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

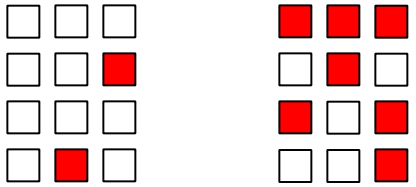
How can we use haplotypes to map causal variants?

Haplotype-based association analysis

What is association?

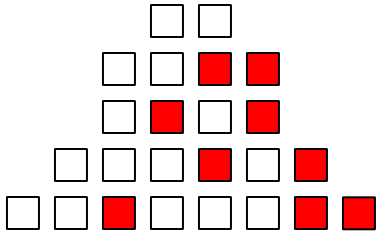
Categorical traits

disease susceptibility genes

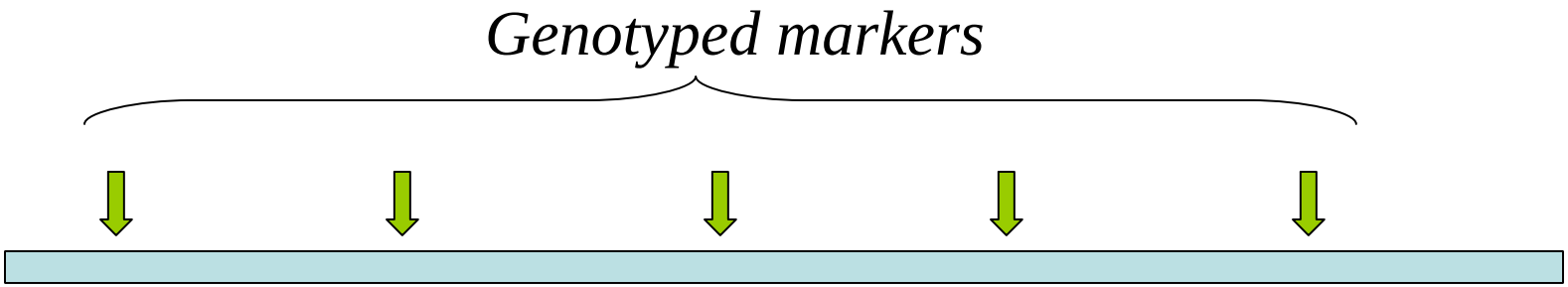


Continuous traits

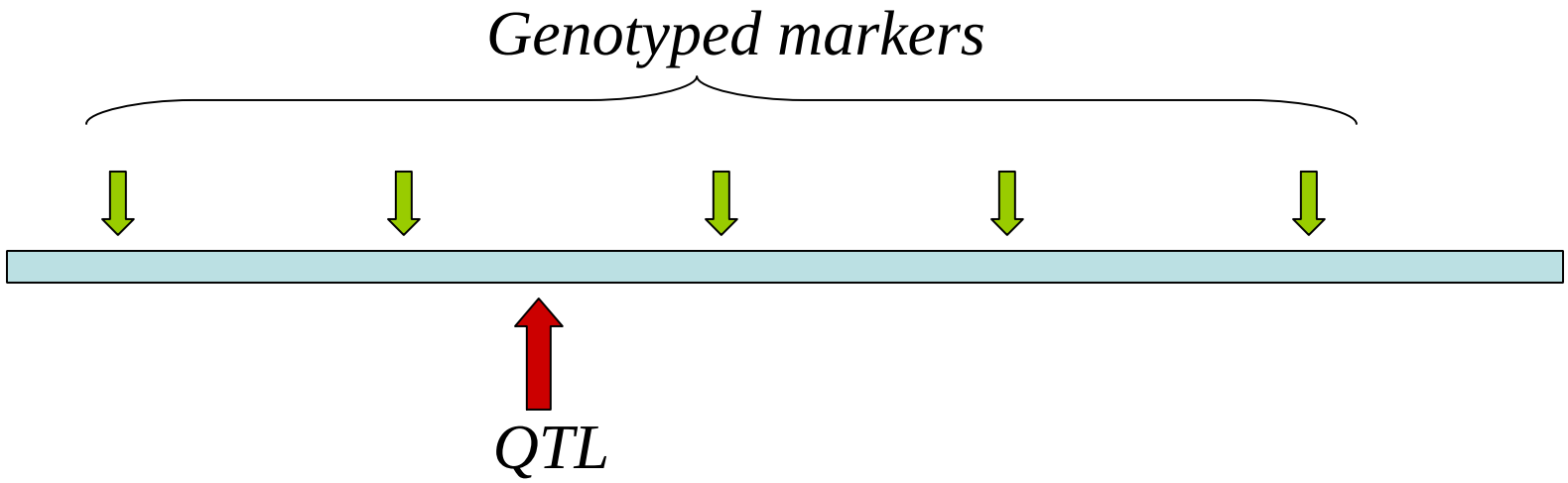
quantitative trait loci, QTL



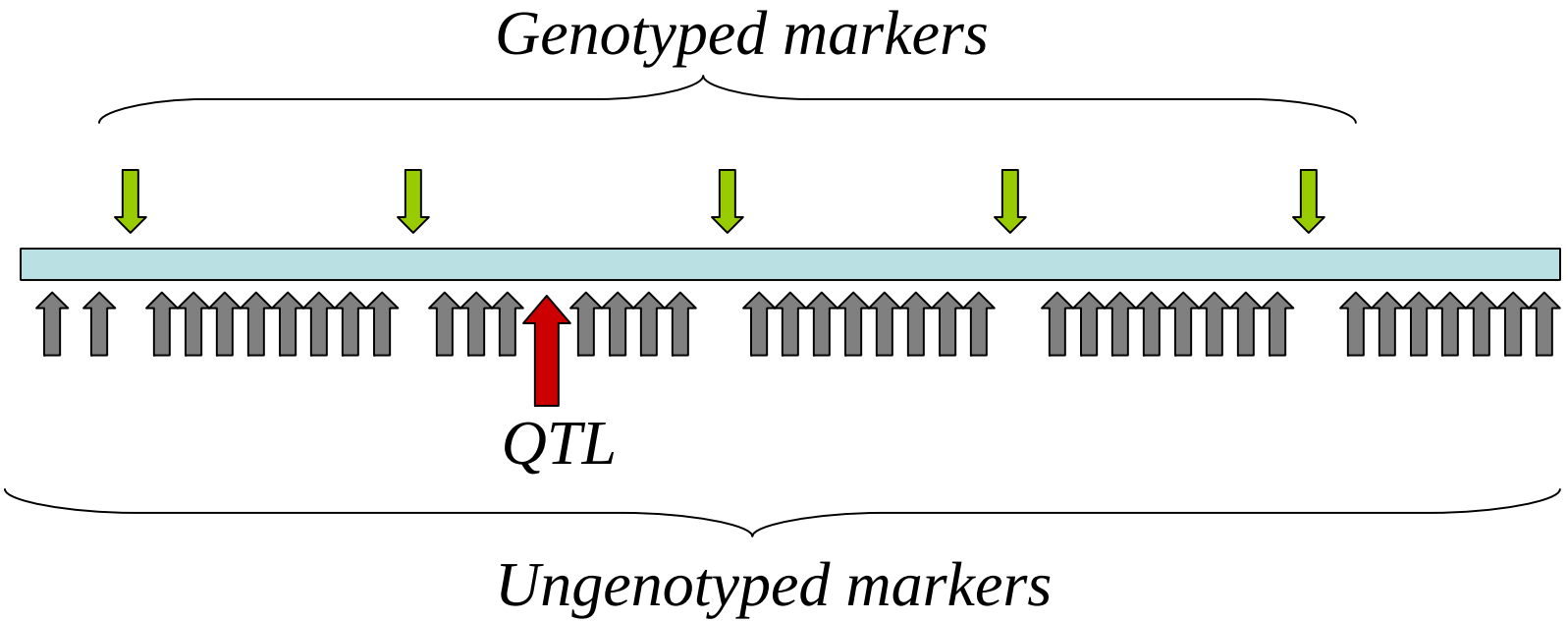
Linkage disequilibrium mapping



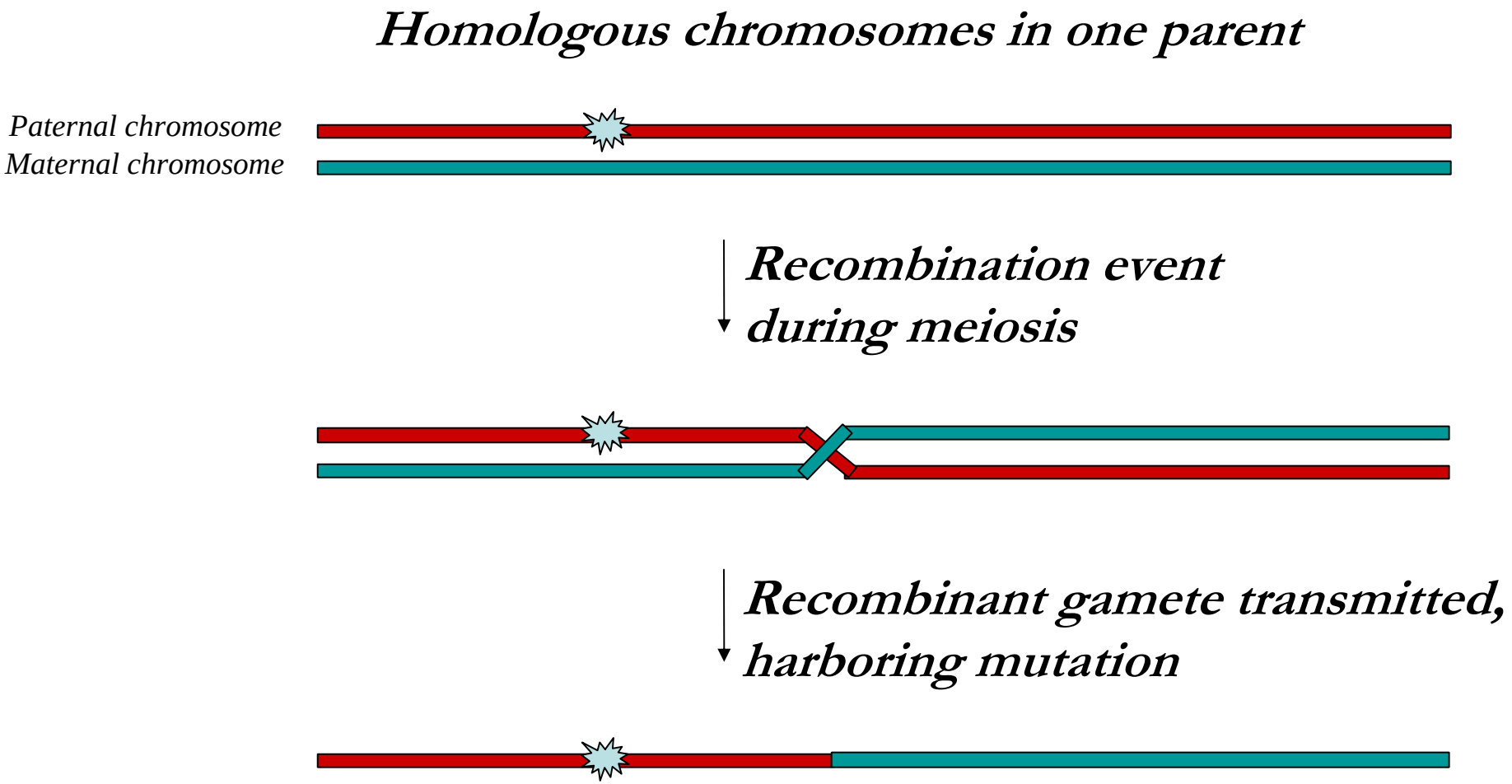
Linkage disequilibrium mapping



Linkage disequilibrium mapping



Recombination



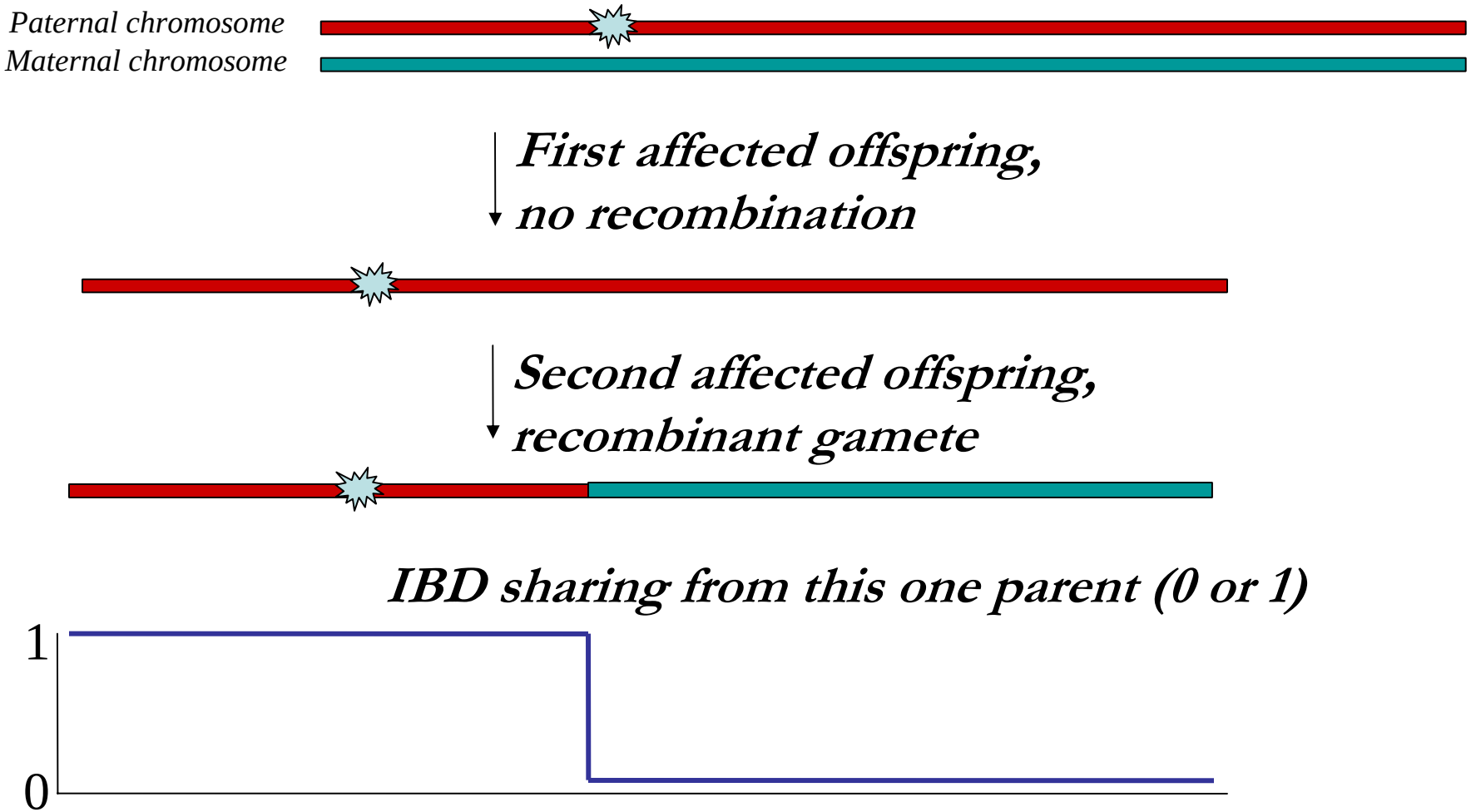
Homologous chromosomes in one parent

Paternal chromosome
Maternal chromosome

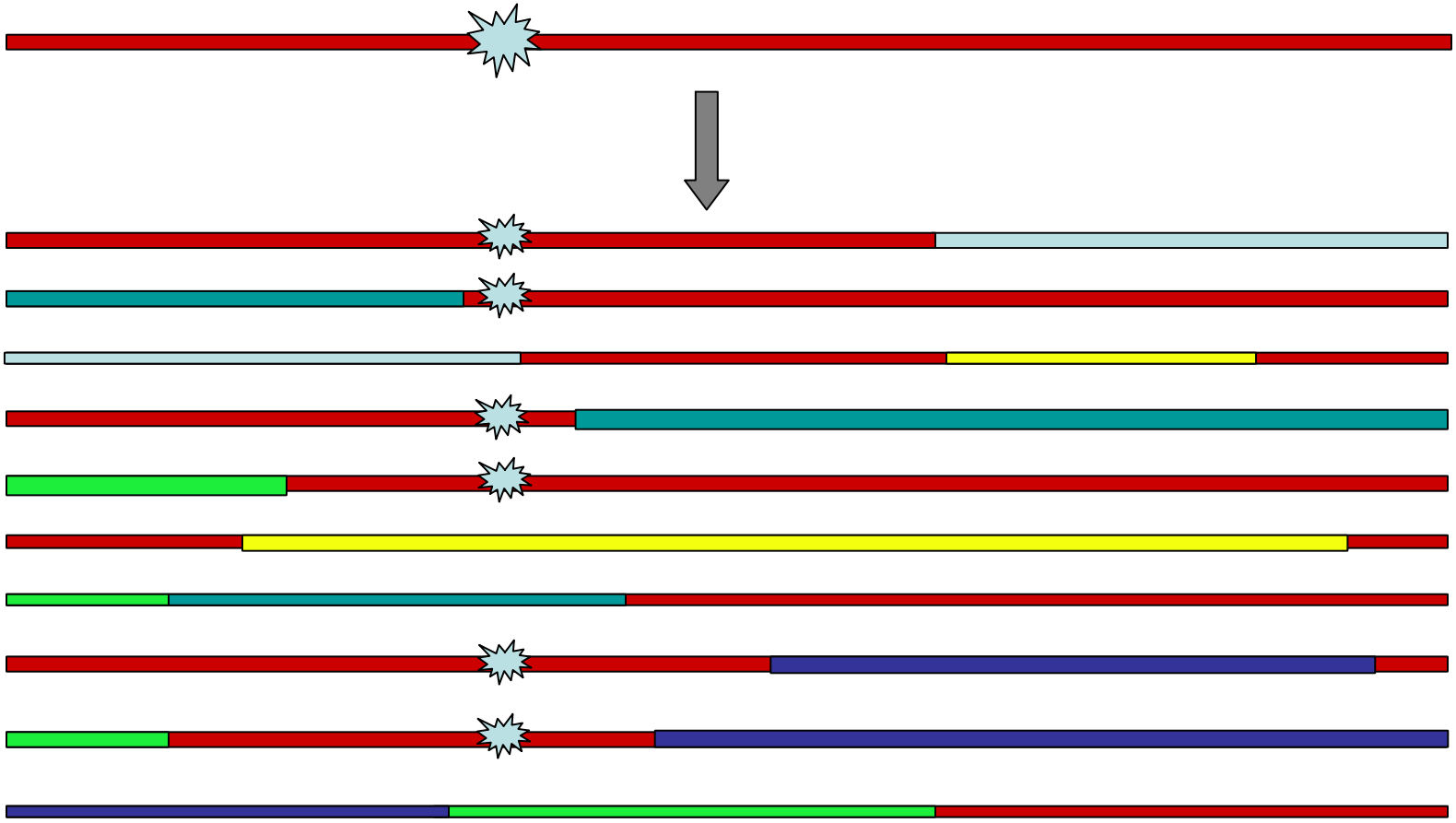
No recombination event during meiosis

Nonrecombinant gamete transmitted not harboring mutation

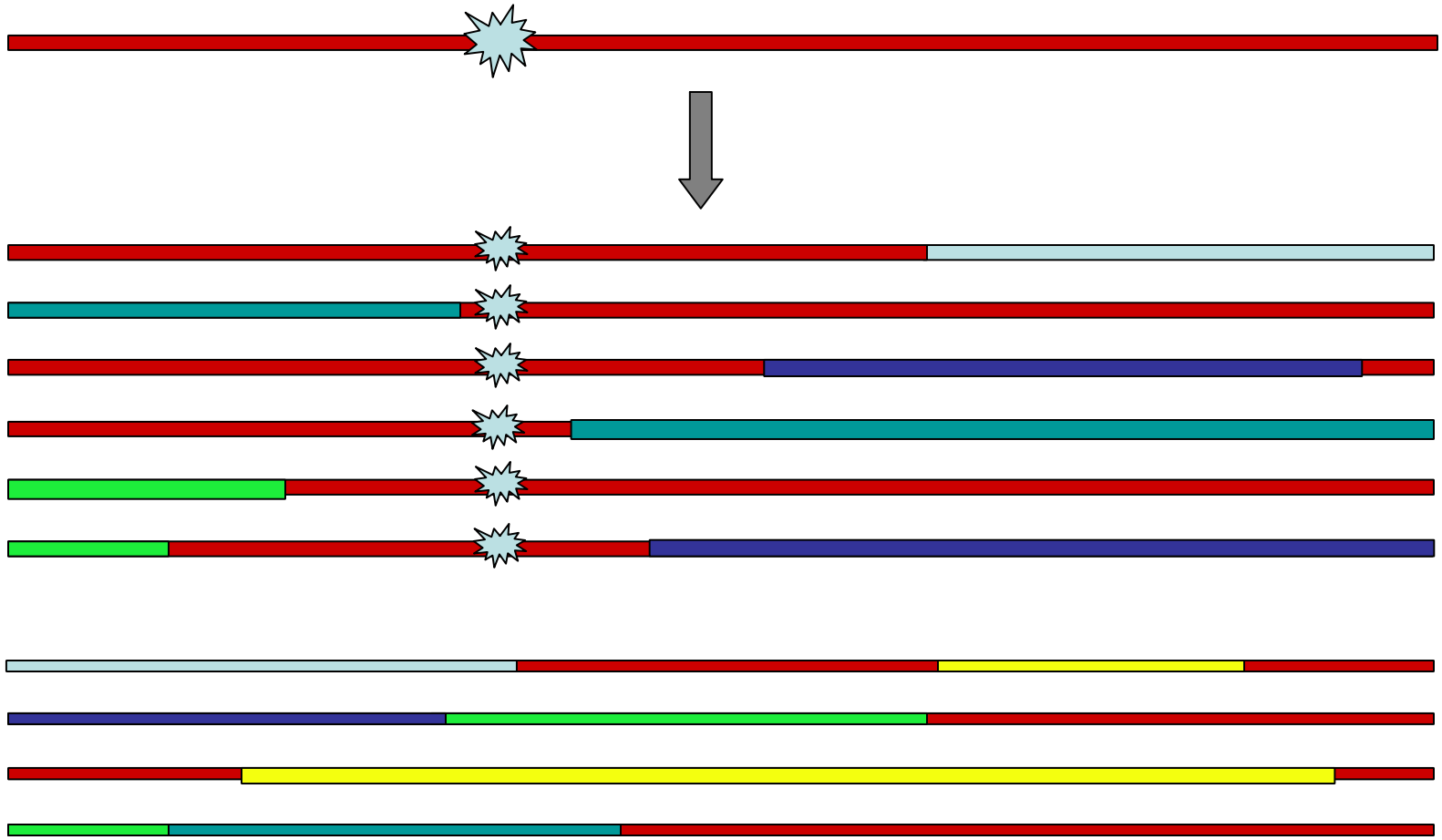
Linkage: affected sib pairs



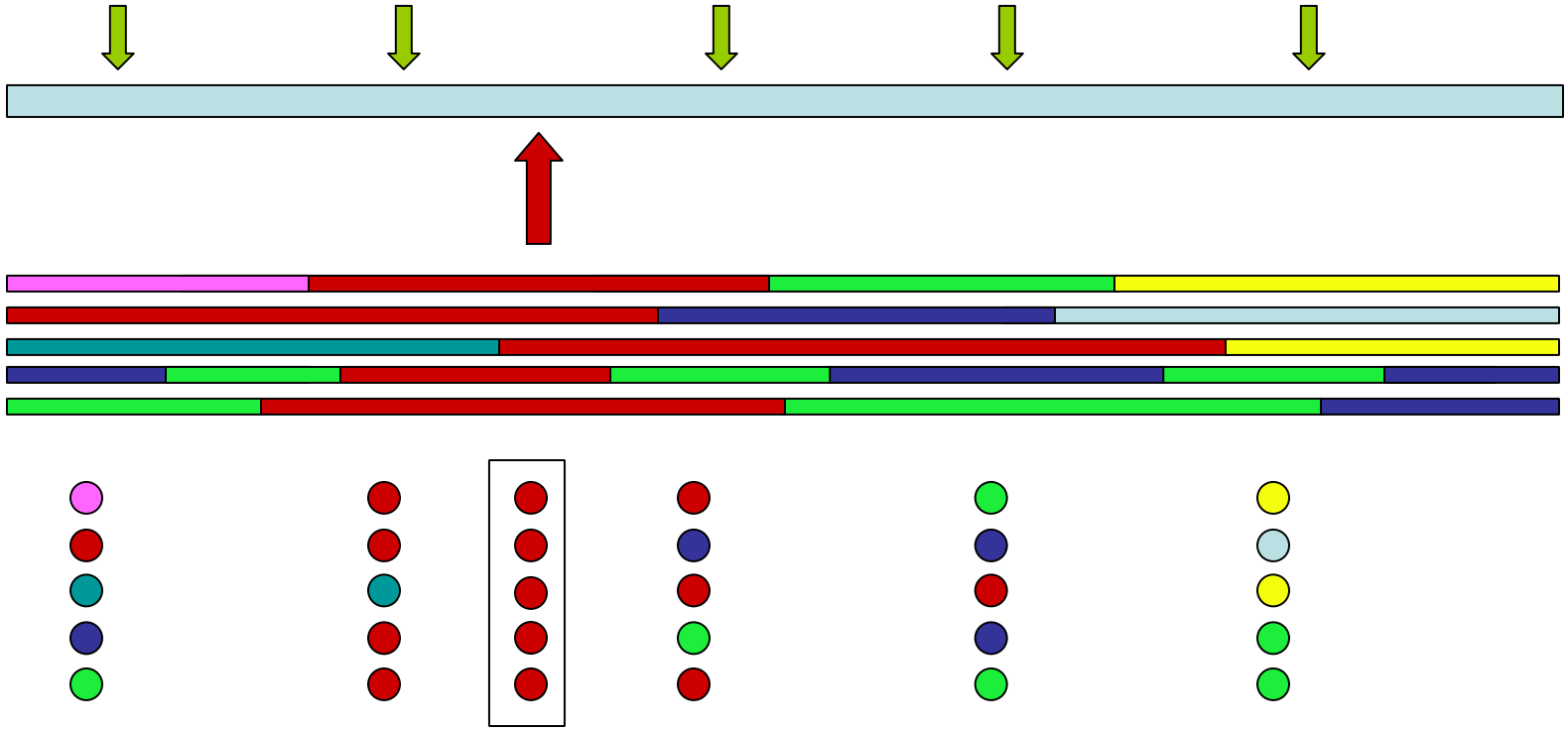
Mutation occurs on a ‘red’ chromosome



Mutation occurs on a ‘red’ chromosome

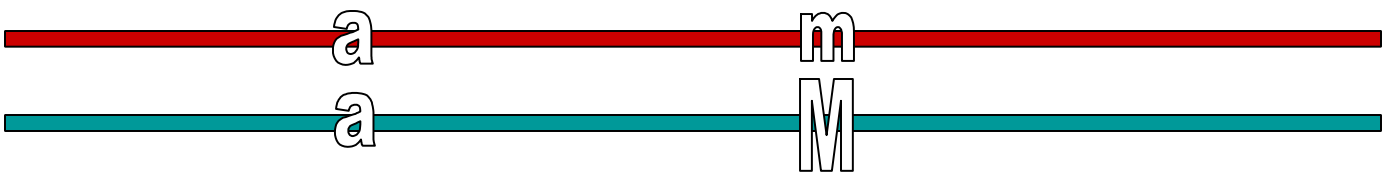


Association due to `linkage disequilibrium'



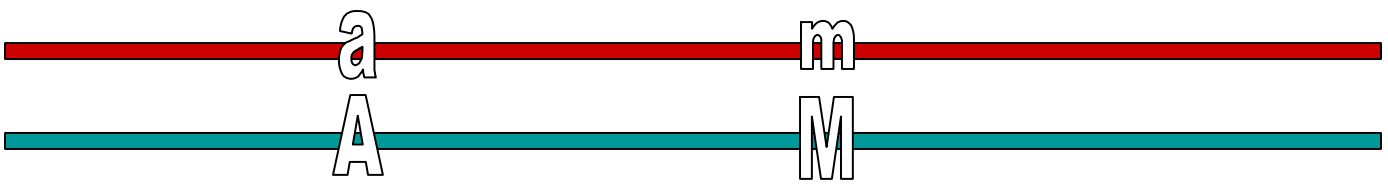
Haplotypes

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



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

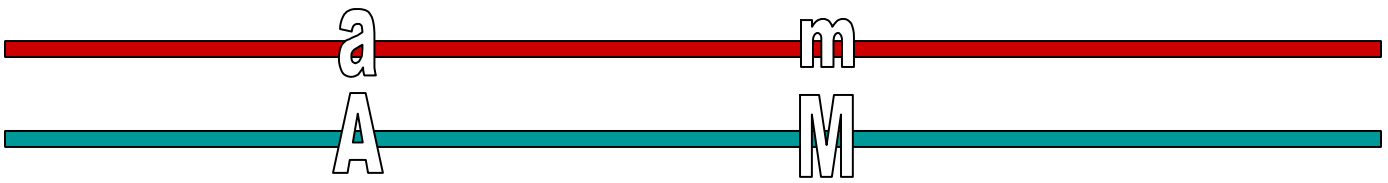
	A	***a***
M	***AM***	***aM***
m		***am***



This individual has ***Aa*** and ***Mm*** genotype

and ***AM*** and ***am*** haplotypes…

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



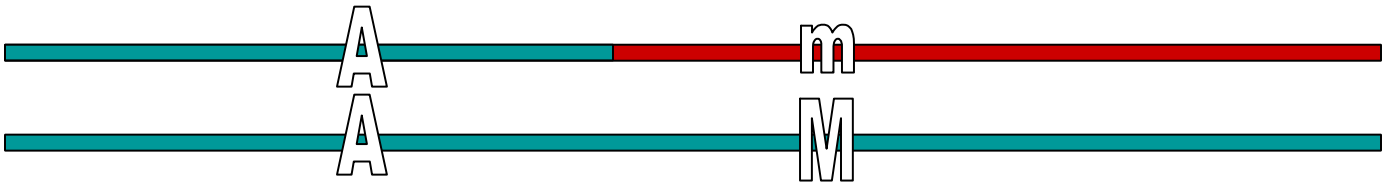
This individual has ***Aa*** and ***Mm*** genotype

and ***AM*** and ***am*** haplotypes…

<u>but</u> given only genotype data,

consistent with ***Am/aM*** as well as ***AM/am***

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



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

Haplotype analysis

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

Haplotype	Freq.	Odds Ratio
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 :

$AB / ab : 2 \times P_{AB} \times P_{ab} = 2 \times 0.25 \times 0.25 = 0.125$

$= 0.125/(0.125+0.125) = 0.50$

$Ab / aB : 2 \times P_{Ab} \times P_{aB} = 2 \times 0.25 \times 0.25 = 0.125$

$= 0.125/(0.125+0.125) = 0.50$

