Calculation of IBD State Probabilities

Gonçalo Abecasis
University of Michigan

Human Genome

- Multiple chromosomes
  - Each one is a DNA double helix
  - 22 autosomes
    - Present in 2 copies
    - One maternal, one paternal
  - 1 pair of sex chromosomes
    - Females have two X chromosomes
    - Males have one X chromosome and one Y chromosome
- Total of ~3 x 10^9 bases

Human Variation

- When two chromosomes are compared most of their sequence is identical
  - Consensus sequence
- About 1 per 1,000 bases differs between pairs of chromosomes in the population
  - In the same individual
  - In the same geographic location
  - Across the world

Aim of Gene Mapping Experiments

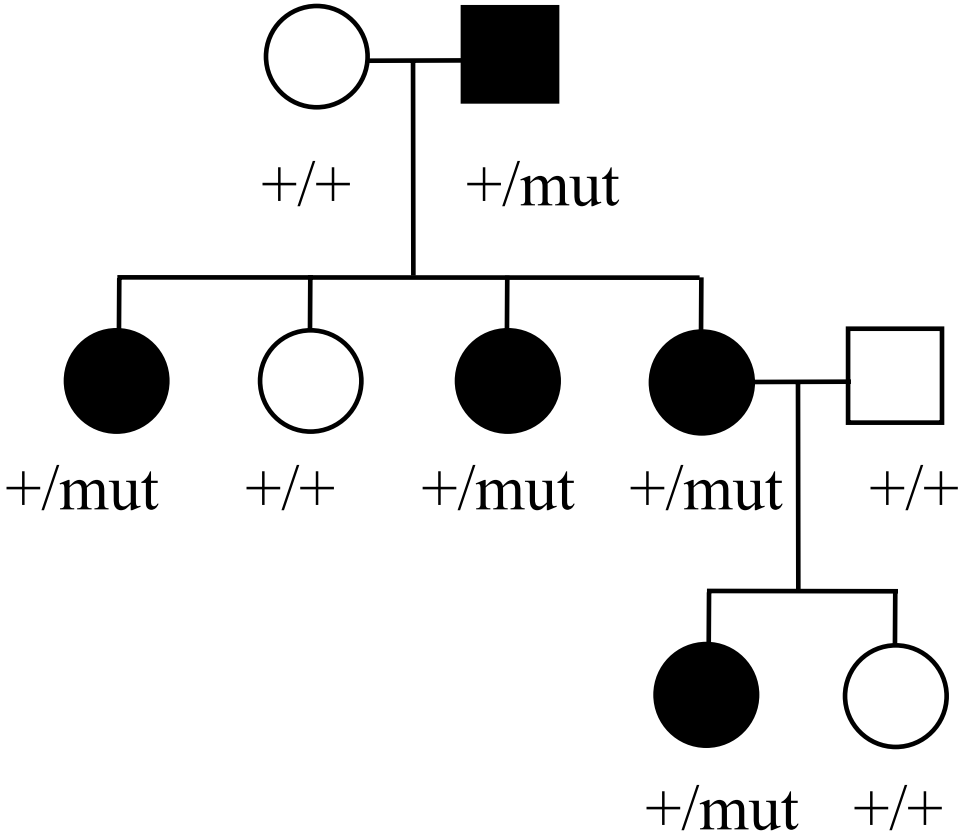
- Identify variants that control interesting traits
  - Susceptibility to human disease
  - Phenotypic variation in the population
- The hypothesis
  - Individuals sharing these variants will be more similar for traits they control
- The difficulty…
  - Testing over 4 million variants is impractical…

Identity-by-Descent (IBD)

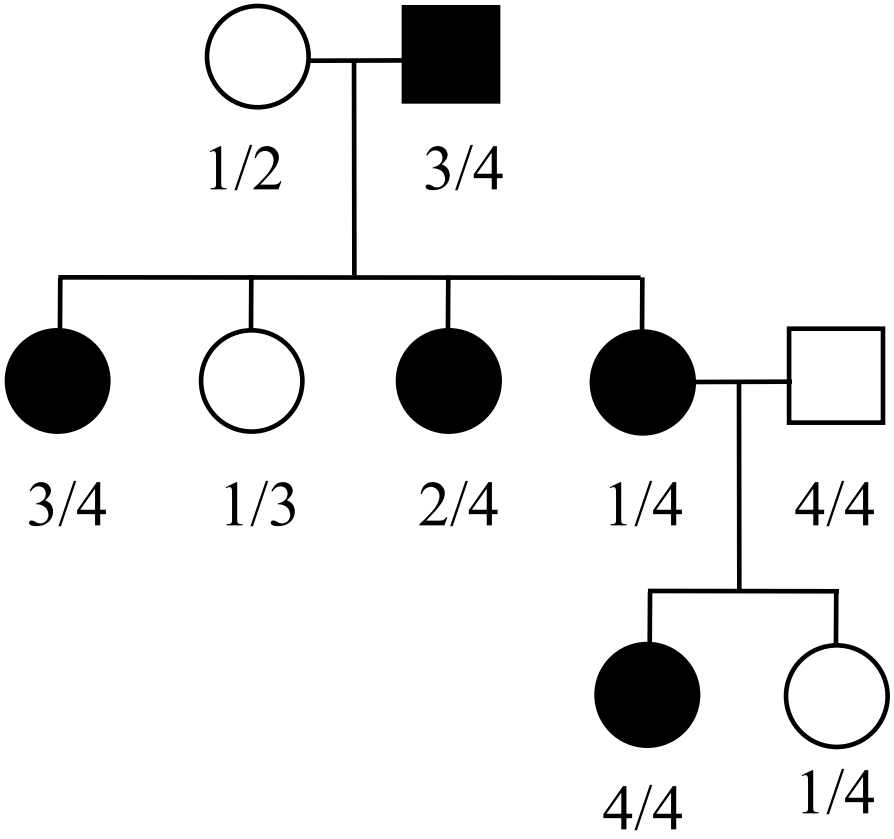
- A property of chromosome stretches that descend from the same ancestor

- Allows surveys of large amounts of variation even when a few polymorphisms measured
  - If a stretch is IBD among a set of individuals, all variants within it will be shared

