# Haplotype analysis

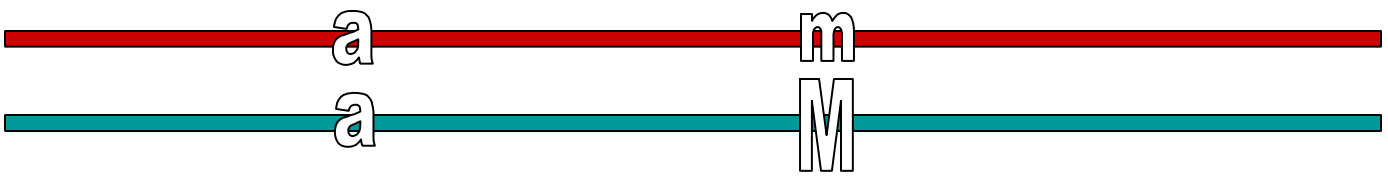
*Shaun Purcell*

shaun@pngu.mgh.harvard.edu

*MGH, Boston*

# Haplotypes

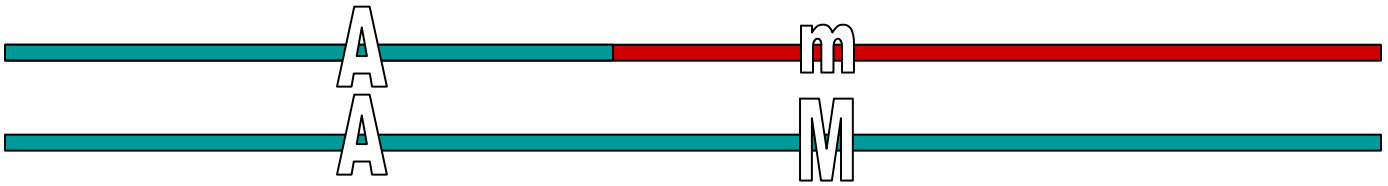
| | ***A*** | ***a*** |
|---|---|---|
| ***M*** | | ***aM*** |
| ***m*** | | ***am*** |




This individual has ***aa*** and ***Mm*** genotypes

and ***am*** and ***aM*** haplotypes

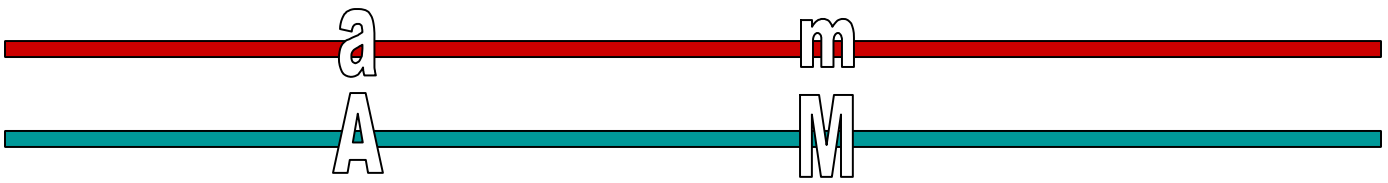
| | *A* | *a* |
|---|---|---|
| *M* | ***AM*** | |
| *m* | ***Am*** | |




This individual has ***AA*** and ***Mm*** genotypes

and ***AM*** and ***Am*** haplotypes

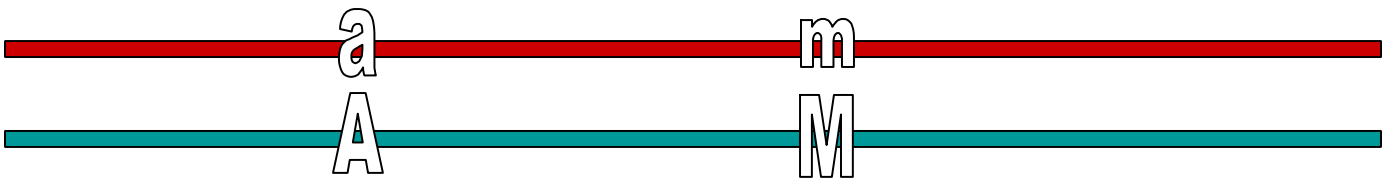
| | | |
|---|---|---|
| | ***A*** | ***a*** |
| ***M*** | ***AM*** | |
| ***m*** | | ***am*** |