E-step

Incomplete data		**Complete data**	**Count**
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

M-step

Incomplete data		**Complete data**	**Count**
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 × 1 + 1 × 0.5 = 2.5

M-step

Incomplete data		**Complete data**	**Count**
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 × 0.5 = 0.5

M-step

Incomplete data		**Complete data**	**Count**
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 × 1 + 1 × 0.5 = 1.5

M-step

Incomplete data		**Complete data**	**Count**
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 × 1 + 1 × 0.5 + 2 × 1 = 3.5

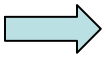
M-step

Haplotype counts, frequencies from complete data

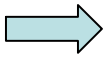
	Count	Freq
AB	2.5	0.3125
aB	0.5	0.0625
Ab	1.5	0.1875
ab	3.5	0.4375
Sum	8.0	1.0000

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 :

$AB / ab : 2 \times P_{AB} \times P_{ab} = 2 \times 0.3125 \times 0.4375 = 0.273$

= 0.273/(0.273+0.023) = 0.92

$Ab / aB : 2 \times P_{Ab} \times P_{aB} = 2 \times 0.1875 \times 0.0625 = 0.023$

= 0.023/(0.273+0.023) = 0.08

back to the M-step…

Incomplete data		Complete data	Count
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 × 1 + 1 × 0.92 = 2.92

back to the M-step…

Incomplete data		**Complete data**	**Count**
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 × 0.08 = 0.08

back to the M-step…

Incomplete data		**Complete data**	**Count**
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 × 1 + 1 × 0.08 = 1.08

back to the M-step…

Incomplete data		**Complete data**	**Count**
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 × 1 + 1 × 0.92 + 2 × 1 = 3.92

back to the M-step…

Haplotype counts, frequencies from complete data

	Count	Freq
AA	2.92	0.365
aB	0.08	0.010
Ab	1.08	0.135
ab	3.92	0.490
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

Genotype		**Phase**	**P(H|G)**
A/A	B/B	AB / AB	1.00
A/a	b/b	Ab / ab	1.00
A/a	B/b	AB / ab	**1.00**
		Ab / aB	**0.00**
a/a	b/b	ab / ab	1.00

Missing genotype data

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

Phase	P(H|G)
A**B**c / A**B**c	$(P_{ABc} \times P_{ABc}) / S$
A**B**c / A**b**c	$(2 \times P_{ABc} \times P_{Abc}) / S$
A**b**c / A**b**c	$(P_{Abc} \times P_{Abc}) / S$