A Segregating Disease Allele

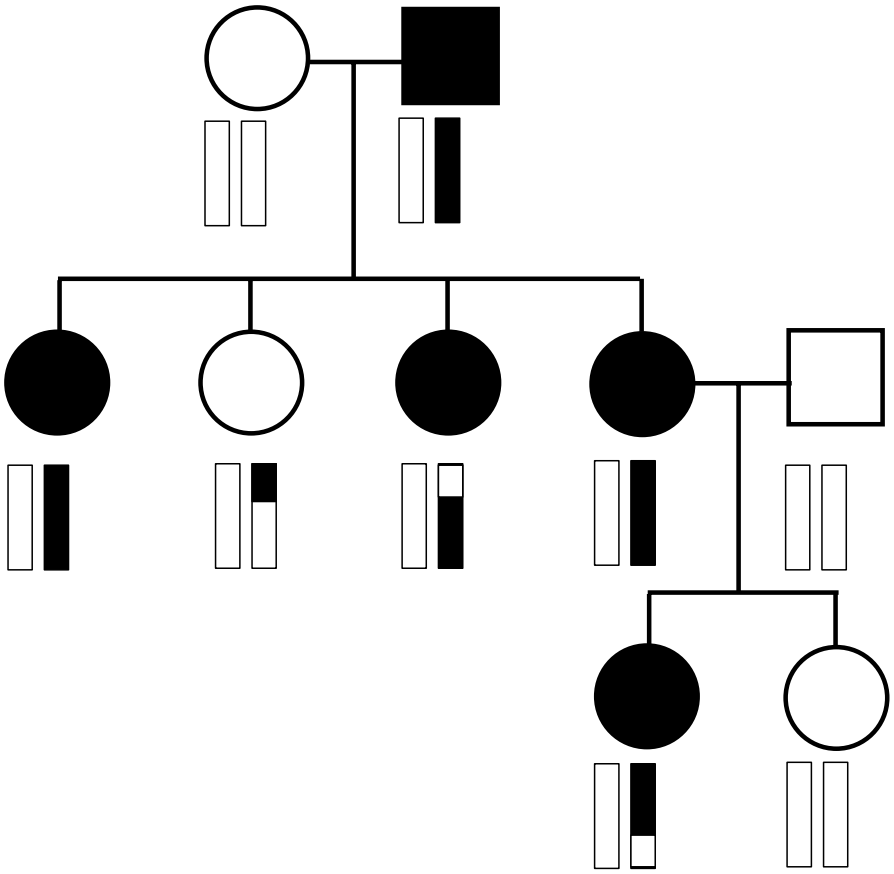


Marker Shared Among Affecteds

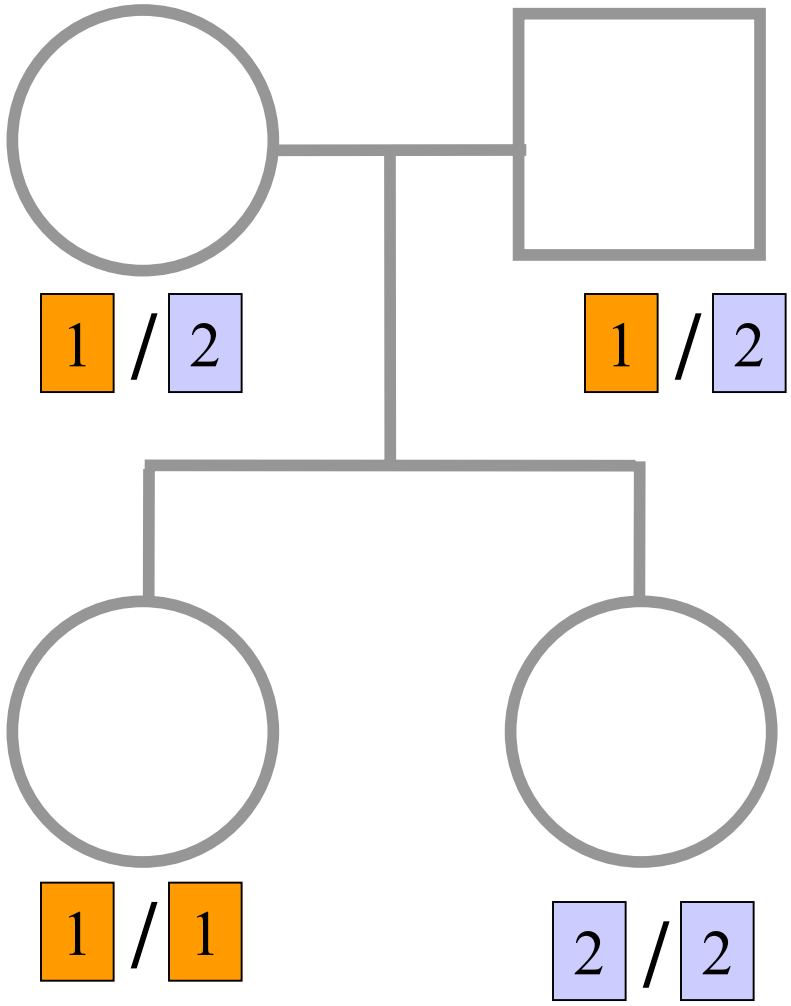


Genotypes for a marker with alleles {1,2,3,4}

Segregating Chromosomes

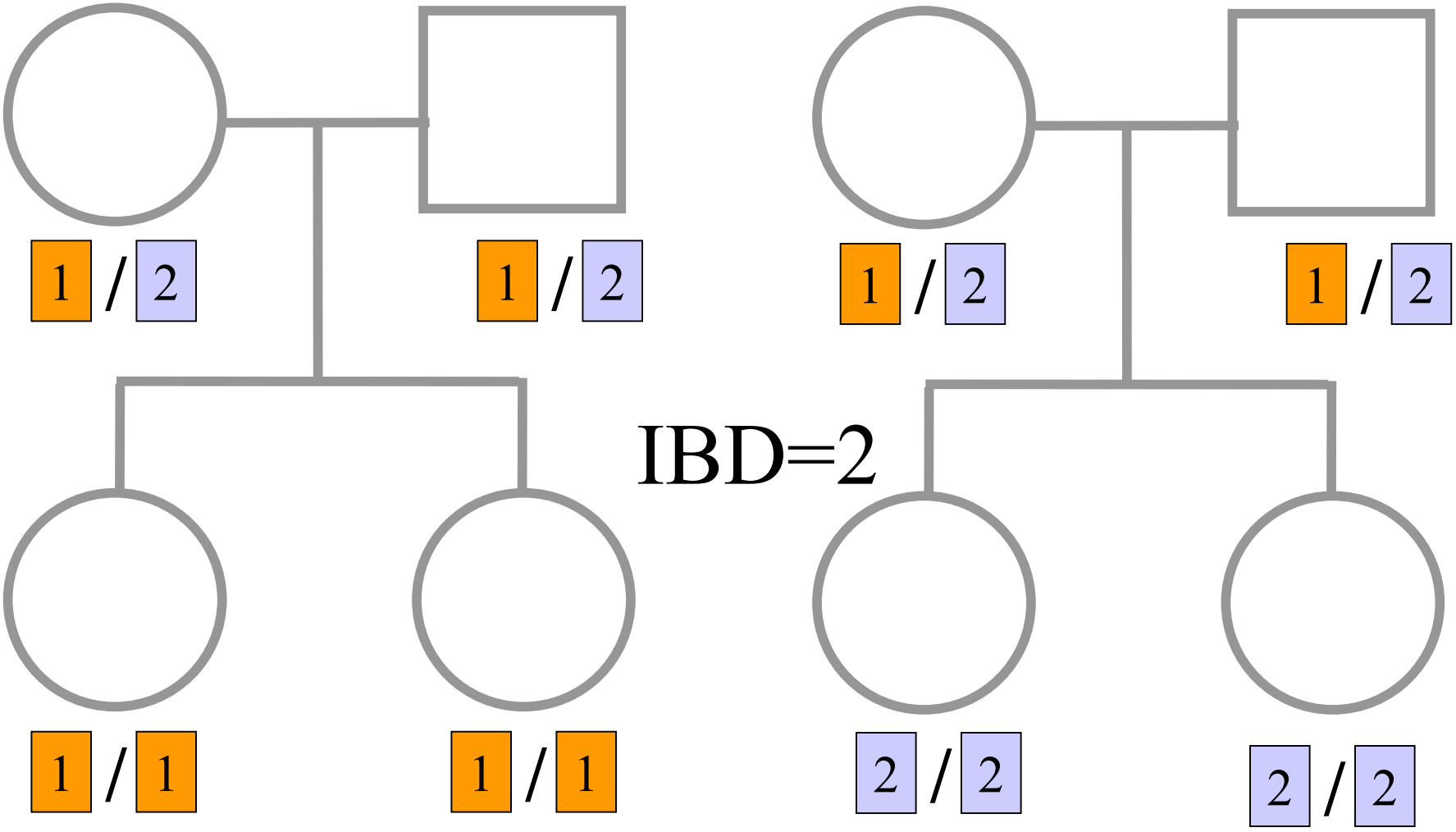


IBD can be trivial…

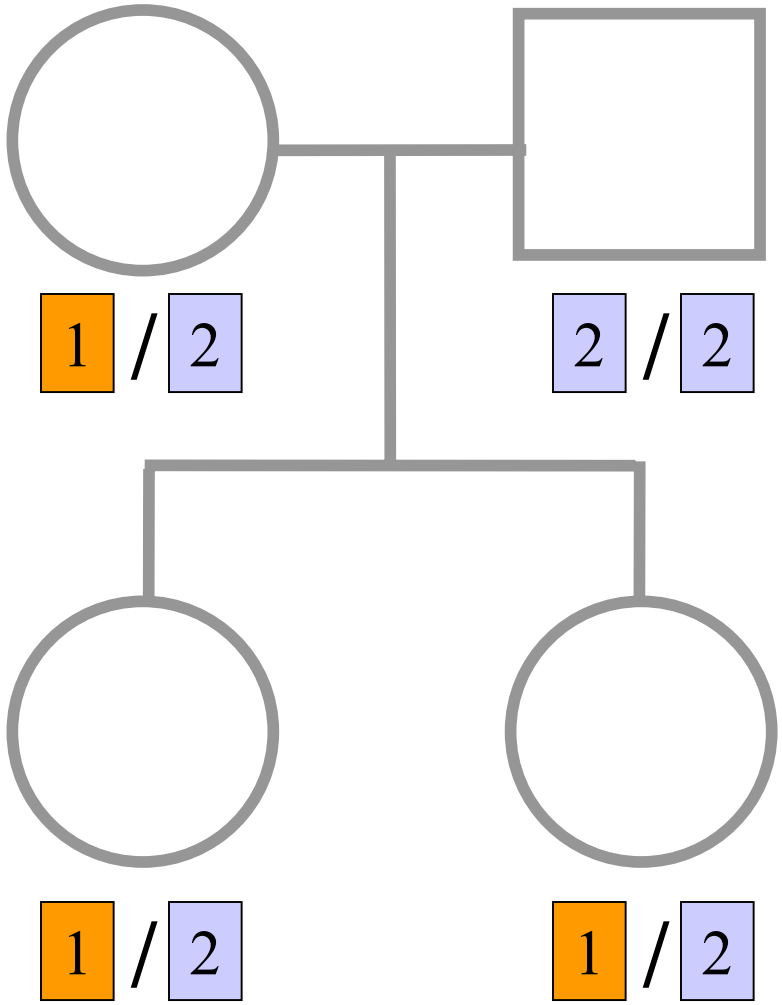


IBD=0

Two Other Simple Cases…



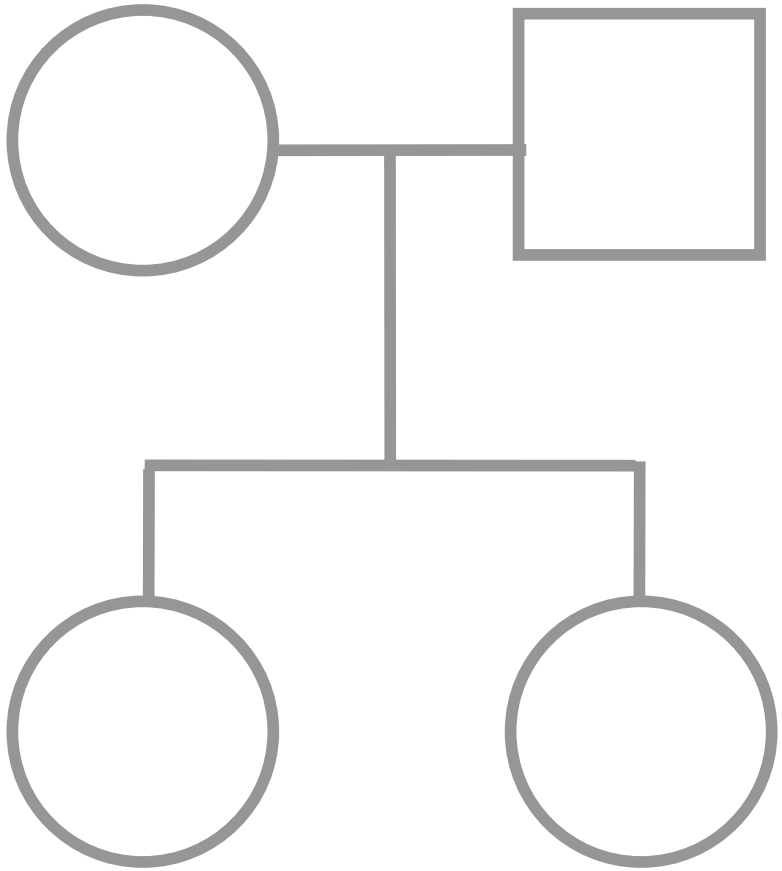
A little more complicated…



IBD=1
(50% chance)

IBD=2
(50% chance)

And even more complicated…



IBD=?

1 / 1

1 / 1

Bayes Theorem for IBD Probabilities

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

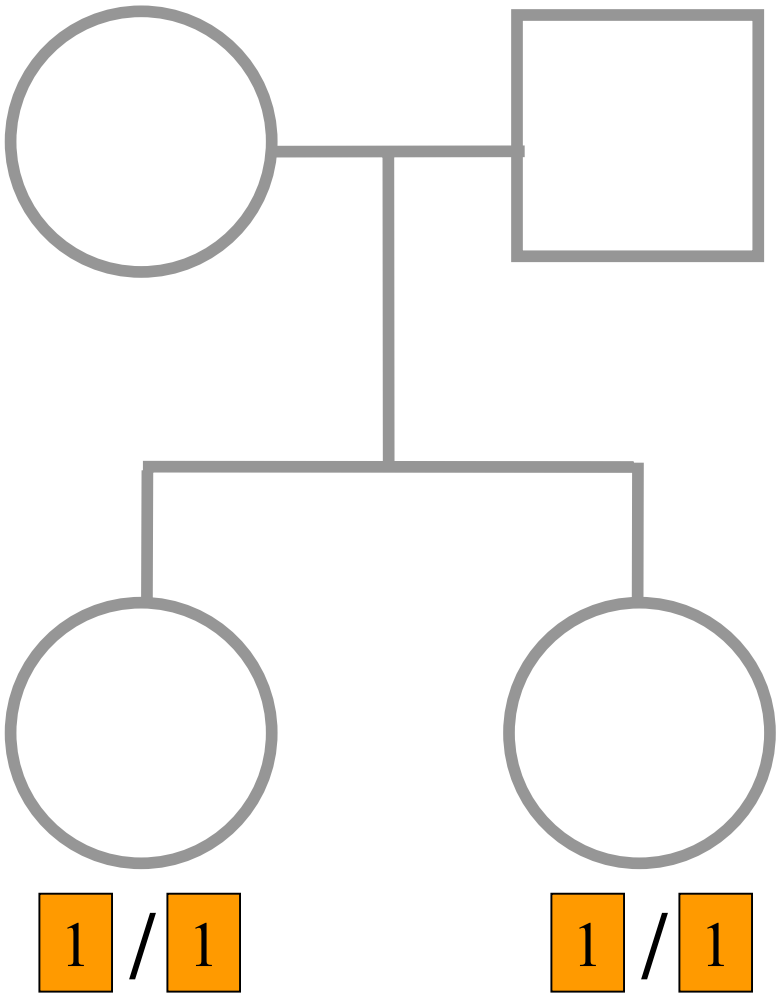