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

and ***AM*** and ***am*** haplotypes…

| | | |
|---|---|---|
| | ***A*** | ***a*** |
| ***M*** | ***AM*** | |
| ***m*** | | ***am*** |




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***

# 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

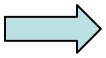
# M-step

Haplotype counts, frequencies from complete data

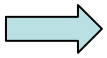
| | Count | Freq |
|---|---|---|
| AB | 2.5 | 0.3125 |
| aB | 0.5 | 0.0625 |
| Ab | 1.5 | 0.1875 |
| ab | 3.5 | 0.4375 |
| Sum | 8.0 | 1.0000 |

# 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…

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

Bayes Rule

$$P(H \mid G) = \frac{P(G \mid H)P(H)}{\sum_{H} P(G \mid H)P(G \mid H)}$$

# Posterior Probabilities

Example:
Genotype AaBb
Haplotype frequencies

| AB | aB | Ab | ab |
|---|---|---|---|
| 0.375 | 0 | 0.125 | 0.5 |

$$P(AB/ab \mid AaBb) = \frac{P(AaBb \mid AB/ab)P(AB/ab)}{P(AaBb \mid AB/ab)P(AB/ab) + P(AaBb \mid Ab/aB)P(Ab/aB)}$$

$$= \frac{1 \times 0.375 \times 0.5}{1 \times 0.375 \times 0.5 + 1 \times 0.125 \times 0}$$