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

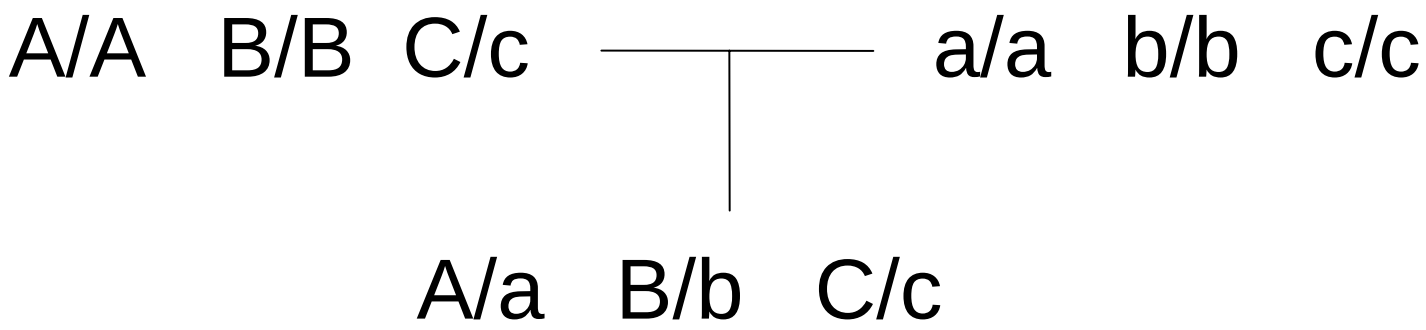
Using parental genotypes

Can often help to resolve phase

A/a B/b C/c

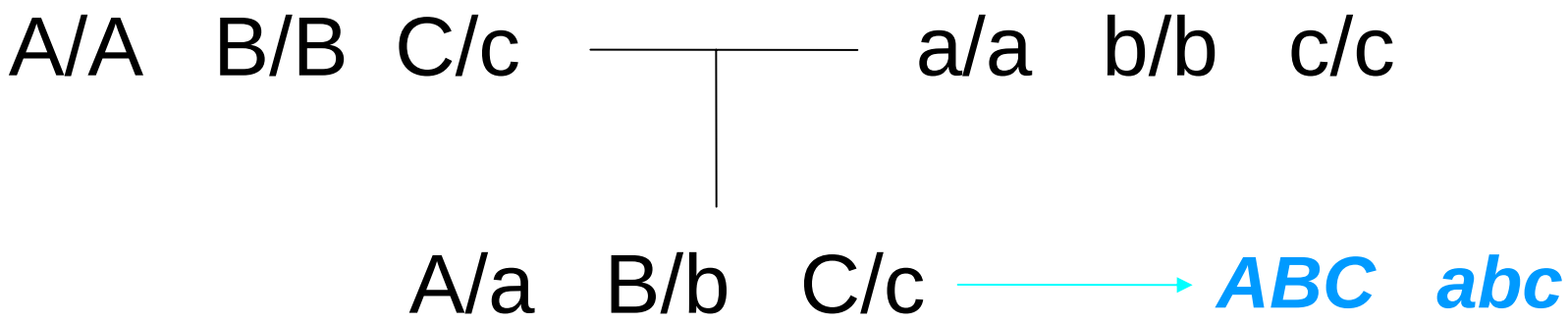
Using parental genotypes

Can often help to resolve phase



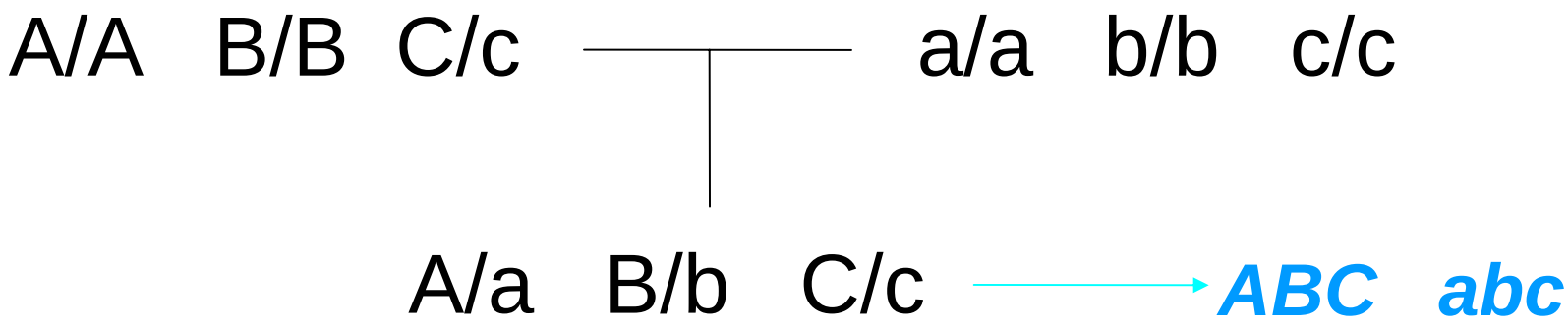
Using parental genotypes

Can often help to resolve phase

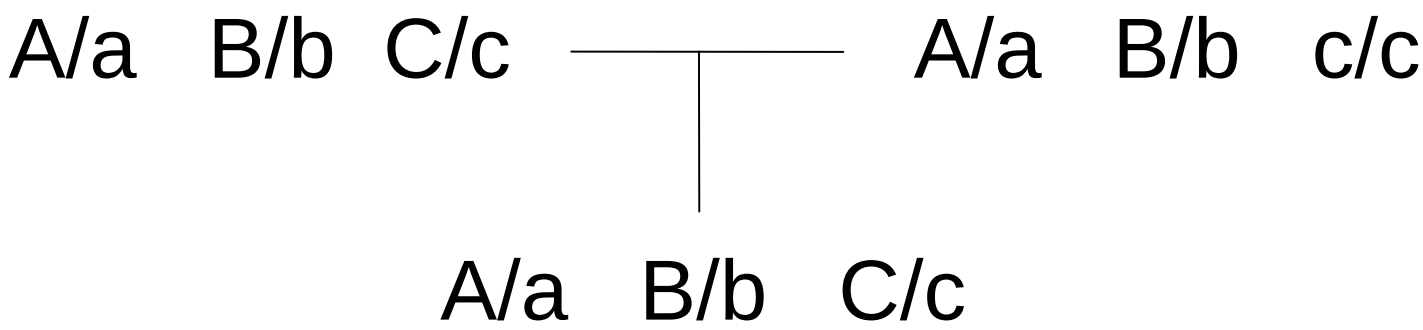


Using parental genotypes

Can often help to resolve phase



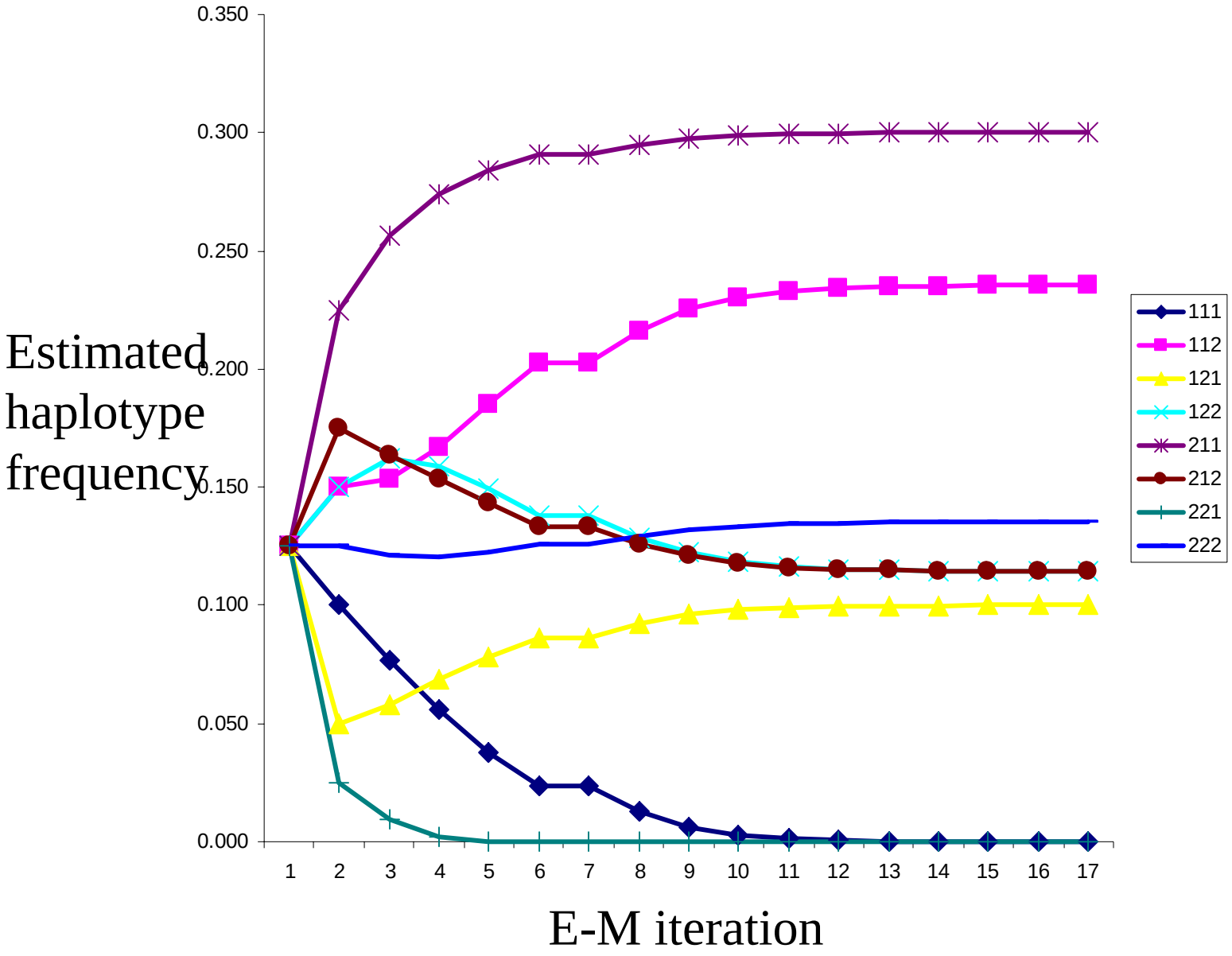
… but not always



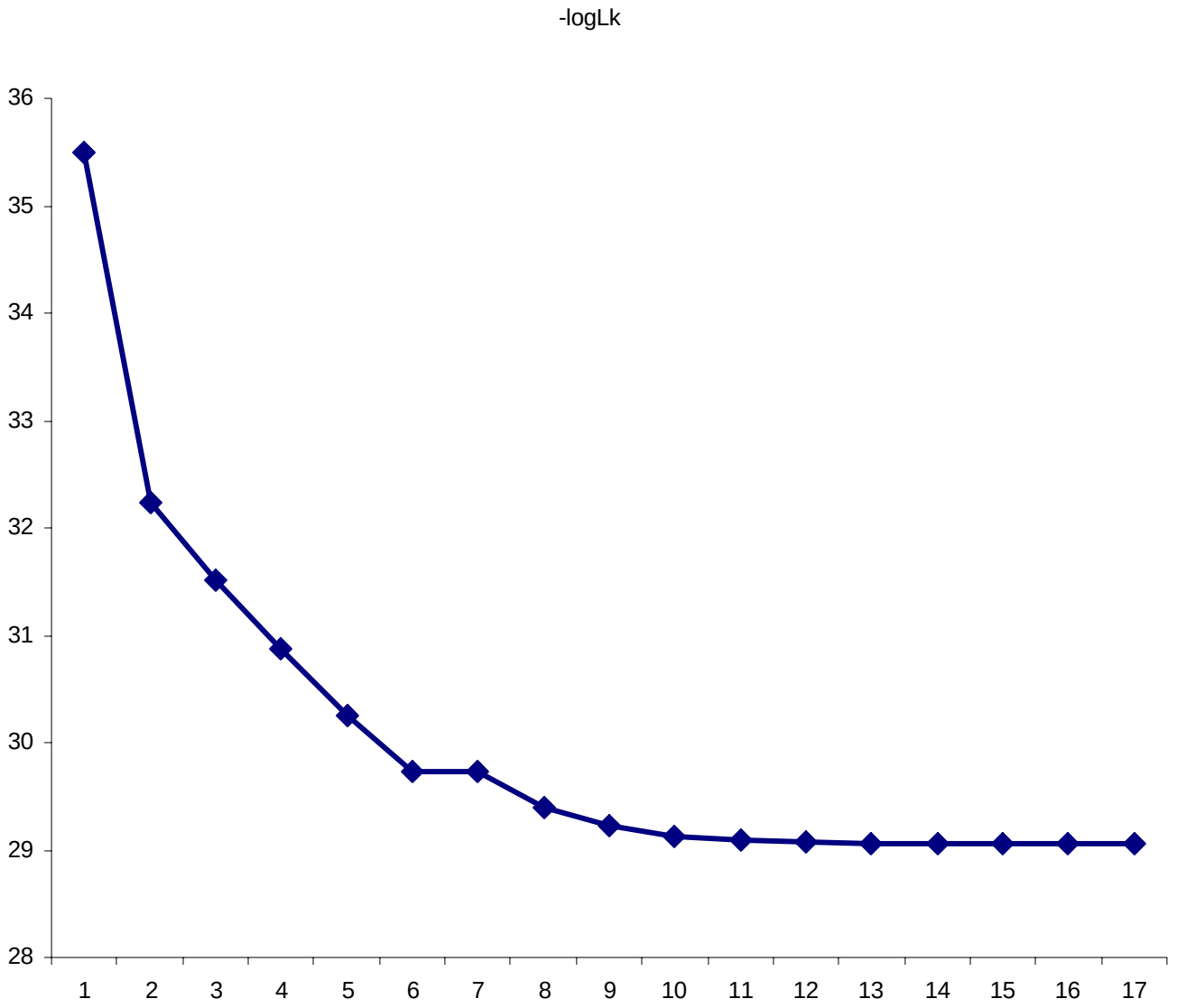
A (slightly) less trivial example

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



Haplotype frequencies

H	P(H)
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

	A	***a***	
M	*pr*	*ps*	*p*
m	*qr*	*qs*	*q*
	r	*s*	

Linkage disequilibrium

	A	*a*	
M	*pr + D*	*ps - D*	*p*
m	*qr - D*	*qs + D*	*q*
	r	*s*	

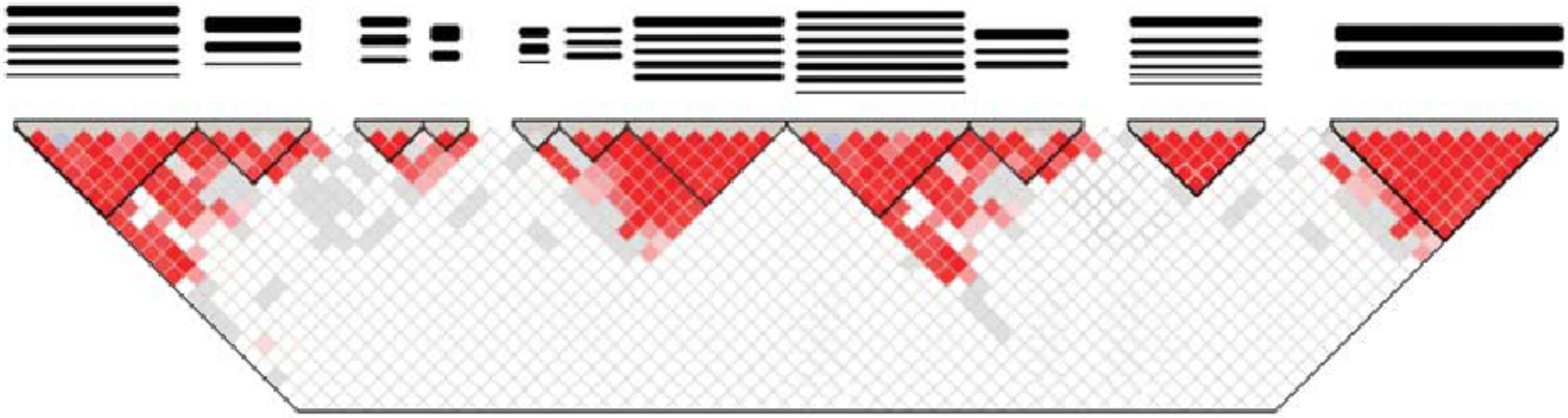
$D_{MAX} = Min(qs, pr)$

$D' = D / D_{MAX}$

$P(A)P(M)$

e.g $D = P(AM) -$

$r^2 = D^2 / pqrs$



Practical sessions

- Visualising data and testing for association in `Haploview`
- Detecting haplotpe 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 C  G G  A A
2_B 1 0 0 1  1  A C  C C  A C  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
```

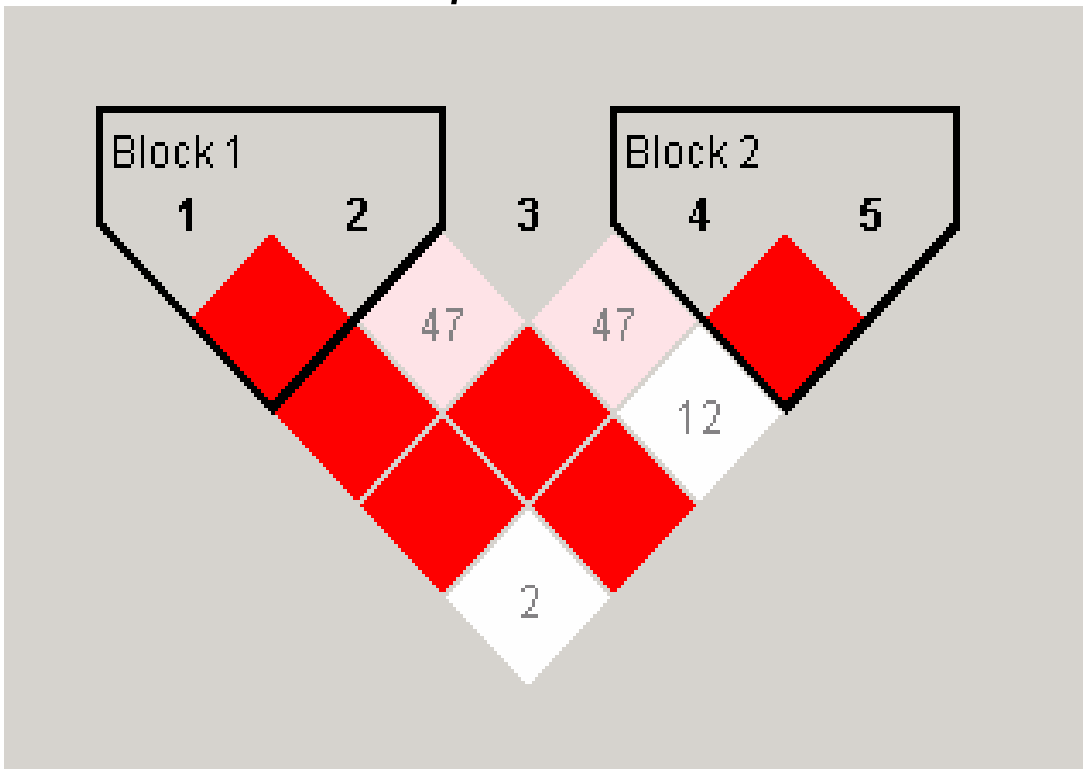
dataACGT.map

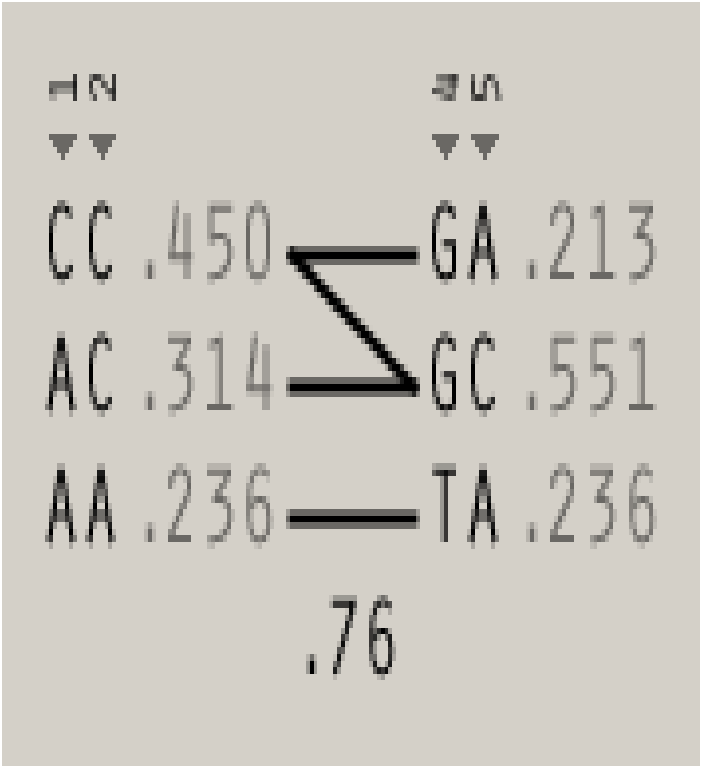
```
1 snp1 0 1
1 snp2 0 2
1 snp3 0 3
1 snp4 0 4
1 snp5 0 5
```

pedstats -p data1234.ped -d data123

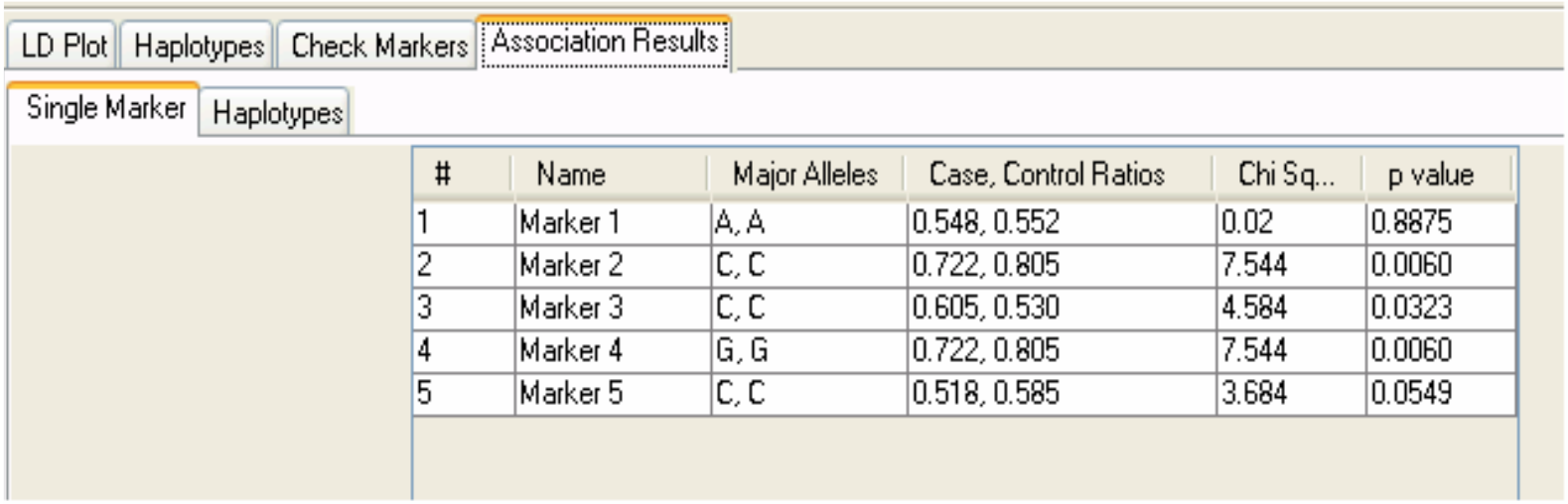
LD, block structure

Based on default "Gabriel blocks"





Single SNP association



#	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

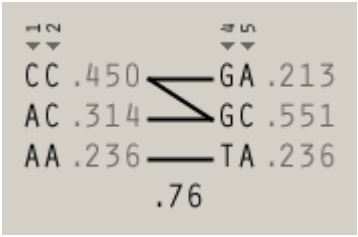
Haplotype	Freq.	Case, Control Ratios	Chi Square	p value
Haplotype Associations				
Block 1				
CC	0.450	0.452, 0.448	0.02	0.8875
AC	0.314	0.270, 0.358	7.112	0.0077
AA	0.236	0.278, 0.195	7.544	0.0060
Block 2				
GC	0.551	0.518, 0.585	3.684	0.0549
TA	0.236	0.278, 0.195	7.544	0.0060
GA	0.213	0.205, 0.220	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



CC	GA
AC	GC
AA	**TA**

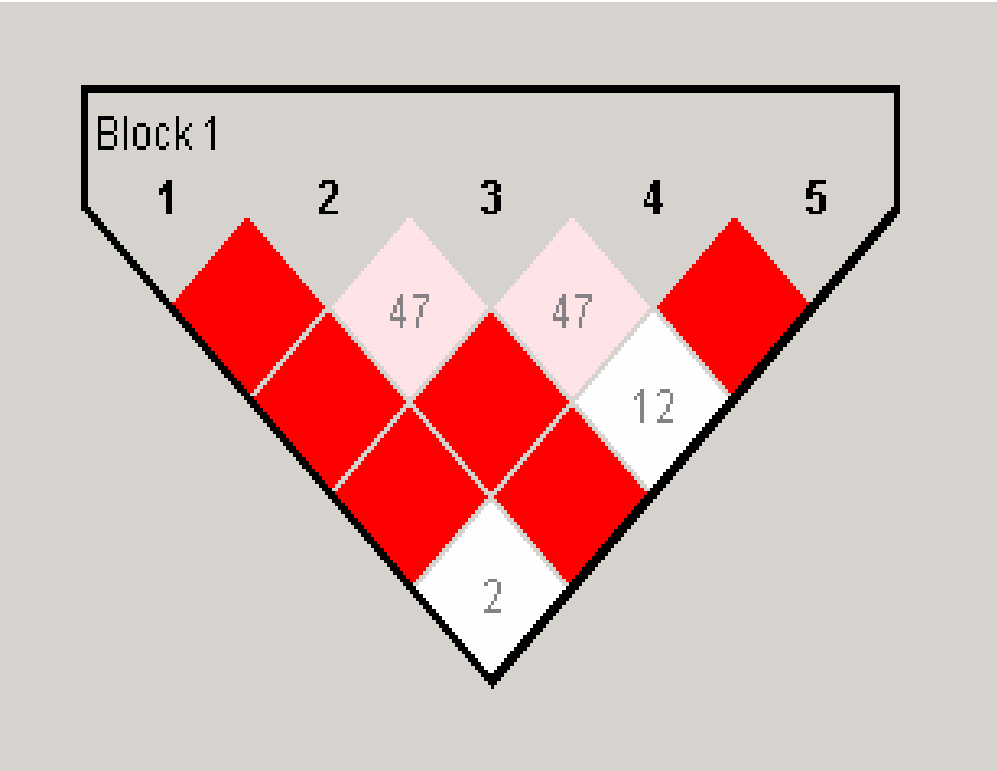
AA	**A**	**TA**	AAATA
AC	**A**	GC	ACAGC
CC	**C**	GA	CCCGA
CC	**C**	GC	CCCGC
AA	**C**	**TA**	**AACTA**
AC	**C**	GC	ACCGC

True model

Significantly associated with increased risk

Significantly associated with decreased risk

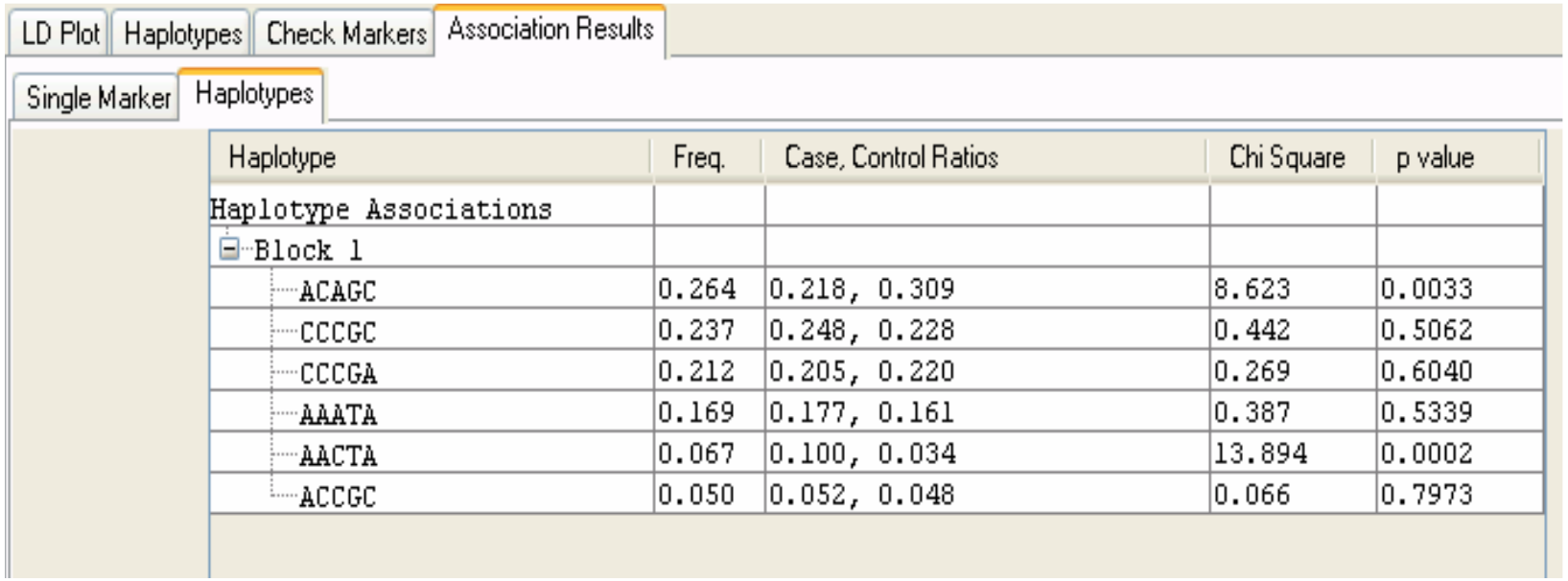
Manually specifying the 'block'



1 2 3 4 5

ACAGC .264
CCCGC .237
CCCGA .212
AAATA .169
AACTA .067
ACCGC .050

Results with 5-SNP block



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**