$$= \frac{P(IBD = i)P(G \mid IBD = i)}{P(G)}$$

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

P(Marker Genotype|IBD State)

		IBD		
Sib	**CoSib**	0	1	2
(a,b)	(c,d)	$p_a p_b p_c p_d$	0	0
(a,a)	(b,c)	$p_a^2 p_b p_c$	0	0
(a,a)	(b,b)	$p_a^2 p_b^2$	0	0
(a,b)	(a,c)	$p_a^2 p_b p_c$	$p_a p_b p_c$	0
(a,a)	(a,b)	$p_a^3 p_b$	$p_a^2 p_b$	0
(a,b)	(a,b)	$p_a^2 p_b^2$	$p_a p_b^2 + p_a^2 p_b$	$p_a p_b$
(a,a)	(a,a)	p_a^4	p_a^3	p_a^2
Prior Probability		¼	½	¼

Worked Example



$$p_1 = 0.5$$

$$P(G \mid IBD=0) = p_1^4 = 1/16$$

$$P(G \mid IBD=1) = p_1^3 = 1/8$$

$$P(G \mid IBD=2) = p_1^2 = 1/4$$

$$P(G) = 1/4\, p_1^4 + 1/2\, p_1^3 + 1/4\, p_1^2 = 9/64$$

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

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

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

The Recombination Process

- The recombination fraction θ is a measure of distance between two loci
  - Probability that different alleles from different grand-parents are inherited at some locus

- It implies the probability of change in IBD state for a pair of chromosomes in siblings:

$$\psi = (1-\theta)^2 + \theta^2$$

Transition Matrix for IBD States

- Allows calculation of IBD probabilities at arbitrary location conditional on linked marker
  - Depends on recombination fraction θ

