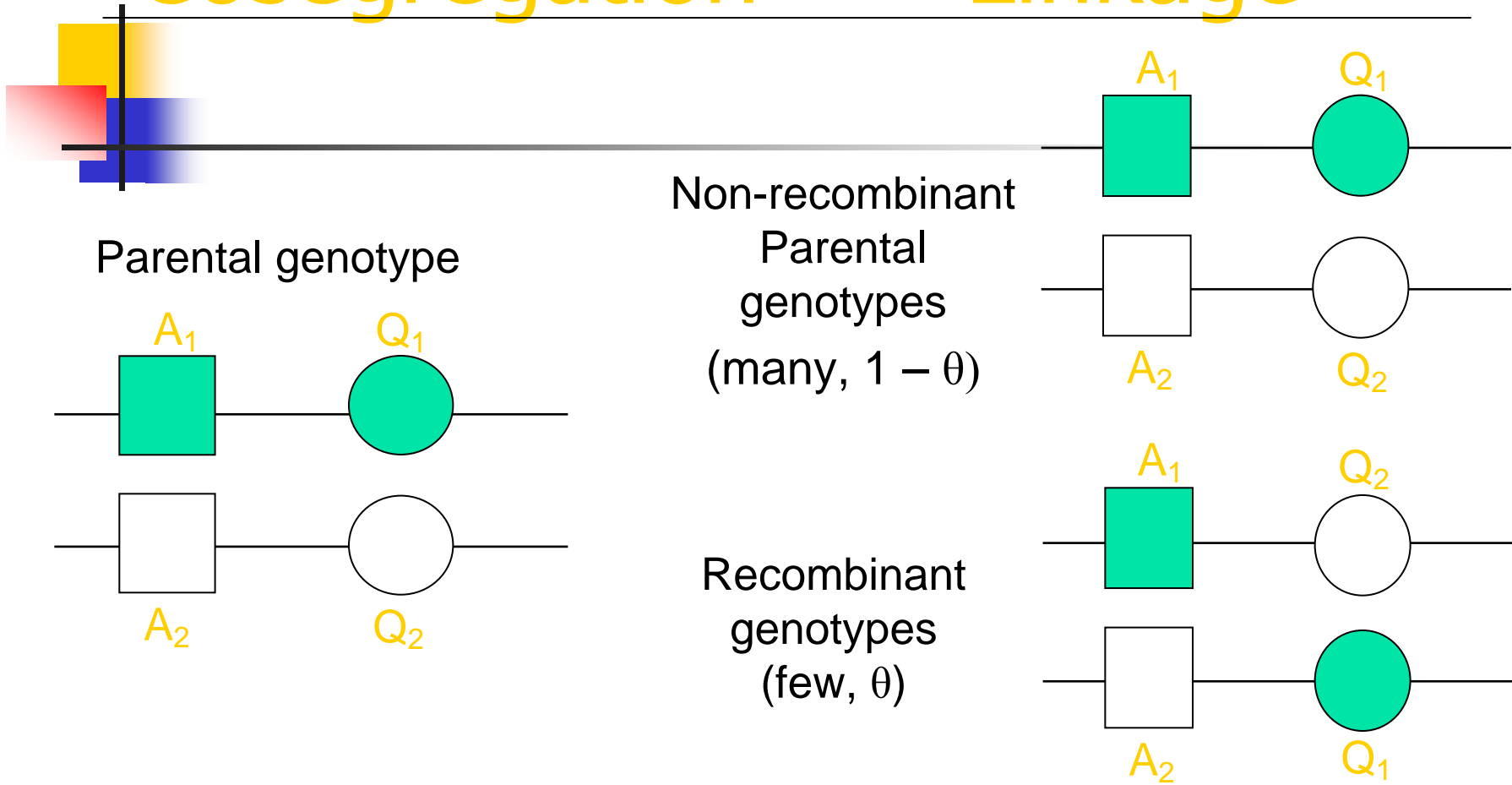# Why is IBD Sharing Important?

- 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

(a)   (b)   (c)   (d)

Cosegregation => Linkage

Parental genotype

A_1 Q_1

A_2 Q_2

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

A_1 Q_1

A_2 Q_2

Recombinant
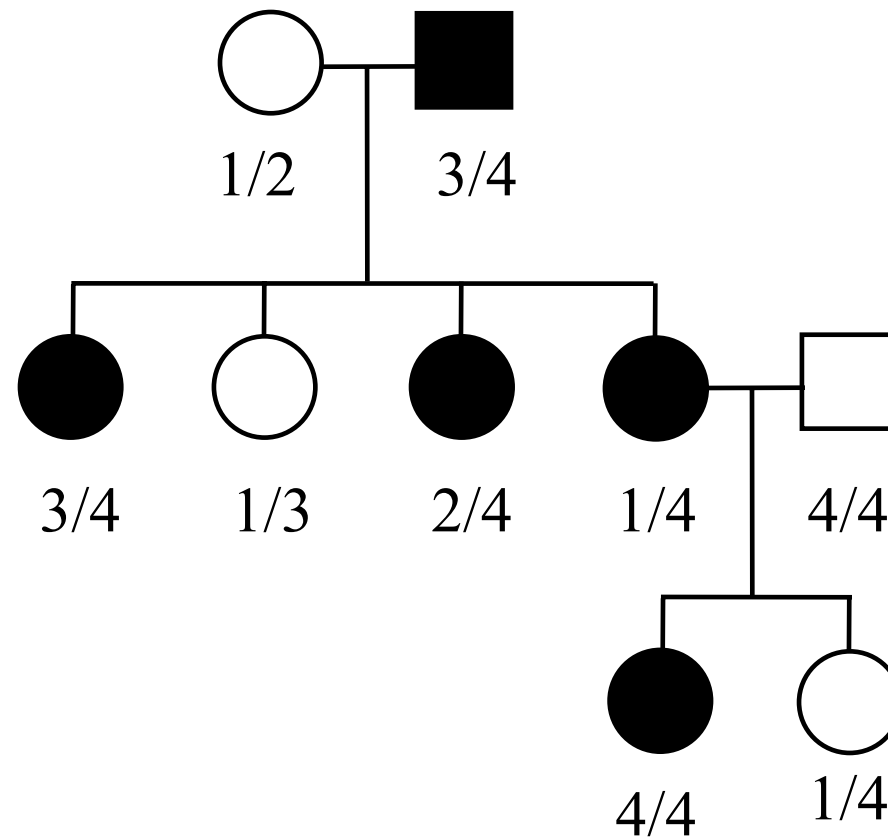genotypes
(few, θ)

A_1 Q_2

A_2 Q_1

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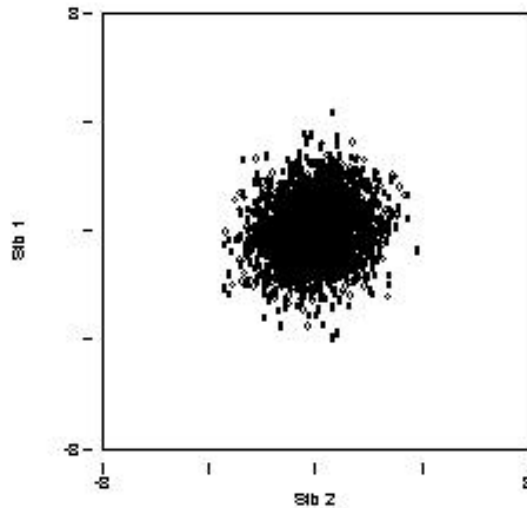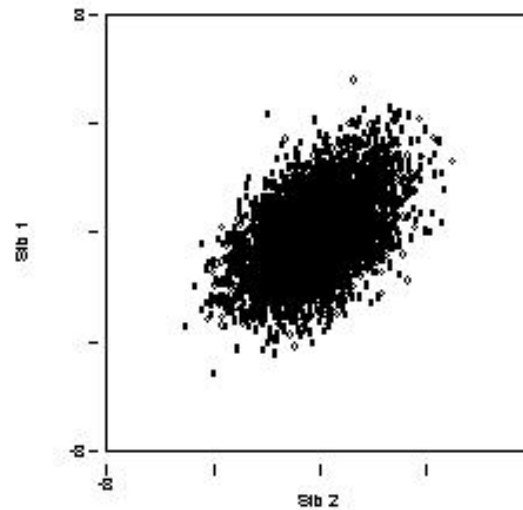
Segregating Chromosomes

