



# Haplotype analysis

*Shaun Purcell*

`spurcell@pngu.mgh.harvard.edu`

*MGH, Boston*



---

# Overview

What are haplotypes?

*Recombination and linkage disequilibrium*

How do we measure haplotypes?

*Estimating haplotype phase and frequency*

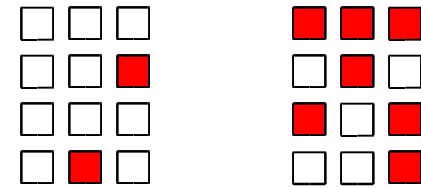
How can we use haplotypes to map causal variants?

*Haplotype-based association analysis*

# What is association?

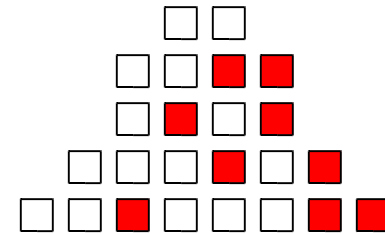
Categorical traits

disease susceptibility genes

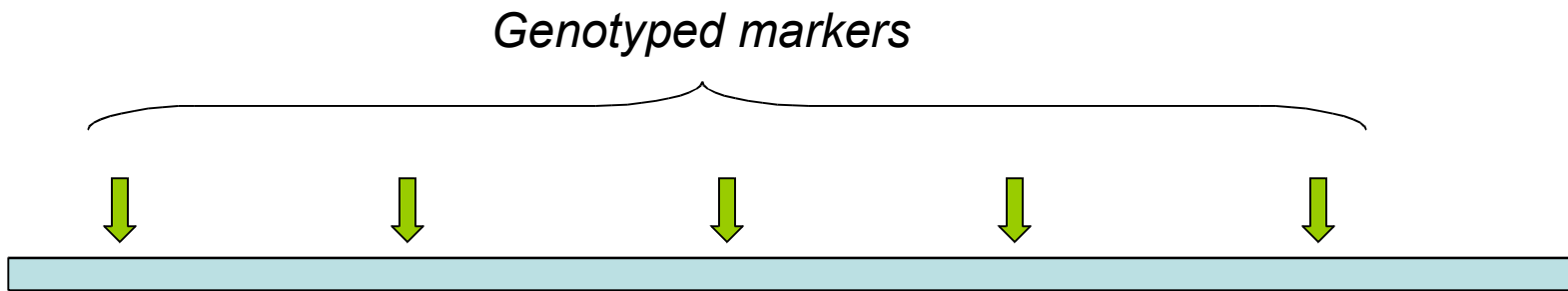


Continuous traits

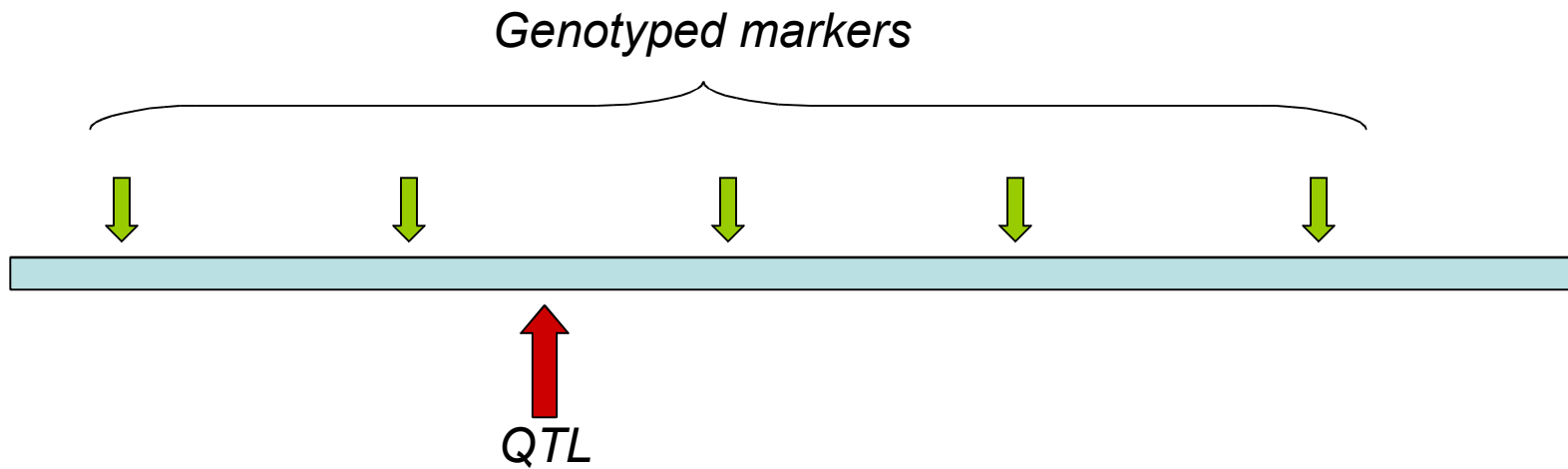
quantitative trait loci, QTL



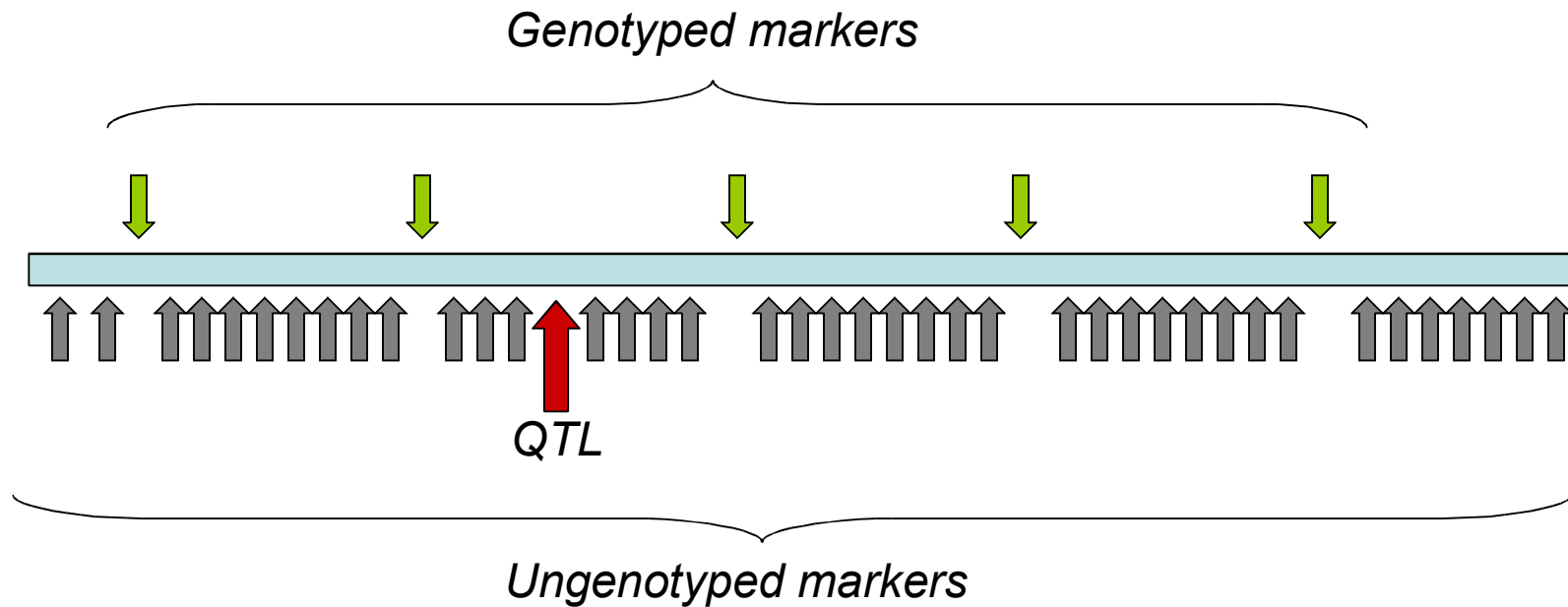
# Linkage disequilibrium mapping



# Linkage disequilibrium mapping



# Linkage disequilibrium mapping

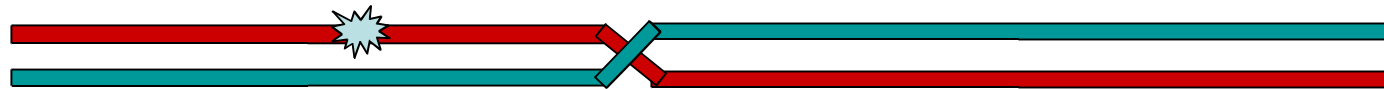


# Recombination

*Homologous chromosomes in one parent*



↓ *Recombination event  
during meiosis*



↓ *Recombinant gamete transmitted,  
harboring mutation*



*Homologous chromosomes in one parent*



↓ *No recombination event  
during meiosis*



↓ *Nonrecombinant gamete transmitted,  
not harboring mutation*





# Linkage: affected sib pairs

Paternal chromosome   
Maternal chromosome 

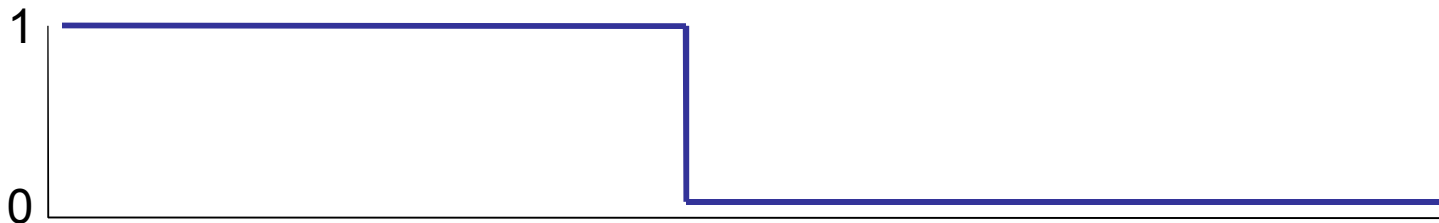
*First affected offspring,  
no recombination*



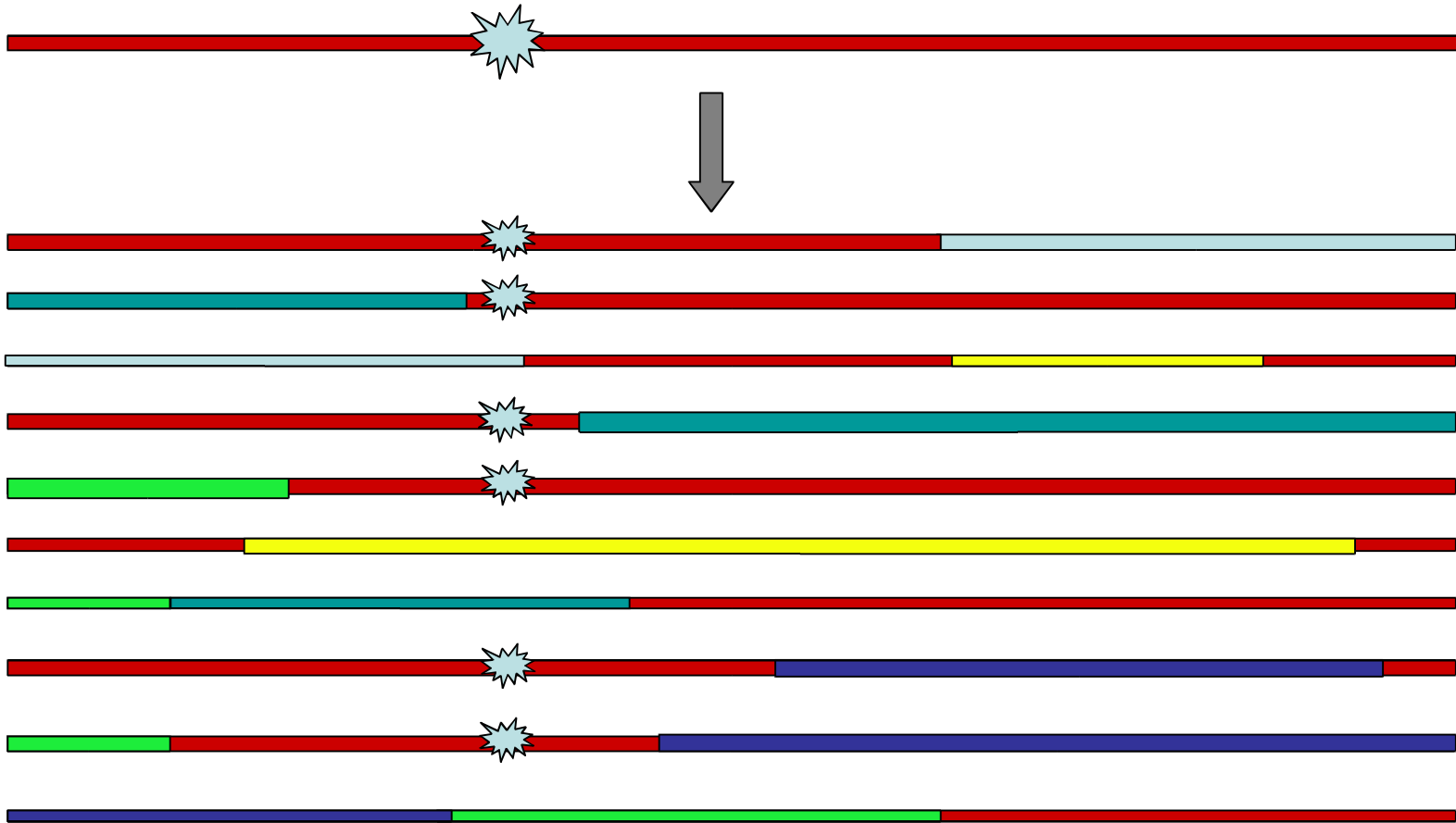
*Second affected offspring,  
recombinant gamete*



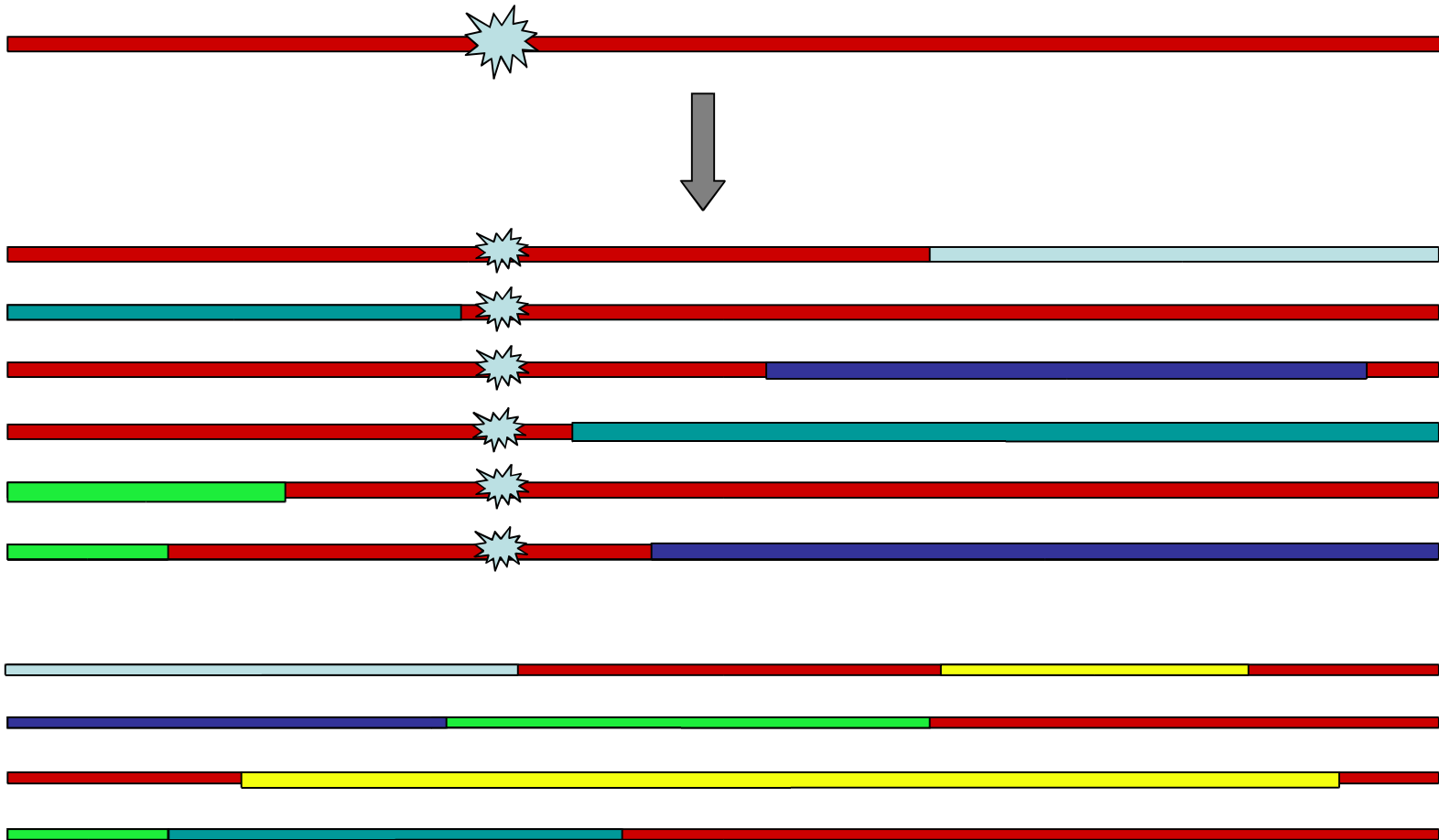
*IBD sharing from this one parent (0 or 1)*