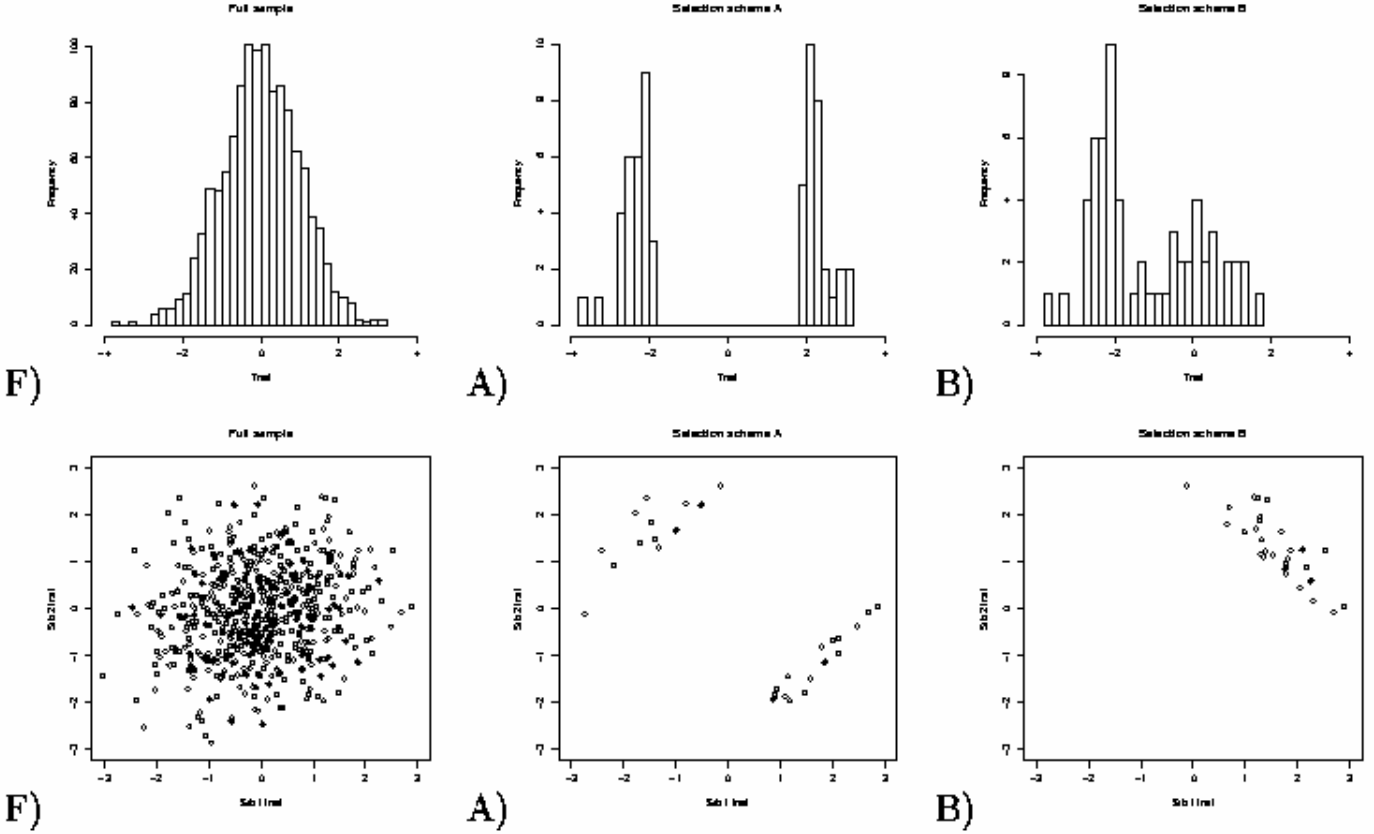
- Examples
  - AA from AAxAA W = 0
  - Aa from AAxAa W = -0.5
  - Aa from AaxAa W = 0

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

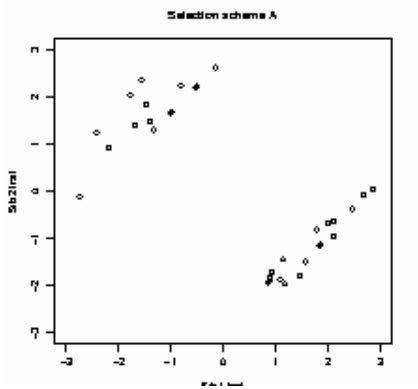
- Model likelihood of observing genotype conditional on trait value

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

  - Singletons:
    - **G** = { AA, Aa, aa }
  - Pairs:
    - **G** = { AA/AA, AA/Aa, AA/aa, … }
  - With parents:
    - **G** = { AA | AAxAA, AA | AAxAa, … }
    - **G** = { AA/AA | AAxAA, AA/AA | AAxAa, … }

Robust in selected samples

- Type I error rates
  - Sib pairs
  - 10% extreme selection
  - Within sibship test



	L(X|G)	L(G|X)
Full sample		
No parents	5.4	5.4
Parents	5.2	5.0
Selected sample		
No parents	26.7	5.3
Parents	13.8	5.0

Extension to haplotype analysis

- Probabilistic haplotype reconstruction via E-M algorithm

AA BB cc Dd

ABcD / ABcd $P(P_1) = 1$

AA Bb cc Dd

ABcD / Abcd $P(P_1) = 0$

ABcd / AbcD $P(P_2) = 0$

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
- β [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

$$\mathbf{g} = \begin{bmatrix} 1 & 1 & 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 2 & 1 & 2 \\ 1 & 1 & 2 & 2 & 2 & 1 \\ 2 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{bmatrix} = \begin{bmatrix} 2b_1 \end{bmatrix} \qquad \mathbf{L(G=g)} \begin{bmatrix} 1.0 \end{bmatrix}$$

B matrix

Vector of regression coefficients

Example **B** matrix

Individual *j* Genotypes: `1/1  1/2  1/2`

Haplotypes: `112 / 121` P() = 0.8

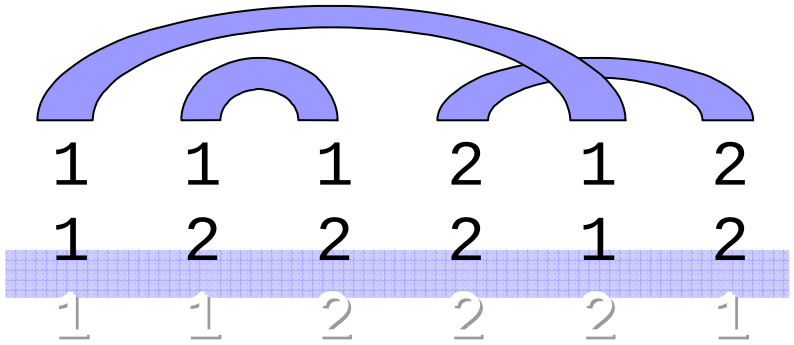
`111 / 122` P() = 0.2

$$\begin{matrix} 1 & 1 & 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 2 & 1 & 2 \\ 1 & 1 & 2 & 2 & 2 & 1 \end{matrix}$$

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

Testing nested hypotheses

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



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

$$b_1 = b_5$$
$$b_2 = b_3$$
$$b_4 = b_6$$

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 $P(P_C \mid P_{P,M})$
  - 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

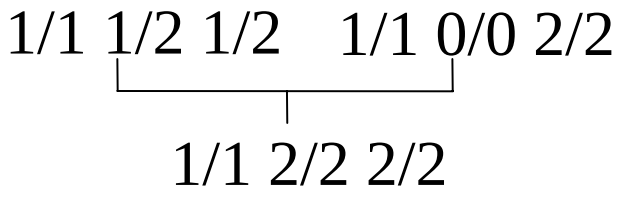
Between/within partitioning

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 ]
```

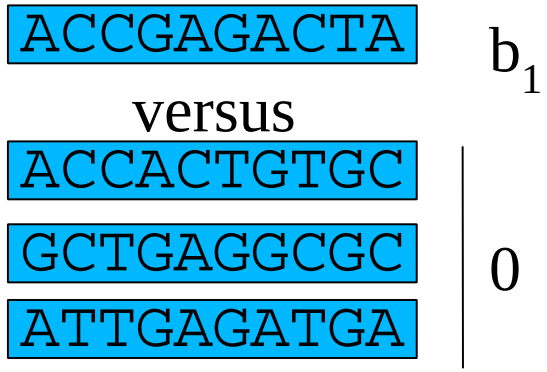
B **W**

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*

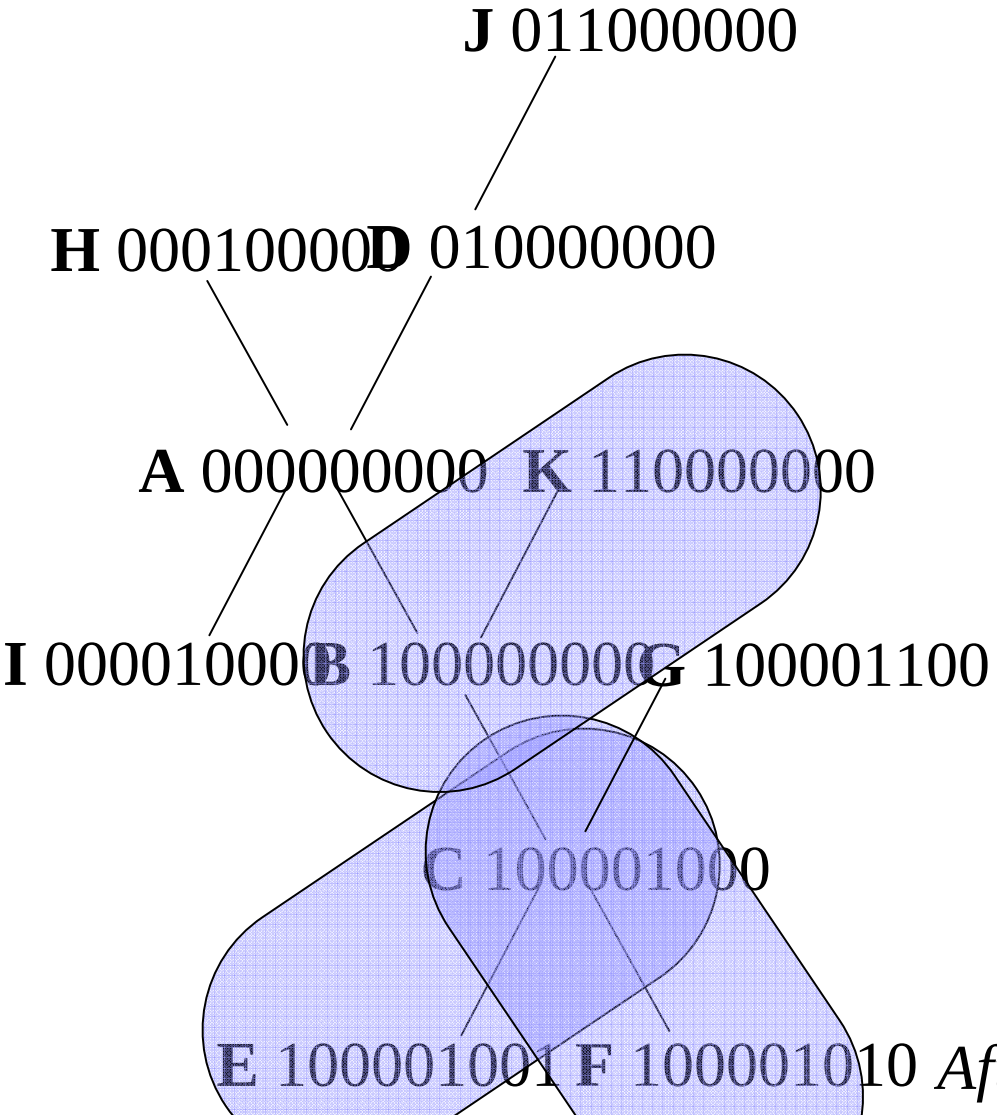


- *Omnibus test*
  - *single test with H-1 df*
  - *compare each haplotype against an (arbitrary) reference haplotype*
  - *built-in correction for multiple tests*