MARKER

DISEASE GENE
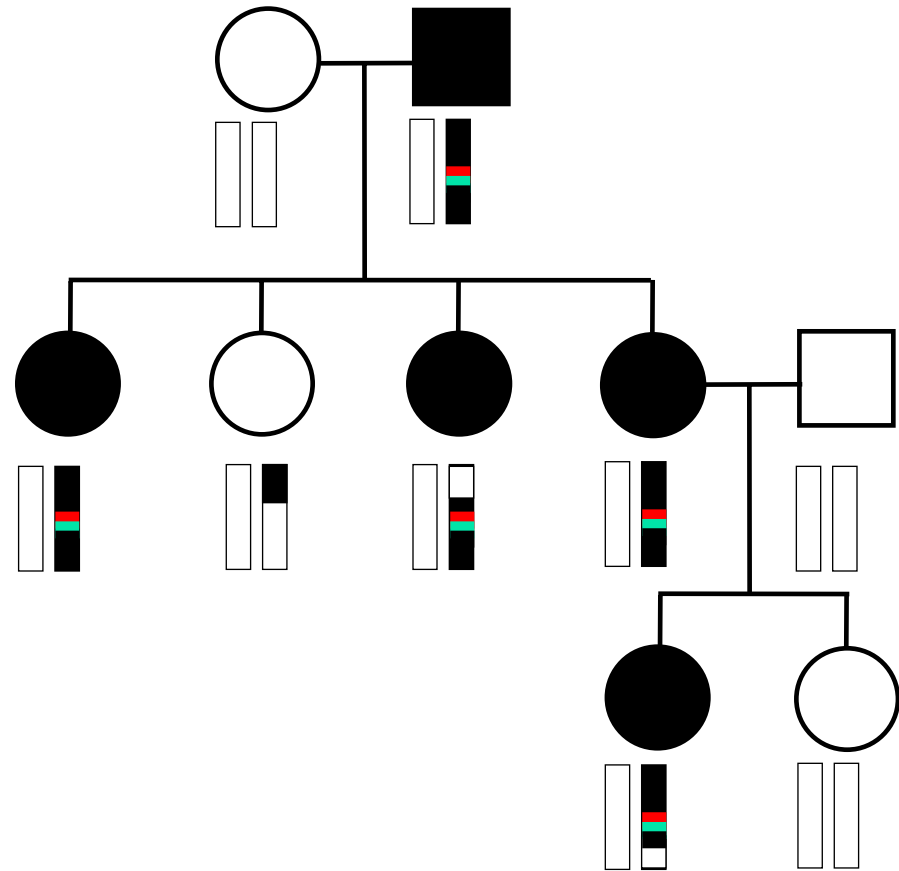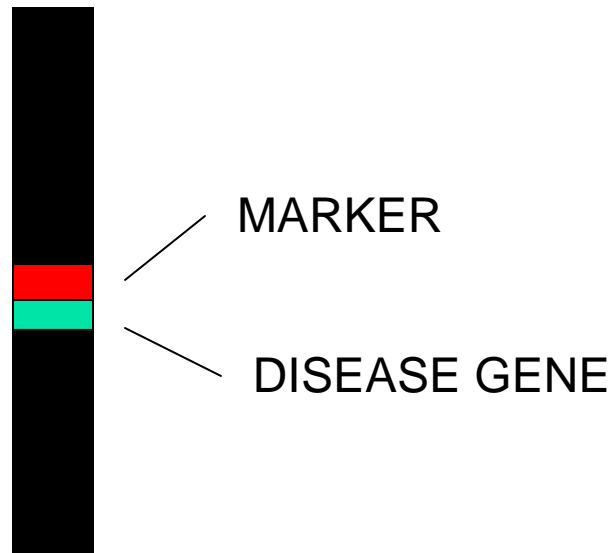
Marker Shared Among Affecteds

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

Linkage between QTL and marker

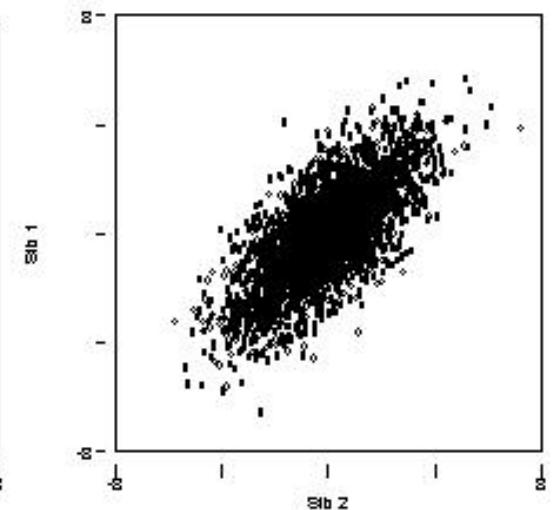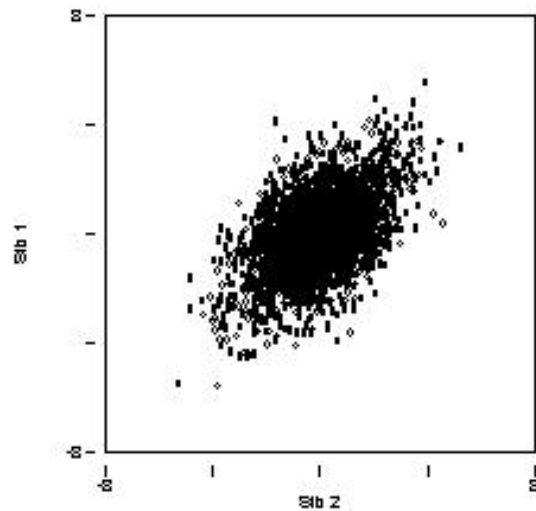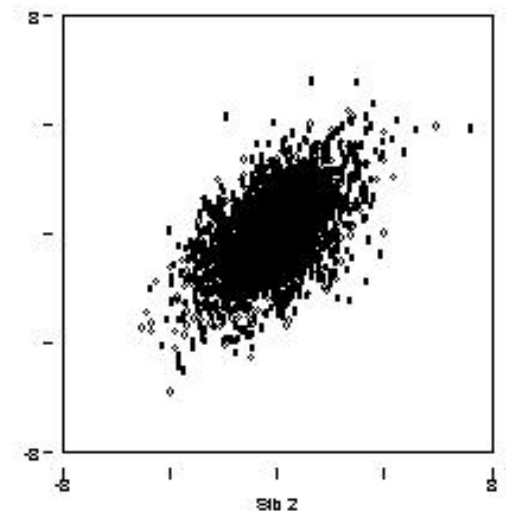| QTL	IBD 0	IBD 1	IBD 2
Marker	IBD 0	IBD 1	IBD 2

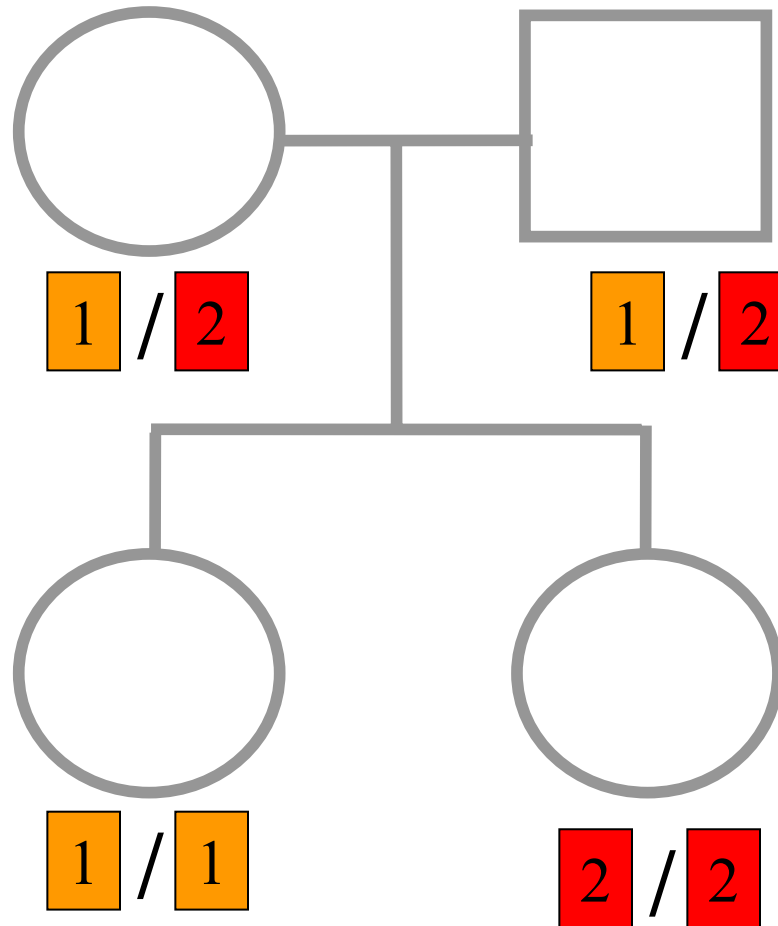# NO Linkage between QTL and marker
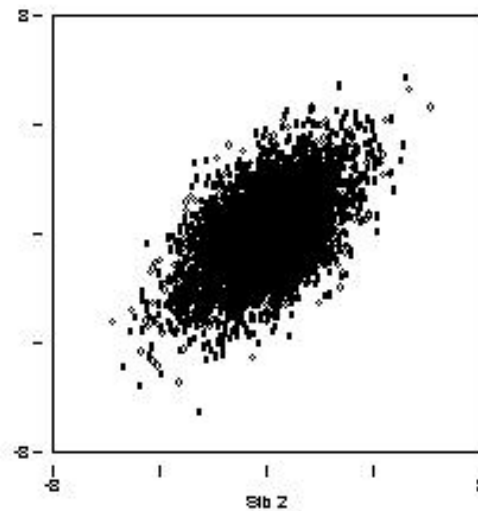
Marker