		Conditional IBD Probabilities at distance θ		
		0	1	2
Known IBD State	0	$(1-\psi)^2$	$2\psi(1-\psi)$	ψ^2
	1	$\psi(1-\psi)$	$(1-\psi)^2+\psi^2$	$\psi(1-\psi)$
	2	ψ^2	$2\psi(1-\psi)$	$(1-\psi)^2$

$$\psi = (1-\theta)^2 + \theta^2$$

Moving along chromosome

- Input
  - Vector v of IBD probabilities at location A
  - Matrix T of transition probabilities A→B
- Output
  - Vector v' of probabilities at location B
    - Conditional on probabilities at location A
- For k IBD states, requires k^2 operations

$$L(\mathbf{v}'_i \mid \mathbf{v}) = \sum_j L(\mathbf{v}_j) T(\mathbf{v}_i \rightarrow \mathbf{v}'_j, \theta)$$

Combining Information From Multiple Markers

$P(G_1 \mid IBD_1 = 0)$	$P(G_1 \mid IBD_1 = 1)$	$P(G_1 \mid IBD_2 = 2)$
	* **T**	
$P(G_1 \mid IBD_2 = 0, \theta_{1,2})$	$P(G_1 \mid IBD_2 = 1, \theta_{1,2})$	$P(G_1 \mid IBD_2 = 2, \theta_{1,2})$
	o	
$P(G_2 \mid IBD_2 = 0)$	$P(G_2 \mid IBD_2 = 1)$	$P(G_2 \mid IBD_2 = 2)$
	=	
$P(G_1,G_2 \mid IBD_2 = 0, \theta_{1,2})$	$P(G_1,G_2 \mid IBD_2 = 2, \theta_{1,2})$	$P(G_1,G_2 \mid IBD_2 = 2, \theta_{1,2})$

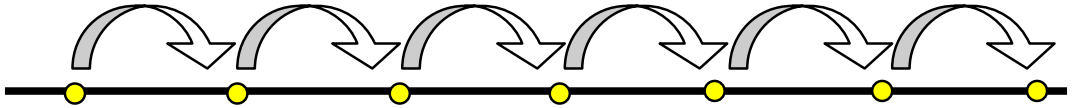
Baum Algorithm

- Markov Model for IBD
  - Vectors $\mathbf{v}_\ell$ of probabilities at each location
  - Transition matrix $\mathbf{T}$ between locations
- Key equations…
  - $\mathbf{v}_{\ell|1..\ell} = \mathbf{v}_{\ell-1|1..\ell-1}\,\mathbf{T} \circ \mathbf{v}_\ell$
  - $\mathbf{v}_{\ell|\ell..m} = \mathbf{v}_{\ell+1|\ell+1..m}\,\mathbf{T} \circ \mathbf{v}_\ell$
  - $\mathbf{v}_{\ell|1..m} = (\mathbf{v}_{1..\ell-1}\,\mathbf{T}) \circ \mathbf{v}_\ell \circ (\mathbf{v}_{\ell+1..1}\,\mathbf{T})$

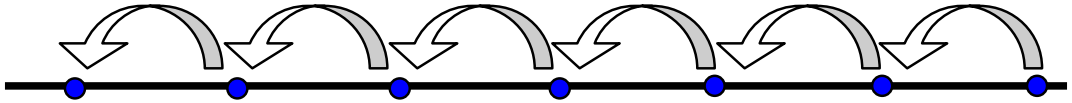
Pictorial Representation

- Single Marker

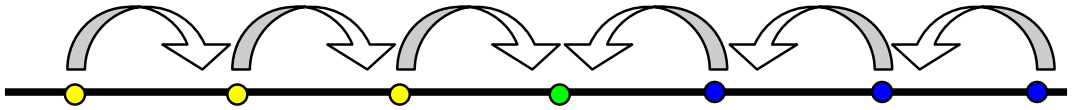
- Left Conditional



- Right Conditional



- Full Likelihood



Complexity of the Problem in Larger Pedigrees

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

Elston-Stewart Algorithm

- Factorize likelihood by individual
  - Each step assigns phase
    - for all markers
    - for one individual
  - Complexity $\propto n \cdot e^m$
- Small number of markers
- Large pedigrees
  - With little inbreeding

Lander-Green Algorithm

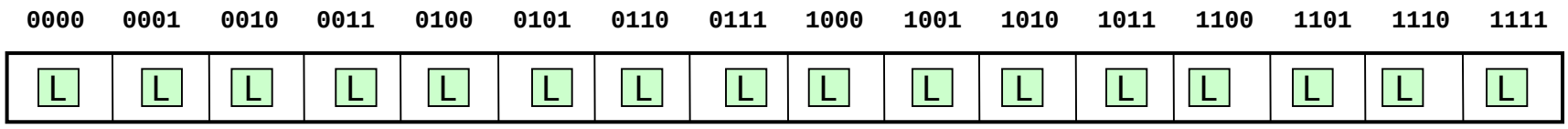
- Factorize likelihood by marker
  - Each step assigns phase
    - For one marker
    - For all individuals in the pedigree
  - Complexity $\propto m \cdot e^n$
- Strengths
  - Large number of markers
  - Relatively small pedigrees
- Natural extension of Baum algorithm

Other methods

- Number of MCMC methods proposed
  - Simulated annealing, Gibbs sampling
  - ~Linear on # markers
  - ~Linear on # people
- Hard to guarantee convergence on very large datasets
  - Many widely separated local minima

Lander-Green inheritance vector

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



Lander-Green Markov Model

- Transition matrix $\mathbf{T}^{\otimes 2n}$

$$\mathbf{T} = \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}$$

- $\mathbf{v}_{\ell|1..\ell} = \mathbf{v}_{\ell-1|1..\ell-1}\, \mathbf{T}^{\otimes 2n} \circ \mathbf{v}_{\ell}$
- $\mathbf{v}_{\ell|\ell..m} = \mathbf{v}_{\ell+1|\ell+1..m}\, \mathbf{T}^{\otimes 2n} \circ \mathbf{v}_{\ell}$
- $\mathbf{v}_{\ell|1..m} = (\mathbf{v}_{1..\ell-1}\, \mathbf{T}^{\otimes 2n}) \circ \mathbf{v}_{\ell} \circ (\mathbf{v}_{\ell+1..1}\, \mathbf{T}^{\otimes 2n})$