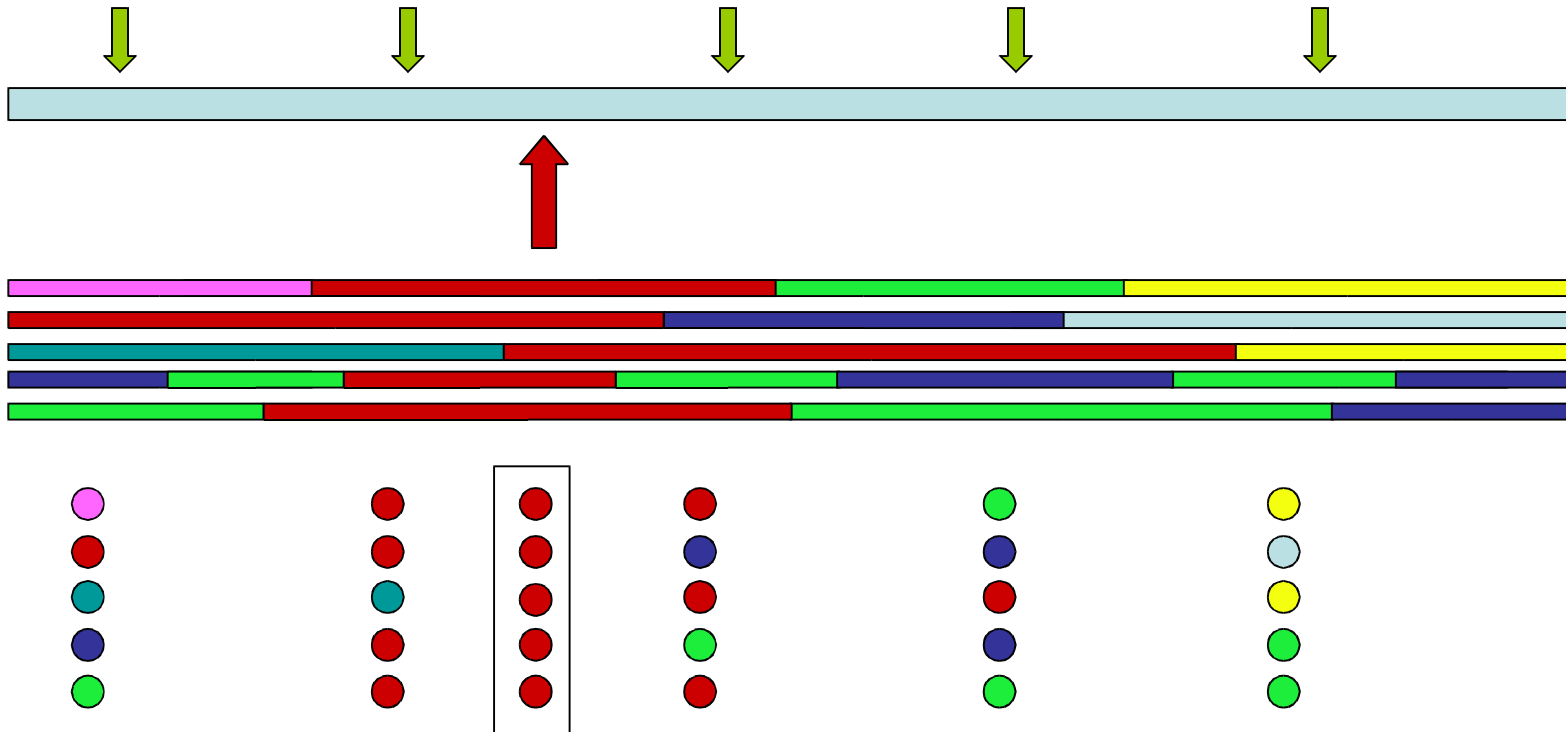
# Mutation occurs on a 'red' chromosome



# Mutation occurs on a 'red' chromosome

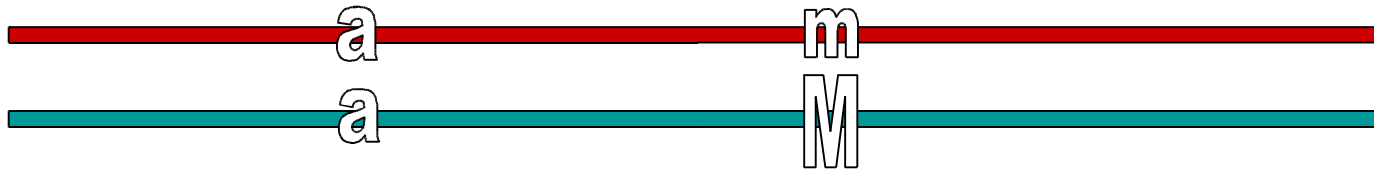


# Association due to 'linkage disequilibrium'



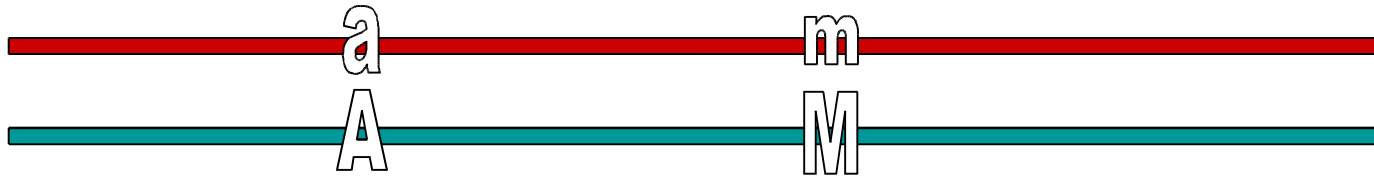
# Haplotypes

|          | <i>A</i> | <i>a</i>  |
|----------|----------|-----------|
| <i>M</i> |          | <i>aM</i> |
| <i>m</i> |          | <i>am</i> |



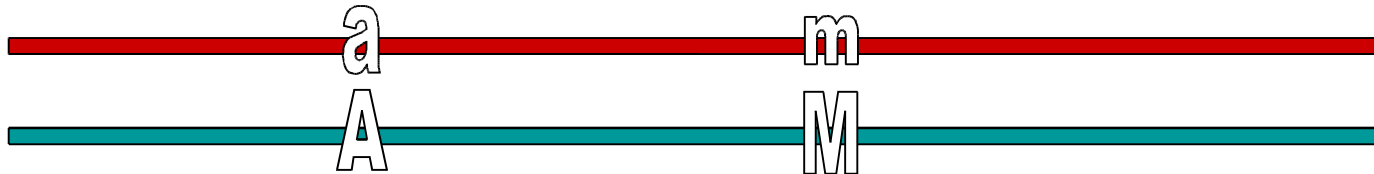
This individual has *aa* and *Mm* genotypes  
and *am* and *aM* haplotypes

*M*      *A*      *a*  
*m*      *AM*      *aM*  
         *m*      *am*



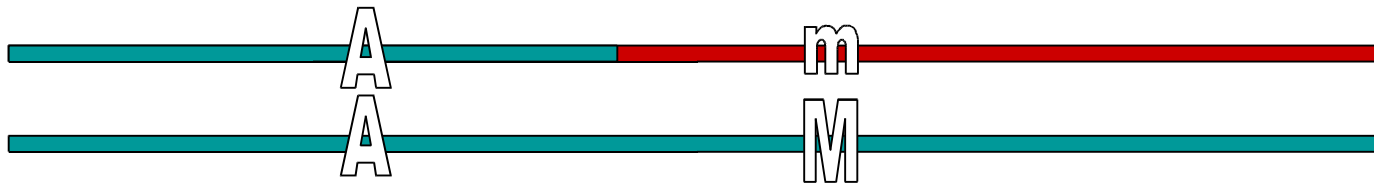
This individual has *Aa* and *Mm* genotype  
and *AM* and *am* haplotypes...

|          |           |           |
|----------|-----------|-----------|
|          | <i>A</i>  | <i>a</i>  |
| <i>M</i> | <i>AM</i> | <i>aM</i> |
| <i>m</i> |           | <i>am</i> |



This individual has *Aa* and *Mm* genotype  
 and *AM* and *am* haplotypes...  
but given only genotype data,  
 consistent with *Am/aM* as well as *AM/am*

|          |           |           |
|----------|-----------|-----------|
|          | <i>A</i>  | <i>a</i>  |
| <i>M</i> | <i>AM</i> | <i>aM</i> |
| <i>m</i> | <i>Am</i> | <i>am</i> |



This individual has ***AA*** and ***Mm*** genotypes  
and ***AM*** and ***Am*** haplotypes





# Haplotype analysis

1. Estimate haplotypes from genotypes
2. Associate haplotypes with trait

| <u>Haplotype</u> | <u>Freq.</u> | <u>Odds Ratio</u> |
|------------------|--------------|-------------------|
| AAGG             | 40%          | 1.00*             |
| AAGT             | 30%          | 2.21              |
| CGCG             | 25%          | 1.07              |
| AGCT             | 5%           | 0.92              |

\* baseline, fixed to 1.00



# Measuring haplotypes

Expectation – Maximisation algorithm

Applicable in situations where there are more categories than can be distinguished

i.e. ‘incomplete data problems’

*Complete data = ( Observed data , Missing data )*

*Haplotype data = ( Genotype data , Phase data )*



# Measuring haplotypes

---

Genotypes

**A/A**   **B/b**   **C/c**

Haplotypes

**ABC** / **Abc**

or

**ABc** / **AbC**

Phases



# E-M algorithm

---

1. Guess haplotype frequencies
2. (**E**) Use those frequencies to replace ambiguous genotypes with fractional haplotype counts
3. (**M**) Estimate frequency of each haplotype by counting
4. Repeat (2) and (3) until convergence



# Dataset to be phased

---

4 individuals genotyped for 2 diallelic markers

|     |     |     |
|-----|-----|-----|
| ID1 | A/A | B/B |
| ID2 | A/a | b/b |
| ID3 | A/a | B/b |
| ID4 | a/a | b/b |



# Dataset to be phased

4 individuals genotyped for 2 diallelic markers

|     |     |     |                   |
|-----|-----|-----|-------------------|
| ID1 | A/A | B/B | AB / AB           |
| ID2 | A/a | b/b | Ab / ab           |
| ID3 | A/a | B/b | AB / ab ? Ab / aB |
| ID4 | a/a | b/b | ab / ab           |



# E-step

---

Replace ambiguous  $A/a B/b$  genotype with :

$AB / ab$  :

$Ab / aB$  :



# E-step

$$P_{AB} = 0.25$$

$$P_{aB} = 0.25$$

$$P_{Ab} = 0.25$$

$$P_{ab} = 0.25$$

Replace ambiguous  $A/a$   $B/b$  genotype with :

$$AB / ab : 2 \times P_{AB} \times P_{ab}$$

$$Ab / aB : 2 \times P_{Ab} \times P_{aB}$$





# E-step

$$P_{AB} = 0.25$$

$$P_{aB} = 0.25$$

$$P_{Ab} = 0.25$$

$$P_{ab} = 0.25$$

Replace ambiguous A/a B/b genotype with :

$$\begin{aligned} AB / ab : 2 \times P_{AB} \times P_{ab} &= 2 \times 0.25 \times 0.25 = 0.125 \\ &= 0.125 / (0.125 + 0.125) = 0.50 \end{aligned}$$

$$\begin{aligned} Ab / aB : 2 \times P_{Ab} \times P_{aB} &= 2 \times 0.25 \times 0.25 = 0.125 \\ &= 0.125 / (0.125 + 0.125) = 0.50 \end{aligned}$$



# E-step

---

## Incomplete data

A/A      B/B

A/a      b/b

A/a      B/b

a/a      b/b

## Complete data

AB / AB

Ab / ab

AB / ab

Ab / aB

ab / ab

## Count

1.00

1.00

0.50

0.50

1.00

# M-step

| <u>Incomplete data</u> |     | <u>Complete data</u> | <u>Count</u> |
|------------------------|-----|----------------------|--------------|
| A/A                    | B/B | AB / AB              | 1.00         |
| A/a                    | b/b | Ab / ab              | 1.00         |
| A/a                    | B/b | AB / ab              | 0.50         |
|                        |     | Ab / aB              | 0.50         |
| a/a                    | b/b | ab / ab              | 1.00         |

**Counting AB haplotype** =  $2 \times 1 + 1 \times 0.5 = 2.5$



# M-step

| <u>Incomplete data</u> |     | <u>Complete data</u> | <u>Count</u> |
|------------------------|-----|----------------------|--------------|
| A/A                    | B/B | AB / AB              | 1.00         |
| A/a                    | b/b | Ab / ab              | 1.00         |
| A/a                    | B/b | AB / ab              | 0.50         |
|                        |     | Ab / aB              | 0.50         |
| a/a                    | b/b | ab / ab              | 1.00         |

**Counting aB haplotype** =  $1 \times 0.5 = 0.5$



# M-step

| <u>Incomplete data</u> |     | <u>Complete data</u> | <u>Count</u> |
|------------------------|-----|----------------------|--------------|
| A/A                    | B/B | AB / AB              | 1.00         |
| A/a                    | b/b | Ab / ab              | 1.00         |
| A/a                    | B/b | AB / ab              | 0.50         |
|                        |     | Ab / aB              | 0.50         |
| a/a                    | b/b | ab / ab              | 1.00         |

**Counting Ab haplotype** =  $1 \times 1 + 1 \times 0.5 = 1.5$



# M-step

| <u>Incomplete data</u> |     | <u>Complete data</u> | <u>Count</u> |
|------------------------|-----|----------------------|--------------|
| A/A                    | B/B | AB / AB              | 1.00         |
| A/a                    | b/b | Ab / ab              | 1.00         |
| A/a                    | B/b | AB / ab              | 0.50         |
|                        |     | Ab / aB              | 0.50         |
| a/a                    | b/b | ab / ab              | 1.00         |

**Counting ab haplotype** =  $1 \times 1 + 1 \times 0.5 + 2 \times 1 = 3.5$



# M-step

---

Haplotype counts, frequencies from complete data

|     | <u>Count</u> | <u>Freq</u>   |
|-----|--------------|---------------|
| AB  | 2.5          | 0.3125        |
| aB  | 0.5          | 0.0625        |
| Ab  | 1.5          | 0.1875        |
| ab  | <u>3.5</u>   | <u>0.4375</u> |
| Sum | 8.0          | 1.0000        |

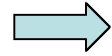
# back to the E-step....

$$P_{AB} = 0.25$$

$$P_{aB} = 0.25$$

$$P_{Ab} = 0.25$$

$$P_{ab} = 0.25$$



are now replaced with  
the updated estimates



$$P_{AB} = 0.3125$$

$$P_{aB} = 0.0625$$

$$P_{Ab} = 0.1875$$

$$P_{ab} = 0.4375$$



# back to the E-step....

$$P_{AB} = 0.25$$

$$P_{aB} = 0.25$$

$$P_{Ab} = 0.25$$

$$P_{ab} = 0.25$$



are now replaced with  
the updated estimates



$$P_{AB} = 0.3125$$

$$P_{aB} = 0.0625$$

$$P_{Ab} = 0.1875$$

$$P_{ab} = 0.4375$$

Replace ambiguous A/a B/b genotype with :

$$\begin{aligned} AB / ab : 2 \times P_{AB} \times P_{ab} &= 2 \times 0.3125 \times 0.4375 = 0.273 \\ &= 0.273 / (0.273 + 0.023) = 0.92 \end{aligned}$$

$$\begin{aligned} Ab / aB : 2 \times P_{Ab} \times P_{aB} &= 2 \times 0.1875 \times 0.0625 = 0.023 \\ &= 0.023 / (0.273 + 0.023) = 0.08 \end{aligned}$$

# back to the M-step...

| <u>Incomplete data</u> |     | <u>Complete data</u> | <u>Count</u> |
|------------------------|-----|----------------------|--------------|
| A/A                    | B/B | AB / AB              | 1.00         |
| A/a                    | b/b | Ab / ab              | 1.00         |
| A/a                    | B/b | AB / ab              | 0.92         |
|                        |     | Ab / aB              | 0.08         |
| a/a                    | b/b | ab / ab              | 1.00         |

**Counting AB haplotype** =  $2 \times 1 + 1 \times 0.92 = 2.92$

# back to the M-step...

| <u>Incomplete data</u> |     | <u>Complete data</u> | <u>Count</u> |
|------------------------|-----|----------------------|--------------|
| A/A                    | B/B | AB / AB              | 1.00         |
| A/a                    | b/b | Ab / ab              | 1.00         |
| A/a                    | B/b | AB / ab              | 0.92         |
|                        |     | Ab / <b>aB</b>       | 0.08         |
| a/a                    | b/b | ab / ab              | 1.00         |

**Counting aB haplotype** =  $1 \times 0.08 = 0.08$

# back to the M-step...

| <u>Incomplete data</u> |     | <u>Complete data</u> | <u>Count</u> |
|------------------------|-----|----------------------|--------------|
| A/A                    | B/B | AB / AB              | 1.00         |
| A/a                    | b/b | Ab / ab              | 1.00         |
| A/a                    | B/b | AB / ab              | 0.92         |
|                        |     | Ab / aB              | 0.08         |
| a/a                    | b/b | ab / ab              | 1.00         |

**Counting Ab haplotype** =  $1 \times 1 + 1 \times 0.08 = 1.08$

# back to the M-step...

| <u>Incomplete data</u> |     | <u>Complete data</u> | <u>Count</u> |
|------------------------|-----|----------------------|--------------|
| A/A                    | B/B | AB / AB              | 1.00         |
| A/a                    | b/b | Ab / ab              | 1.00         |
| A/a                    | B/b | AB / ab              | 0.92         |
|                        |     | Ab / aB              | 0.08         |
| a/a                    | b/b | ab / ab              | 1.00         |

**Counting ab haplotype** =  $1 \times 1 + 1 \times 0.92 + 2 \times 1 = 3.92$



# back to the M-step...

Haplotype counts, frequencies from complete data

|     | <u>Count</u> | <u>Freq</u>  |
|-----|--------------|--------------|
| AA  | 2.92         | 0.365        |
| aB  | 0.08         | 0.010        |
| Ab  | 1.08         | 0.135        |
| ab  | <u>3.92</u>  | <u>0.490</u> |
| Sum | 8.0          | 1.0000       |



---

and back, again, to the E-step...

and back, again, to the M-step...

and back, again, to the E-step...

and back, again, to the M-step...

and back, again, to the E-step...

and back, again, to the M-step...

.....