IBD 0   IBD 1   IBD 2

IBD can be trivial...

IBD=0

Two Other Simple Cases...

IBD=2

1 / 2   1 / 2   1 / 2   1 / 2

1 / 1   1 / 1   2 / 2   2 / 2
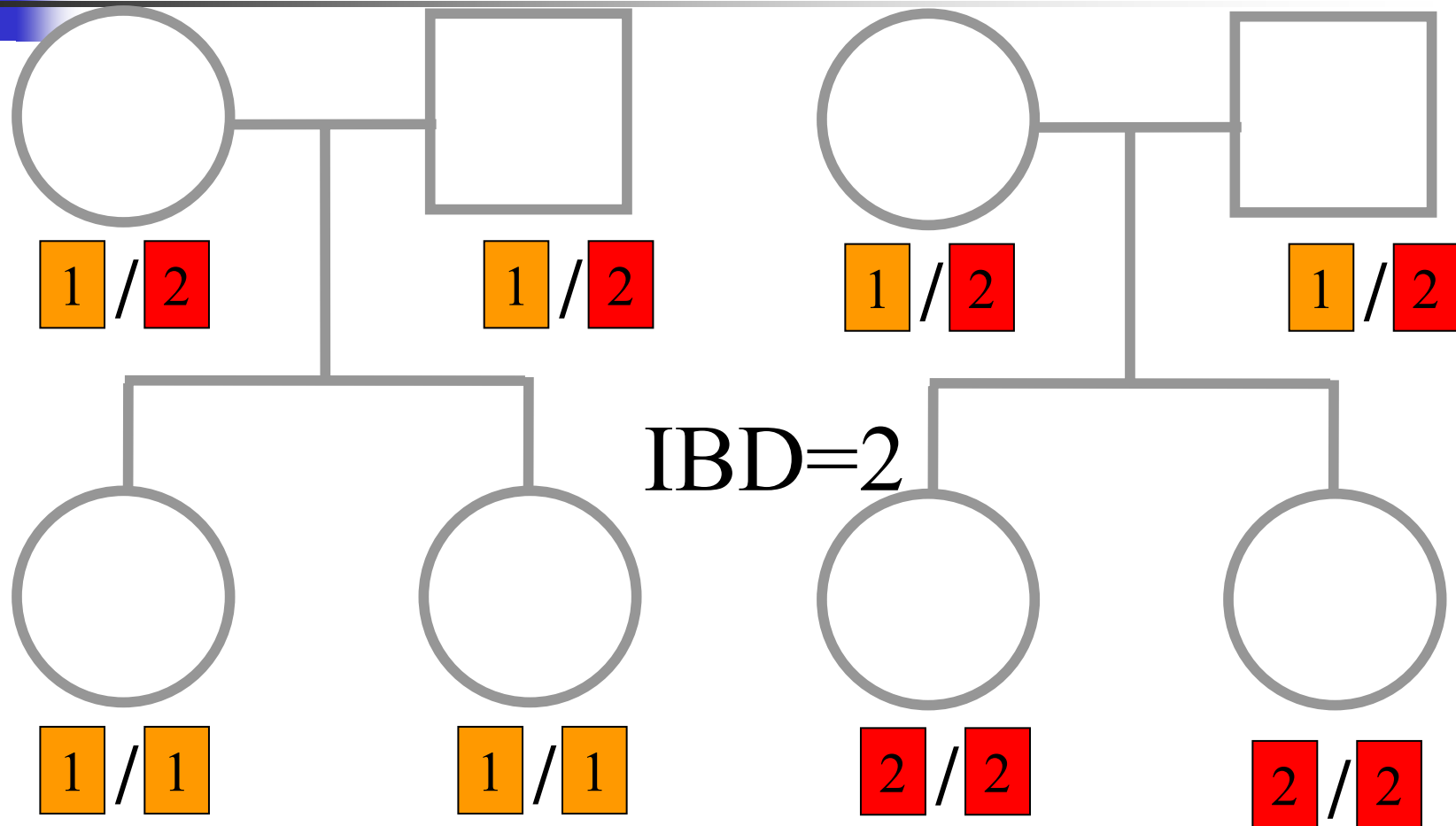
A little more complicated...
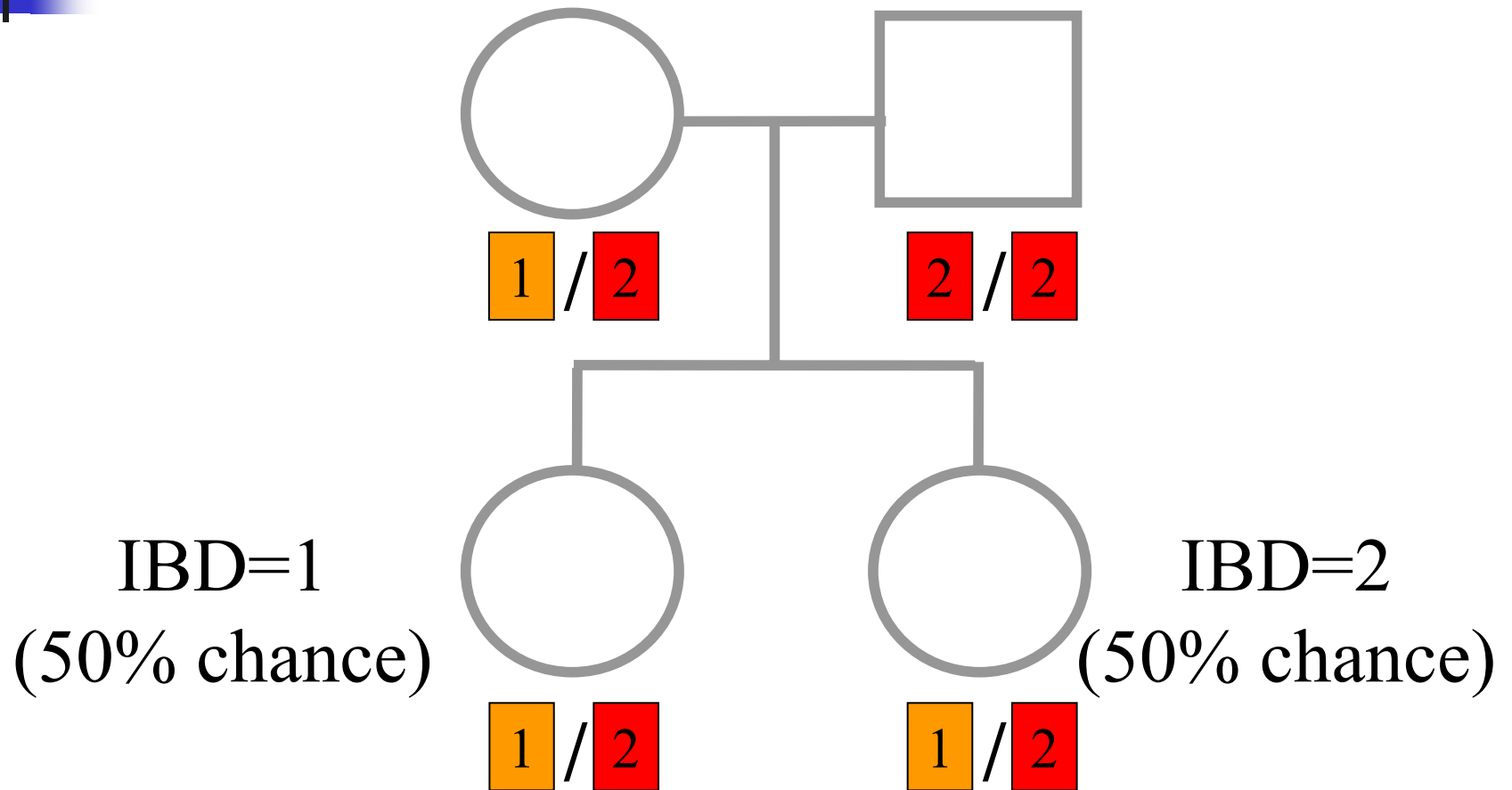
And even more complicated…