$$= 1$$

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

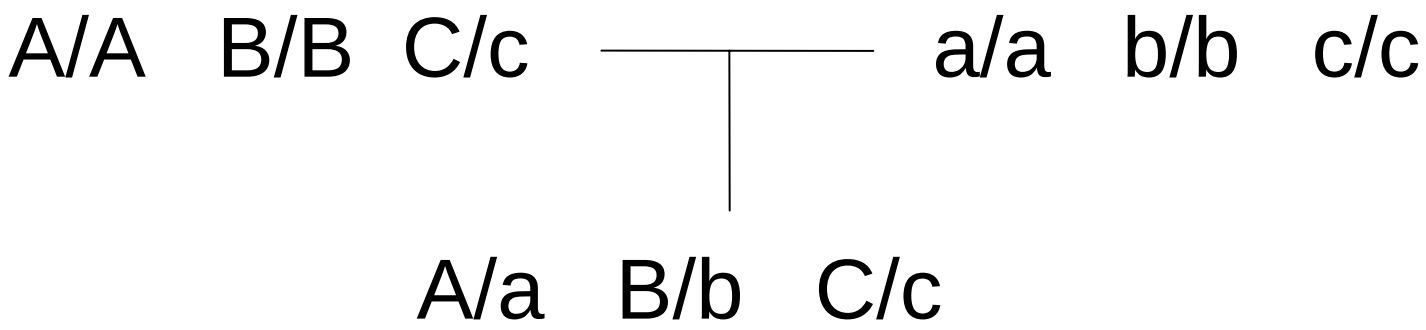
# Using parental genotypes

Can often help to resolve phase

A/a  B/b  C/c

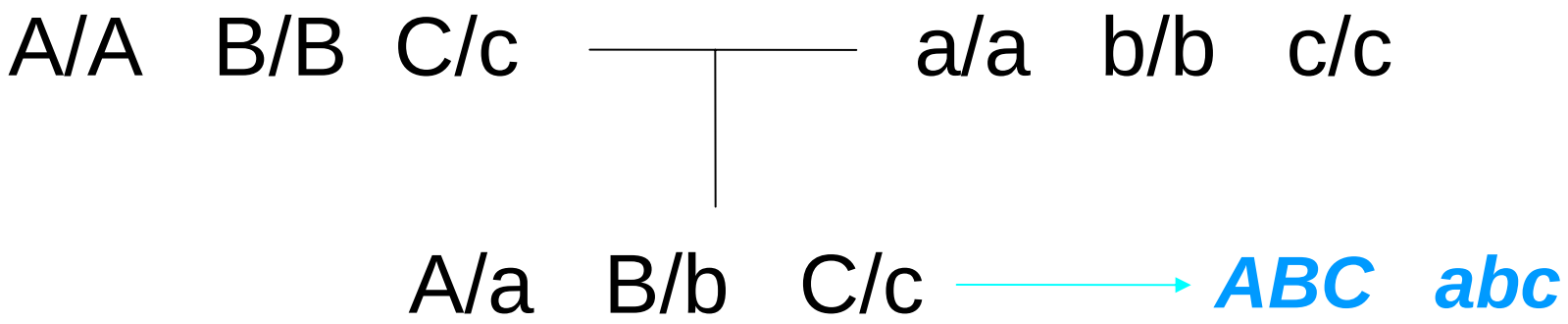
# Using parental genotypes

Can often help to resolve phase



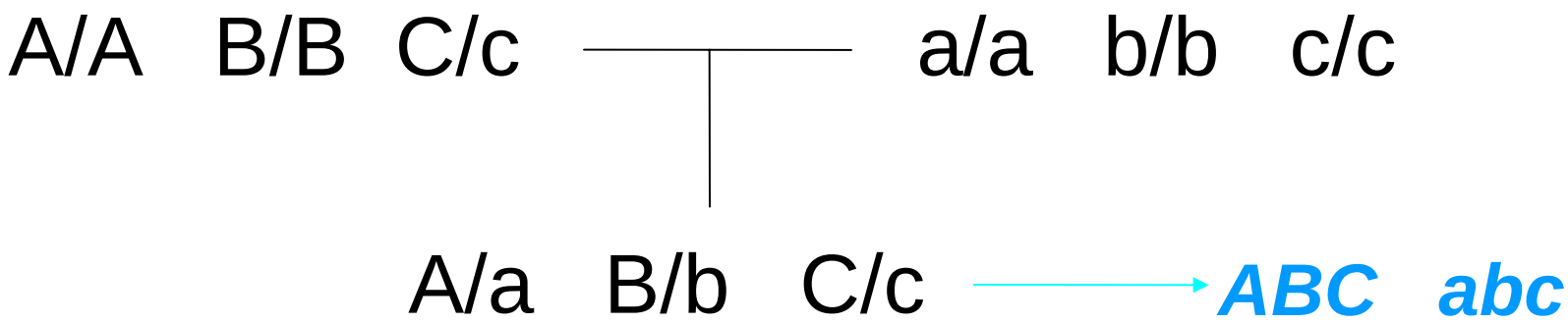