Secondary analysis

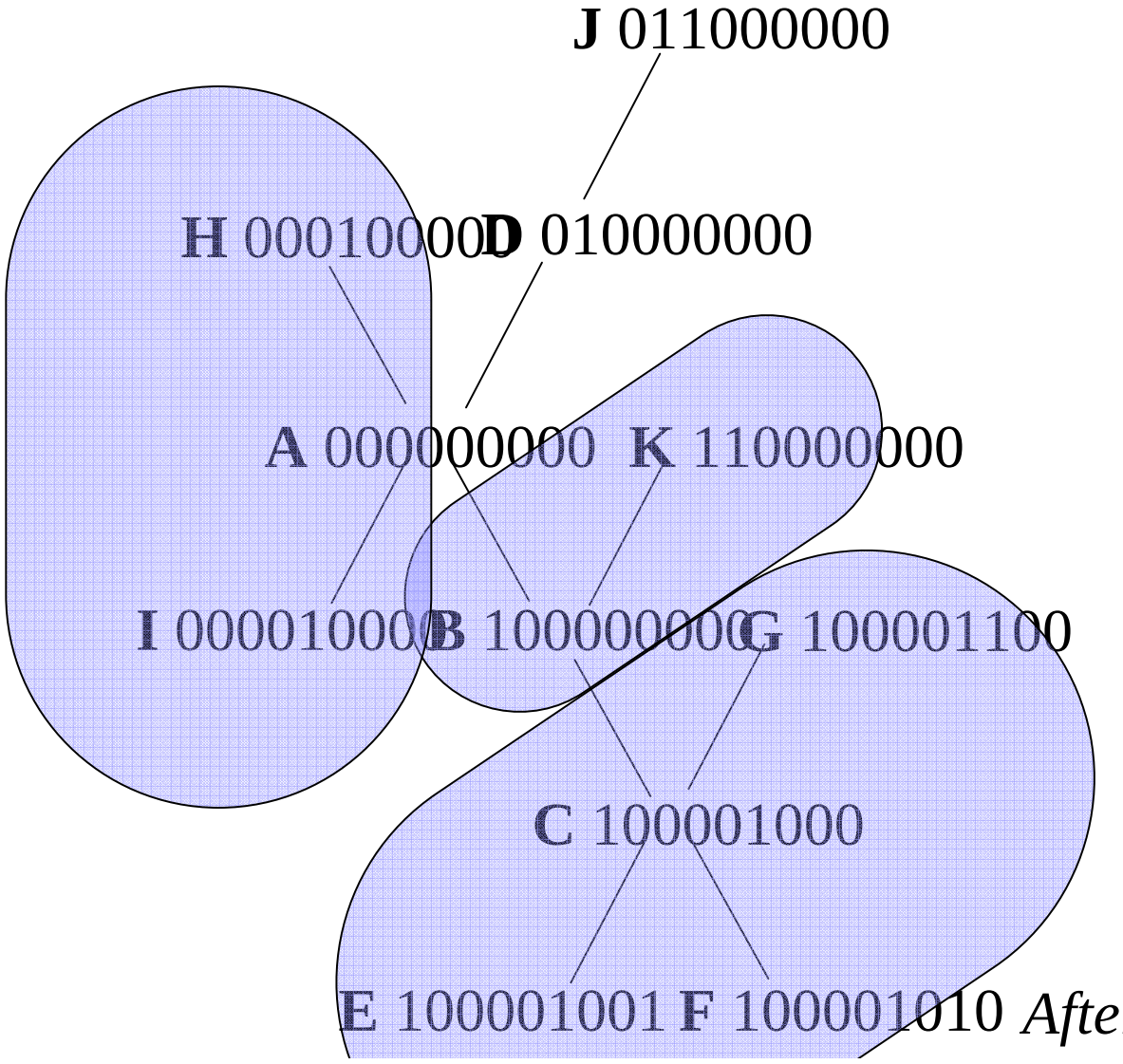
- *H* haplotypes will have *H*-1 coefficients
  - Reduces power of test – high degrees of freedom
- More similar haplotypes should have more similar effects

Cladogram-collapsing



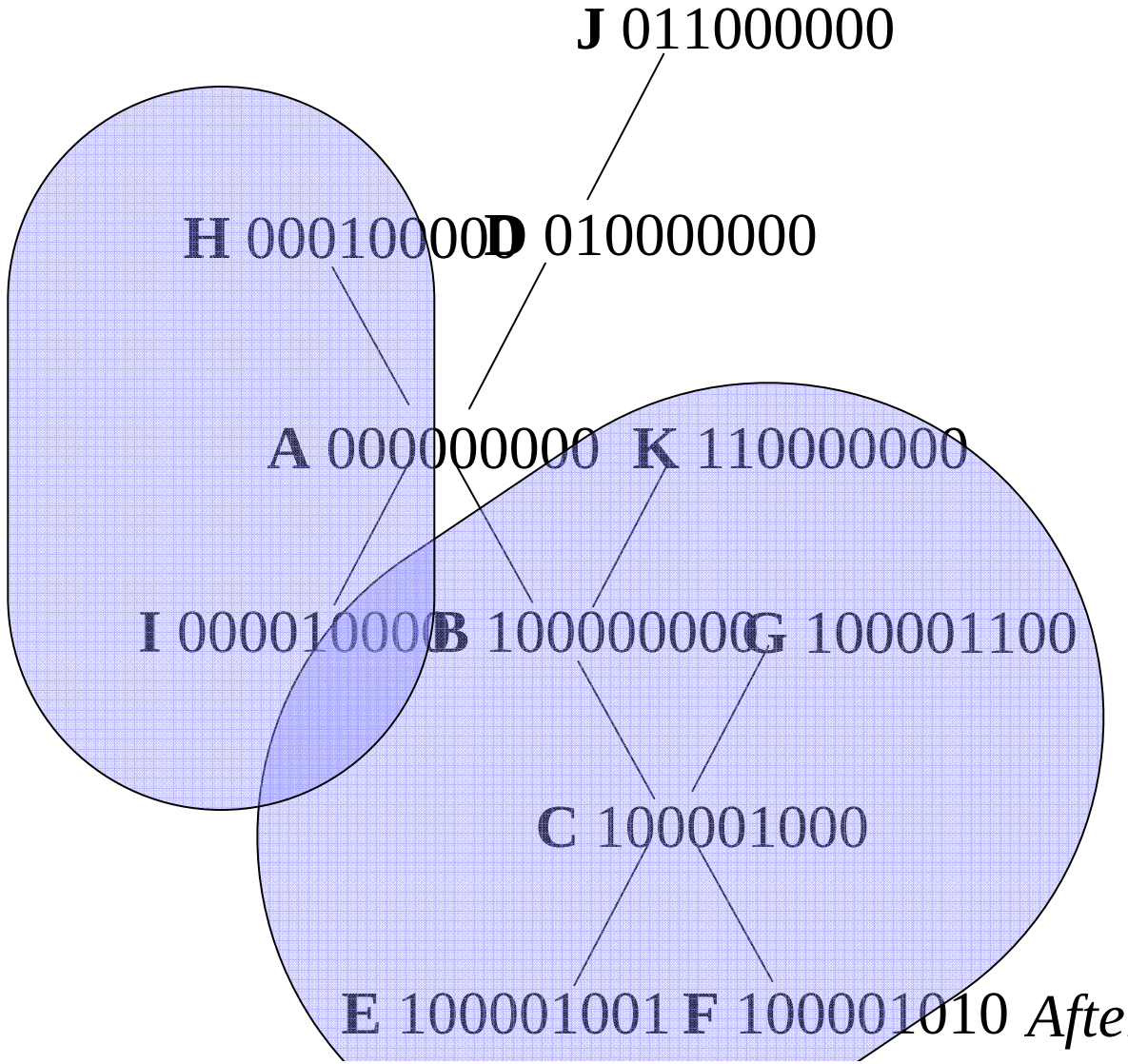
After Seltman et al (

Cladogram-collapsing



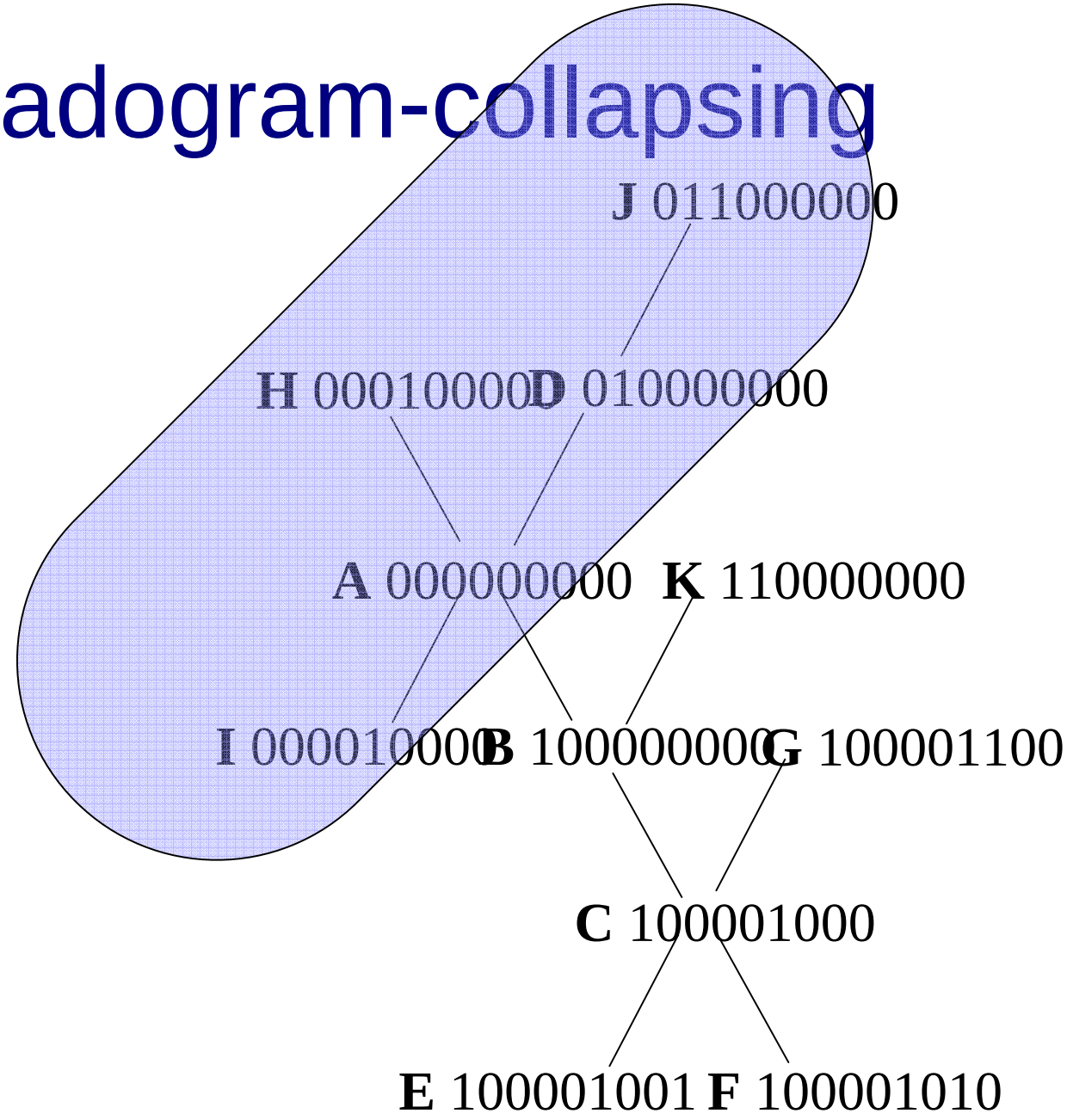
After Seltman et al (

Cladogram-collapsing

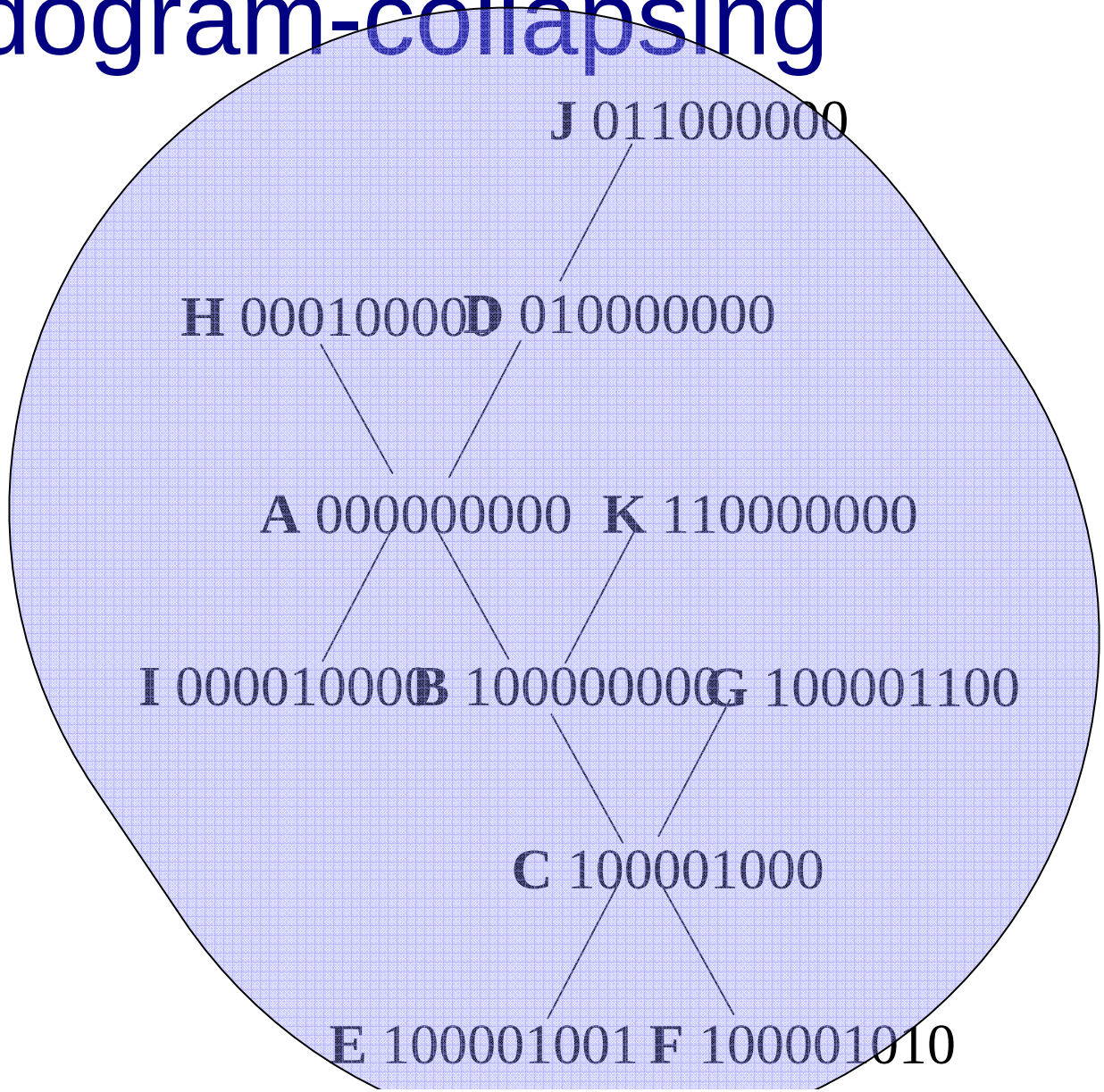


After Seltman et al (

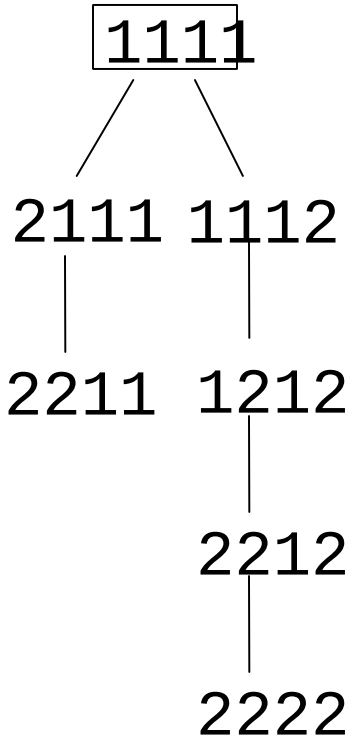
Cladogram-collapsing

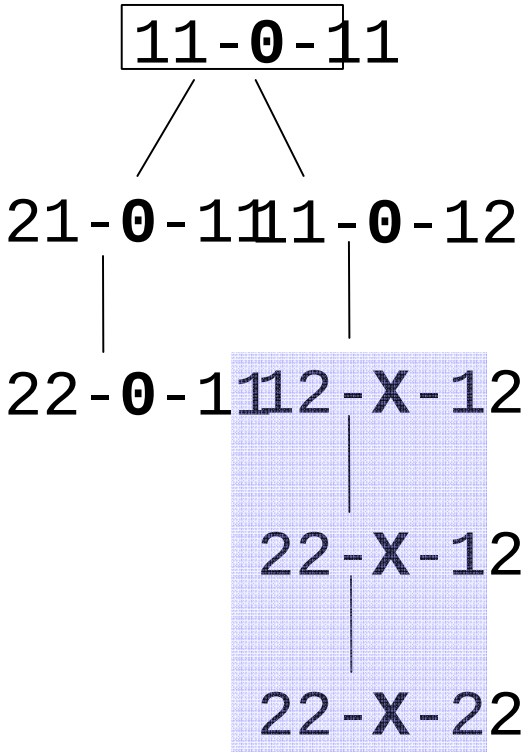


Cladogram-collapsing



Secondary analysis



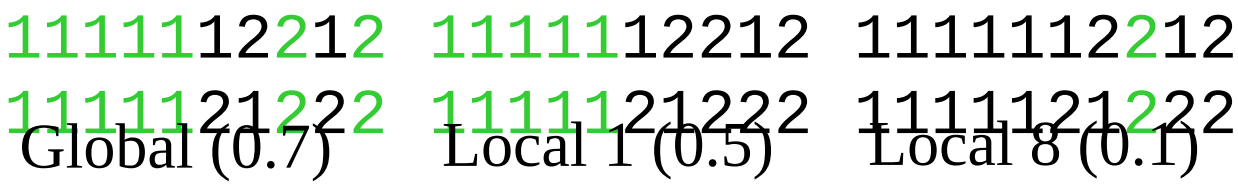


Secondary analysis

Haplotype	Estimated coefficients
2211	*0.000*
2111	-0.092
1111	0.102
1112	-0.234
1212	0.634
2212	0.332
2222	0.865