IBD=? 1 / 1   1 / 1
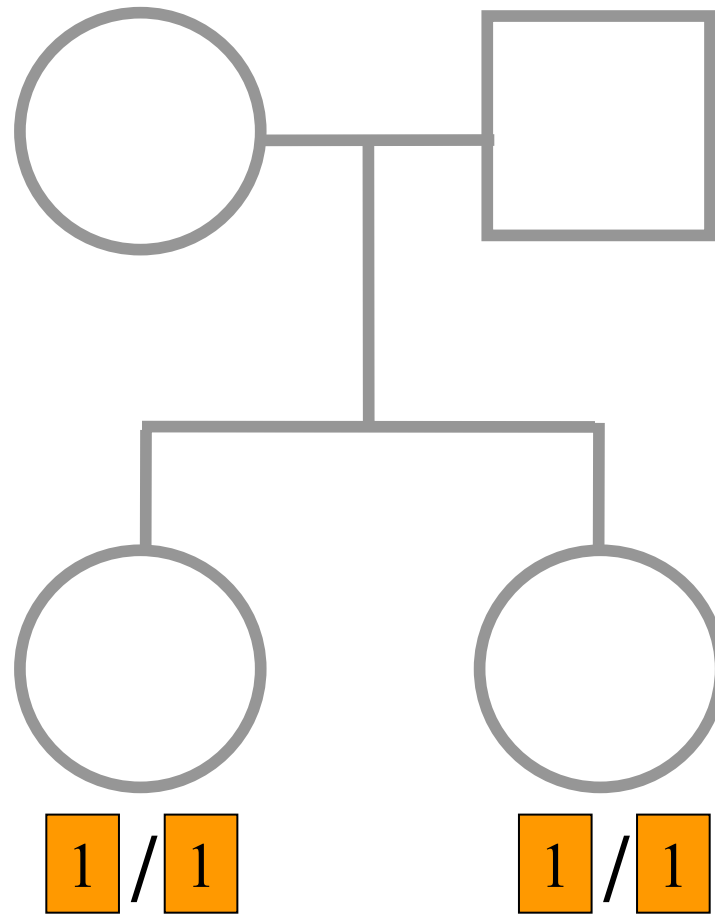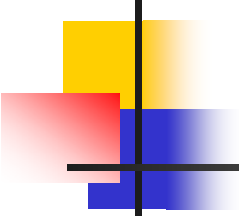
Bayes Theorem for IBD Probabilities

$$P(IBD = i \mid G) = \frac{P(IBD = i, G)}{P(G)}$$

$$= \frac{P(IBD = i)P(G \mid IBD = i)}{P(G)}$$

$$= \frac{P(IBD = i)P(G \mid IBD = i)}{\sum_{j=0,1,2} P(IBD = j)P(G \mid IBD = j)}$$

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
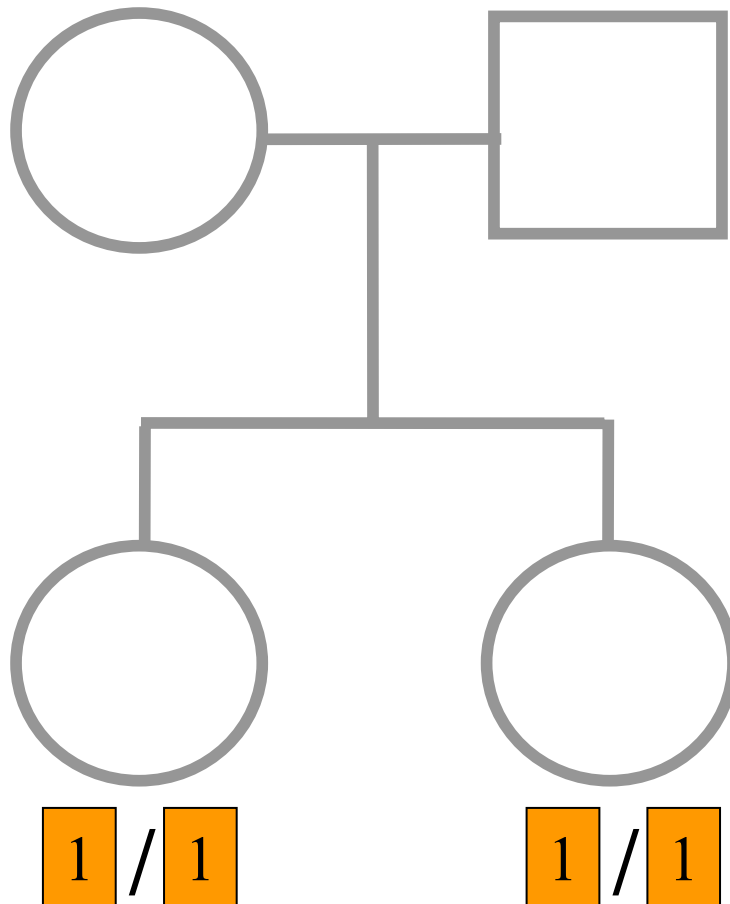
$p_1 = 0.5$