# Haplotype frequency estimates

|       | AB    | aB     | Ab     | ab     |
|-------|-------|--------|--------|--------|
| $i_0$ | 0.250 | 0.250  | 0.250  | 0.250  |
| $i_1$ | 0.315 | 0.0625 | 0.1875 | 0.4375 |
| $i_2$ | 0.365 | 0.010  | 0.135  | 0.490  |
| ...   | ...   | ...    | ...    | ...    |
| $i_N$ | 0.375 | 0.000  | 0.125  | 0.500  |





# Posterior probabilities

| <u>Genotype</u> |     | <u>Phase</u> | <u>P(H G)</u> |
|-----------------|-----|--------------|---------------|
| A/A             | B/B | AB / AB      | 1.00          |
| A/a             | b/b | Ab / ab      | 1.00          |
| A/a             | B/b | AB / ab      | <b>1.00</b>   |
|                 |     | Ab / aB      | <b>0.00</b>   |
| a/a             | b/b | ab / ab      | 1.00          |

# Missing genotype data

A/A 0/0 c/c consistent with 3 phases

| <u>Phase</u> | <u>P(H G)</u>                             |
|--------------|---|
| ABc / ABc    | $( P_{ABc} \times P_{ABc} ) / S$          |
| ABc / Abc    | $( 2 \times P_{ABc} \times P_{Abc} ) / S$ |
| Abc / Abc    | $( P_{Abc} \times P_{Abc} ) / S$          |

where  $S = P_{ABc} \times P_{ABc} + 2 \times P_{ABc} \times P_{Abc} + P_{Abc} \times P_{Abc}$



---

# Using parental genotypes

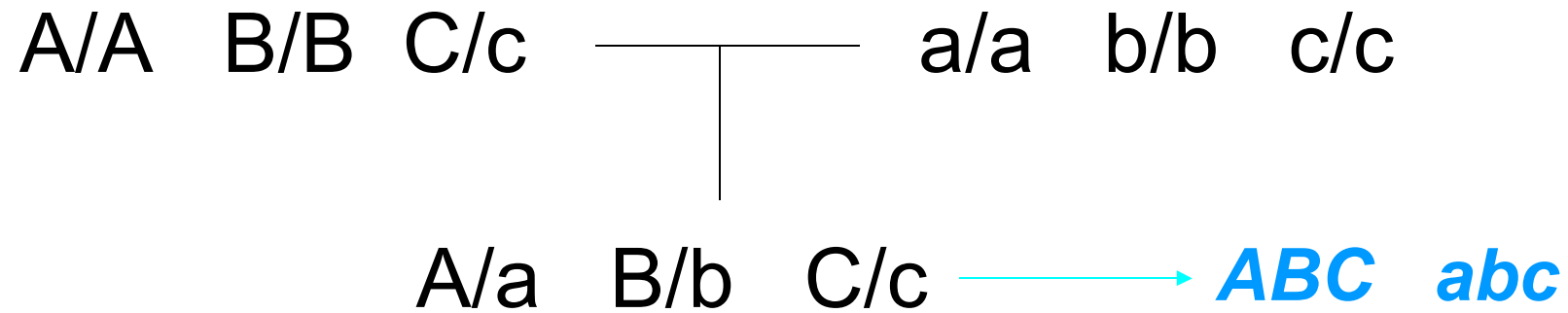
Can often help to resolve phase

$A/a$     $B/b$     $C/c$



# Using parental genotypes

Can often help to resolve phase

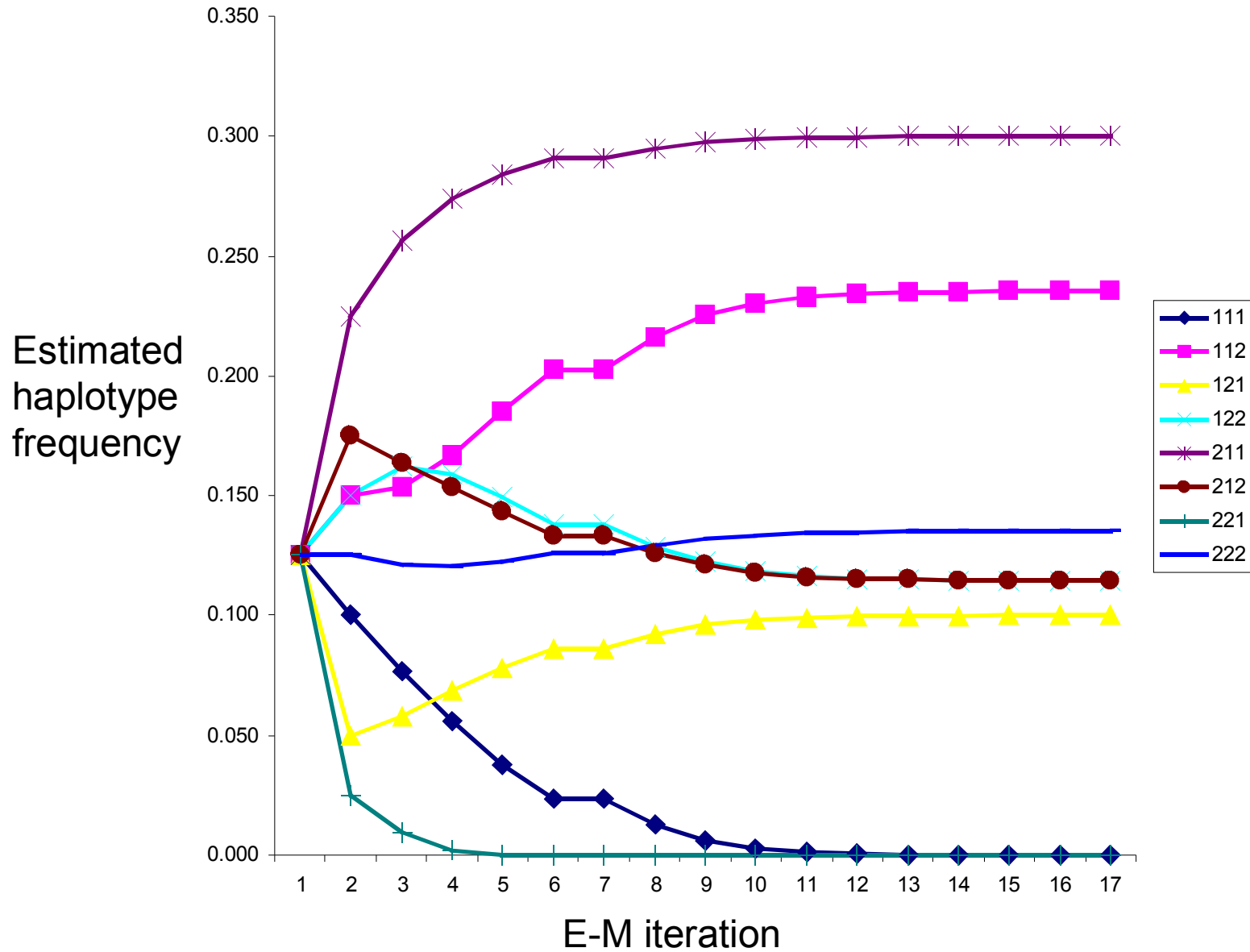




# A (slightly) less trivial example

|    |     |     |     |           |
|----|-----|-----|-----|-----------|
| 1  | 1 1 | 1 2 | 1 2 | ?         |
| 2  | 1 2 | 1 1 | 1 2 | ?         |
| 3  | 2 2 | 1 1 | 1 2 | 211 / 212 |
| 4  | 1 2 | 1 2 | 1 1 | ?         |
| 5  | 1 2 | 1 1 | 1 2 | ?         |
| 6  | 1 1 | 2 2 | 2 2 | 122 / 122 |
| 7  | 1 2 | 1 1 | 2 2 | 112 / 212 |
| 8  | 2 2 | 1 1 | 1 1 | 211 / 211 |
| 9  | 1 2 | 1 2 | 2 2 | ?         |
| 10 | 2 2 | 2 2 | 2 2 | 222 / 222 |

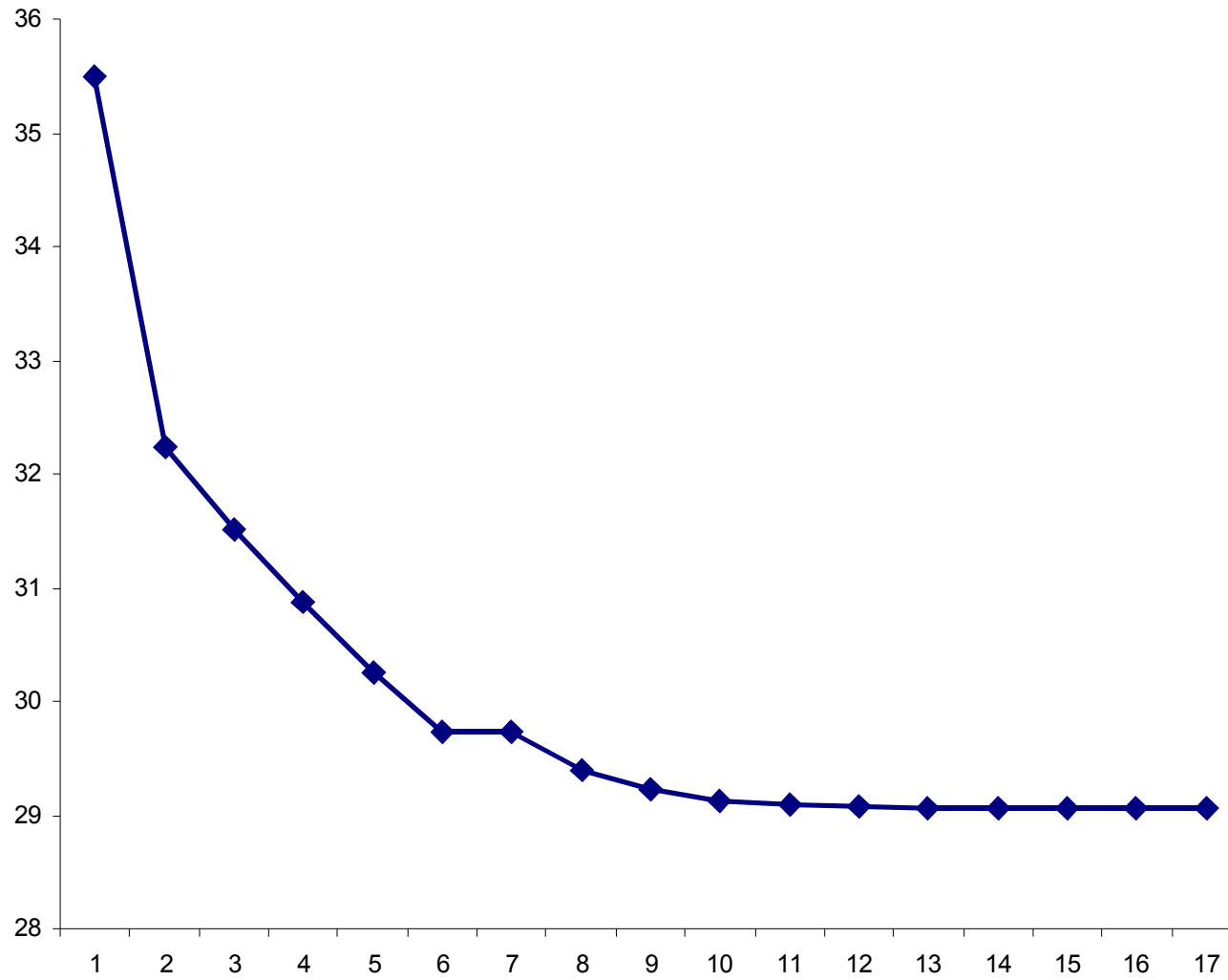
# haplotype frequencies





# log-likelihood

-logLk





# Haplotype frequencies

| <u>H</u> | <u>P(H)</u> |
|----------|-------------|
| 211      | 0.299996    |
| 112      | 0.235391    |
| 222      | 0.135402    |
| 122      | 0.114604    |
| 212      | 0.114602    |
| 121      | 0.099994    |
| 111      | 0.000010    |
| 221      | 0.000000    |



| ID | chr | Hap | P(H G)    |
|----|-----|-----|-----------|
| 1  | 1   | 111 | 0.0001234 |
| 1  | 2   | 122 | 0.0001234 |
| 1  | 1   | 112 | 0.9998766 |
| 1  | 2   | 121 | 0.9998766 |
| 2  | 1   | 111 | 0.0000411 |
| 2  | 2   | 212 | 0.0000411 |
| 2  | 1   | 112 | 0.9999589 |
| 2  | 2   | 211 | 0.9999589 |
| 3  | 1   | 211 | 1.0000000 |
| 3  | 2   | 212 | 1.0000000 |
| 4  | 1   | 111 | 0.0000000 |
| 4  | 2   | 221 | 0.0000000 |
| 4  | 1   | 121 | 1.0000000 |
| 4  | 2   | 211 | 1.0000000 |
| 5  | 1   | 111 | 0.0000411 |
| 5  | 2   | 212 | 0.0000411 |
| 5  | 1   | 112 | 0.9999589 |
| 5  | 2   | 211 | 0.9999589 |

| ID | chr | Hap | P(H G)    |
|----|-----|-----|-----------|
| 6  | 1   | 122 | 1.0000000 |
| 6  | 2   | 122 | 1.0000000 |
| 7  | 1   | 112 | 1.0000000 |
| 7  | 2   | 212 | 1.0000000 |
| 8  | 1   | 211 | 1.0000000 |
| 8  | 2   | 211 | 1.0000000 |
| 9  | 1   | 112 | 0.7080343 |
| 9  | 2   | 222 | 0.7080343 |
| 9  | 1   | 122 | 0.2919657 |
| 9  | 2   | 212 | 0.2919657 |
| 10 | 1   | 222 | 1.0000000 |
| 10 | 2   | 222 | 1.0000000 |

# A (slightly) less trivial example

|    |     |     |     |                       |
|----|-----|-----|-----|-----------------------|
| 1  | 1 1 | 1 2 | 1 2 | 112 / 121             |
| 2  | 1 2 | 1 1 | 1 2 | 112 / 211             |
| 3  | 2 2 | 1 1 | 1 2 | 211 / 212             |
| 4  | 1 2 | 1 2 | 1 1 | 121 / 211             |
| 5  | 1 2 | 1 1 | 1 2 | 112 / 211             |
| 6  | 1 1 | 2 2 | 2 2 | 122 / 122             |
| 7  | 1 2 | 1 1 | 2 2 | 112 / 212             |
| 8  | 2 2 | 1 1 | 1 1 | 211 / 211             |
| 9  | 1 2 | 1 2 | 2 2 | 112 / 222 ? 122 / 212 |
| 10 | 2 2 | 2 2 | 2 2 | 222 / 222             |



# But it's not always this easy...

For  $m$  SNPs there are...

$2^m$  possible haplotypes

$2^{m-1} (2^m + 1)$  possible haplotype pairs

For  $m = 10$  then

**1,024** possible haplotypes

**524, 800** possible haplotype pairs

# Linkage equilibrium

|          | <i>A</i>  | <i>a</i>  |          |
|----------|-----------|-----------|----------|
| <i>M</i> | <i>pr</i> | <i>ps</i> | <i>p</i> |
| <i>m</i> | <i>qr</i> | <i>qs</i> | <i>q</i> |
|          | <i>r</i>  | <i>s</i>  |          |

# Linkage disequilibrium

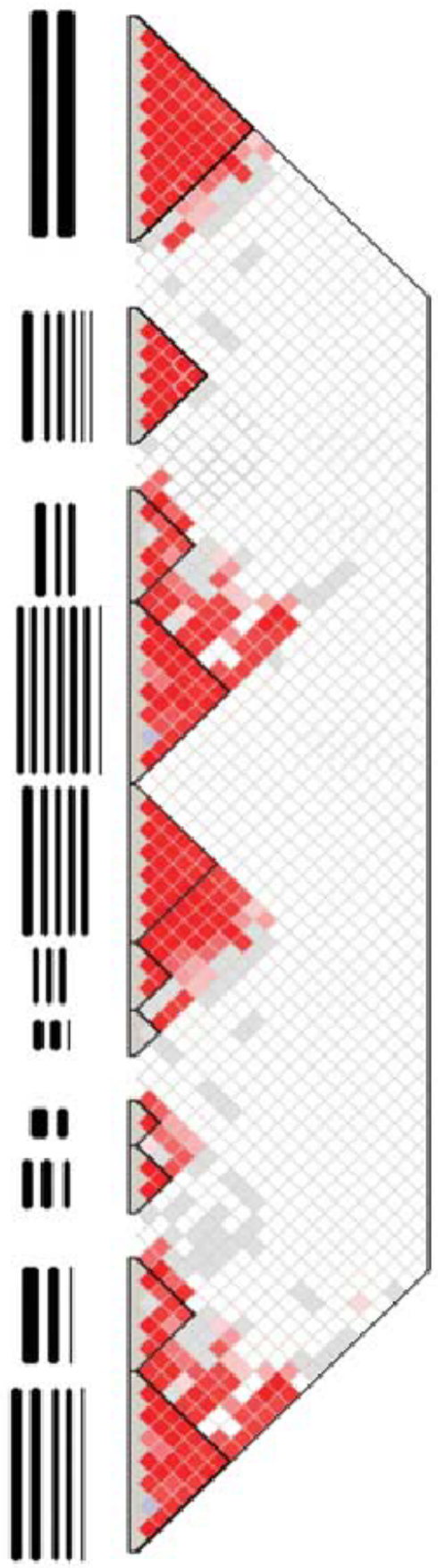
|          | <i>A</i> | <i>a</i> |          |
|----------|----------|----------|----------|
| <i>M</i> | $pr + D$ | $ps - D$ | <i>p</i> |
| <i>m</i> | $qr - D$ | $qs + D$ | <i>q</i> |
|          | <i>r</i> | <i>s</i> |          |

$$D_{MAX} = \text{Min}(qs, pr)$$

$$D' = D / D_{MAX}$$

$$\text{e.g } D = P(AM) - P(A)P(M)$$

$$r^2 = D^2 / pqrs$$







# Practical sessions

- Visualising data and testing for association in `Haploview`
- Detecting haplotype association using `whap`
- Fitting nested model to explore the association using `whap`