# Using parental genotypes

Can often help to resolve phase

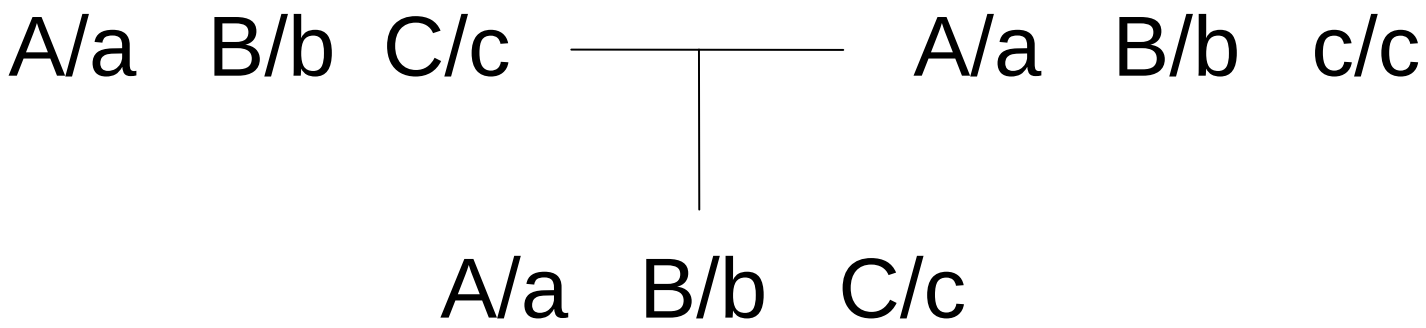


# Using parental genotypes

Can often help to resolve phase




*… but not always*

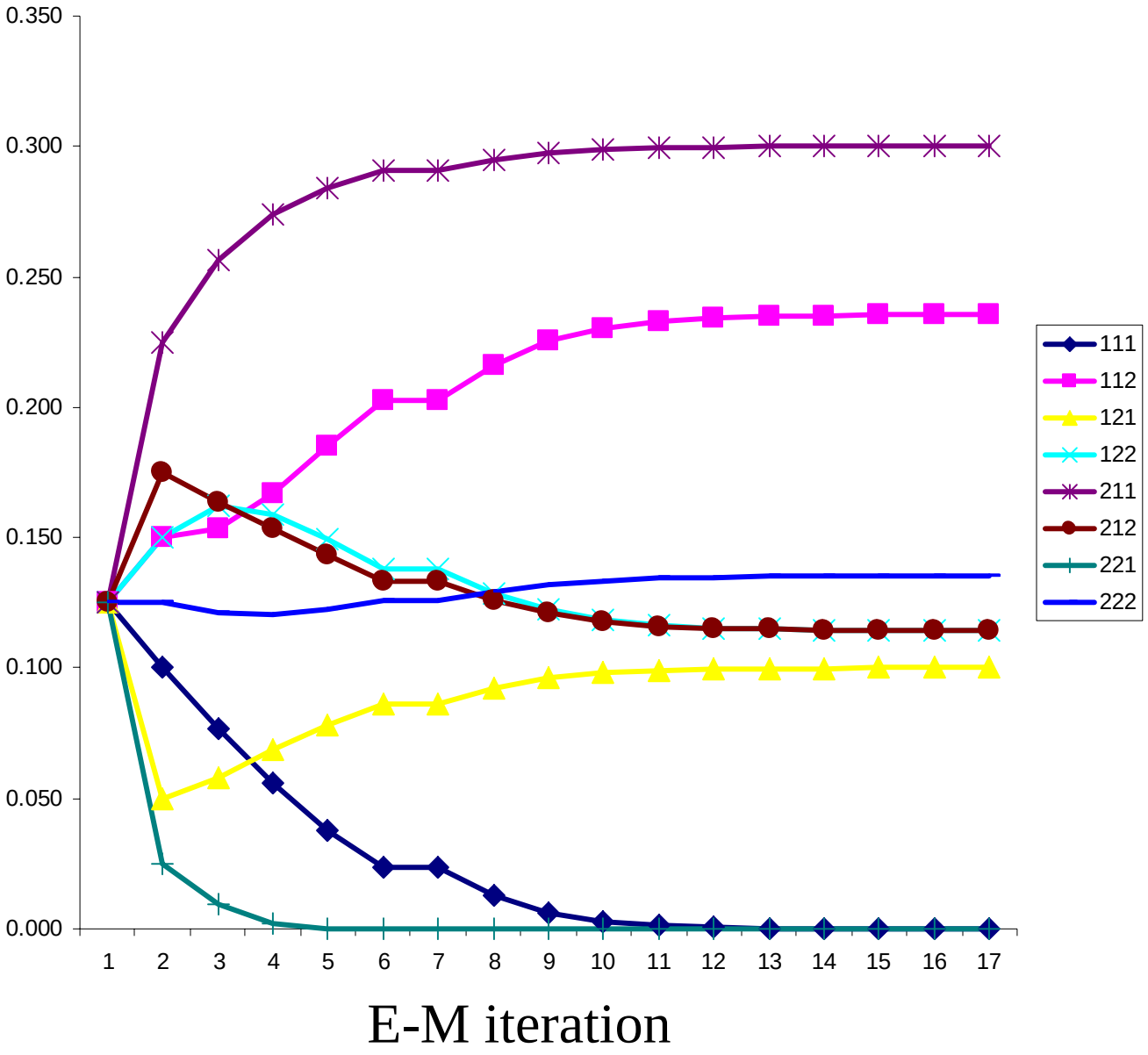


# A (slightly) less trivial example

| | | | | |
|---|---|---|---|---|
| 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 |