Secondary analysis

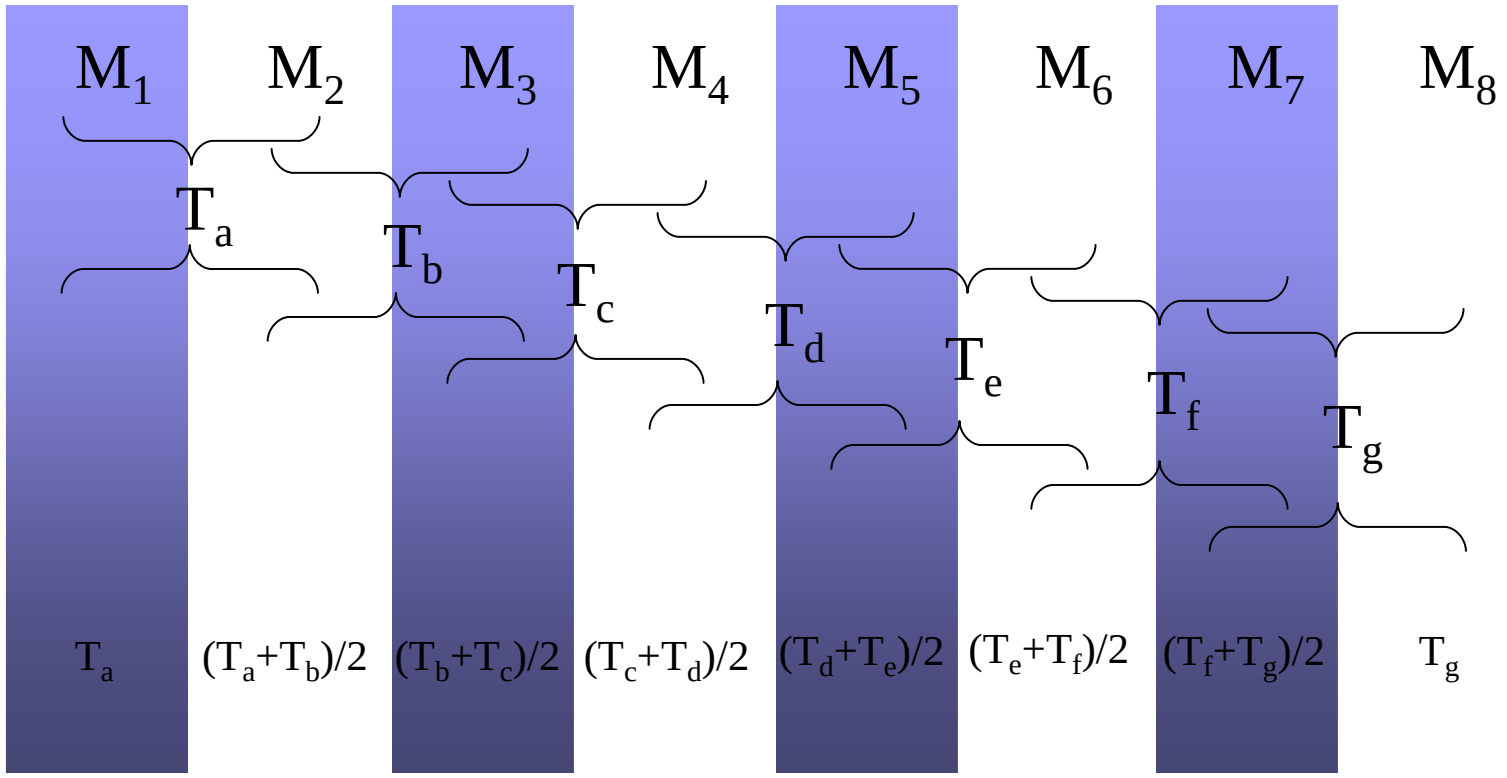
- Haplotype similarity
  - Global and local identity



- 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/~sha

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   [1]      0.000            0.000   [1]
2112121   0.169   -0.249          -0.249  [2]      0.000            0.000   [1]
2221211   0.122   -0.417          -0.417  [3]      0.000            0.000   [1]
2212222   0.115   -0.419          -0.419  [4]      0.000            0.000   [1]
2122222   0.112   0.044           0.044   [5]      0.000            0.000   [1]
1112121   0.099   -0.213          -0.213  [6]      0.000            0.000   [1]
2222221   0.041   0.115           0.115   [7]      0.000            0.000   [1]
2212221   0.029   -0.662          -0.662  [8]      0.000            0.000   [1]
---       -----                   ------                            -------
                                  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
```

	Haplotype	*Freq*	*B & W coeffs*		*Chi-sq*	*p*
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
whap --file dataACGT --phase > phase.out
```

Just print out phases
or send to a file

- Single SNP analysis

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

Analyse 1st SNP
Analyse 5th SNP
Sliding window
+ empirical p-values

- Haplotype analysis

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

Omnibus test
As above
All haplotype-specific t

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

AA**A**TA / AC**C**GC 0.324 AA**C**TA / AC**A**GC 0.676

1_A	1	1	ACCGC ACAGC	1.000
2_A	**1**	**1**	**AACTA ACAGC**	**0.676**
2_A	**1**	**2**	**AAATA ACCGC**	**0.324**
3_A	1	1	ACAGC AAATA	1.000
4_A	1	1	AAATA AACTA	1.000
5_A	**1**	**1**	**ACAGC AACTA**	**0.676**
5_A	**1**	**2**	**ACCGC AAATA**	**0.324**
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

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
```

Empirical p-values, corrected for multiple testing

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
```

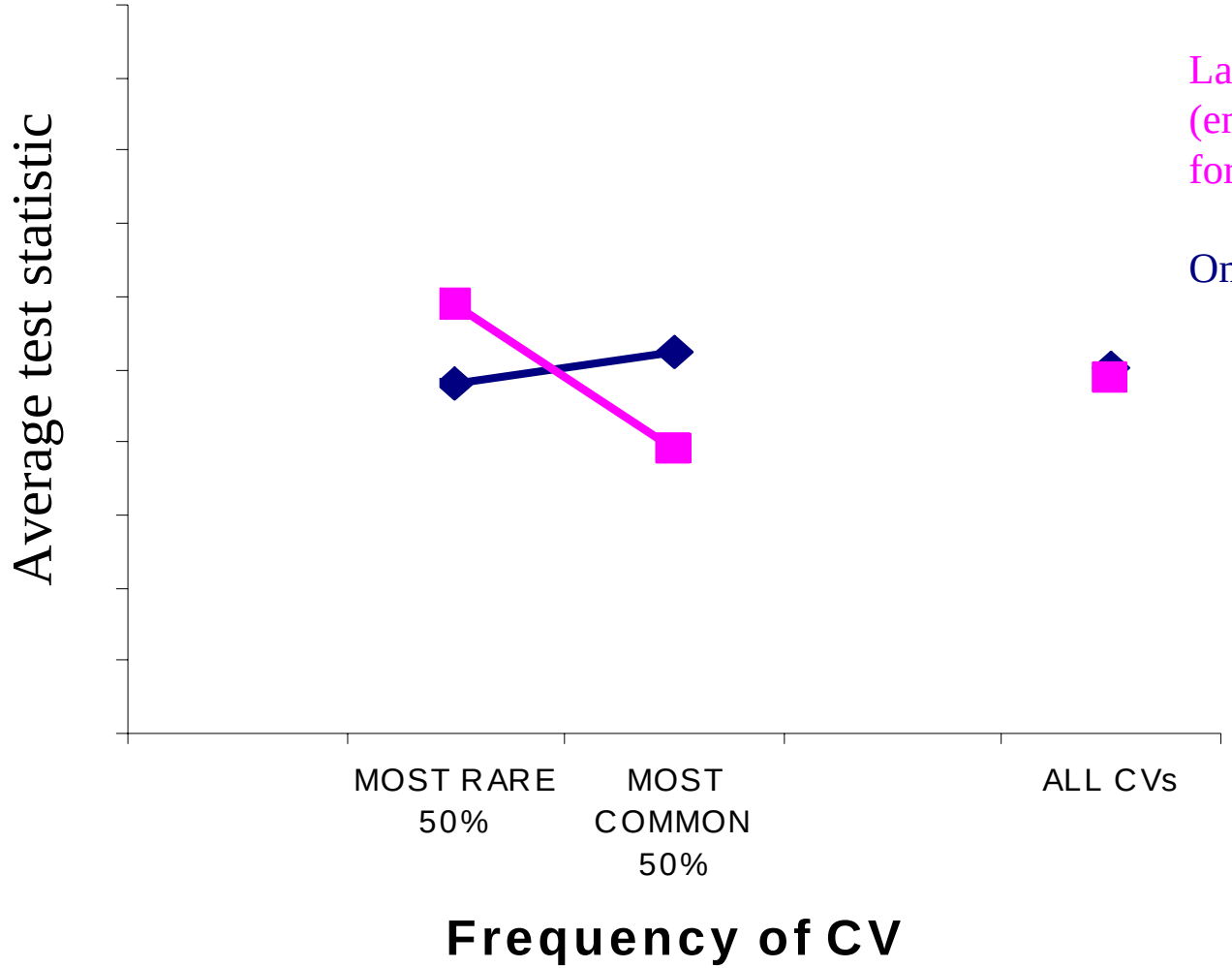
Hap	Freq	Alt(B)	Alt(W)		Null(B)	Null(W)	
ACAGC	0.264	0.000	0.000	[1]	0.000	0.000	[1]
CCCGC	0.237	0.406	0.406	[2]	0.000	0.000	[1]
CCCGA	0.212	0.269	0.269	[3]	0.000	0.000	[1]
AAATA	0.169	0.383	0.383	[4]	0.000	0.000	[1]
AACTA	**0.067**	**1.338**	**1.338**	[5]	0.000	0.000	[1]
ACCGC	0.050	0.424	0.424	[6]	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

Haplotype	*Chi-sq(1df)*	*p-value*	*beta*	*OR*
ACAGC	**8.546**	**0.00346**	**-0.472**	**0.62**
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
AACTA	**13.929**	**0.00019**	**1.128**	**3.08**
ACCGC	0.073	0.787	0.092	1.09

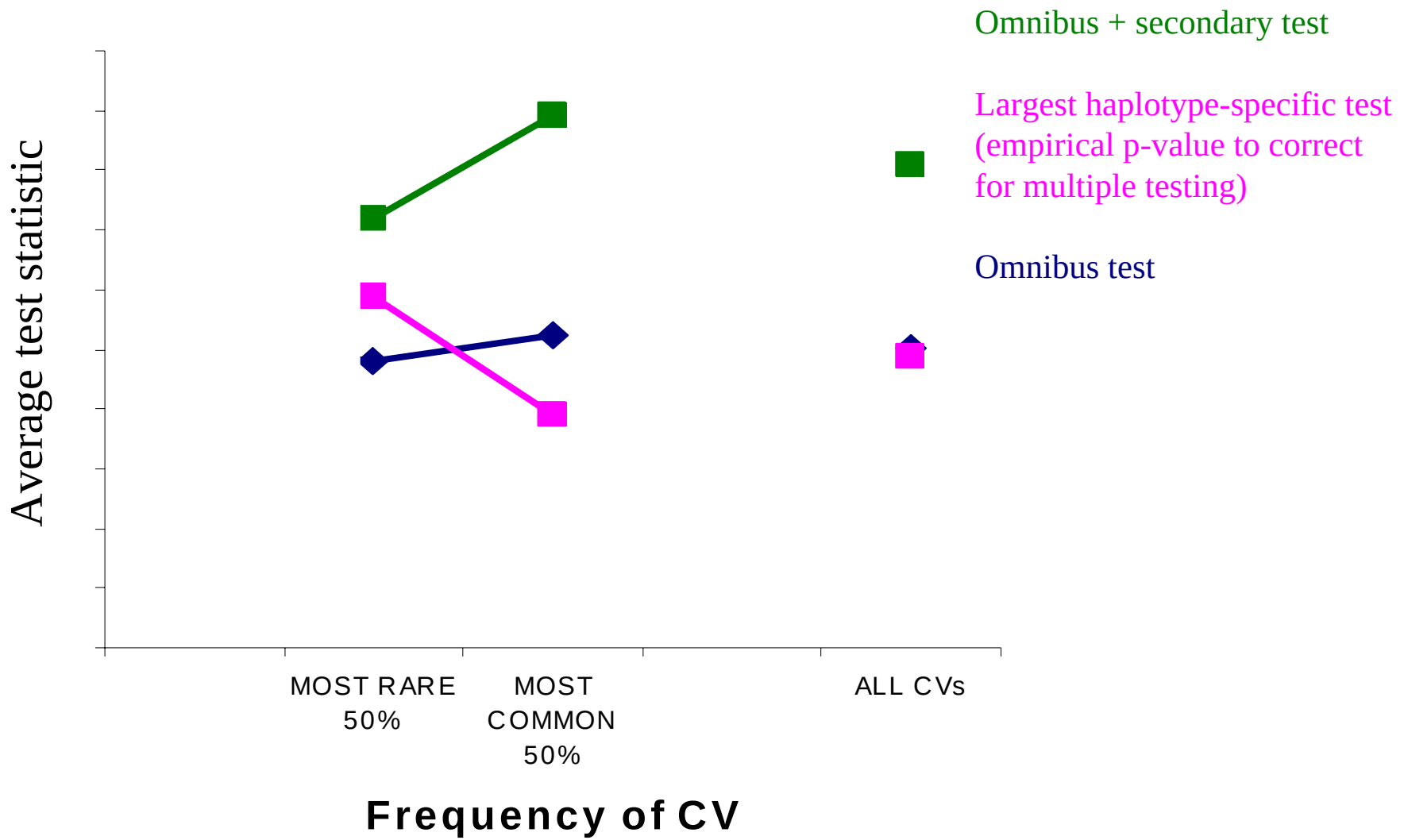
Haplotype-specific or omnibus?



Largest haplotype-specific test (empirical p-value to correct for multiple testing)

Omnibus test

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*?"

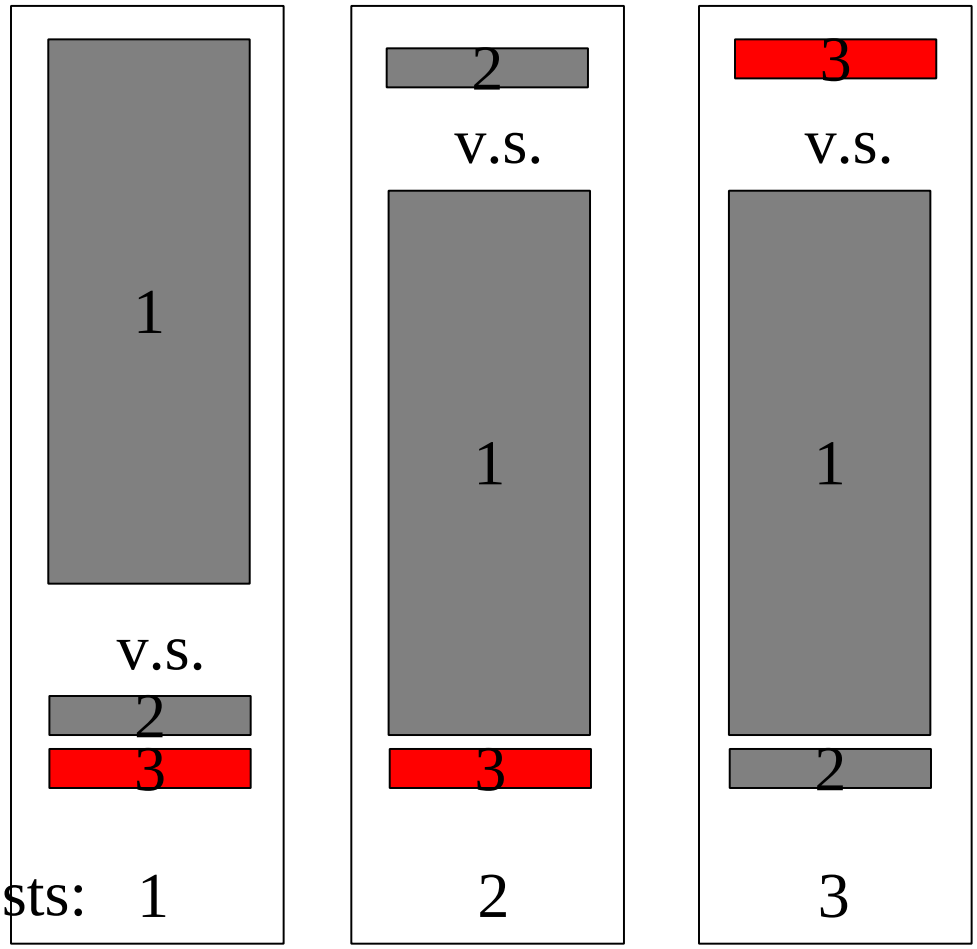
Interpreting effects

True model

1	AACG	90%
2	GGAC	05%
3	AAAC	05%

Looks like

1	AACG	90%
2	GGAC	05%
3	AAAC	05%



Interpreting effects

True model

1	AACG	50%	*strong effect*
2	GGAC	40%	
3	AAAC	10%	*mild effect*

Under an omnibus test

1	AACG	OR = 1.0
2	GGAC	OR = 0.4
3	AAAC	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`

1111 [1]	11**11** [1]
1122 [2]	11**22** [2]
2221 [3]	22**21** [3]
2222 [4]	22**22** [2]
2211 [5]	22**11** [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?

Haplotype	Freq	Odds ratio
AB	0.50	1.00 (fixed)
ab	0.45	2.00
Ab	0.05	?