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

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

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

$P(G) =$

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

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

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

1 / 1   1 / 1

Worked Example
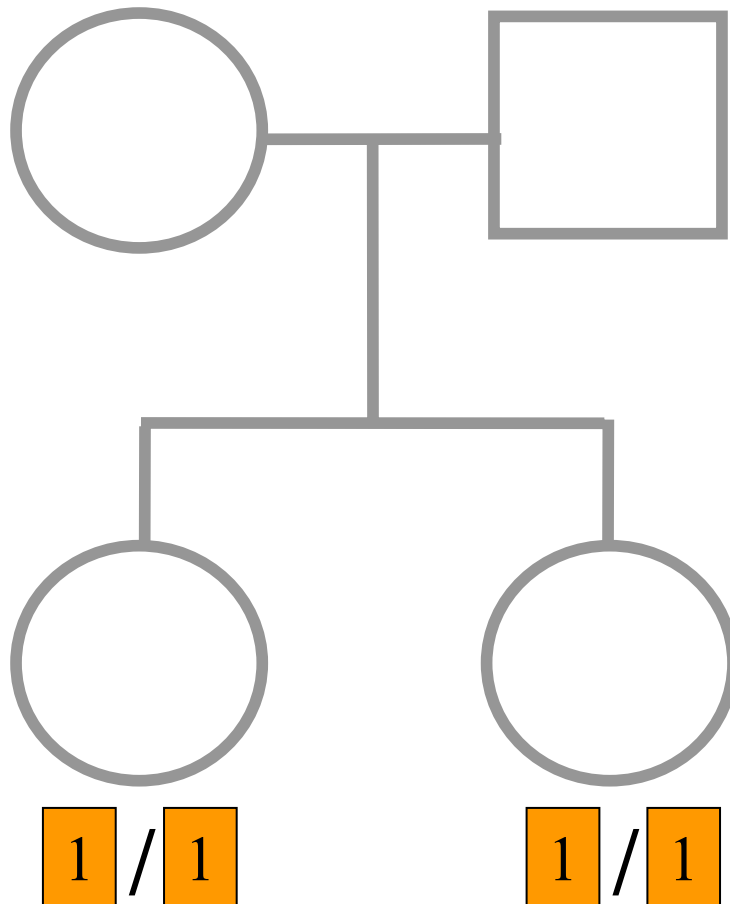
$p_1 = 0.5$

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

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

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

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

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

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

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

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

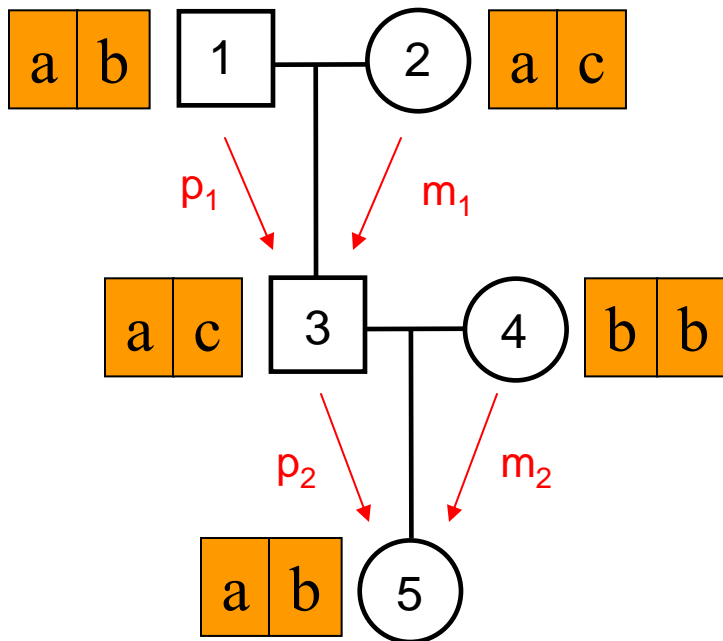p_i (m_i) is 0 if the grandpaternal allele transmitted
p_i (m_i) is 1 if the grandmaternal allele is transmitted
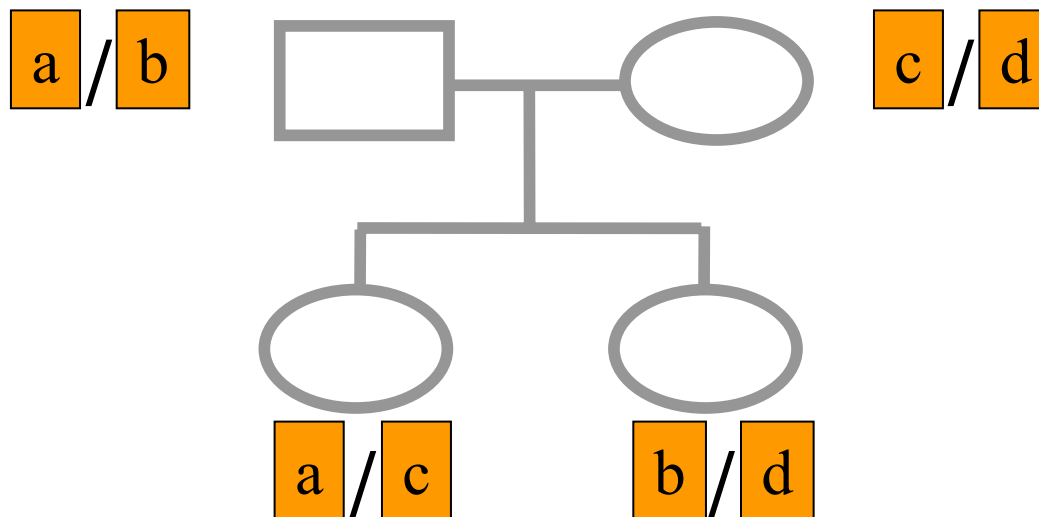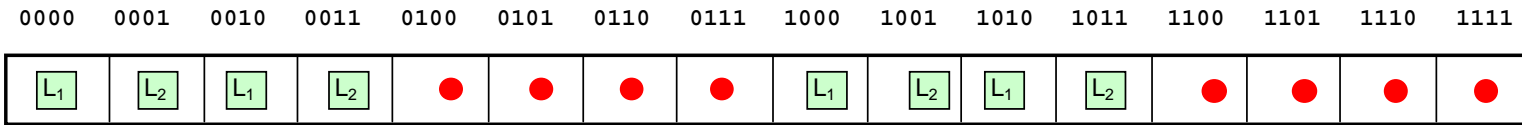

$$v(x) = [0,0,1,1]$$

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
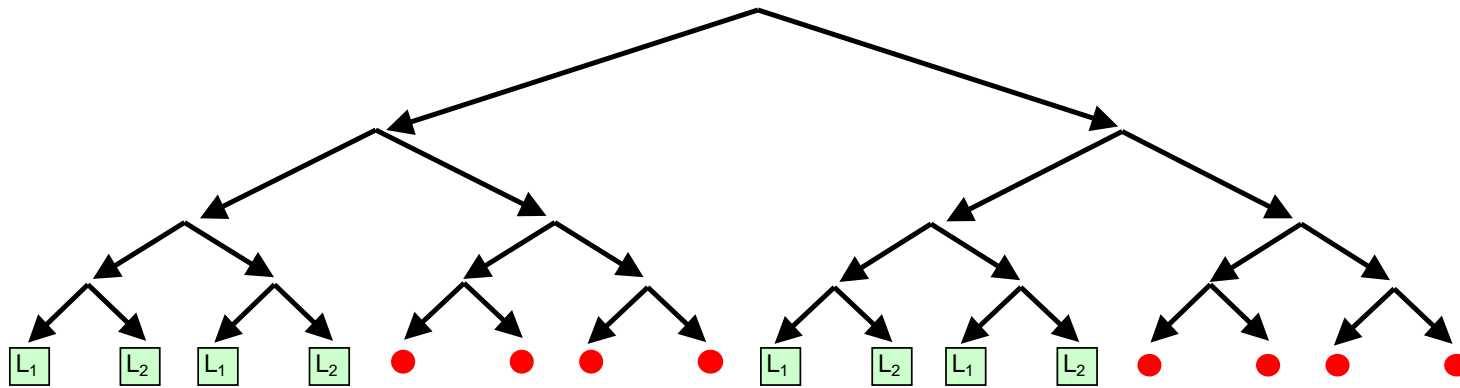

Inheritance vector	Prior	Posterior
0000	1/16	1/8
0001	1/16	1/8
0010	1/16	0
0011	1/16	0
0100	1/16	1/8
0101	1/16	1/8
0110	1/16	0
0111	1/16	0
1000	1/16	1/8
1001	1/16	1/8
1010	1/16	0
1011	1/16	0
1100	1/16	1/8
1101	1/16	1/8
1110	1/16	0
1111	1/16	0

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

0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
L_1	L_2	L_1	L_2	●	●	●	●	L_1	L_2	L_1	L_2	●	●	●	●