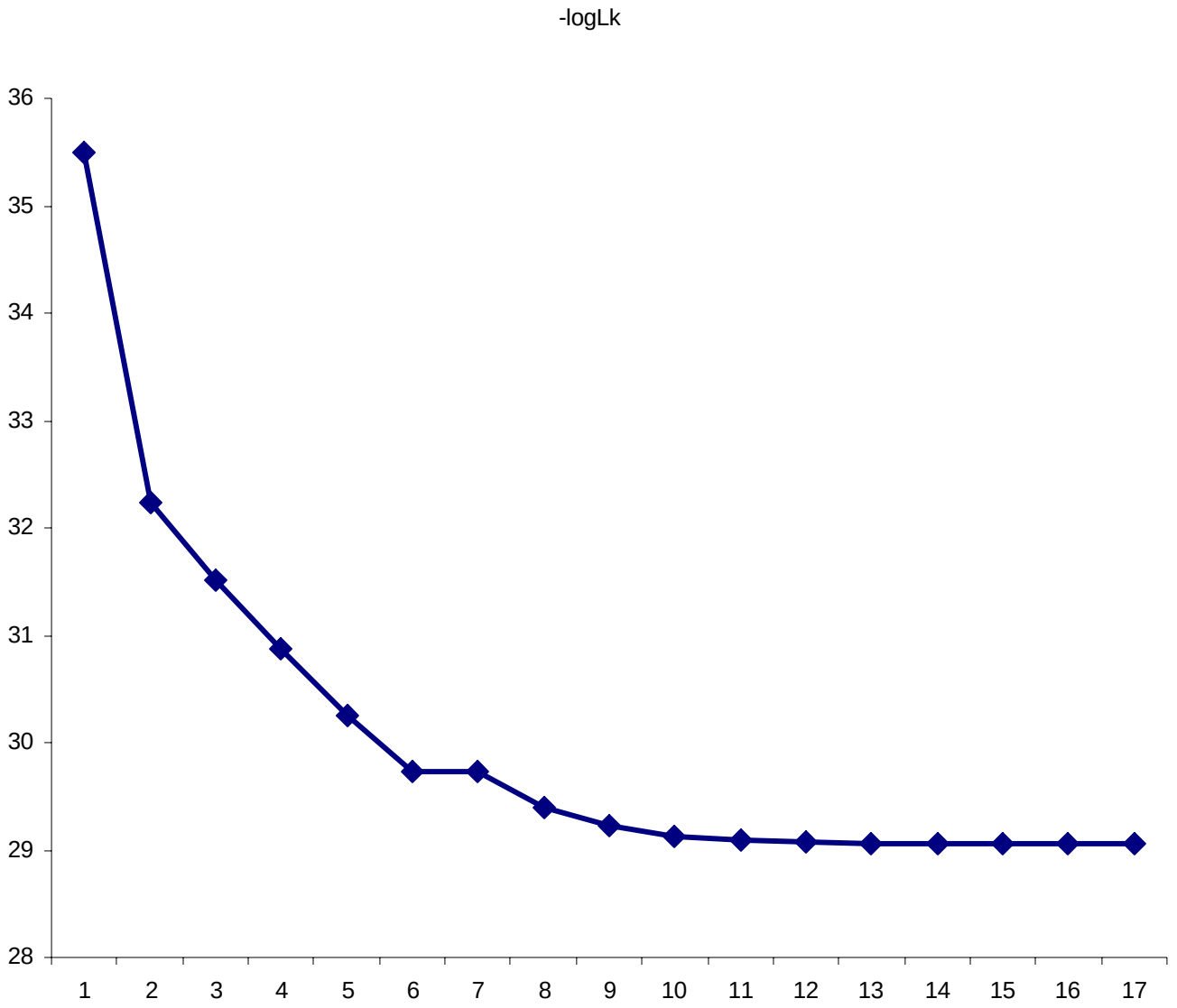
# haplotype frequencies

Estimated haplotype frequency




E-M iteration

# 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 |

# 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

# Haplotype analysis software

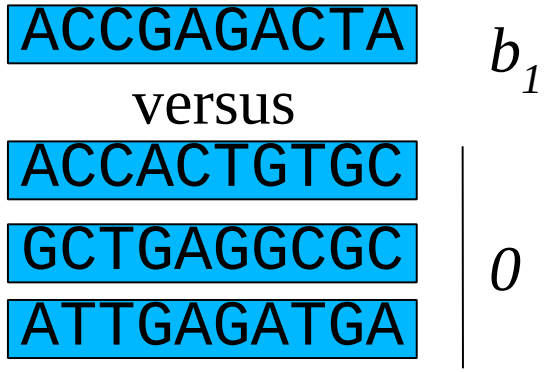
- Many available packages:
  - EH+/Genecouting (Zhao)
  - HaploView (Barrett)
  - PHASE (Stephens)
  - FBAT/HBAT/PBAT (Xu *et al*, Lange)
  - haplo.score (Schaid)
  - eHap (Roeder) / ET-TDT (Seltman)
  - UNPHASED (Dudbridge)
  - PLINK (Purcell *et al*)
  - whap (Purcell & Sham)

# whap

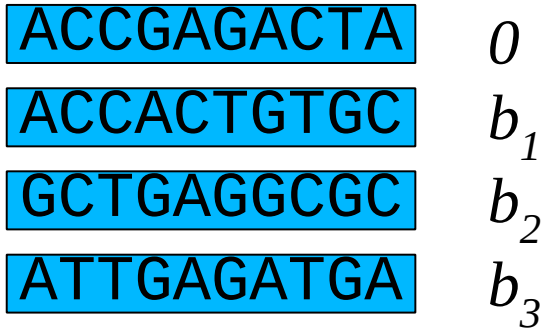
- Numerous recent methods using GLM approach
  - Schaid *et al* (02) *AJHG*
  - Zaykin *et al* (02) *Hum Hered*
- Quantitative and qualitative traits
- Mixture of regressions framework