MERLIN

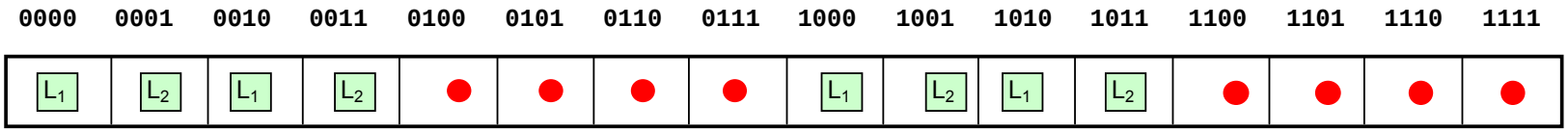
Multipoint Engine for Rapid Likelihood Inference

- Linkage analysis
- Haplotyping
- Error detection
- Simulation
- IBD State Probabilities

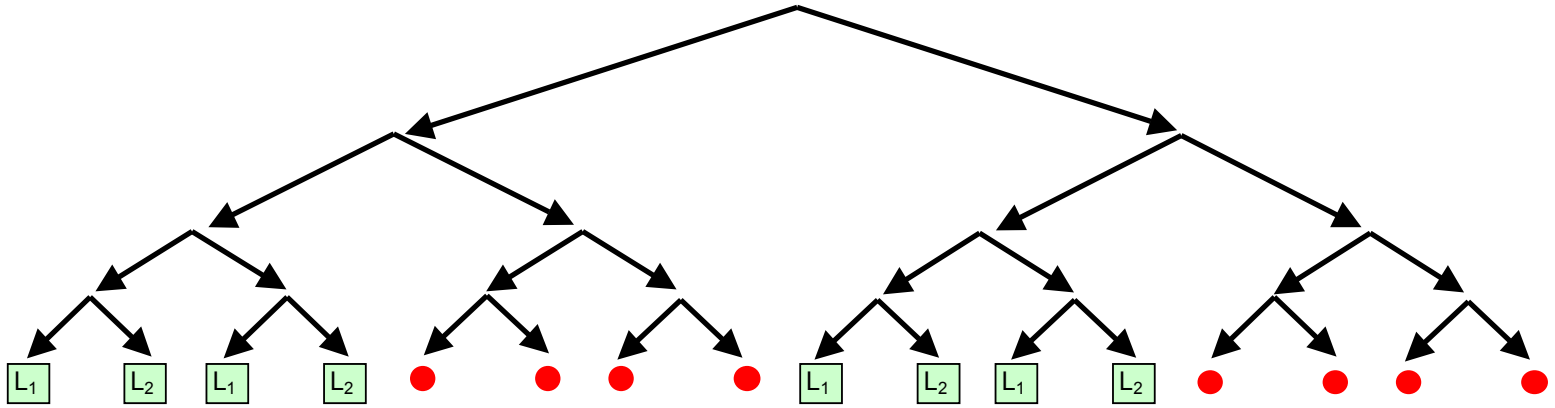
Intuition: $\mathbf{v}_\ell$ has low complexity

- Likelihoods for each element depend on:
  - Is it consistent with observed genotypes?
    - If not, likelihood is zero
  - What founder alleles are compatible?
    - Product of allele frequencies for possible founder alleles
- In practice, much fewer than 2^{2n} outcomes
  - Most elements are zero
  - Number of distinct values is small