b) packed tree

c) sparse tree
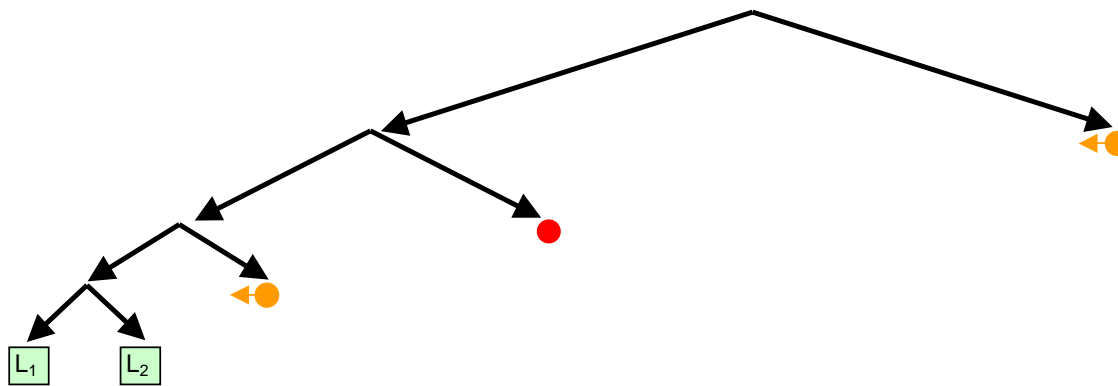
Legend

●   Node with zero likelihood

◀●   Node identical to sibling

L_1 L_2 Likelihood for this branch

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

a / b
1 / 1

c / d
1 / 2

IBD = 0

a / c
1 / 1

b / d
1 / 2

IBD = 0 or IBD = 1?
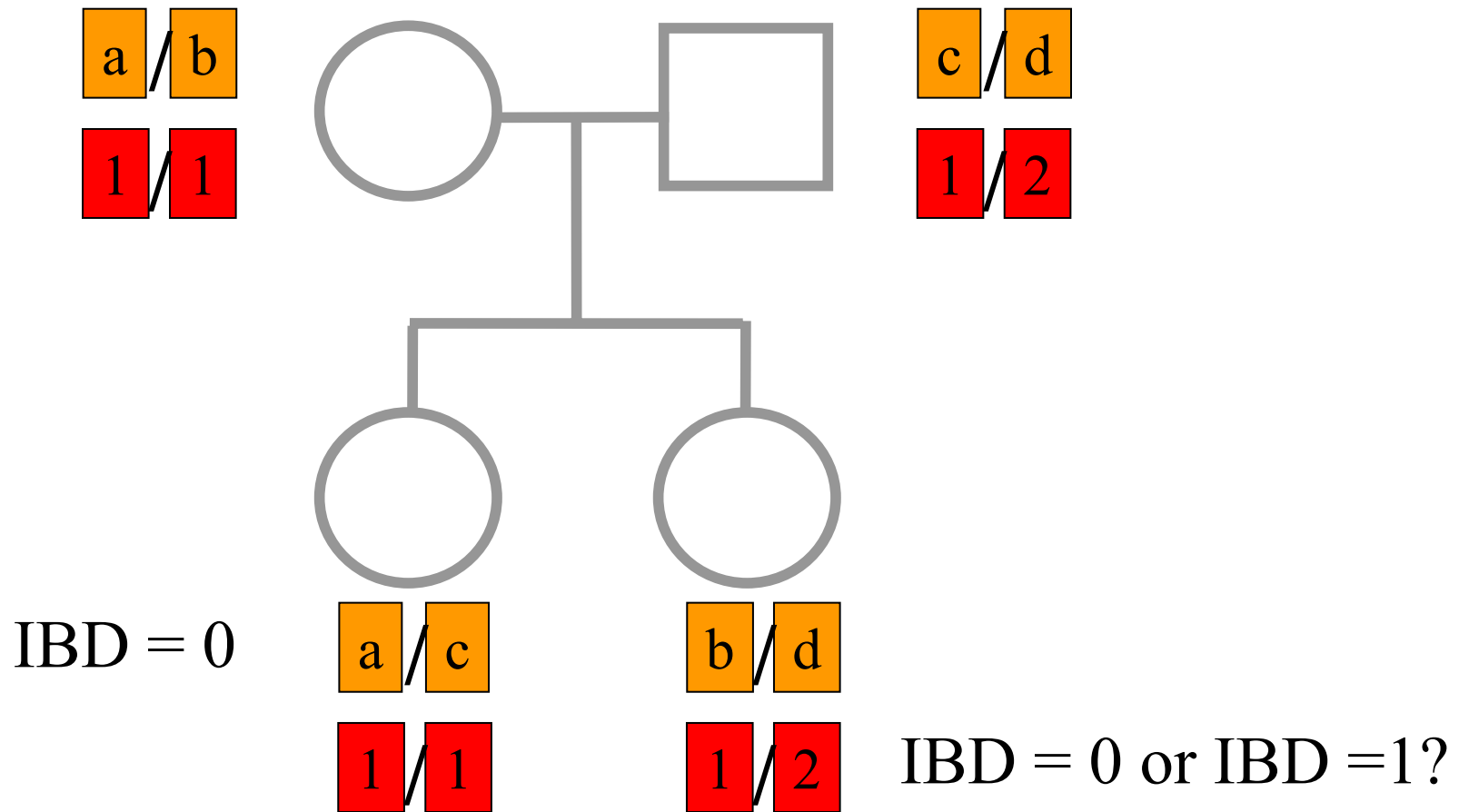
Complexity of the Problem in Larger Pedigrees

- For each person
  - $2n$ 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
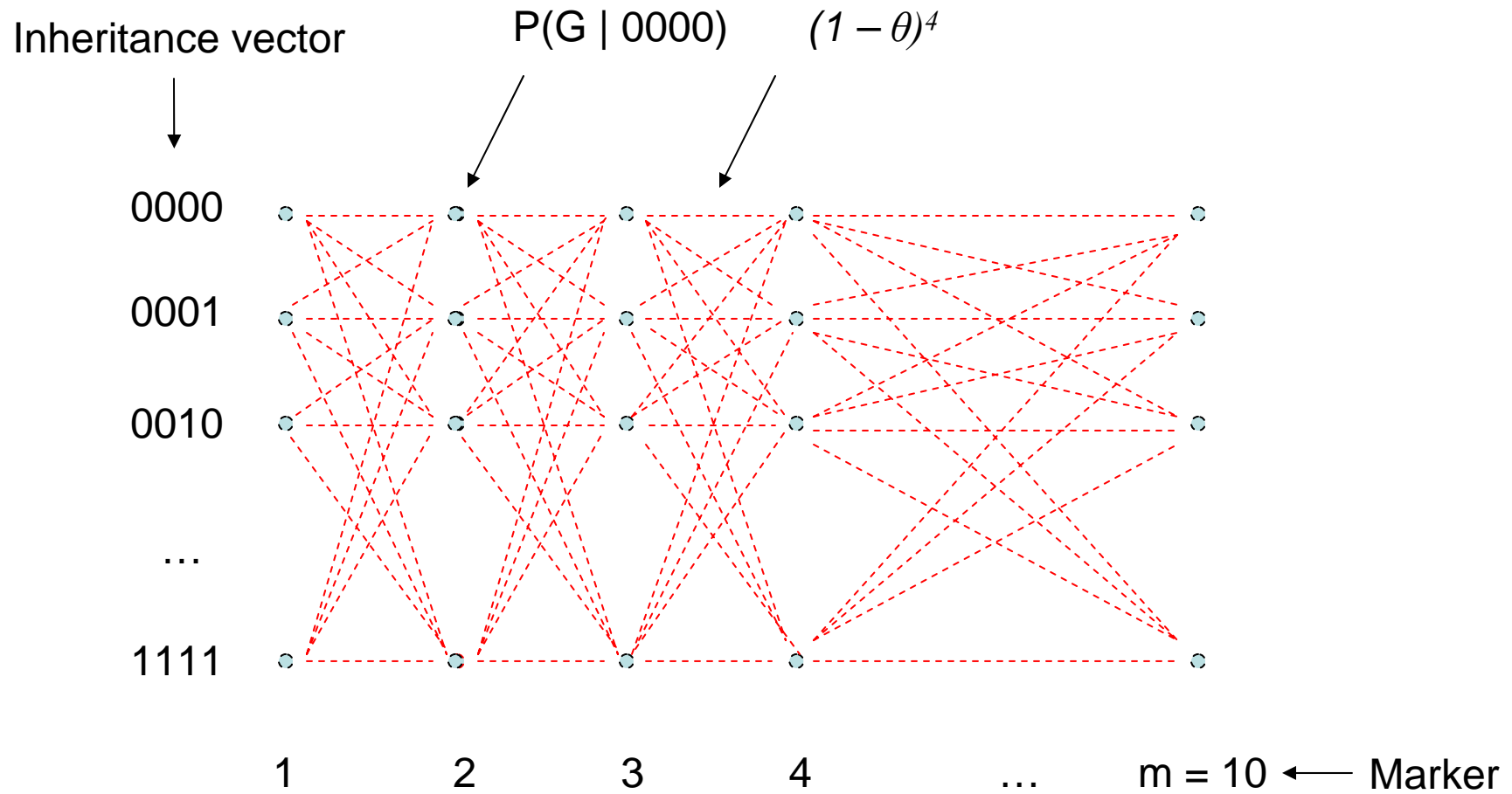
$(2^{2 \times n})^m = (2^{2 \times 2})^{10} =\sim 10^{12}$ possible paths !!!