- Between/within family model
- Model either L(X|G) or L(G|X)
- Covariates and moderators
- Flexible specification of nested submodels

# 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*

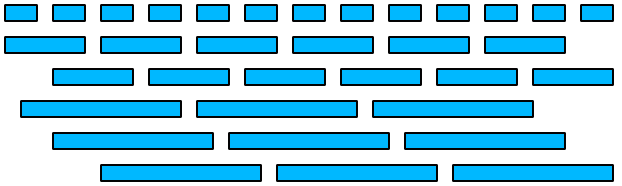



- ***Omnibus test***
  - *single test with H-1 df*
  - *compare each haplotype against an (arbitrary) reference haplotype*
  - *built-in correction for multiple tests*

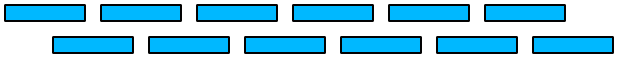


# Covering large genomic areas

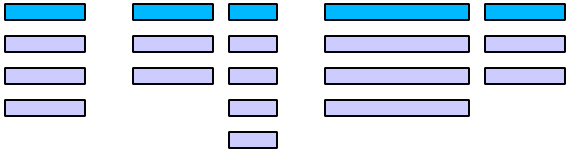
*Exhaustive haplotype approach (ETDT)*



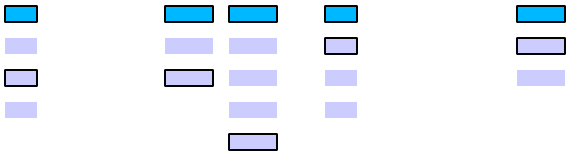
*Sliding window of fixed size (whap)*



*Haplotype-specific block-based tests (HaploView)*



*Specific small multimarker predictors of known, common but otherwise untagged variants (HaploView, plink)*