# Practical 1 : Haploview

- Folder `F : \pshaun\haplotype\`
- Pedigree format: `data1234.ped`
- Case/control sample (N=200+200)
  
- Load data into Haploview
- Examine LD and block structure
- Examine single SNP association
- Examine haplotype-based association

# Sample files

## dataACGT.ped

```
1_A 1 0 0 1 2 A A C C C A G G C C
2_A 1 0 0 1 2 A A A C C A T G A C
...
1_B 1 0 0 1 1 C C C C C G G A A
2_B 1 0 0 1 1 A C C C A G G C A
...
```

## data1234.ped

```
1_A 1 0 0 1 2 1 1 2 2 2 1 3 3 2 2
...
```

## dataACGT.dat

```
A disease
M snp1
M snp2
M snp3
M snp4
M snp5
```

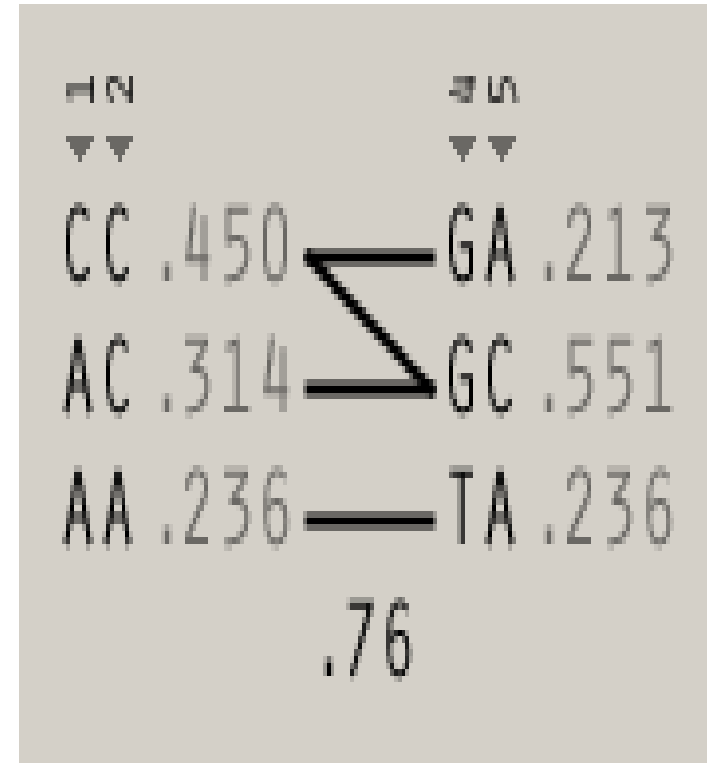
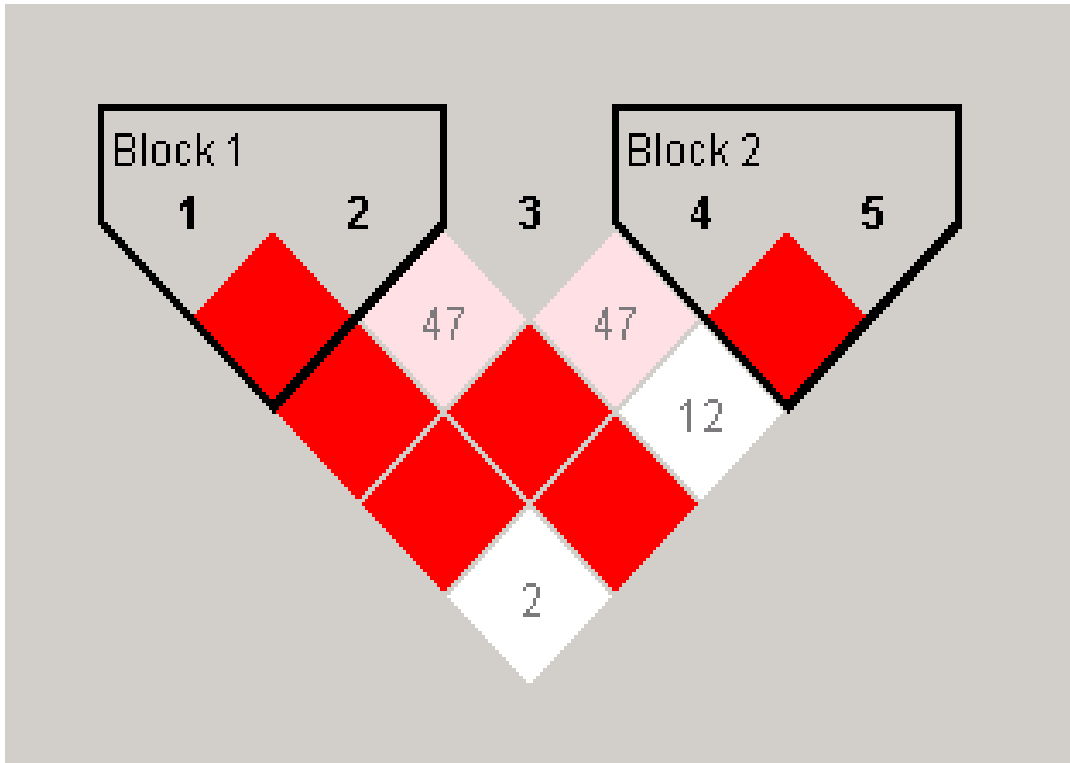
## dataACGT.map

```
1 snp1 0 1
1 snp2 0 2
1 snp3 0 3
1 snp4 0 4
1 snp5 0 5
```

```
pedstats -p data1234.ped -d data1234.dat
```

# LD, block structure

Based on default "Gabriel blocks"



# Single SNP association

| LD Plot Haplotypes Check Markers Association Results |          |               |                      |           |         |  |
|--|----------|---------------|----------------------|-----------|---------|--|
| Single Marker Haplotypes                             |          |               |                      |           |         |  |
| #  | Name     | Major Alleles | Case, Control Ratios | Chi Sq... | p value |  |
| 1  | Marker 1 | A, A          | 0.548, 0.552         | 0.02      | 0.8875  |  |
| 2  | Marker 2 | C, C          | 0.722, 0.805         | 7.544     | 0.0060  |  |
| 3  | Marker 3 | C, C          | 0.605, 0.530         | 4.584     | 0.0323  |  |
| 4  | Marker 4 | G, G          | 0.722, 0.805         | 7.544     | 0.0060  |  |
| 5  | Marker 5 | C, C          | 0.518, 0.585         | 3.684     | 0.0549  |  |

# Block-based haplotype tests

| LD Plot                |       | Haplotypes           |  | Check Markers |         | Association Results |  |
|------------------------|-------|----------------------|--|---------------|---------|---------------------|--|
| Single Marker          |       | Haplotypes           |  |               |         |                     |  |
| Haplotype              | Freq. | Case, Control Ratios |  | Chi Square    | p value |                     |  |
| Haplotype Associations |       |                      |  |               |         |                     |  |
| [-] Block 1            |       |                      |  |               |         |                     |  |
| CC                     | 0.450 | 0.452, 0.448         |  | 0.02          | 0.8875  |                     |  |
| AC                     | 0.314 | 0.270, 0.358         |  | 7.112         | 0.0077  |                     |  |
| AA                     | 0.236 | 0.278, 0.195         |  | 7.544         | 0.0060  |                     |  |
| [-] Block 2            |       |                      |  |               |         |                     |  |
| GC                     | 0.551 | 0.518, 0.585         |  | 3.684         | 0.0549  |                     |  |
| TA                     | 0.236 | 0.278, 0.195         |  | 7.544         | 0.0060  |                     |  |
| GA                     | 0.213 | 0.205, 0.220         |  | 0.269         | 0.6040  |                     |  |



# The true model

General population haplotype frequencies

ACAGC 0.25

CCCGC 0.25

CCCGA 0.20

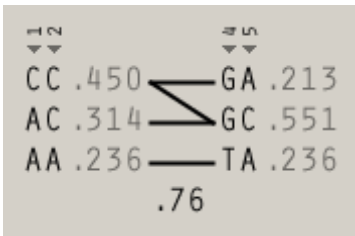
AAATA 0.20

**AACTA 0.05**      *Increases risk for disease*

ACCGC 0.05

*Implied from block definitions*

*True model*



CC      GA  
**AC**      GC  
**AA**      **TA**

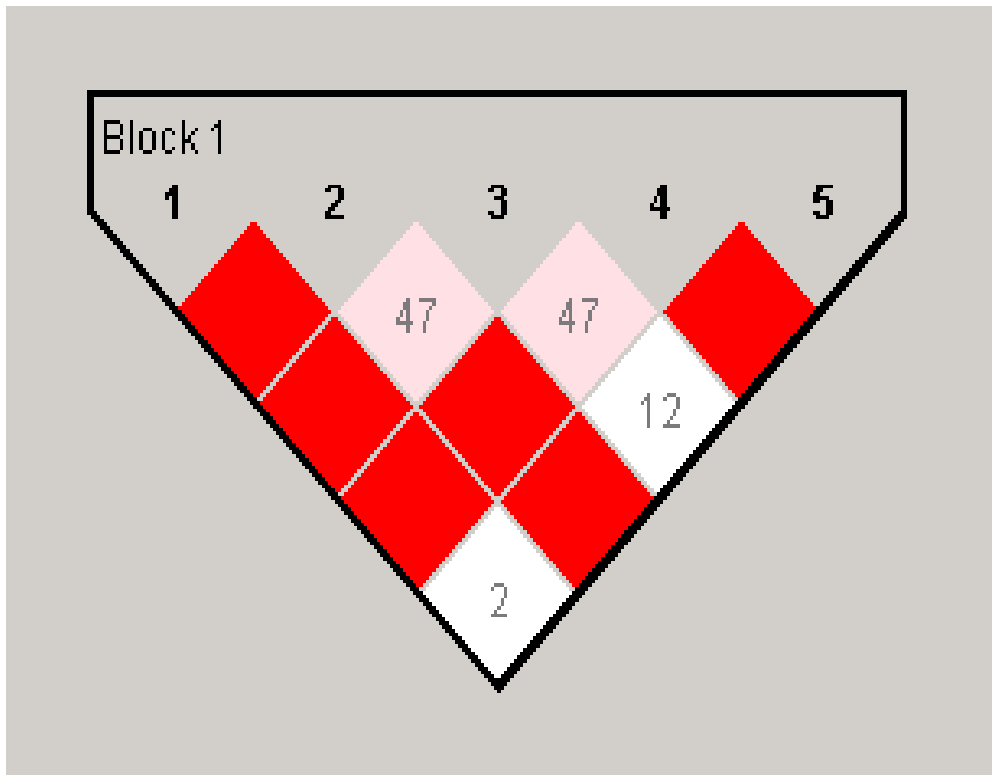
**AA** **A** **TA**  
**AC** **A** GC  
CC **C** GA  
CC **C** GC  
**AA** **C** **TA**  
**AC** **C** GC

AAATA  
ACAGC  
CCCGA  
CCCGC  
**AACTA**  
ACCGC

*Significantly associated with increased risk*  
*Significantly associated with decreased risk*



# Manually specifying the 'block'



|       | 1 | 2 | 3 | 4 | 5    |
|-------|---|---|---|---|------|
| ACAGC |   |   |   |   | .264 |
| CCCGC |   |   |   |   | .237 |
| CCCGA |   |   |   |   | .212 |
| AAATA |   |   |   |   | .169 |
| AACTA |   |   |   |   | .067 |
| ACCGC |   |   |   |   | .050 |

# Results with 5-SNP block

| LD Plot Haplotypes Check Markers Association Results |       |                      |            |         |  |
|--|-------|----------------------|------------|---------|--|
| Single Marker Haplotypes                             |       |                      |            |         |  |
| Haplotype  | Freq. | Case, Control Ratios | Chi Square | p value |  |
| Haplotype Associations                               |       |                      |            |         |  |
| [-] Block 1  |       |                      |            |         |  |
| ACAGC  | 0.264 | 0.218, 0.309         | 8.623      | 0.0033  |  |
| CCCGC  | 0.237 | 0.248, 0.228         | 0.442      | 0.5062  |  |
| CCCGA  | 0.212 | 0.205, 0.220         | 0.269      | 0.6040  |  |
| AAATA  | 0.169 | 0.177, 0.161         | 0.387      | 0.5339  |  |
| AACTA  | 0.067 | 0.100, 0.034         | 13.894     | 0.0002  |  |
| ACCGC  | 0.050 | 0.052, 0.048         | 0.066      | 0.7973  |  |



# whap

- Numerous recent methods using GLM approach
  - Schaid *et al* (02) *AJHG*
  - Zaykin *et al* (02) *Hum Hered*
  - Seltman *et al* (03) *Genet Epi*
- Quantitative and qualitative traits
- Mixture of regressions framework
- Between/within family model
- Model either  $L(X|G)$  or  $L(G|X)$
- Independent secondary test, 1 df
- Flexible specification of nested submodels

# Single locus analysis

- Fulker *et al* (1999)

| $S_1$ | $S_2$ | $S_1$ | $S_2$ | B   | W   | $S_1$ | $S_2$ |
|-------|-------|-------|-------|-----|-----|-------|-------|
| AA    | AA    | 1     | 1     | 1   | 0   | B+W   | B-W   |
| AA    | Aa    | 1     | 0     | 0.5 | 0.5 | B+W   | B-W   |
| AA    | aa    | 1     | -1    | 0   | 1   | B+W   | B-W   |

Note :  $W = S_1 - B$

# Parental genotypes

- Use parental genotypes to generate **B**

- Examples

- AA from AAxAA      $W = 0$

- Aa from AAxAa      $W = -0.5$

- Aa from AaxAa      $W = 0$

| Mat | Pat | B    |
|-----|-----|------|
| 1   | 1   | 1    |
| 1   | 0   | 0.5  |
| 1   | -1  | 0    |
| 0   | 1   | 0.5  |
| 0   | 0   | 0    |
| 0   | -1  | -0.5 |
| -1  | 1   | 0    |
| -1  | 0   | -0.5 |
| -1  | -1  | -1   |

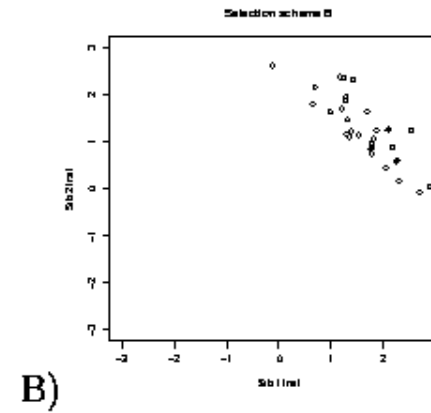
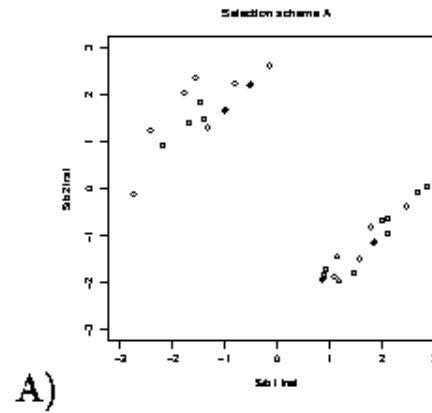
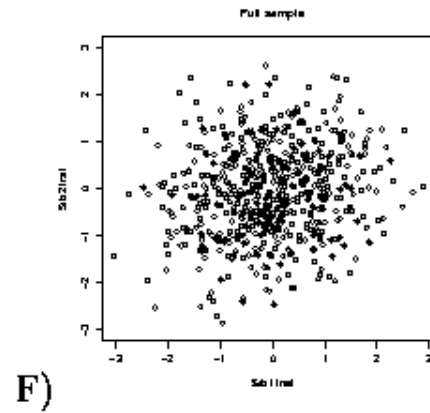
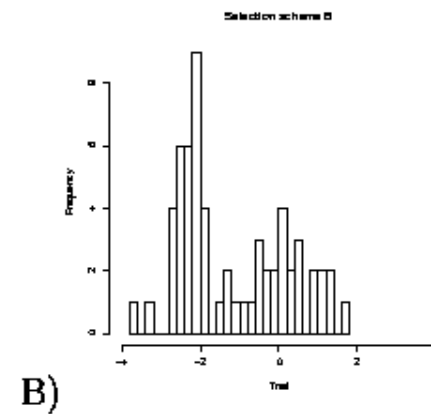
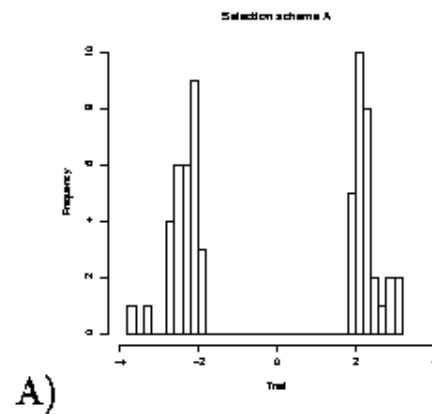
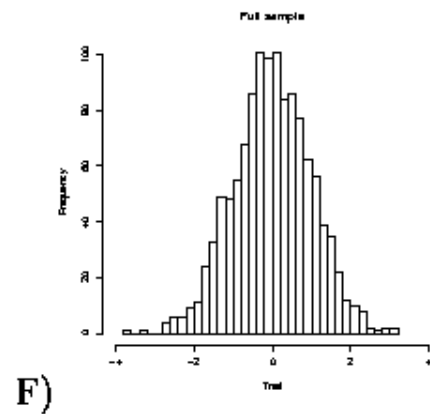


# Available tests

■  $X \sim N( bB + wW , \delta^2 )$

- Basic test
  - $H_A : b = w$
  - $H_0 : b = w = 0$
- Robust test
  - $H_A : b, w$
  - $H_0 : b, w = 0$
- Test for stratification
  - $H_A : b, w$
  - $H_0 : b = w$
- Robust test (2)
  - $H_A : b = 0, w$
  - $H_0 : b = w = 0$