Alt	Null
[1]	[1]
[2]	[2]
[3]	[2]

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

Conditional tests

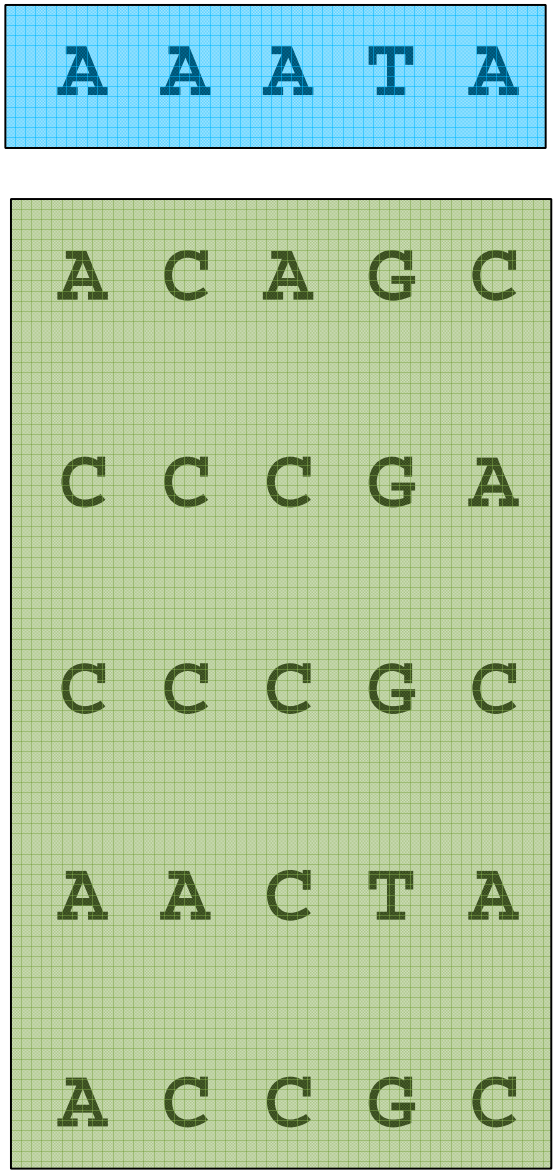
- Assuming significant omnibus test:
  - can we make it go away?

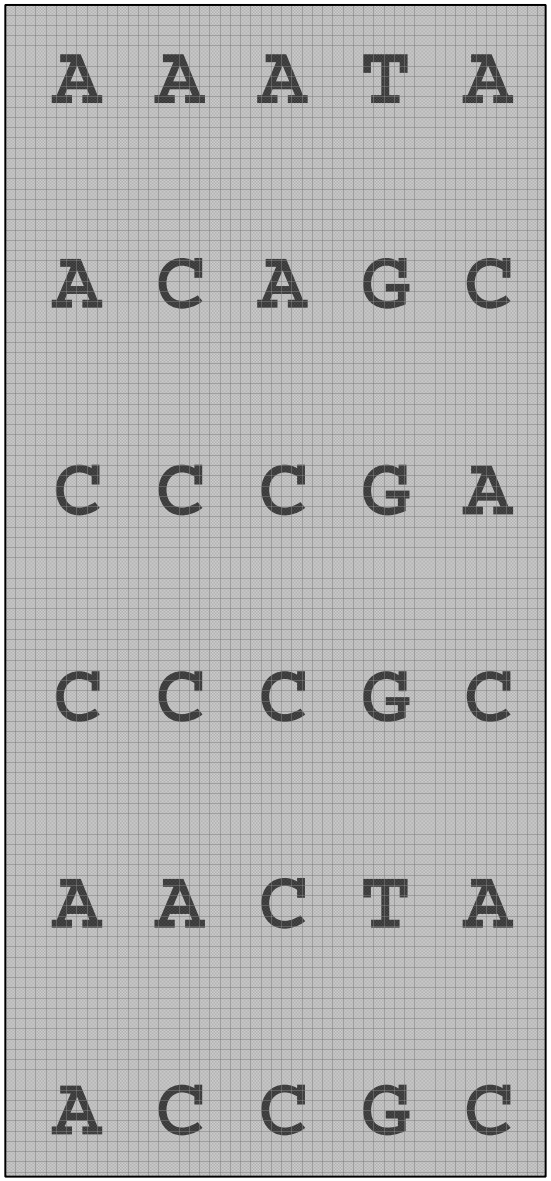
- *X* independently contributes (if signif.)
  - `--alt 1,2,3,4,5  --null 2,3,4,5`
  - "independent effect test"

- *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)

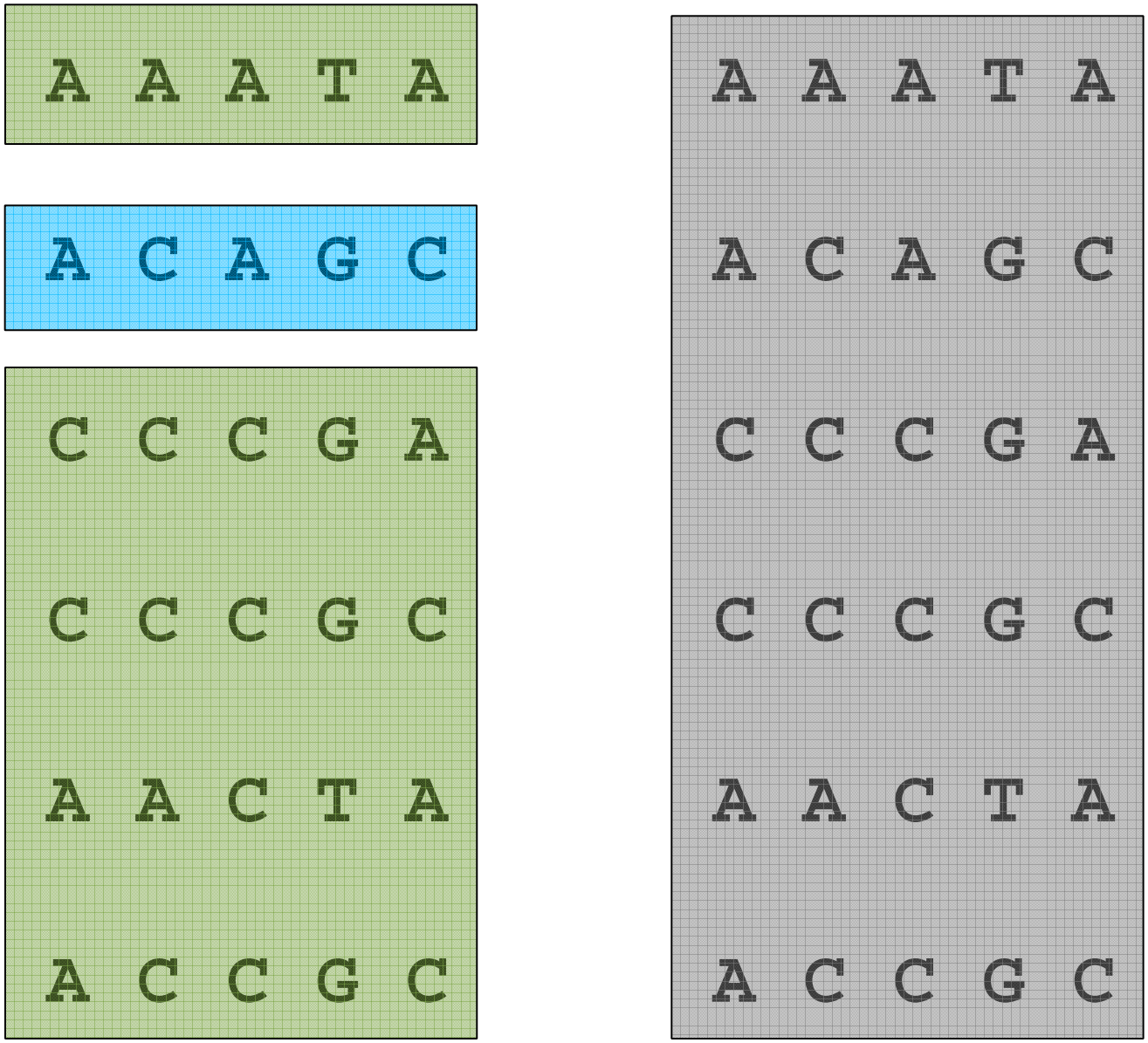
```
--constrain 1,2,2,2,2,2 / 1,
```





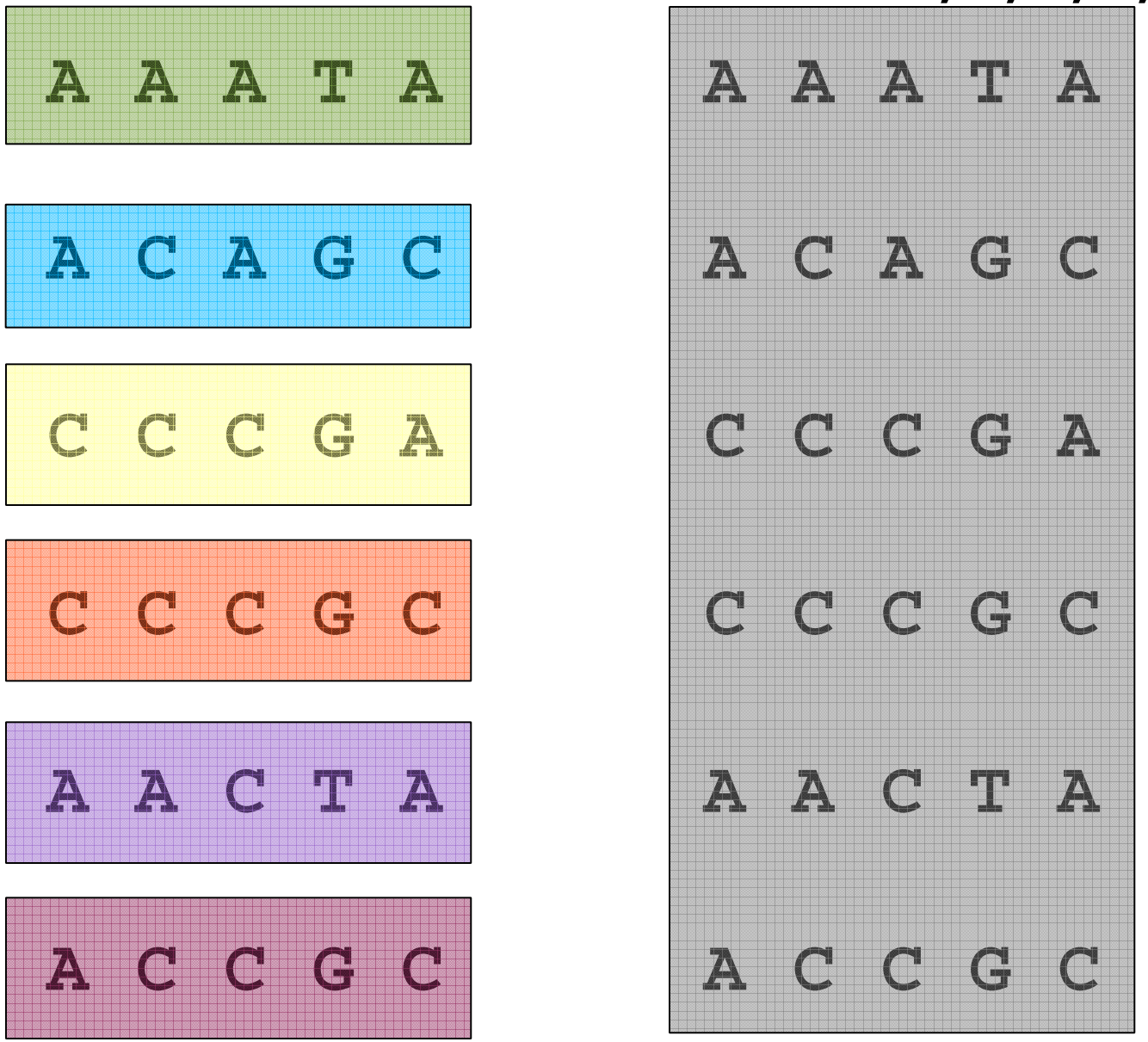
Haplotype-specific test (H2)

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



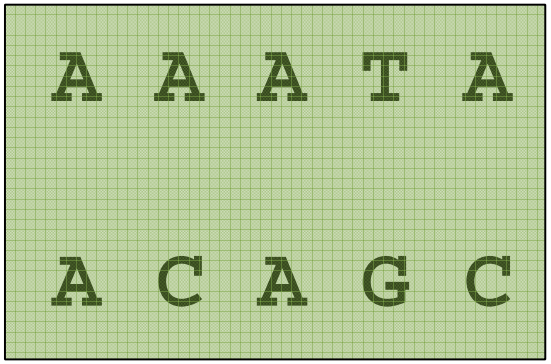
Omnibus test (df=5)

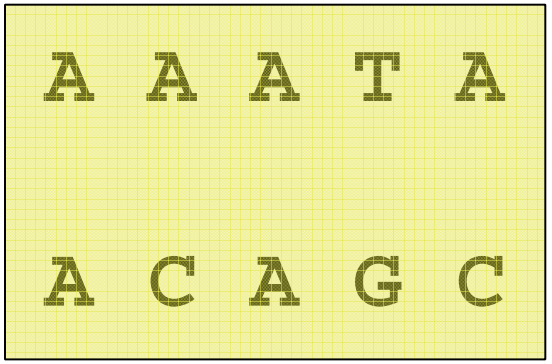
```
--constrain 1,2,3,4,5,6 / 1,1
```

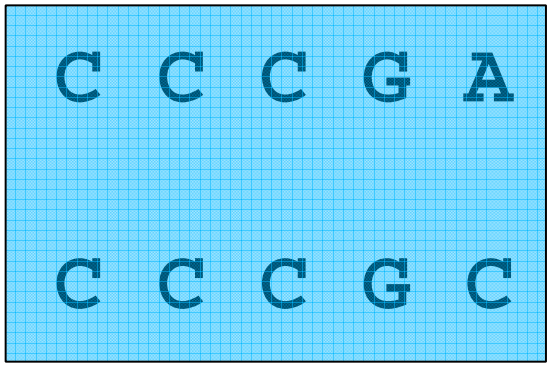


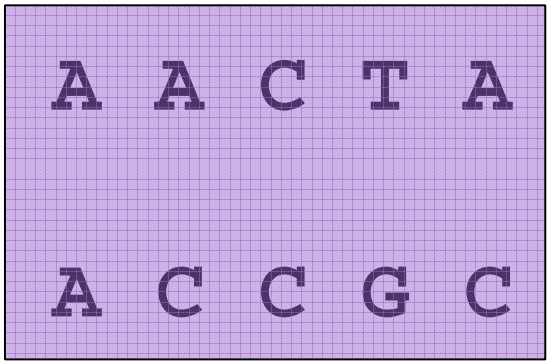
Clade-based homogeneity test (1df)