*For full details: http://pngu.mgh.harvard.edu/purcell/whap/*

# File formats

- Similar to 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 sessions

- Analysis of simulated data
- Detecting haplotype association using `whap`
- Fitting nested model to explore the association using `whap`

# Practical: Simulated data

*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
...
```

*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
```

If `pedstats` program available,
you can check the datafile with:

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

# Practical: 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 | |

# Practical

- 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 1st SNP
whap --file dataACGT --alt 5                    Analyse 5th 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

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

*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

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

| *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 tests

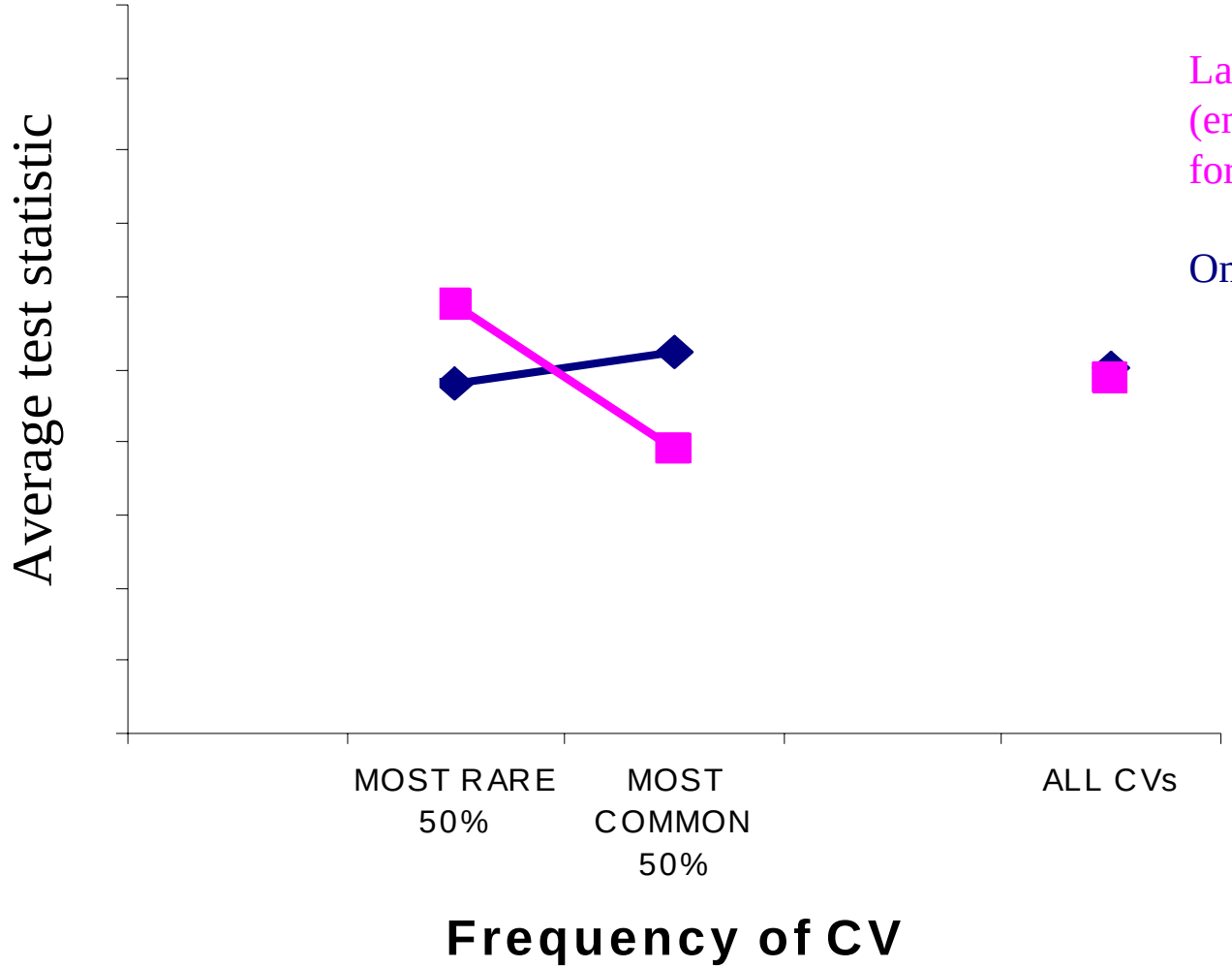
```
whap --file dataACGT --hs
```

| *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 |

From logistic regression OR is calculated by e^(Beta),

where e is 2.718281828459….