# Analysis of selected samples



# Conditioning on trait values

- Model likelihood of observing genotype conditional on trait value

$$L(G|X) = \frac{L(X|G)L(G)}{\sum L(X|G)L(G)}$$

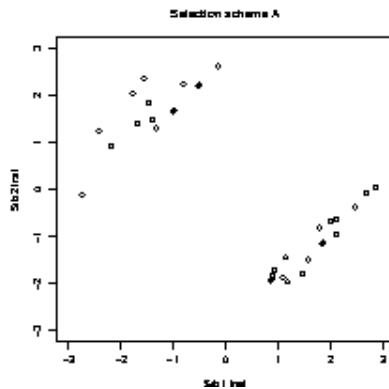
- Singletons:
  - $\mathbf{G} = \{ AA, Aa, aa \}$
- Pairs:
  - $\mathbf{G} = \{ AA/AA, AA/Aa, AA/aa, \dots \}$
- With parents:
  - $\mathbf{G} = \{ AA | AA \times AA, AA | AA \times Aa, \dots \}$
  - $\mathbf{G} = \{ AA/AA | AA \times AA, AA/AA | AA \times Aa, \dots \}$



# Robust in selected samples

- Type I error rates

- Sib pairs
- 10% extreme selection
- Within sibship test



|                        | L(X G) | L(G X) |
|------------------------|--------|--------|
| <b>Full sample</b>     |        |        |
| <i>No parents</i>      | 5.4    | 5.4    |
| <i>Parents</i>         | 5.2    | 5.0    |
| <b>Selected sample</b> |        |        |
| <i>No parents</i>      | 26.7   | 5.3    |
| <i>Parents</i>         | 13.8   | 5.0    |
|                        |        |        |

# Extension to haplotype analysis

- Probabilistic haplotype reconstruction via E-M algorithm

AA BB cc Dd

ABcD / ABcd  $P(P_1) = 1.00$

AA Bb cc Dd

ABcD / Abcd  $P(P_1) = 0.85$

ABcd / AbcD  $P(P_2) = 0.15$

# Weighted likelihood

- Individual  $i$  has  $G$  consistent phases

$$\lambda = \sum_G L(X|G)L(G)$$

Estimated via E-M algorithm





# Quantitative & qualitative traits

- Quantitative traits

$$L(X|G) = \varphi(g_{ip}, s^2)$$

- Qualitative traits

$$L(X|G) = \frac{1}{1 + e^{-g_{ip}}}$$

$$g = B\beta + c$$

- **B** [phase x haplotype] matrix of scores
- $\beta$  [haplotype x 1] vector of regression coefficients
- $c$  is a constant

# Example **B** matrix

Individual  $i$       Genotypes: 1/1    1/1    1/1

                    Haplotypes: 111 / 111     $P() = 1.0$

|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 1 | 2 |
| 1 | 2 | 2 | 2 | 1 | 2 |
| 1 | 1 | 2 | 2 | 2 | 1 |

$$\mathbf{g} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{bmatrix} = \begin{bmatrix} 2b_1 \end{bmatrix} \quad \mathbf{L}(\mathbf{G}=\mathbf{g}) = \begin{bmatrix} 1.0 \end{bmatrix}$$

**B** matrix

Vector of  
regression coefficients

# Example B matrix

Individual  $j$       Genotypes: 1/1    1/2    1/2

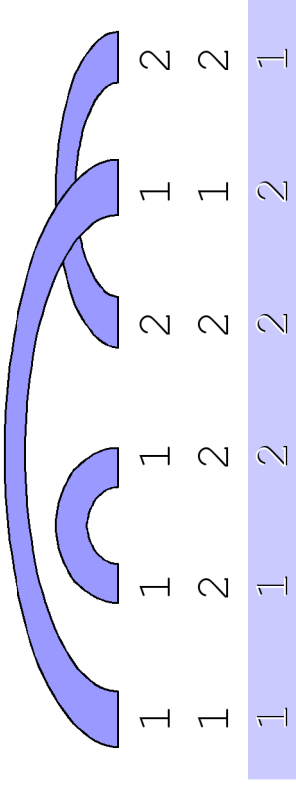
                         Haplotypes: 112 / 121     $P() = 0.8$   
    111 / 122     $P() = 0.2$

|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 1 | 2 |
| 1 | 2 | 2 | 2 | 1 | 2 |
| 1 | 1 | 2 | 2 | 2 | 1 |

$$\mathbf{g} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix} = \begin{pmatrix} b_2 + b_5 \\ b_1 + b_3 \end{pmatrix} \quad \mathbf{L}(\mathbf{G}=\mathbf{g}) = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix}$$

# Testing nested hypotheses

- Test effect of a locus conditional on haplotype background. e.g. drop the 3<sup>rd</sup> locus



$$\mathbf{g} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_2 \\ c_3 \\ c_1 \\ c_3 \end{pmatrix} = \begin{pmatrix} c_2 + c_1 \\ c_1 + c_2 \end{pmatrix}$$

$$\begin{aligned}
 b_1 &= b_5 \\
 b_2 &= b_3 \\
 b_4 &= b_6
 \end{aligned}$$



# Parental genotypes

- Phase parental genotypes via E-M
  - Parental phase  $P(P_{P,M}) = P(P_P) \times P(P_M)$
- For each  $P_{P,M}$  enumerate offspring phases,  $P_C$  consistent with  $G_C$ 
  - Calculate  $P(P_C | P_{P,M})$
  - Can allow for recombination
- Weighted likelihood over all  $P_{P,M}$  and  $P_C$





---

# Between/within partitioning

- **B** matrix depends on parental phase
- **W = G - B**
  
- To calculate **B** for a specific  $P_{P,M}$ 
  - average **all possible**  $P_C$  given  $P_{P,M}$ 
    - i.e. whether or not consistent with  $G_C$

# Between/within partitioning

Individual  $k$                       Genotypes:    1/1    1/2

Parental Genotypes:    1/1    1/1    X    1/2    1/2

Parental Haplotypes:    11 / 11    X    11 / 22  
                                   11 / 11    X    12 / 12

*Consistent with  
offspring genotypes*

*All possible*

Haplotypes | parents 11/11 X 11/22:

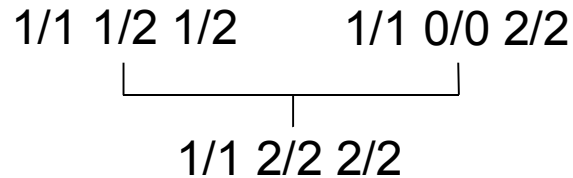
11 / 11  
11 / 22

Haplotypes | parents 11/11 X 12/12:

11 / 12

11 / 12

# Between/within partitioning



Seven haplotypes > 1%  
 212 111 211 112  
 222 122 121

|                      |         | 212     | 111     | 211   | 112   | 222   | 122   | 121     | 212     | 111     | 211    | 112    | 222   | 122   | 121     |         |
|----------------------|---------|---------|---------|-------|-------|-------|-------|---------|---------|---------|--------|--------|-------|-------|---------|---------|
| 122\111 x 112\122 => | 122\122 | [ 0.000 | 0.500   | 0.000 | 0.500 | 0.000 | 1.000 | 0.000 ] | [ 0.000 | -0.500  | 0.000  | -0.500 | 0.000 | 1.000 | 0.000 ] |         |
| 122\111 x 122\112 => | 122\122 | [ 0.000 | 0.500   | 0.000 | 0.500 | 0.000 | 1.000 | 0.000 ] | [ 0.000 | -0.500  | 0.000  | -0.500 | 0.000 | 1.000 | 0.000 ] |         |
| 122\111 x 122\122 => | 122\122 | [ 0.000 | 0.500   | 0.000 | 0.000 | 0.000 | 1.500 | 0.000 ] | [ 0.000 | -0.500  | 0.000  | 0.000  | 0.000 | 0.500 | 0.000 ] |         |
|                      | =>      | 122\122 | [ 0.000 | 0.500 | 0.000 | 0.000 | 0.000 | 1.500   | 0.000 ] | [ 0.000 | -0.500 | 0.000  | 0.000 | 0.000 | 0.500   | 0.000 ] |
| 111\122 x 112\122 => | 122\122 | [ 0.000 | 0.500   | 0.000 | 0.500 | 0.000 | 1.000 | 0.000 ] | [ 0.000 | -0.500  | 0.000  | -0.500 | 0.000 | 1.000 | 0.000 ] |         |
| 111\122 x 122\112 => | 122\122 | [ 0.000 | 0.500   | 0.000 | 0.500 | 0.000 | 1.000 | 0.000 ] | [ 0.000 | -0.500  | 0.000  | -0.500 | 0.000 | 1.000 | 0.000 ] |         |
| 111\122 x 122\122 => | 122\122 | [ 0.000 | 0.500   | 0.000 | 0.000 | 0.000 | 1.500 | 0.000 ] | [ 0.000 | -0.500  | 0.000  | 0.000  | 0.000 | 0.500 | 0.000 ] |         |
|                      | =>      | 122\122 | [ 0.000 | 0.500 | 0.000 | 0.000 | 0.000 | 1.500   | 0.000 ] | [ 0.000 | -0.500 | 0.000  | 0.000 | 0.000 | 0.500   | 0.000 ] |

**B**

**W**

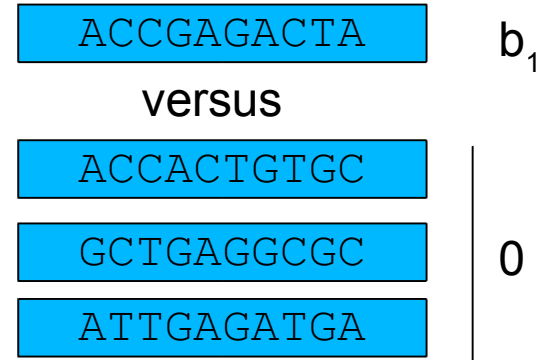
Offspring matrix

[ 0.000 0.000 0.000 0.000 0.000 2.000 0.000 ]

# Two main types of test

## ■ Haplotype-specific tests

- $H$  tests each with 1 df
- compare each haplotype versus all others
- correction for multiple tests not built-in



## ■ Omnibus test

- single test with  $H-1$  df
- compare each haplotype against an (arbitrary) reference haplotype
- built-in correction for multiple tests



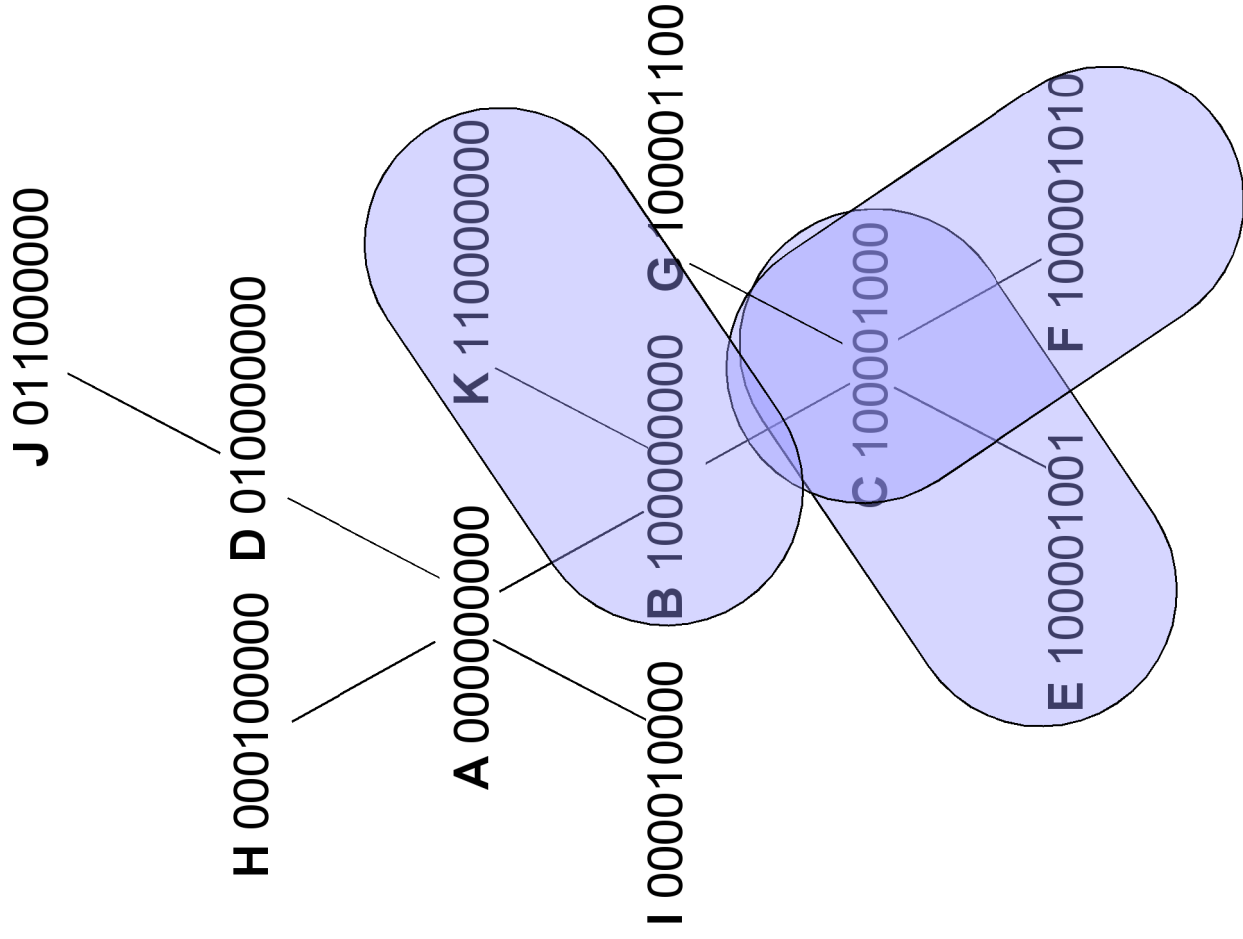


---

# Secondary analysis

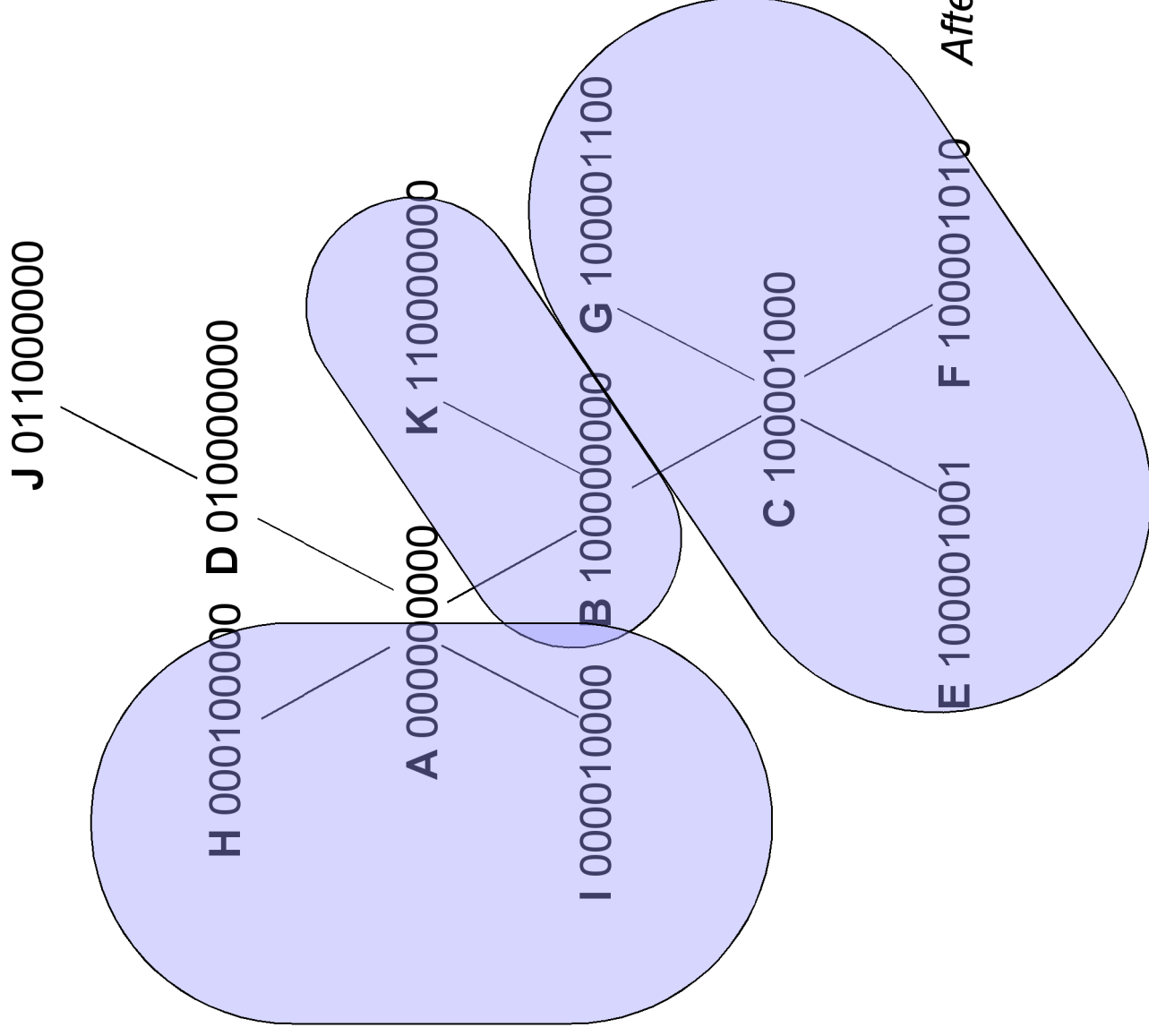
- $H$  haplotypes will have  $H-1$  coefficients
  - Reduces power of test – high degrees of freedom
- More similar haplotypes should have more similar effects