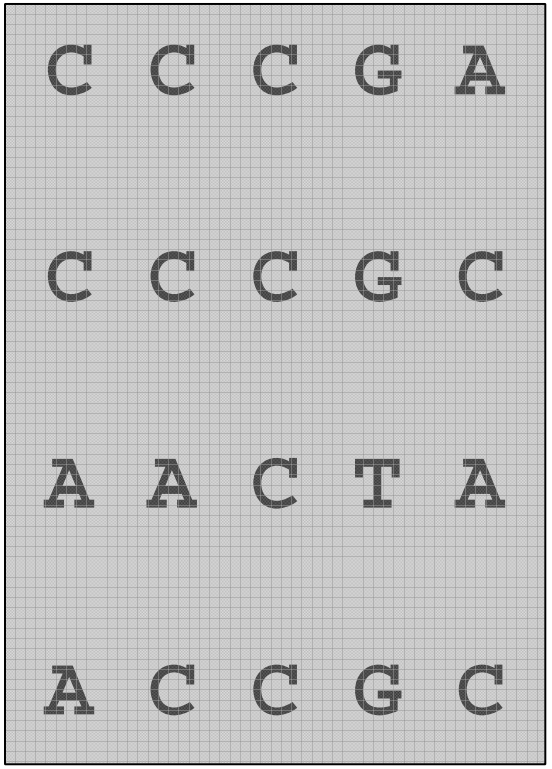
```
--constrain 1,1,2,2,3,3 / 1,1
```



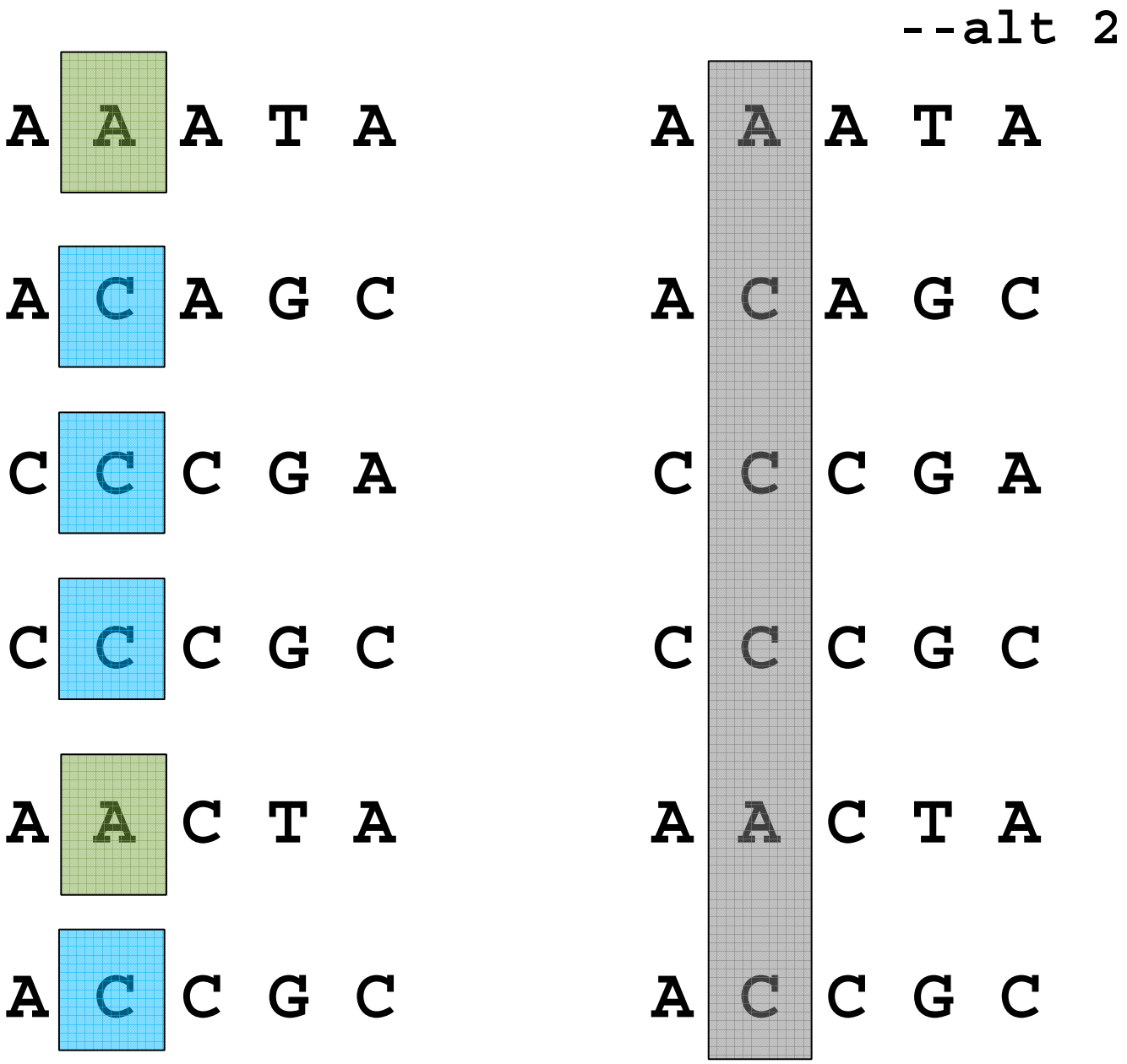




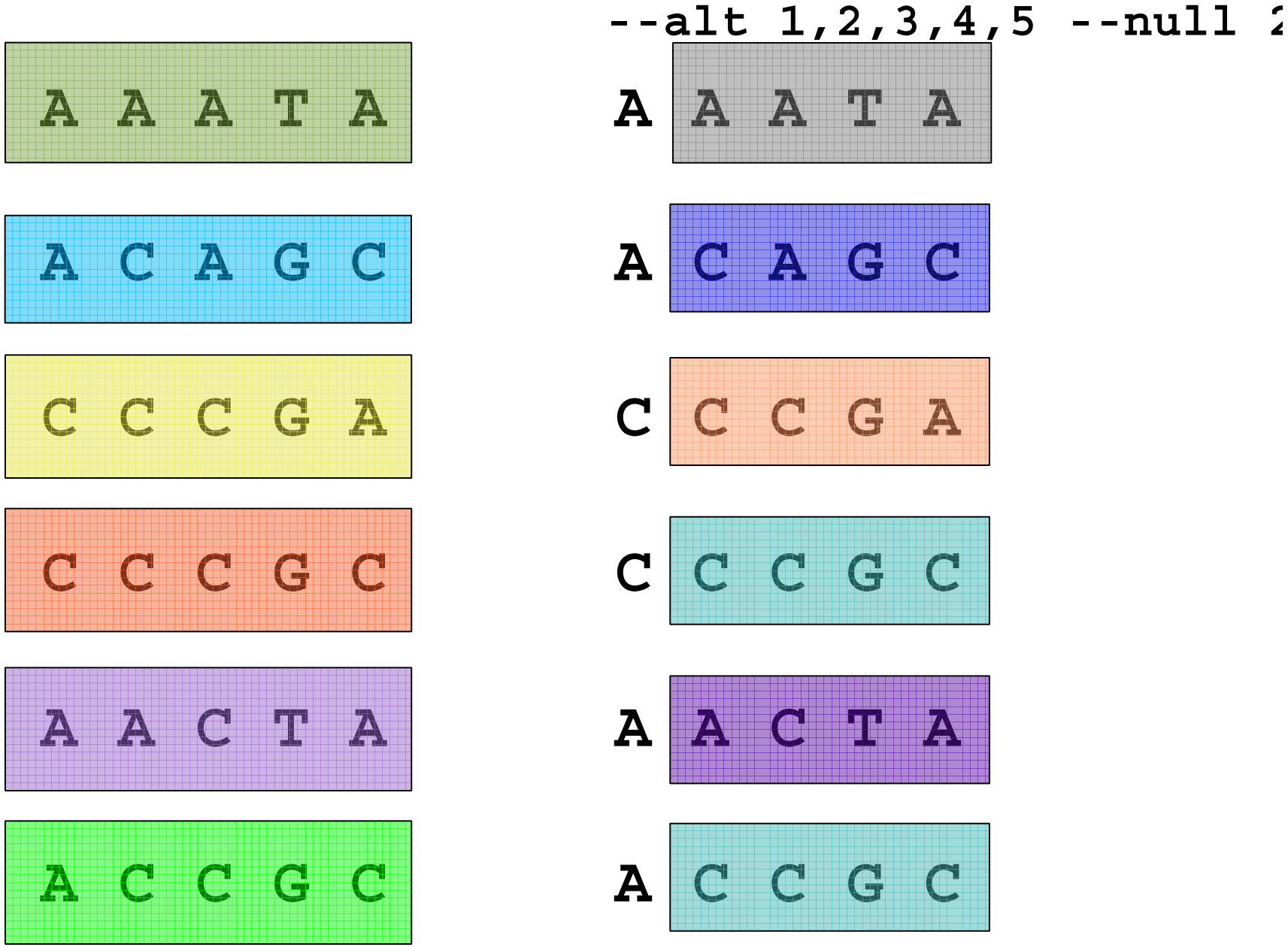




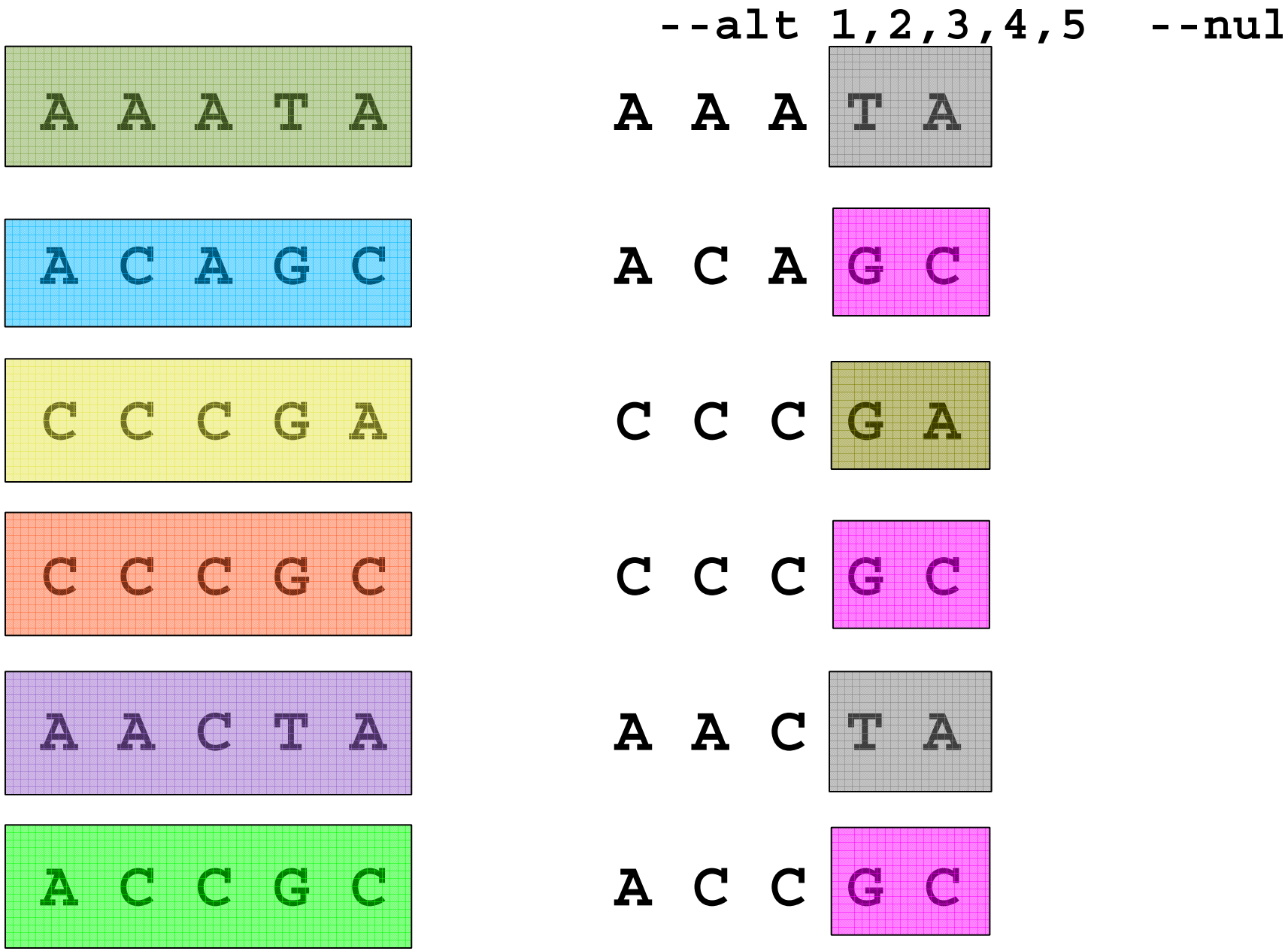
Single SNP test (2nd marker)



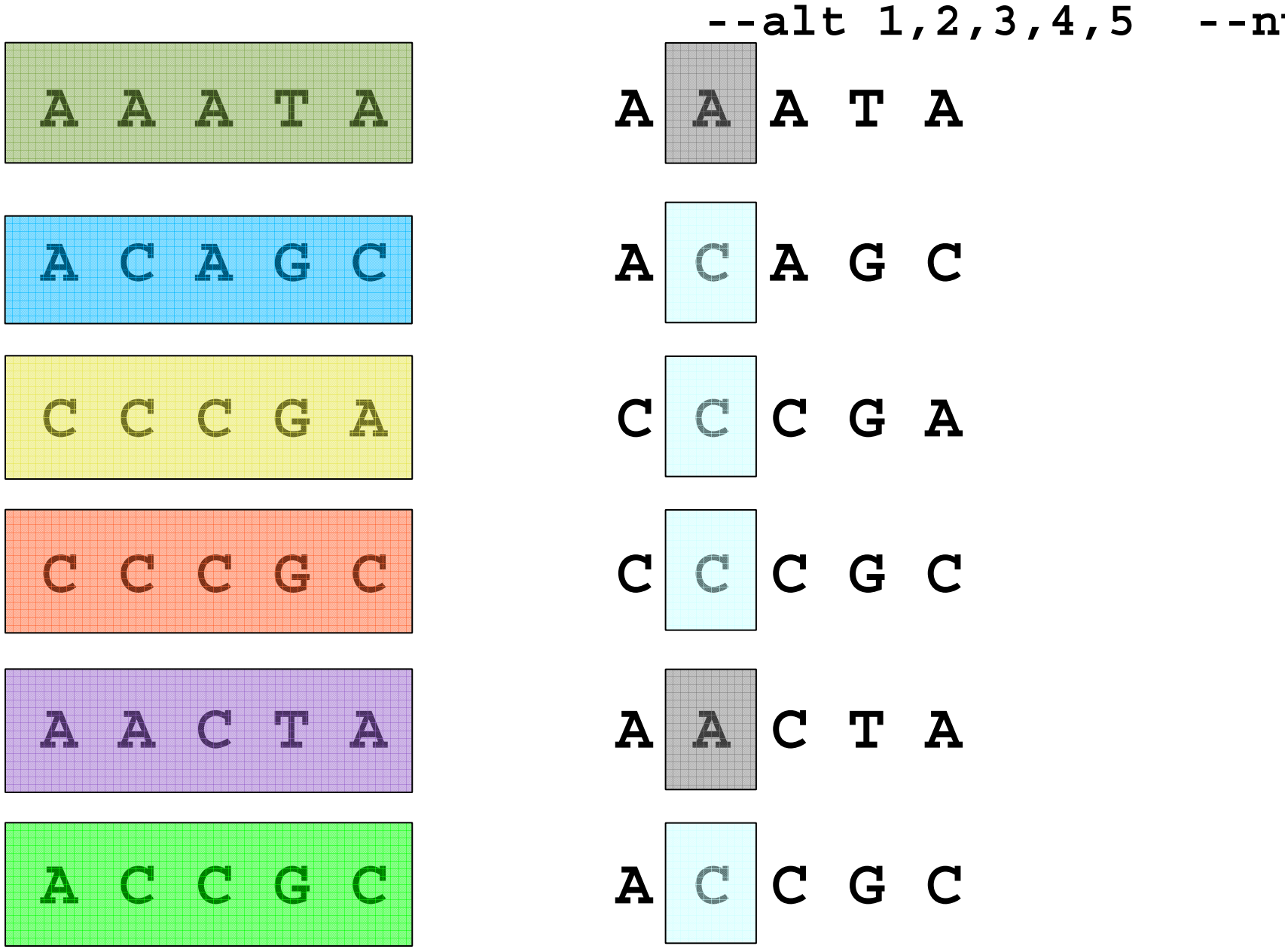
Independent effect test for SNP 1



Independent effect test for SNPs 1, 2 and 3

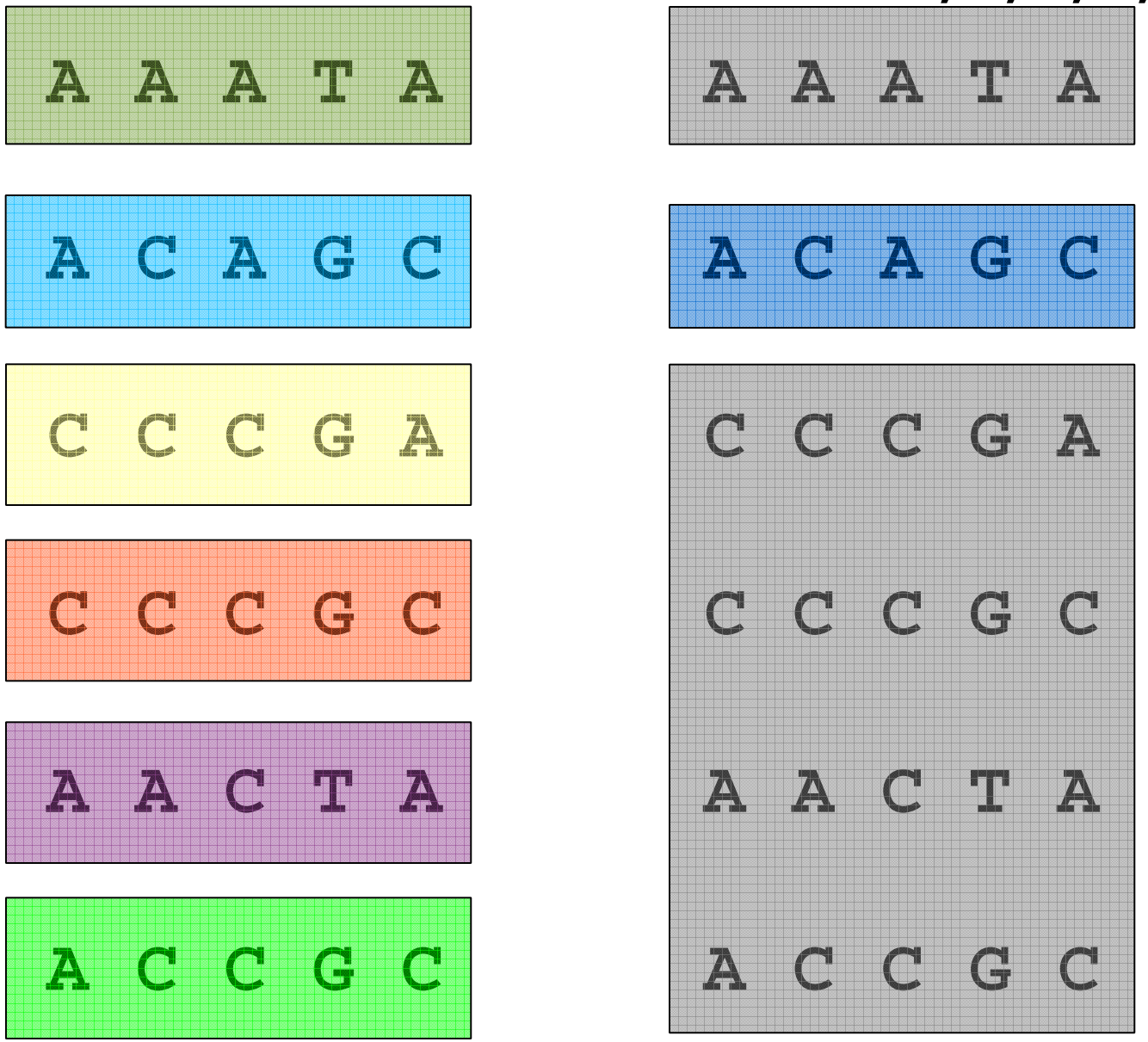


Sole-variant test for 2nd SNP



Sole-variant test for haplotype 2

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



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

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

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

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

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

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

SNP1	19.060	0.000765
SNP2	12.288	0.0153

Sole-variant tests for haplotypes

Standard haplotype-specific tests

Haplotype	Chi-sq(1df)	p-value
ACAGC	**8.546**	**0.00346**
CCCGC	0.428	0.513
CCCGA	0.265	0.607
AAATA	0.381	0.537
AACTA	**13.929**	**0.00019**
ACCGC	0.073	0.787

Sole-variant tests for haplotypes

Haplotype	Chi-sq(4df)	p-value
ACAGC	10.533	0.0323
CCCGC	18.651	0.00092
CCCGA	18.814	0.000855
AAATA	18.698	0.000901
AACTA	**5.150**	**0.272**
ACCGC	10.006	0.00784

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

Including the causal variant

```
AC-C-AGC
CC-C-CGC
CC-C-CGA
AA-C-ATA
AA-T-CTA
AC-C-CGC
```

```
1_A 1 0 0 1 2 A A   C C   C A   G G   C C   C C
2_A 1 0 0 1 2 A A   A C   C A   T G   A C   T C
3_A 1 0 0 1 2 A A   C A   A A   G T   C A   C C
4_A 1 0 0 1 2 A A   A A   A C   T T   A A   C T
5_A 1 0 0 1 2 A A   C A   A C   G T   C A   C T
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
```

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
```

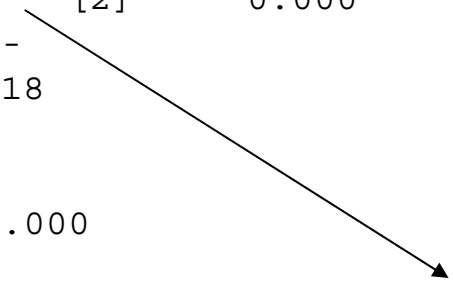
Single locus test of the CV

```
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) ~ 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]
CCCCGC	0.237	0.411	0.411	[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]
AATCTA	**0.065**	**1.317**	**1.317**	**[5]**	0.000	0.000	[1]
ACCCGC	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

CV --alt 1,2,3,4,5,6 --null 3 LRT = 5.901 df = 4 p = 0.207

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]
AATCTA	**0.065**	**1.317**	**1.317**	**[5]**	**1.065**	**1.065**	**[2]**
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”