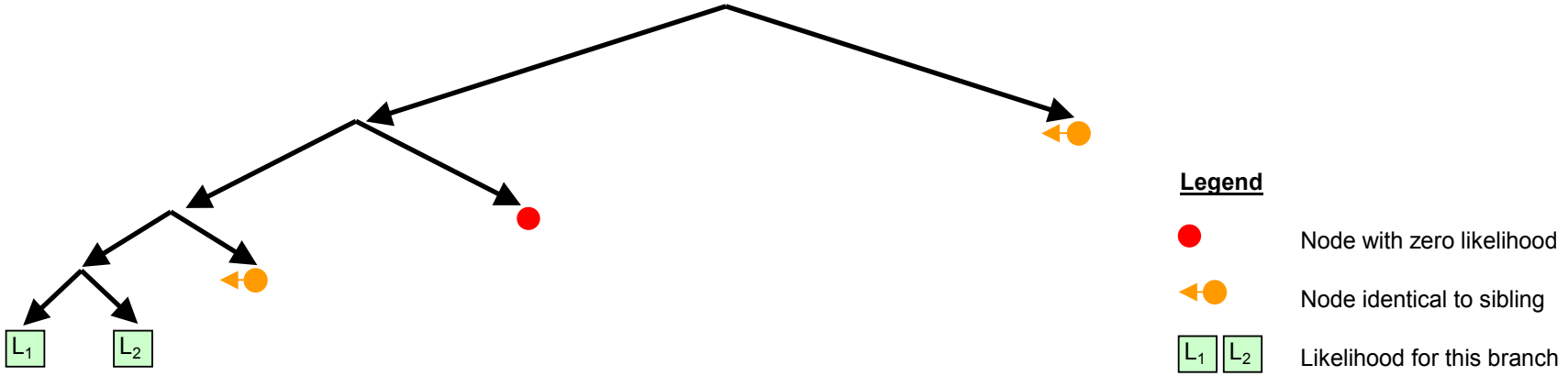
a) **bit-indexed array**



b) **packed tree**



c) **sparse tree**



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

Tree Complexity: Microsatellite

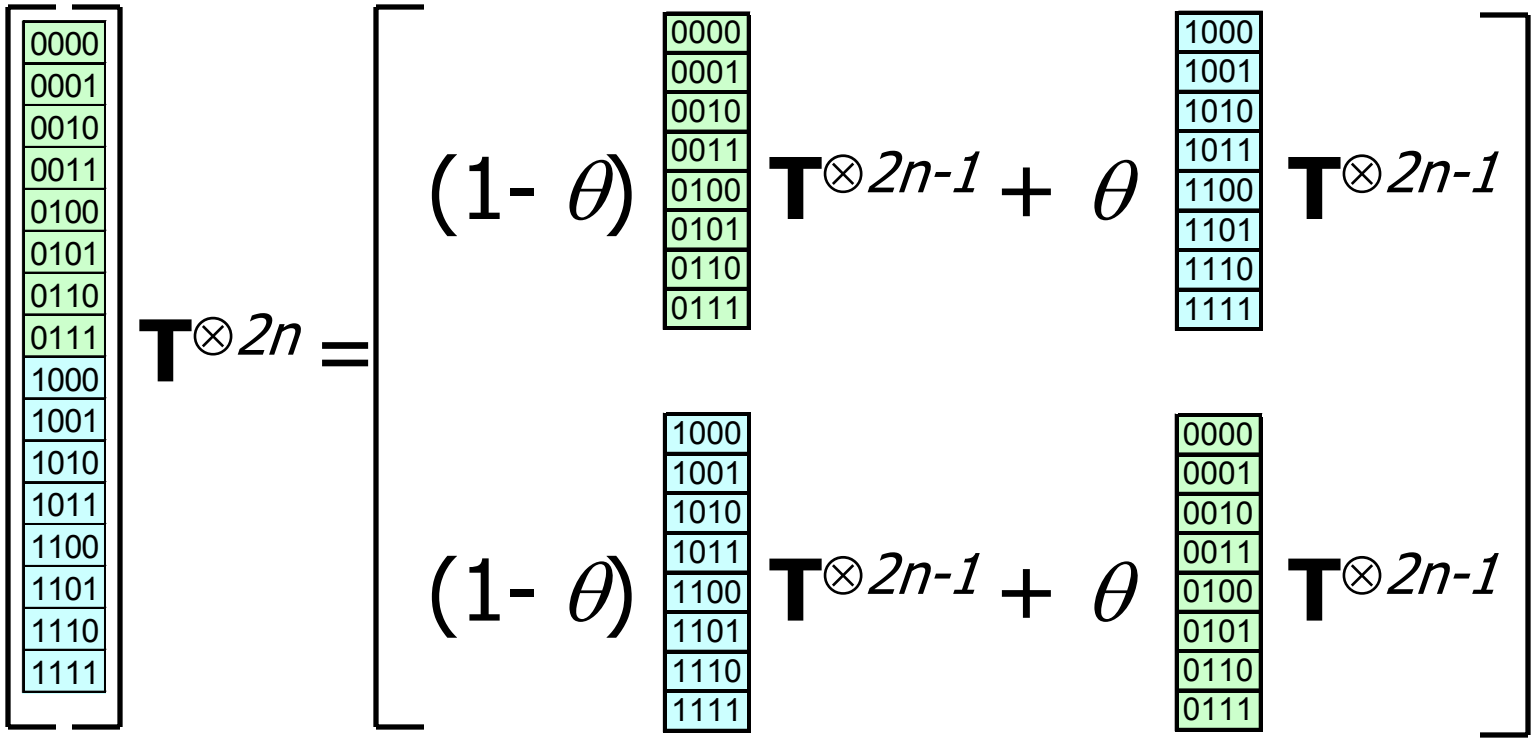
Missing Genotypes	Info	Total Nodes			Leaf Nodes
		Mean	Median	95% C.I.	
4-allele marker with equifrequent alleles					
-	0.72	154.7	72	64 – 603	5.2
5%	0.68	245.2	122	64 – 1166	9.9
10%	0.64	446.3	171	65 – 2429	24.1
20%	0.55	1747.4	405	69 – 15943	107.3
50%	0.28	19880.6	2882	154 –140215	2574.5

(Simulated pedigree with 28 individuals, 40 meioses, requiring 2^{32} = ~4 billion likelihood evaluations using conventional schemes)

Intuition: Trees speedup convolution

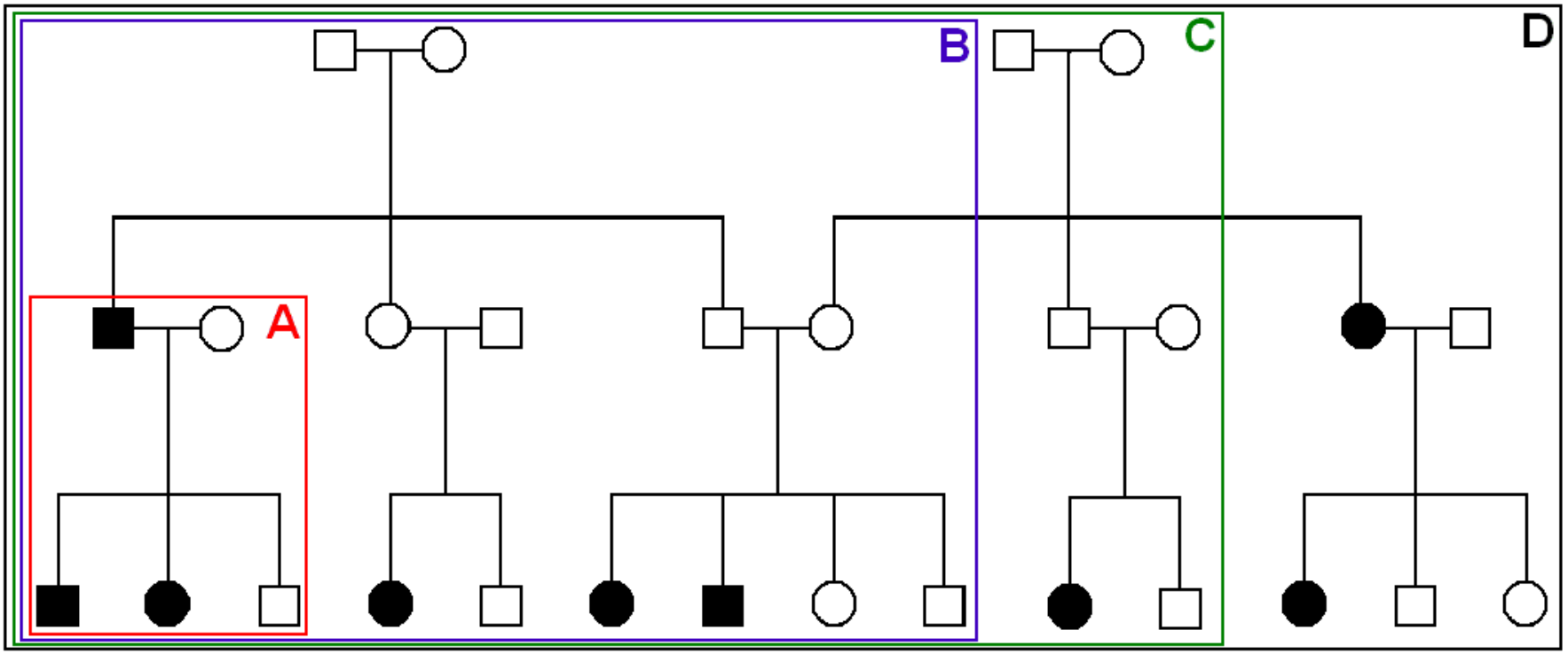
- Trees summarize redundant information
  - Portions of vector that are repeated
  - Portions of vector that are constant or zero
- Speeding up convolution
  - Use sparse-matrix by vector multiplication
  - Use symmetries in divide and conquer algorithm