# Cladogram-collapsing



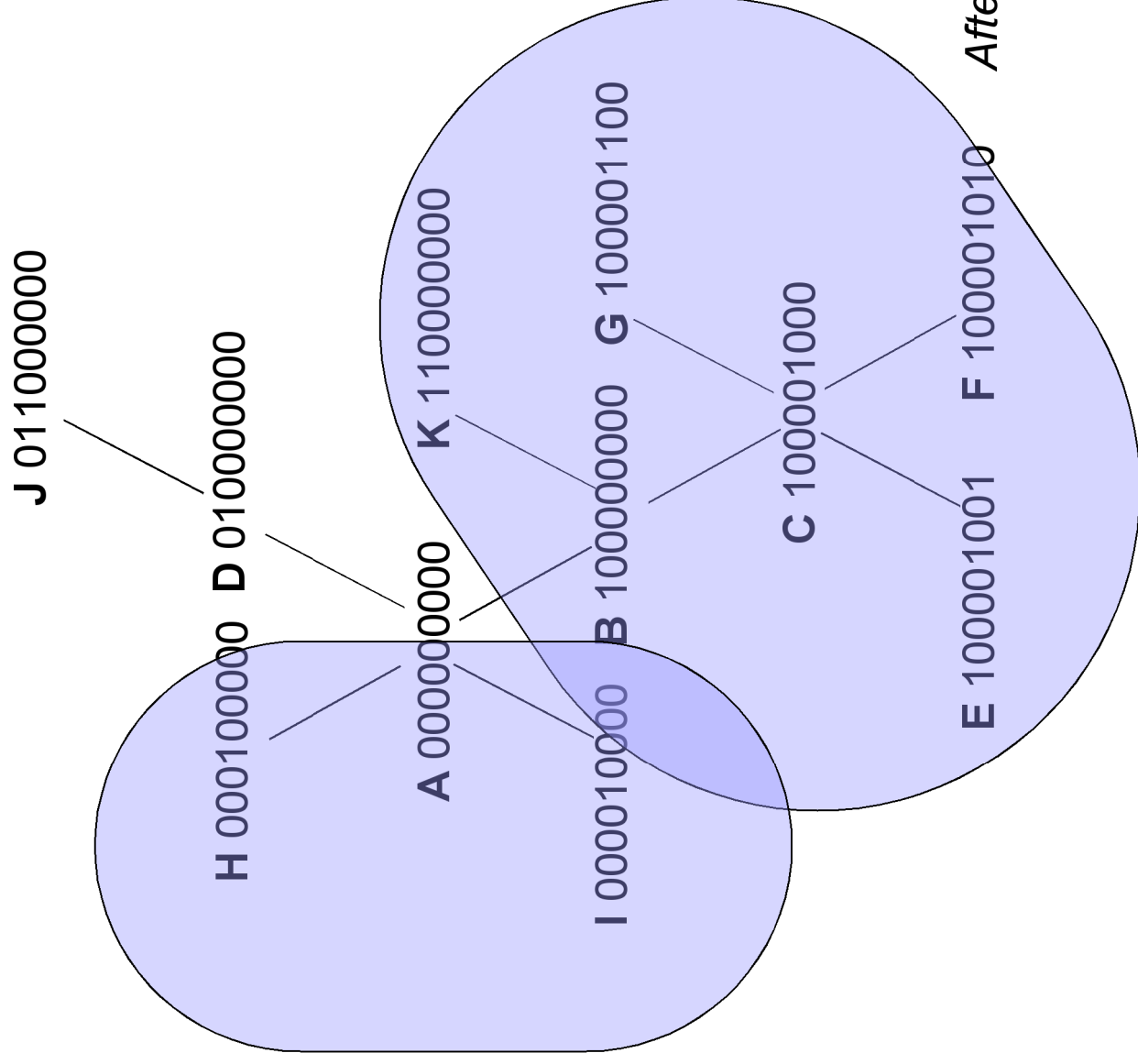
*After Seltman et al (03)*

# Cladogram-collapsing



*After Seltman et al (03)*

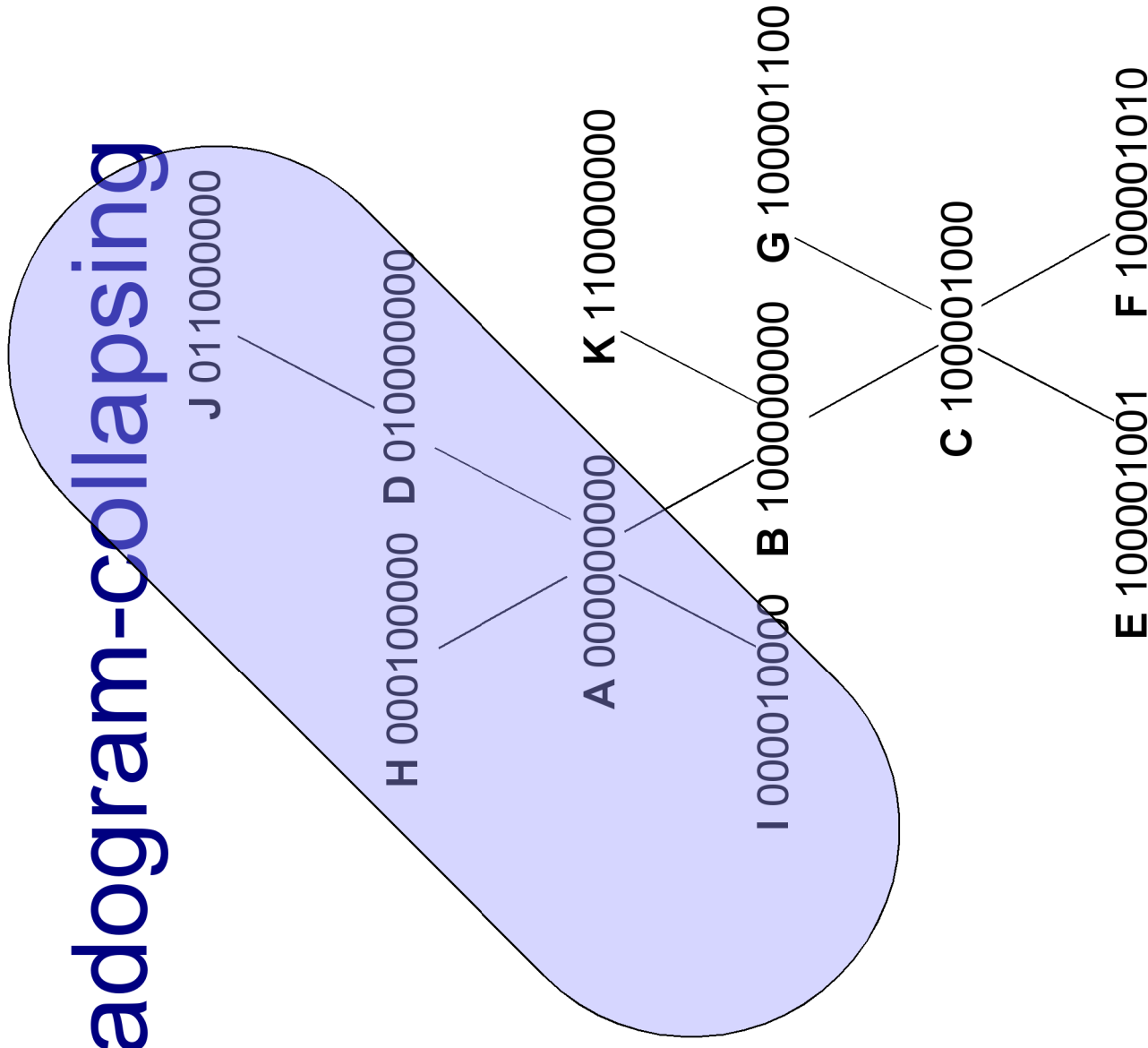
# Cladogram-collapsing



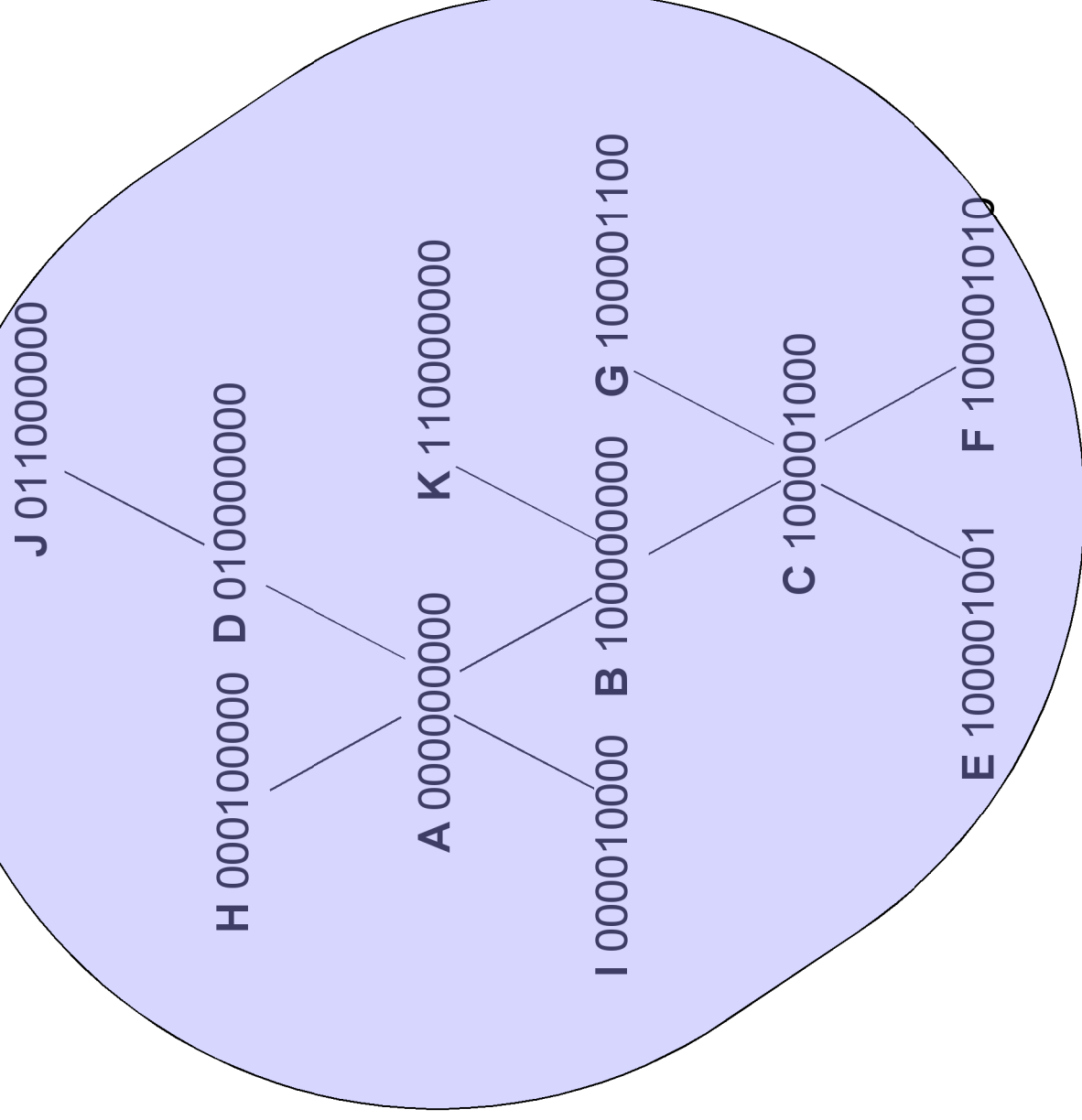
*After Seltman et al (03)*



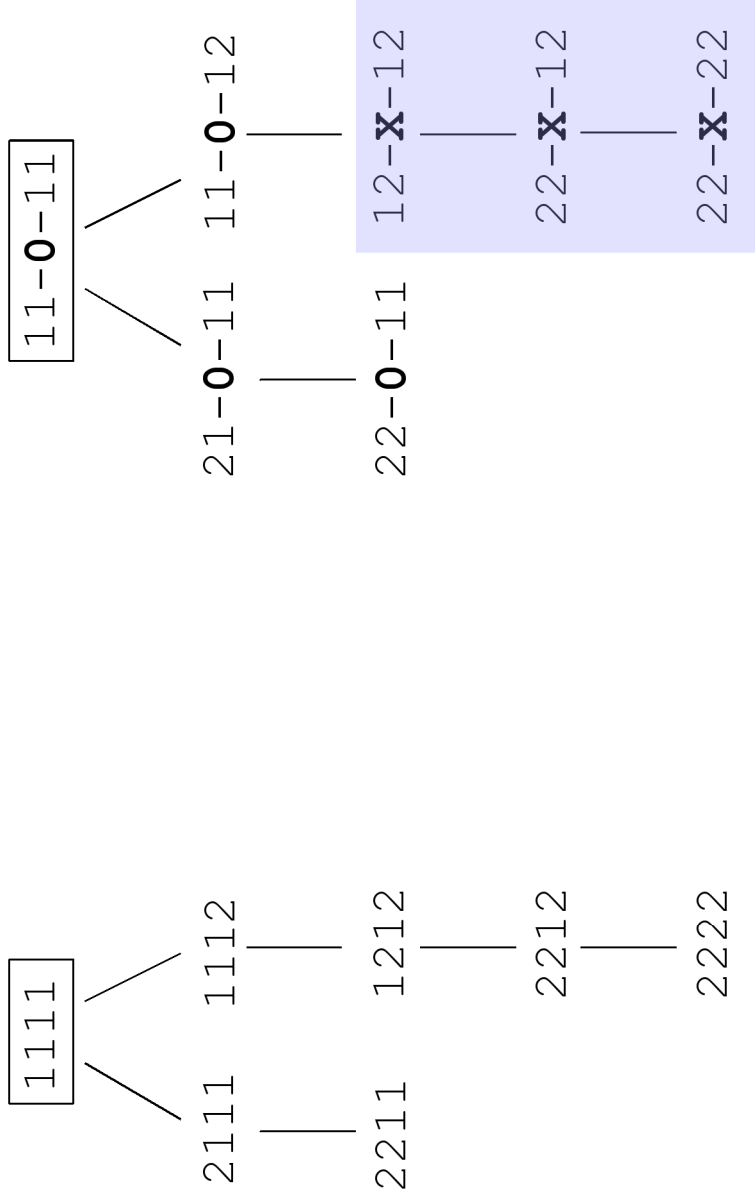
# Cladogram-collapsing



# Cladogram-collapsing



# Secondary analysis



# Secondary analysis

| <u>Haplotype</u> | <u>Estimated coefficients</u> |
|------------------|-------------------------------|
| 2211             | 0.000                         |
| 2111             | -0.092                        |
| 1111             | 0.102                         |
| 1112             | -0.234                        |
| 1212             | 0.634                         |
| 2212             | 0.332                         |
| 2222             | 0.865                         |

# Secondary analysis

## ■ Haplotype similarity

### □ Global and local identity

1111112212

1111121222

Global (0.7)

1111112212

1111121222

Local 1 (0.5)

1111112212

1111121222

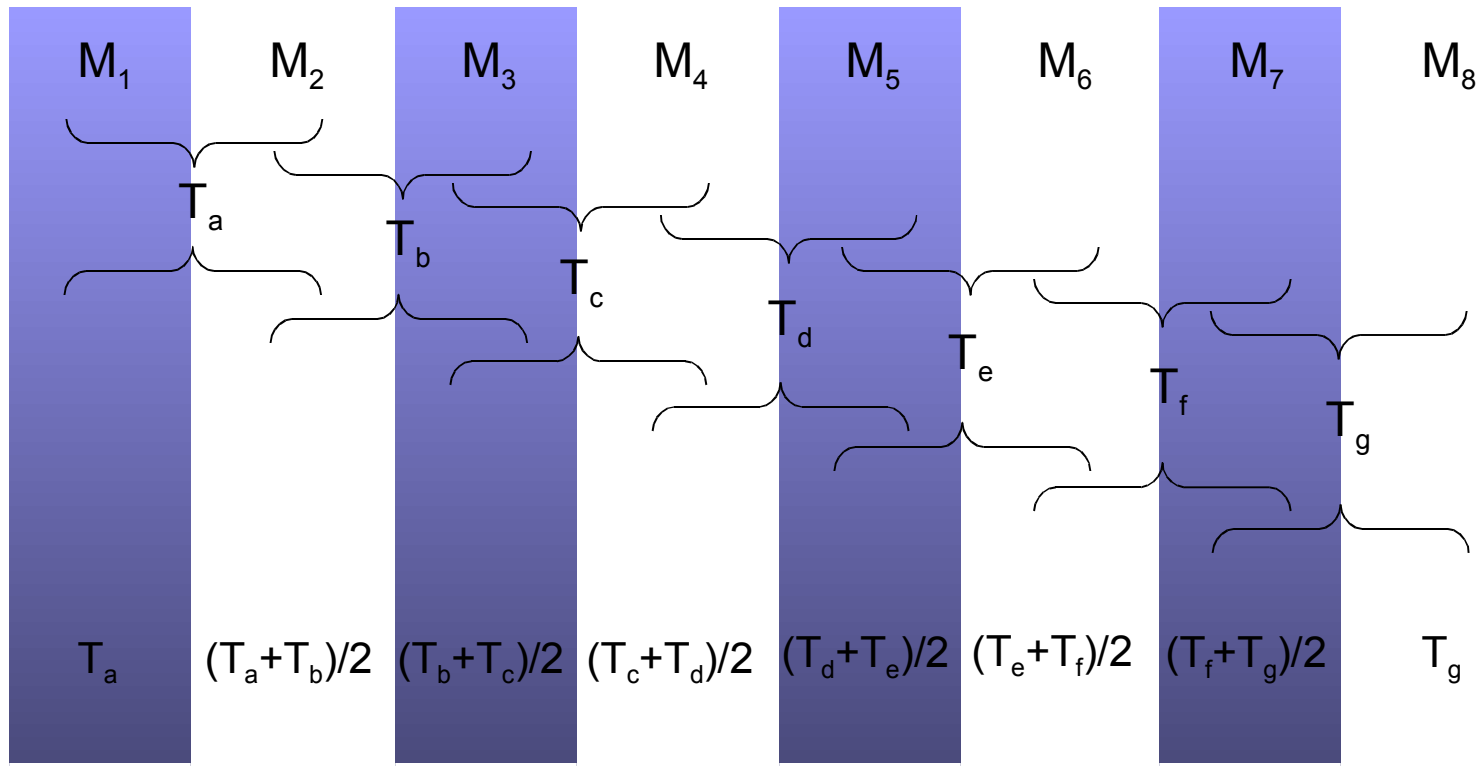
Local 8 (0.1)