P(IBD) = 2 at Marker Three

IBD   Inheritance vector

(2)   0000

(1)   0001

(1)   0010

…

(2)   1111

1   2   3   4   …   m = 10 ← Marker
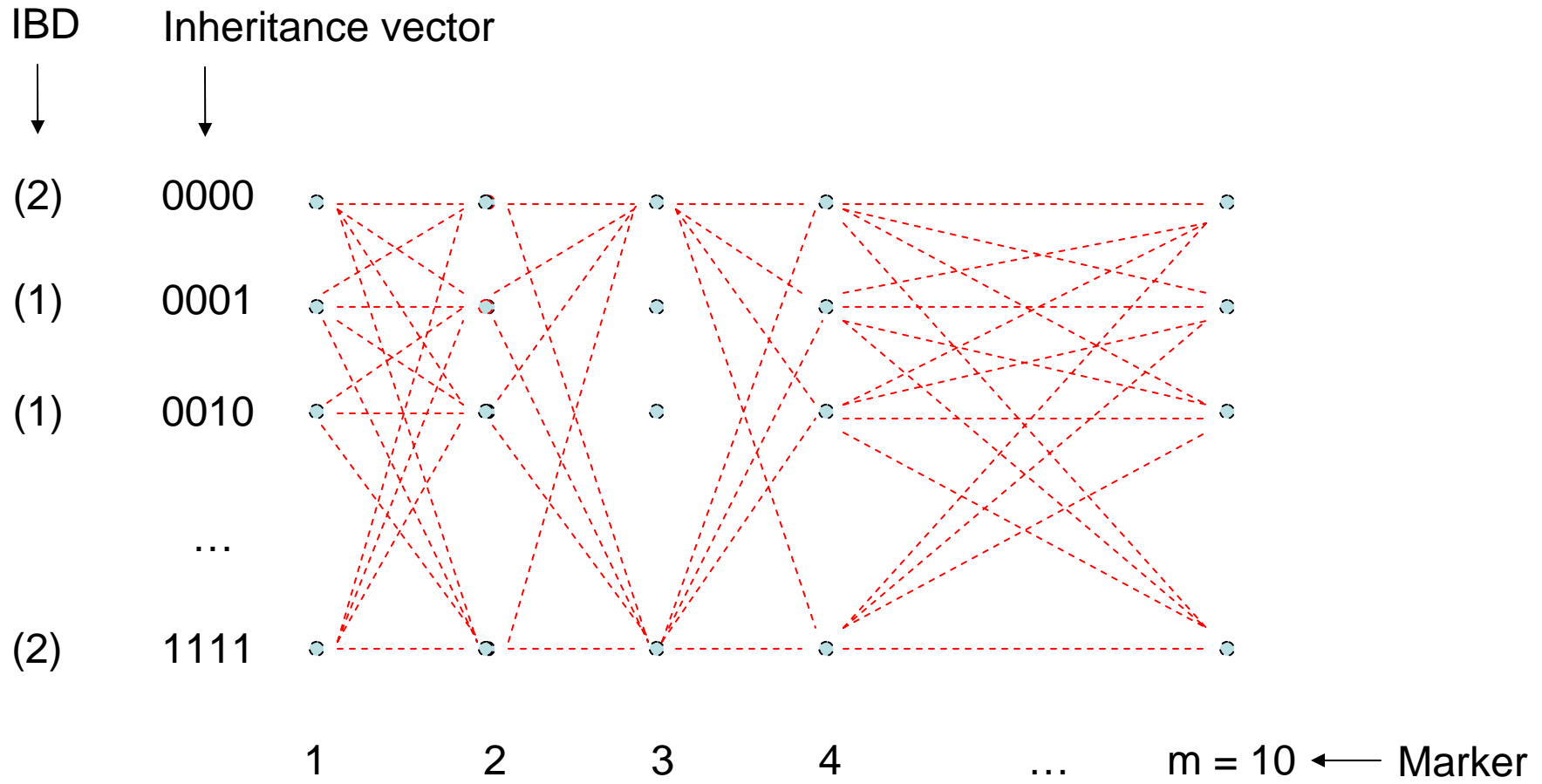
(L[0000] + L[0101] + L[1010] + L[1111]) / L[ALL]

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

Inheritance vector

0000
0001
0010
…
1111

1   2   3   4   …   m = 10 ← Marker
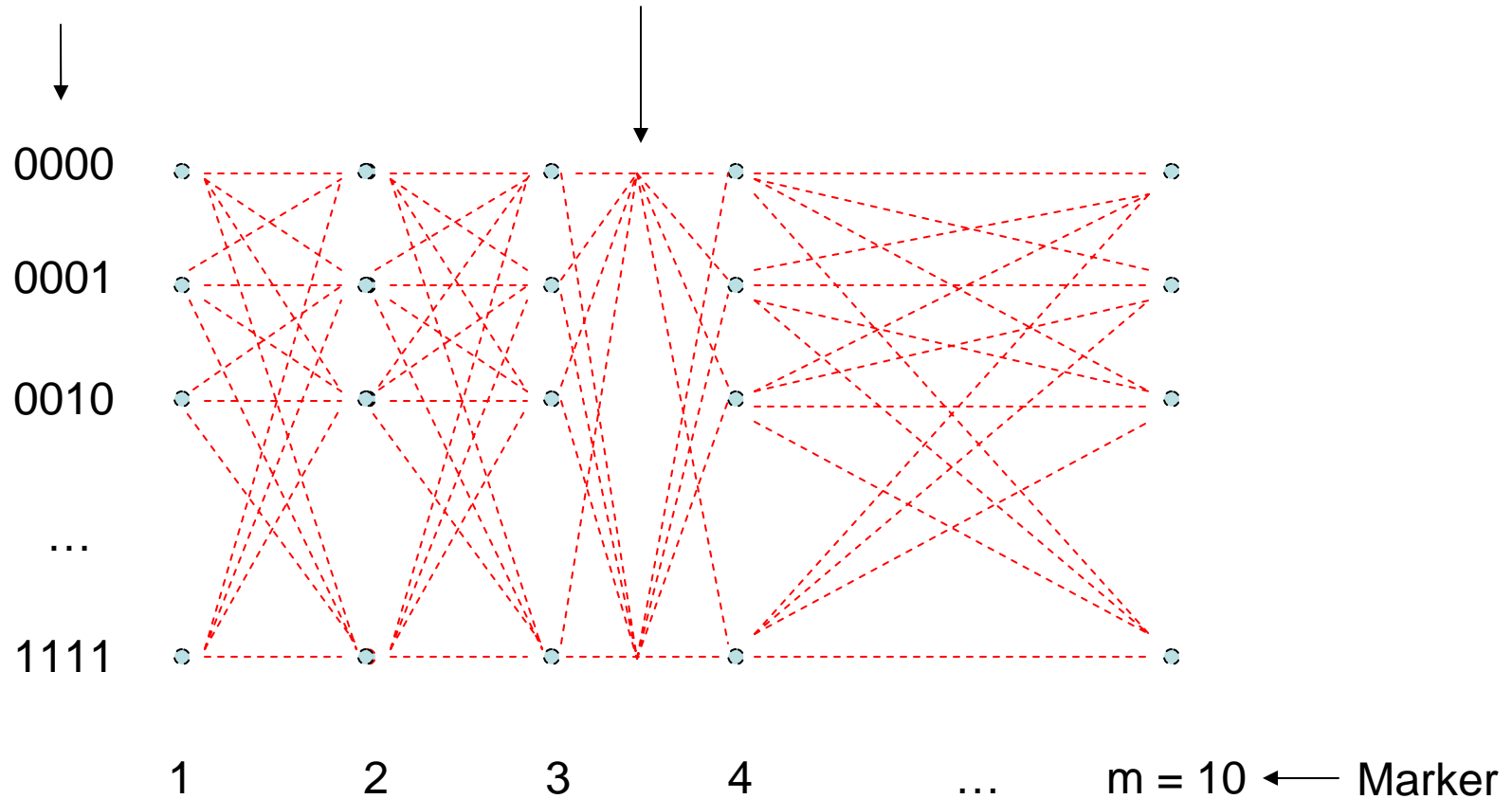
(L[0000] + L[0101] + L[1010] + L[1111]) / L[ALL]

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")

- The conditional probability of an inheritance vector v_{i+1} at locus $i+1$, given the inheritance vector v_i at locus i is $\theta_i^j(1-\theta_i)^{2n-j}$ where θ is the recombination fraction and j is the number of changes in elements of the inheritance vector

Example:

Locus 1   Locus 2

[0000]   [0001]