Elston-Idury Algorithm



Uses divide-and-conquer to carry out matrix-vector multiplication in $O(N \log N)$ operations, instead of $O(N^2)$

Test Case Pedigrees



Timings – Marker Locations

	Top Generation Genotyped			
	A (x1000)	B	C	D
Genehunter	38s	37s	18m16s	*
Allegro	18s	2m17s	3h54m13s	*
Merlin	11s	18s	13m55s	*

	Top Generation Not Genotyped			
	A (x1000)	B	C	D
Genehunter	45s	1m54s	*	*
Allegro	18s	1m08s	1h12m38s	*
Merlin	13s	25s	15m50s	*

Intuition: Approximate Sparse **T**

- Dense maps, closely spaced markers
- Small recombination fractions θ
- Reasonable to set θ^k with zero
  - Produces a very sparse transition matrix
- Consider only elements of **v** separated by $<k$ recombination events
  - At consecutive locations

Additional Speedup…

	Time	Memory
Exact	40s	100 MB
No recombination	<1s	4 MB
≤1 recombinant	2s	17 MB
≤2 recombinants	15s	54 MB
Genehunter 2.1	16min	1024MB

Keavney et al (1998) ACE data, 10 SNPs within gene,
4-18 individuals per family

Capabilities

- Linkage Analysis
  - QTL
  - Variance Components
- Haplotypes
  - Most likely
  - Sampling
  - All
- Error Detection
  - Most SNP typing errors are Mendelian consistent
- Recombination
  - No. of recombinants per family per interval can be controlled
- Others: pairwise and larger IBD sets, info content, …

MERLIN Website

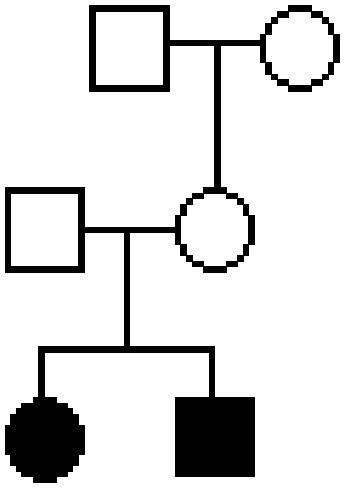
www.sph.umich.edu/csg/abecasis/Merlin

- 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

Describing Relationships



FAMILY	PERSON	FATHER	MOTHER	SEX
example	granpa	unknown	unknown	m
example	granny	unknown	unknown	f
example	father	unknown	unknown	m
example	mother	granny	granpa	f
example	sister	mother	father	f
example	brother	mother	father	m

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	Zigosity.
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>
```

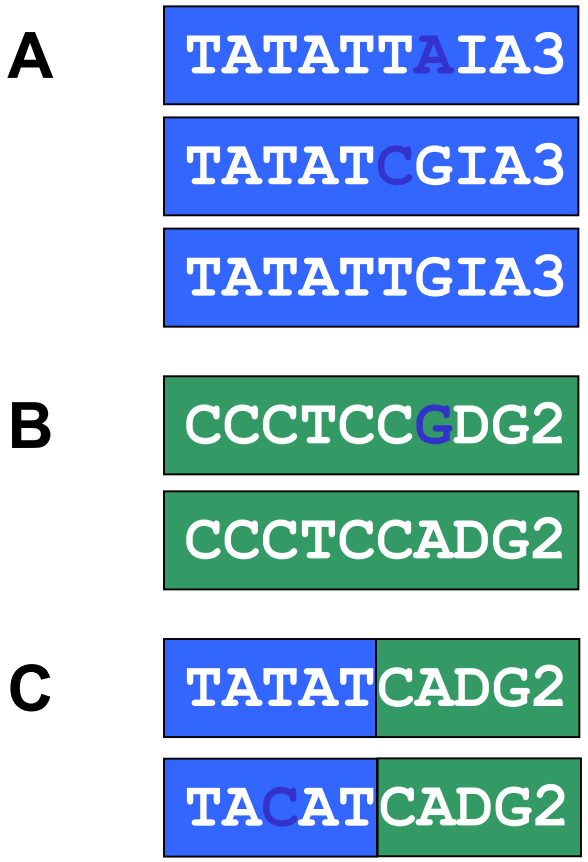
Indicates location of individual markers, necessary to derive recombination fractions between them

Example Data Set: Angiotensin-1

- British population
- Circulating ACE levels
  - Normalized separately for males / females
- 10 di-allelic polymorphisms
  - 26 kb
  - Common
  - In strong linkage disequilibrium
- Keavney et al, HMG, 1998

Haplotype Analysis

- 3 clades
  - All common haplotypes
  - >90% of all haplotypes
- "B" = "C"
  - Equal phenotypic effect
  - Functional variant on right
- Keavney et al (1998)



Objectives of Exercise

- Verify contents of input files
- Calculate IBD information using Merlin
- Time permitting, conduct simple linkage analysis

Things to think about…

- Allele Sharing Among Large Sets
  - The basis of non-parametric linkage statistics
- Parental Sex Specific Allele Sharing
  - Explore the effect of imprinting
- Effect of genotyping error
  - Errors in genotype data lead to erroneous IBD