## ■ Haplotype effect similarity

### □ Squared difference in MLE regression coefficients

$$(b_1 - b_2)^2 = (0.405 - 0.620)^2 = 0.462$$

# Sliding window analysis



For full details: <http://www.broad.mit.edu/~shaun/whap/>

# File formats

## ■ QTDT/Merlin input format

data.ped

```
1 1 0 0 1 -9 1 2 A A
1 2 0 0 2 -9 2 2 C C
1 3 1 2 1 -0.23 1 2 A C
```

data.dat

```
T quant1
M rs000001
M rs000002
```

data.map

```
14 rs000001 0 123232
14 rs000002 0 123887
```

## ■ Example command lines

```
whap --file data --alt 5,6,7 --null 5,7
```

```
whap --file data --alt 1,2,3 --at 5 --sec --perm 5000
```

```
whap --file data --alt 1,2 --window --cond --prev 0.02 --model w --wperm 5000
```

# Omnibus test

```
whap --file data --alt 5,6,7,8,9,10,11 --at 2
```

300 individuals w/out parents. 0 individuals with parents.  
275 of 300 individuals are informative

| Hap     | Freq  | Alt (B) | Alt (W) | Null (B) | Null (W) |
|---------|-------|---------|---------|----------|----------|
| 2122221 | 0.313 | 0.000   | 0.000   | 0.000    | 0.000    |
| 2112121 | 0.169 | -0.249  | -0.249  | 0.000    | 0.000    |
| 2221211 | 0.122 | -0.417  | -0.417  | 0.000    | 0.000    |
| 2212222 | 0.115 | -0.419  | -0.419  | 0.000    | 0.000    |
| 2122222 | 0.112 | 0.044   | 0.044   | 0.000    | 0.000    |
| 1112121 | 0.099 | -0.213  | -0.213  | 0.000    | 0.000    |
| 2222221 | 0.041 | 0.115   | 0.115   | 0.000    | 0.000    |
| 2212221 | 0.029 | -0.662  | -0.662  | 0.000    | 0.000    |

-----  
766.078  
-----  
787.673  
-----

Proportion of haplotypes covered = 0.955

LRT = 21.595

df = 7

p = 0.00298



# Haplotype-specific tests

```
whap --file data --alt 1,2,3 --at 2 --hs
```

|   | <i>Haplotype</i> | <i>Freq</i> | <i>B &amp; W coeffs</i> |        | <i>Chi-sq</i> | <i>p</i> |
|---|------------------|-------------|-------------------------|--------|---------------|----------|
| 1 | AGC              | 0.525       | -0.472                  | -0.472 | 8.546         | 0.00346  |
| 2 | CGC              | 0.220       | 0.107                   | 0.107  | 0.428         | 0.513    |
| 3 | CGA              | 0.180       | -0.088                  | -0.088 | 0.265         | 0.606    |
| 4 | ATA              | 0.075       | 0.116                   | 0.116  | 0.381         | 0.537    |



# Practical 2

- Use whap to phase dataACGT.ped

```
whap --file dataACGT --phase
```

*Just print out phases*

```
whap --file dataACGT --phase > probs.txt
```

*...or send to a file*

- Single SNP analysis

```
whap --file dataACGT --alt 1
```

*Analyse 1<sup>st</sup> SNP*

```
whap --file dataACGT --alt 5
```

*Analyse 5<sup>th</sup> SNP*

```
whap --file dataACGT --window --perm 50
```

*Sliding window*

*+ empirical p-values*

- Haplotype analysis

```
whap --file dataACGT
```

*Omnibus test*

```
whap --file dataACGT --alt 1,2,3,4,5
```

*As above*

```
whap --file dataACGT --hs
```

*All haplotype-specific tests*

# Performance of phasing

Of 400 individuals, 16 could not be assigned phase with (near) certainty: all 16 had the same genotypes: AA AC AC GT AC

**AAATA / ACCGC 0.324**

**AAATA / ACAGC 0.676**

|            |          |          |              |              |              |
|------------|----------|----------|--------------|--------------|--------------|
| 1_A        | 1        | 1        | ACCGC        | ACAGC        | 1.000        |
| <b>2_A</b> | <b>1</b> | <b>1</b> | <b>AACTA</b> | <b>ACAGC</b> | <b>0.676</b> |
| <b>2_A</b> | <b>1</b> | <b>2</b> | <b>AAATA</b> | <b>ACCGC</b> | <b>0.324</b> |
| 3_A        | 1        | 1        | ACAGC        | AAATA        | 1.000        |
| 4_A        | 1        | 1        | AAATA        | AACTA        | 1.000        |
| <b>5_A</b> | <b>1</b> | <b>1</b> | <b>ACAGC</b> | <b>AACTA</b> | <b>0.676</b> |
| <b>5_A</b> | <b>1</b> | <b>2</b> | <b>ACCGC</b> | <b>AAATA</b> | <b>0.324</b> |
| 6_A        | 1        | 1        | ACAGC        | ACAGC        | 1.000        |
| 7_A        | 1        | 1        | AAATA        | CCCGC        | 1.000        |
| 8_A        | 1        | 1        | CCCGC        | ACCGC        | 1.000        |
| 9_A        | 1        | 1        | ACCGC        | ACAGC        | 1.000        |
| ...        |          |          |              |              |              |
| ...        |          |          |              |              |              |

# Single SNP analysis

```
whap --file data --window --perm 500
```

```
Global permutation tests
```

```
-----
```

```
P_MAX = 6.791      p = 0.0279
```

```
P_SUM = 21.618    p = 0.0119
```

← Empirical p-values, corrected  
for multiple testing

```
Local permutation tests
```

```
-----
```

```
>> snp1 1 P_1= 0.019      p= 0.8822
```

```
>> snp2 2 P_2= 6.791      p= 0.0119
```

```
>> snp3 3 P_3= 4.412      p= 0.0199
```

```
>> snp4 4 P_4= 6.791      p= 0.0119
```

```
>> snp5 5 P_5= 3.605      p= 0.0518
```

# Omnibus test

```
whap --file dataACGT --alt 1,2,3,4,5
```

```
WHAP! | v2.04 | 05/09/03 | S. Purcell, P. Sham | purcell@wi.mit.edu  
400 individuals w/out parents. 0 individuals with parents. Binary trait:
```

```
400 of 400 individuals/trios are informative
```

| Hap          | Freq         | Alt (B)      | Alt (W)      |     | Null (B) | Null (W) |     |
|--------------|--------------|--------------|--------------|-----|----------|----------|-----|
| ---          | -----        | -----        | -----        |     | -----    | -----    |     |
| ACAGC        | 0.264        | 0.000        | 0.000        | [1] | 0.000    | 0.000    | [1] |
| CCCGC        | 0.237        | 0.406        | 0.406        | [2] | 0.000    | 0.000    | [1] |
| CCCGA        | 0.212        | 0.269        | 0.269        | [3] | 0.000    | 0.000    | [1] |
| AAATA        | 0.169        | 0.383        | 0.383        | [4] | 0.000    | 0.000    | [1] |
| <b>AACTA</b> | <b>0.067</b> | <b>1.338</b> | <b>1.338</b> | [5] | 0.000    | 0.000    | [1] |
| ACCGC        | 0.050        | 0.424        | 0.424        | [6] | 0.000    | 0.000    | [1] |
| ---          | -----        |              | -----        |     |          | -----    |     |
|              |              |              | 535.439      |     |          | 554.518  |     |

```
Proportion of haplotypes covered = 1.000
```

```
LRT = 19.079
```

```
df = 5
```

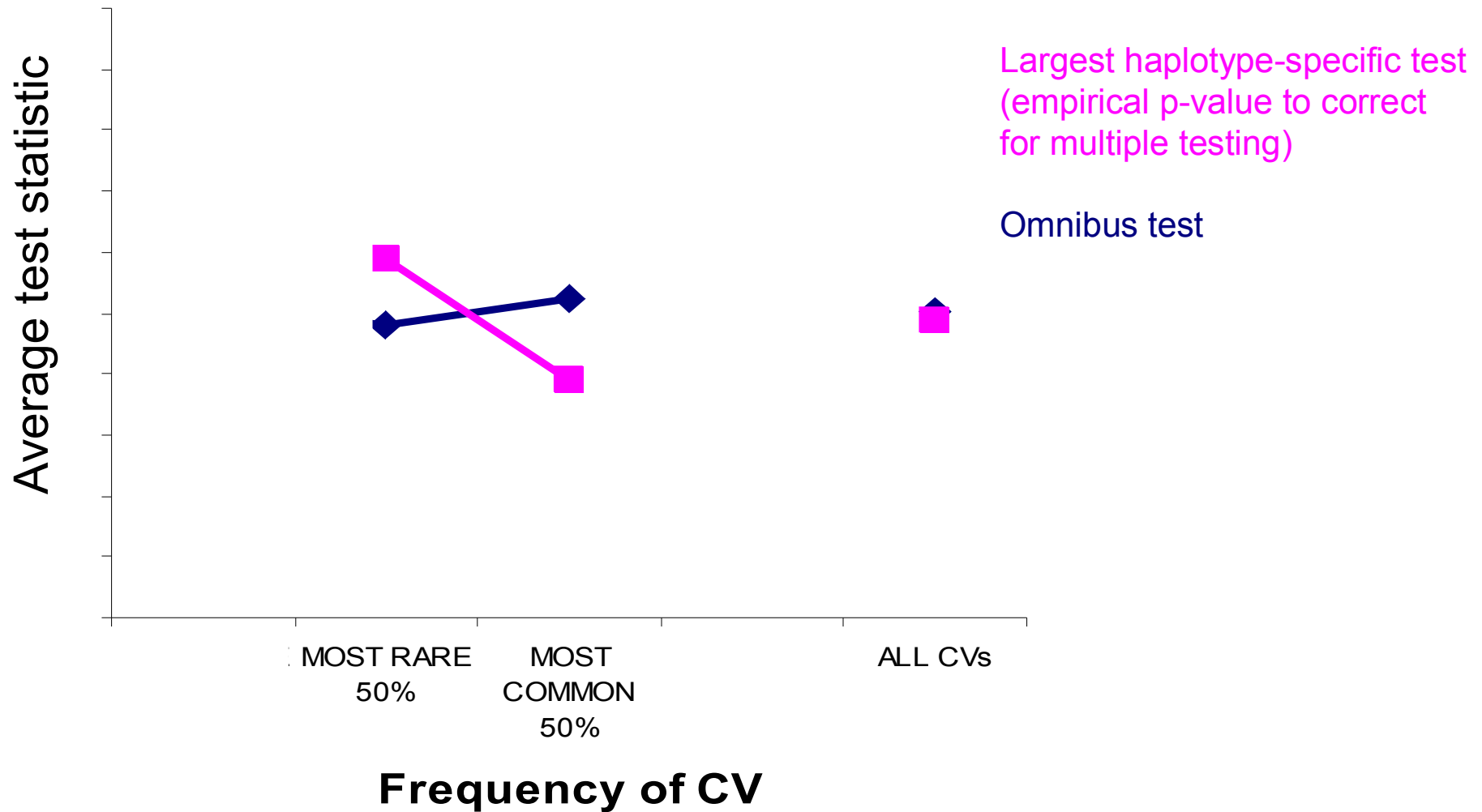
```
p = 0.00186
```



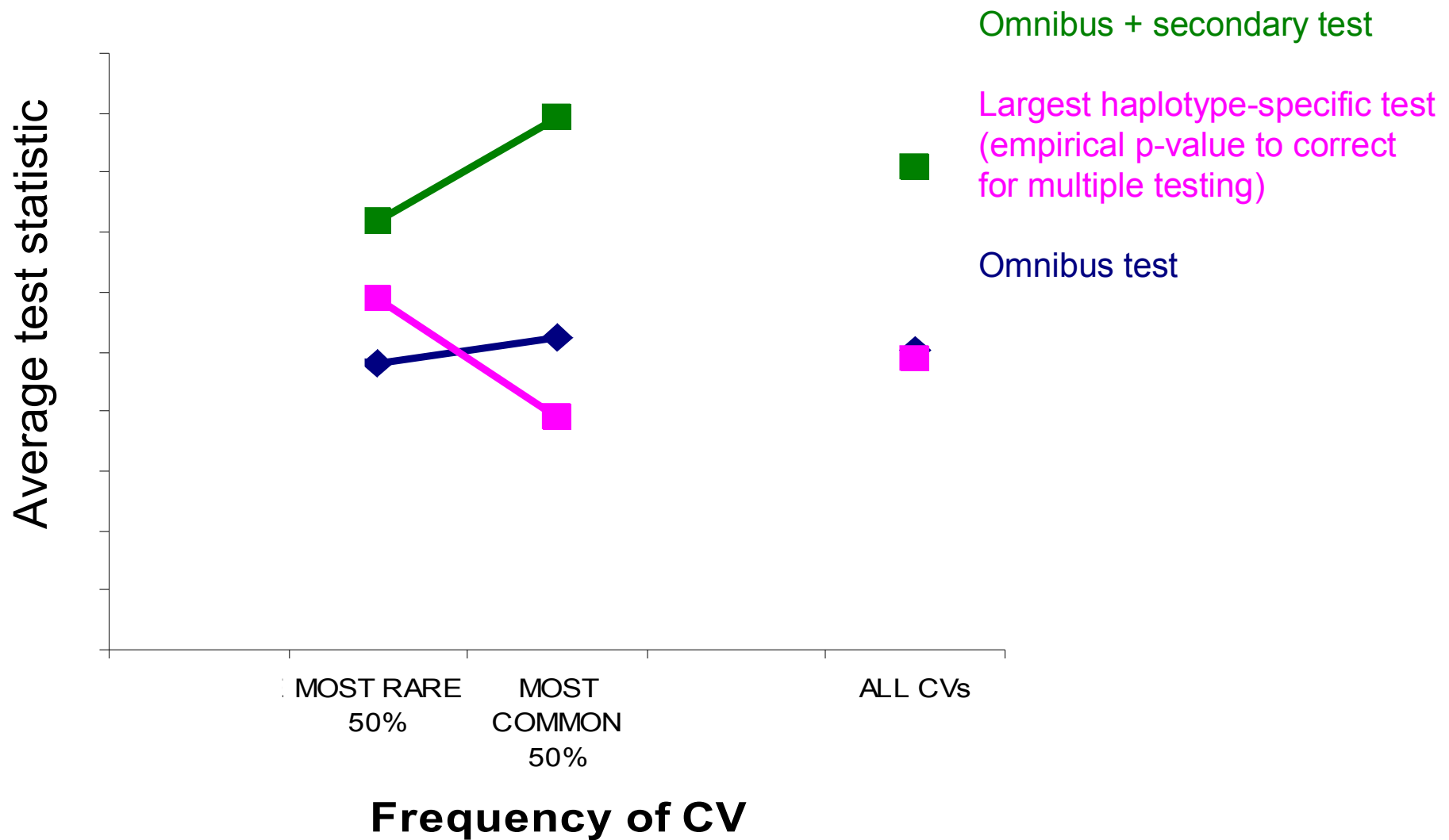
# Haplotype-specific tests

| <i>Haplotype</i> | <i>Chi-sq(1df)</i> | <i>p-value</i> | <i>beta</i>   | <i>OR</i>   |
|------------------|--------------------|----------------|---------------|-------------|
| <b>ACAGC</b>     | <b>8.546</b>       | <b>0.00346</b> | <b>-0.472</b> | <b>0.62</b> |
| CCCGC            | 0.428              | 0.513          | 0.107         | 1.11        |
| CCCGA            | 0.265              | 0.607          | -0.088        | 0.91        |
| AAATA            | 0.381              | 0.537          | 0.116         | 1.23        |
| <b>AACTA</b>     | <b>13.929</b>      | <b>0.00019</b> | <b>1.128</b>  | <b>3.08</b> |
| ACCGC            | 0.073              | 0.787          | 0.092         | 1.09        |

# Haplotype-specific or omnibus?



# Haplotype-specific or omnibus?







# Practical 3 : exploring the effect

## ■ Detection

- single SNP
- haplotype-specific
- omnibus test

## ■ “Is $X$ associated with my phenotype?”

- where  $X$  is either an allele, genotype, haplotype or set of haplotypes



# Practical 3 : exploring the effect

- Exploring the nature of an association
  - i.e. assuming there is an association, where is it coming from?
  - a single haplotype or multiple haplotype effects?
  - a single variant explains the entire effect?
- “Is  $X$  associated with my phenotype independent of  $Y$ ?”