Conditional probability $= (1-\theta)^3\theta$

Lander-Green Algorithm

Inheritance vector

$\downarrow$

0000

0001

0010

…

1111

    1   2   3   4   …   m = 10 ← Marker
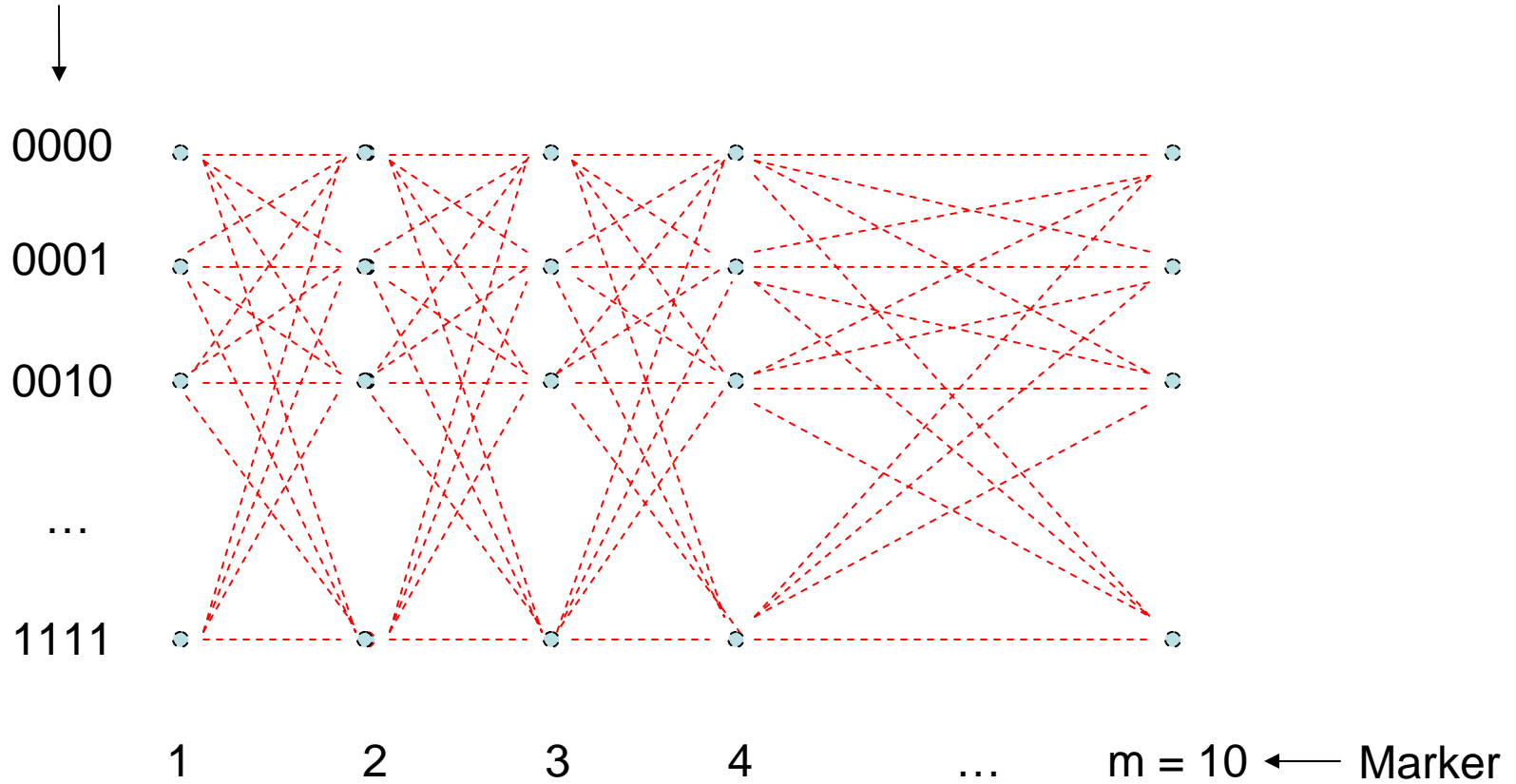
$$M(2^{2n})^2 = 10 \times 16^2 = 2560 \text{ calculations}$$
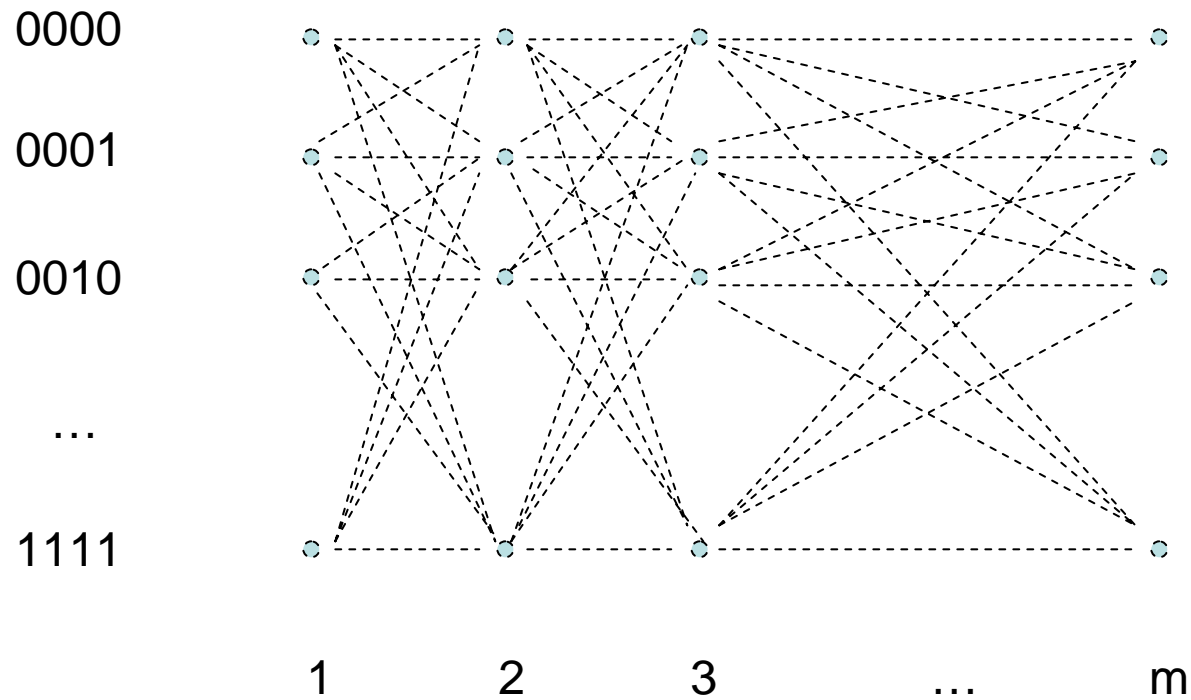
$$\text{Total Likelihood} = \mathbf{1'Q_1T_1Q_2T_2\ldots T_{m-1}Q_m1}$$

$$Q_i = \begin{matrix} P(G|[0000]) & 0 & 0 & 0 \\ 0 & P(G|[0001]) & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & P(G|[1111]) \end{matrix}$$

2^{2n} x 2^{2n} **diagonal matrix of single locus probabilities at locus *i***

$$T_i = \begin{matrix} (1-\theta)^4 & (1-\theta)^3\theta & \ldots & \theta^4 \\ (1-\theta)^3\theta & (1-\theta)^4 & \ldots & (1-\theta)\theta^3 \\ \ldots & \ldots & \ldots & \ldots \\ \theta^4 & (1-\theta)\theta^3 & \ldots & (1-\theta)^4 \end{matrix}$$

2^{2n} x 2^{2n} **matrix of transitional probabilities between locus *i* and locus *i+1***

$\sim m(2^{2n})^2$ operations = 2560 for this case !!!

Further speedups…

- Trees summarize redundant information
    - Portions of vector that are repeated
    - Portions of vector that are constant or zero
- Speeding up convolution
    - Use sparse-matrix by vector multiplication
    - Use symmetries in divide and conquer algorithm (Idury & Elston, 1997)

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

©1998 Jeff Bucchino

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
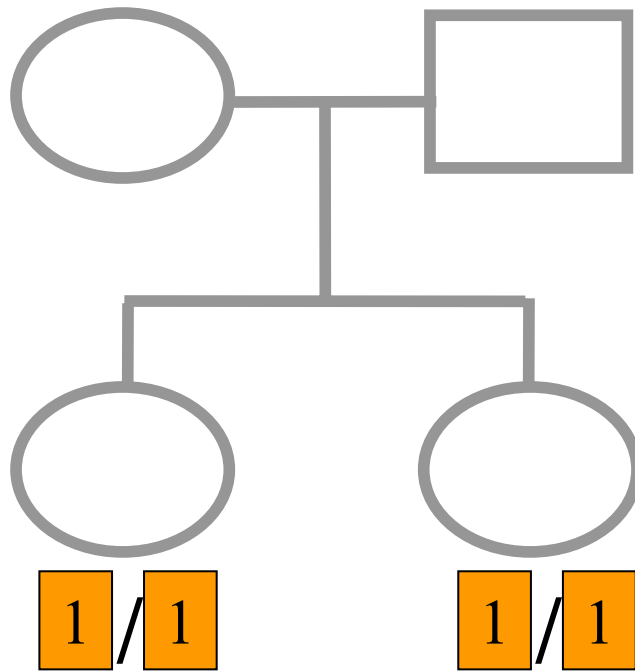
$$p_1 = 0.5$$

$$P(IBD = 0 \mid G) = \tfrac{1}{9}$$

$$P(IBD = 1 \mid G) = \tfrac{4}{9}$$

$$P(IBD = 2 \mid G) = \tfrac{4}{9}$$

| 1 / 1 | | 1 / 1 |

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 ith 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