# Interpreting effects

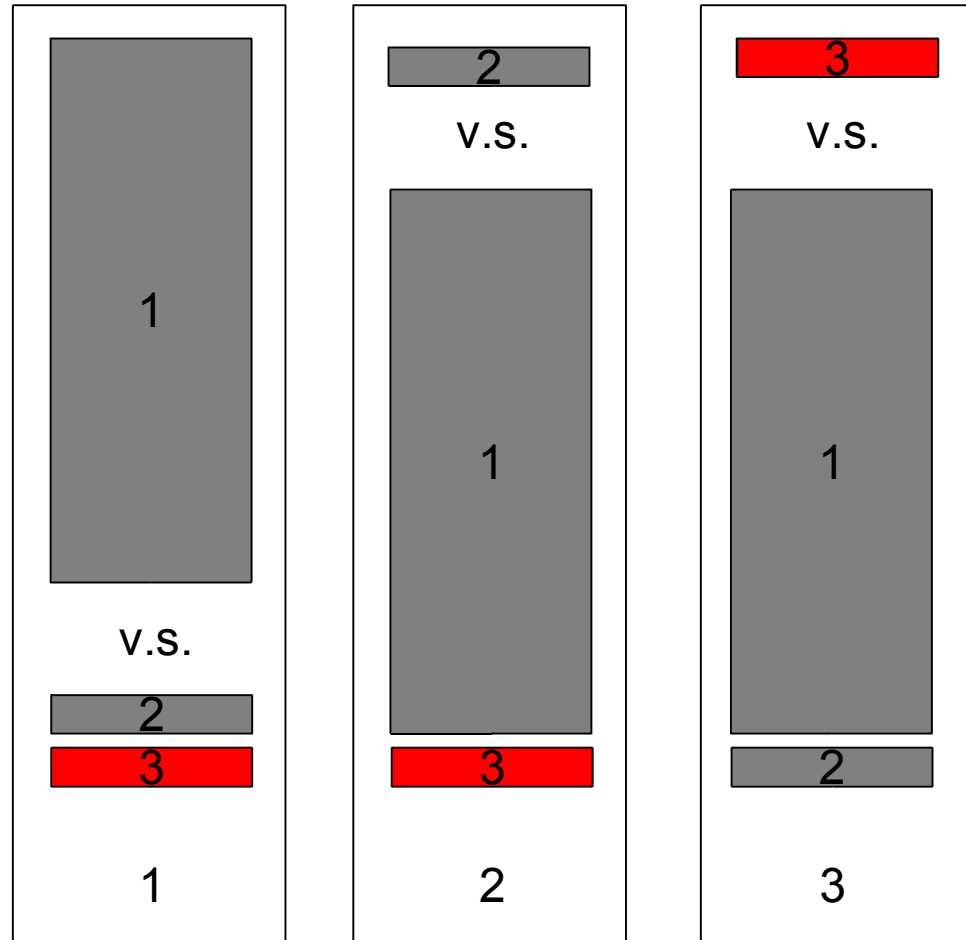
*True model*

|   |             |     |
|---|-------------|-----|
| 1 | <b>AACG</b> | 90% |
| 2 | <b>GGAC</b> | 05% |
| 3 | <b>AAAC</b> | 05% |

*Looks like*

|   |             |     |
|---|-------------|-----|
| 1 | <b>AACG</b> | 90% |
| 2 | <b>GGAC</b> | 05% |
| 3 | <b>AAAC</b> | 05% |

Haplotype-specific tests:





# Interpreting effects

*True model*

- |   |             |     |                      |
|---|-------------|-----|----------------------|
| 1 | <b>AACG</b> | 50% | <i>strong effect</i> |
| 2 | <b>GGAC</b> | 40% |                      |
| 3 | <b>AAAC</b> | 10% | <i>mild effect</i>   |

*Under an omnibus test*

- |   |             |          |
|---|-------------|----------|
| 1 | <b>AACG</b> | OR = 1.0 |
| 2 | <b>GGAC</b> | OR = 0.4 |
| 3 | <b>AAAC</b> | OR = 0.9 |

# Specifying the model in whap

- Specify markers to form haplotypes from under the alternate and null

- `--alt 1,2,3,4 --null 3,4`

|             |     |              |     |
|-------------|-----|--------------|-----|
| <b>1111</b> | [1] | 11 <b>11</b> | [1] |
| <b>1122</b> | [2] | 11 <b>22</b> | [2] |
| <b>2221</b> | [3] | 22 <b>21</b> | [3] |
| <b>2222</b> | [4] | 22 <b>22</b> | [2] |
| <b>2211</b> | [5] | 22 <b>11</b> | [1] |



# Specifying the model in whap

## ■ Equate haplotypes directly

□ `--constrain 1,2,3,4,5/1,2,3,2,1`

|      |     |      |     |
|------|-----|------|-----|
| 1111 | [1] | 1111 | [1] |
| 1122 | [2] | 1122 | [2] |
| 2221 | [3] | 2221 | [3] |
| 2222 | [4] | 2222 | [2] |
| 2211 | [5] | 2211 | [1] |

*Note: first haplotype always has to have parameter [1]*

*Must specify as many parameters as there are haplotypes*

# Conditional tests

- Two SNPs both individually predict the phenotype
  - Do they have independent effects?
  - Or can one explain the other?

| <u>Haplotype</u> | <u>Freq</u> | <u>Odds ratio</u> | <u>Alt</u> | <u>Null</u> |
|------------------|-------------|-------------------|------------|-------------|
| AB               | 0.50        | 1.00 (fixed)      | [1]        | [1]         |
| ab               | 0.45        | 2.00              | [2]        | [2]         |
| Ab               | 0.05        | ?                 | [3]        | [2]         |

```
--alt 1,2 --null 2
```



# Conditional tests

- Assuming significant omnibus test:
  - can we make it go away?
- $X$  independently contributes (if signif.)
  - `--alt 1,2,3,4,5 --null 2,3,4,5`
  - “independent effect test”
- $X$  is necessary and sufficient (if test n.signif.)
  - `--alt 1,2,3,4,5 --null 1`
  - `--constrain 1,2,3,4,5,6/1,2,1,1,1,1`
  - “sole variant test”



# Haplotype-specific test (H1)

--constrain 1,2,2,2,2,2 / 1,1,1,1,1,1,1

A A T A

A C A G C

C C C G A

C C C G C

A A C T A

A C C G C

A A A T A

A C A G C

C C C G A

C C C G C

A A C T A

A C C G C

# Haplotype-specific test (H2)

--constrain 1,2,1,1,1,1 / 1,1,1,1,1,1

A A T A

A C A G C

C C G A

C C G C

A A C T A

A C C G C

A A T A

A C A G C

C C G A

C C G C

A A C T A

A C C G C

Omnibus test (df=5)

--constrain 1,2,3,4,5,6 / 1,1,1,1,1,1

A A T A

A C A G C

C C C G A

C C C G C

A A C T A

A C C G C

A A A T A

A C A G C

C C C G A

C C C G C

A A C T A

A C C G C

# Clade-based homogeneity test (1df)

--constrain 1,1,2,2,3,3 / 1,1,2,2,2,2

A A T A  
A C A G C

A A T A  
A C A G C

C C G A  
C C C G C

C C G A  
C C C G C  
A A C T A  
A C C G C

A A C T A  
A C C G C

Single SNP test (2<sup>nd</sup> marker)

--alt 2

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | A | T | A | A | T | A |   |
| A | C | A | G | C | A | G | C |
| C | C | C | G | A | C | G | A |
| C | C | C | G | C | C | G | C |
| A | A | C | T | A | C | T | A |
| A | C | C | G | C | C | G | C |

# Independent effect test for SNP 1

--alt 1,2,3,4,5 --null 2,3,4,5

A A T A

A A T A

A C A G C

A C A G C

C C C G A

C C G A

C C C G C

C C G C

A A C T A

A C T A

A C C G C

C C G C

# Independent effect test for SNPs 1, 2 and 3

--alt 1,2,3,4,5 --null 4,5

A A T A

A A T A

A C A G C

A C A G C

C C C G A

C C C G A

C C C G C

C C C G C

A A C T A

A A C T A

A C C G C

A C C G C

# Sole-variant test for 2<sup>nd</sup> SNP

--alt 1,2,3,4,5 --null 2

A A T A

A A T A

A C A G C

A C A G C

C C C G A

C C C G A

C C C G C

C C C G C

A A C T A

A A C T A

A C C G C

A C C G C



# Sole-variant test for haplotype 2

--constrain 1,2,3,4,5,6 / 1,2,1,1,1,1

A A T A

A A T A

A C A G C

A C A G C

C C C G A

C C C G A

C C C G C

C C C G C

A A C T A

A A C T A

A C C G C

A C C G C



# Practical: conditional tests

- For each SNP, perform an independent effects and a “sole-variant” test. Compare these to the standard single SNP and haplotype-specific tests. What do they tell you?

- Independent effect tests, e.g.

- `whap --file dataACGT --alt 1,2,3,4,5 --null 2,3,4,5`

- Sole-variant SNP tests, e.g.

- `whap --file dataACGT --alt 1,2,3,4,5 --null 1`

- Sole-variant haplotype tests, e.g.

- `--constrain 1,2,3,4,5,6/1,2,2,2,2,2`

- `--constrain 1,2,3,4,5,6/1,2,1,1,1,1`



**Standard SNP test (df=1) (chi-sq, p-value)**

--alt 1

|      |       |         |
|------|-------|---------|
| SNP1 | 0.019 | 0.89    |
| SNP2 | 6.791 | 0.00916 |
| SNP3 | 4.412 | 0.0357  |
| SNP4 | 6.791 | 0.00916 |
| SNP5 | 3.605 | 0.0576  |

**Independent effect test (df=1) (chi-sq, p-value)**

--alt 1,2,3,4,5 --null 2,3,4,5

|      |       |        |
|------|-------|--------|
| SNP1 | 0.003 | 0.959  |
| SNP2 | n/a   | n/a    |
| SNP3 | 8.954 | 0.0114 |
| SNP4 | n/a   | n/a    |
| SNP5 | 0.408 | 0.523  |

**Sole-variant test (df=4) (chi-sq, p-value)**

--alt 1,2,3,4,5 --null 1

|      |        |          |
|------|--------|----------|
| SNP1 | 19.060 | 0.000765 |
| SNP2 | 12.288 | 0.0153   |
| SNP3 | 14.667 | 0.00544  |
| SNP4 | 12.289 | 0.0153   |
| SNP5 | 15.474 | 0.00381  |

# Sole-variant tests for haplotypes

## Standard haplotype-specific tests

| Haplotype    | Chi-sq(1df)   | p-value        |                                     |
|--------------|---------------|----------------|-------------------------------------|
| <b>ACAGC</b> | <b>8.546</b>  | <b>0.00346</b> | 1, 2, 2, 2, 2, 2 / 1, 1, 1, 1, 1, 1 |
| CCCGC        | 0.428         | 0.513          | 1, 2, 1, 1, 1, 1 / 1, 1, 1, 1, 1, 1 |
| CCCGA        | 0.265         | 0.607          | 1, 1, 2, 1, 1, 1 / 1, 1, 1, 1, 1, 1 |
| AAATA        | 0.381         | 0.537          | 1, 1, 1, 2, 1, 1 / 1, 1, 1, 1, 1, 1 |
| <b>AACTA</b> | <b>13.929</b> | <b>0.00019</b> | 1, 1, 1, 1, 2, 1 / 1, 1, 1, 1, 1, 1 |
| ACCGC        | 0.073         | 0.787          | 1, 1, 1, 1, 1, 2 / 1, 1, 1, 1, 1, 1 |

## Sole-variant tests for haplotypes

| Haplotype    | Chi-sq(4df)  | p-value      |                                     |
|--------------|--------------|--------------|-------------------------------------|
| ACAGC        | 10.533       | 0.0323       | 1, 2, 3, 4, 5, 6 / 1, 2, 2, 2, 2, 2 |
| CCCGC        | 18.651       | 0.00092      | 1, 2, 3, 4, 5, 6 / 1, 2, 1, 1, 1, 1 |
| CCCGA        | 18.814       | 0.000855     | 1, 2, 3, 4, 5, 6 / 1, 1, 2, 1, 1, 1 |
| AAATA        | 18.698       | 0.000901     | 1, 2, 3, 4, 5, 6 / 1, 1, 1, 2, 1, 1 |
| <b>AACTA</b> | <b>5.150</b> | <b>0.272</b> | 1, 2, 3, 4, 5, 6 / 1, 1, 1, 1, 2, 1 |
| ACCGC        | 19.006       | 0.000784     | 1, 2, 3, 4, 5, 6 / 1, 1, 1, 1, 1, 2 |

# Including the causal variant

|                 |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|-----------------|-----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AC-C-AGC        | 1_A | 1 | 0 | 0 | 1 | 2 | A | A | C | C | C | A | G | G | C | C | C | C |
| CC-C-CGC        | 2_A | 1 | 0 | 0 | 1 | 2 | A | A | A | C | C | A | T | G | A | C | T | C |
| CC-C-CGA        | 3_A | 1 | 0 | 0 | 1 | 2 | A | A | C | A | A | A | G | T | C | A | C | C |
| AA-C-ATA        | 4_A | 1 | 0 | 0 | 1 | 2 | A | A | A | A | A | C | T | T | A | A | C | T |
| <b>AA-T-CTA</b> | 5_A | 1 | 0 | 0 | 1 | 2 | A | A | C | A | A | C | G | T | C | A | C | T |
| AC-C-CGC        | 6_A | 1 | 0 | 0 | 1 | 2 | A | A | C | C | A | A | G | G | C | C | C | C |

A disease

M snp1

M snp2

M snp3

M snp4

M snp5

M cv

AC-C-CGC  
↑

## Files

cvACGT.\*

cv1234.\*

1 snp1 0 1

1 snp2 0 2

1 cv 0 3

1 snp3 0 4

1 snp4 0 5

1 snp5 0 6

# Single locus test of the CV

```
whap --file data-cv --alt 3
```

```
WHAP! | v2.04 | 05/09/03 | S. Purcell, P. Sham | purcell@wi.mit.edu  
400 individuals w/out parents. 0 individuals with parents. Binary trait:
```

```
400 of 400 individuals/trios are informative
```

| Hap | Freq  | Alt(B) | Alt(W)    | Null(B) | Null(W)   |
|-----|-------|--------|-----------|---------|-----------|
| C   | 0.935 | 0.000  | 0.000 [1] | 0.000   | 0.000 [1] |
| T   | 0.065 | 1.064  | 1.064 [2] | 0.000   | 0.000 [1] |
|     |       |        | 541.518   |         |           |
|     |       |        |           | 554.518 |           |

```
Proportion of haplotypes covered = 1.000
```

```
LRT = 13.000
```

```
df = 1
```

```
p = 0.000311
```

$\exp(1.064) \sim \text{OR } 2.9$

# Omnibus test with CV included

```
whap --file sim-cv --alt 1,2,3,4,5,6
```

```
WHAP! | v2.04 | 05/09/03 | S. Purcell, P. Sham | purcell@wi.mit.edu
```

```
400 individuals w/out parents. 0 individuals with parents. Binary trait:
```

```
400 of 400 individuals/trios are informative
```

| Hap           | Freq         | Alt(B)       | Alt(W)       |            | Null(B) | Null(W) |     |
|---------------|--------------|--------------|--------------|------------|---------|---------|-----|
| ---           | -----        | -----        | -----        |            | -----   | -----   |     |
| ACCAGC        | 0.261        | 0.000        | 0.000        | [1]        | 0.000   | 0.000   | [1] |
| CCCGC         | 0.237        | 0.411        | 0.411        | [2]        | 0.000   | 0.000   | [1] |
| CCCGA         | 0.212        | 0.276        | 0.276        | [3]        | 0.000   | 0.000   | [1] |
| AACATA        | 0.171        | 0.406        | 0.406        | [4]        | 0.000   | 0.000   | [1] |
| <b>AATCTA</b> | <b>0.065</b> | <b>1.317</b> | <b>1.317</b> | <b>[5]</b> | 0.000   | 0.000   | [1] |
| ACCGC         | 0.052        | 0.482        | 0.482        | [6]        | 0.000   | 0.000   | [1] |
| ---           | -----        |              | -----        |            |         | -----   |     |
|               |              |              | 535.616      |            |         | 554.518 |     |

```
Proportion of haplotypes covered = 1.000
```

```
LRT = 18.901
```

```
df = 5
```

```
p = 0.00201
```



# Sole-variant SNP tests

|           |                          |                 |                    |               |                  |
|-----------|--------------------------|-----------------|--------------------|---------------|------------------|
| SNP1      | --alt 1,2,3,4,5,6        | --null 1        | LRT = 18.882       | df = 4        | p = 0.000829     |
| SNP2      | --alt 1,2,3,4,5,6        | --null 2        | LRT = 12.111       | df = 4        | p = 0.0165       |
| <b>CV</b> | <b>--alt 1,2,3,4,5,6</b> | <b>--null 3</b> | <b>LRT = 5.901</b> | <b>df = 4</b> | <b>p = 0.207</b> |
| SNP3      | --alt 1,2,3,4,5,6        | --null 4        | LRT = 14.489       | df = 4        | p = 0.0295       |
| SNP4      | --alt 1,2,3,4,5,6        | --null 5        | LRT = 12.111       | df = 4        | p = 0.0165       |
| SNP5      | --alt 1,2,3,4,5,6        | --null 6        | LRT = 15.296       | df = 4        | p = 0.00413      |



# Sole-variant test of the CV

```
whap --file cvACGT --alt 1,2,3,4,5,6 --null 3
```

```
WHAP! | v2.06 | 13/Dec/04 | S. Purcell, P. Sham | spurcell@pngu.mgh.harvard.edu  
400 individuals w/out parents. 0 individuals with parents. Binary trait:
```

```
400 of 400 individuals/trios are informative
```

| Hap           | Freq         | Alt (B)      | Alt (W)      |            | Null (B)     | Null (W)     |            |
|---------------|--------------|--------------|--------------|------------|--------------|--------------|------------|
| ---           | -----        | -----        | -----        |            | -----        | -----        |            |
| ACCAGC        | 0.261        | 0.000        | 0.000        | [1]        | 0.000        | 0.000        | [1]        |
| CCCCGC        | 0.237        | 0.412        | 0.412        | [2]        | 0.000        | 0.000        | [1]        |
| CCCCGA        | 0.212        | 0.276        | 0.276        | [3]        | 0.000        | 0.000        | [1]        |
| AACATA        | 0.171        | 0.406        | 0.406        | [4]        | 0.000        | 0.000        | [1]        |
| <b>AATCTA</b> | <b>0.065</b> | <b>1.317</b> | <b>1.317</b> | <b>[5]</b> | <b>1.065</b> | <b>1.065</b> | <b>[2]</b> |
| ACCCGC        | 0.052        | 0.483        | 0.483        | [6]        | 0.000        | 0.000        | [1]        |
| ---           | -----        |              | -----        |            |              | -----        |            |
|               |              |              | 535.616      |            |              | 541.518      |            |

```
Proportion of haplotypes covered = 1.000
```

```
LRT = 5.901
```

```
df = 4
```

```
p = 0.207
```

# Single SNP vs “sole-variant”