# Haplotype-specific or omnibus?




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

Omnibus test

# Detection of associations

- Detection test
  - single SNP
  - haplotype-specific
  - omnibus test

- “Is *X* associated with my phenotype?”
  - where *X* is either an allele, genotype, haplotype or set of haplotypes

# Dissection of an association

- Assuming a haplotypic association, explores the nature of the association, e.g.
  - single or multiple haplotype effects?
  - a single SNP explains the entire effect?
- “Is *X* associated with my phenotype independent of *Y*?”

# Interpreting effects

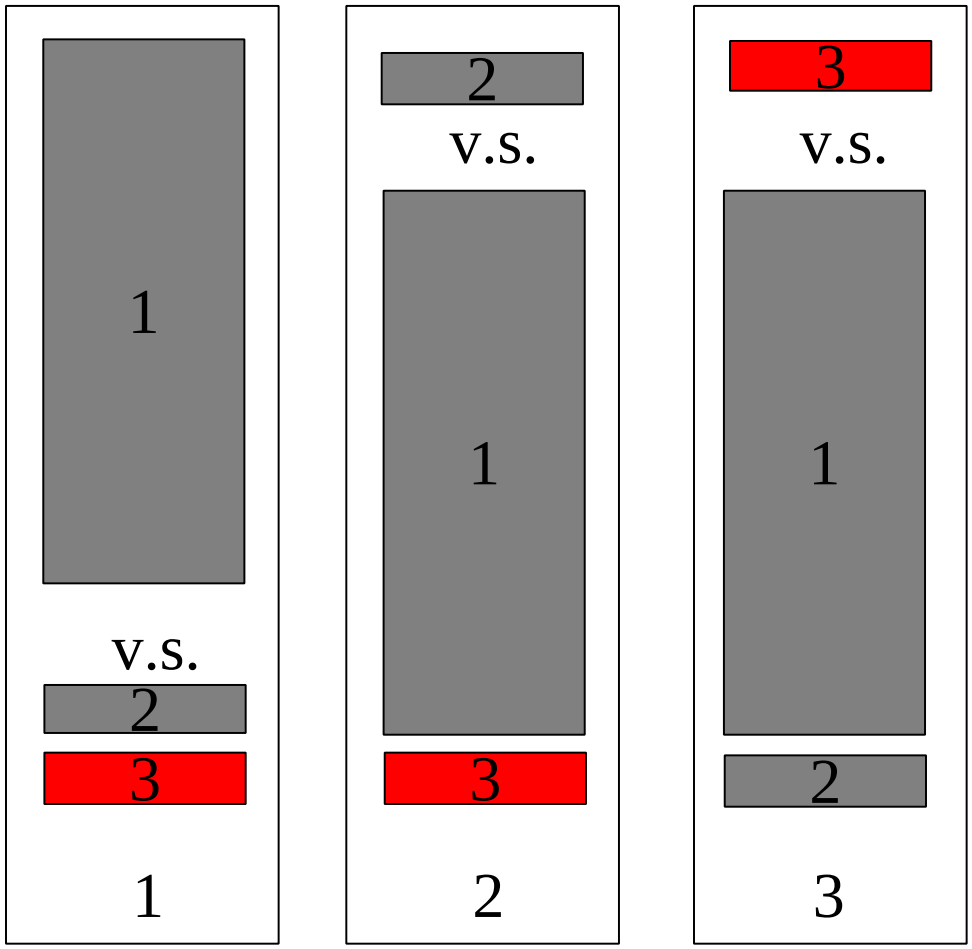
*True model*

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

*Looks like*

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

*Haplotype-specific tests:*



# 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] |

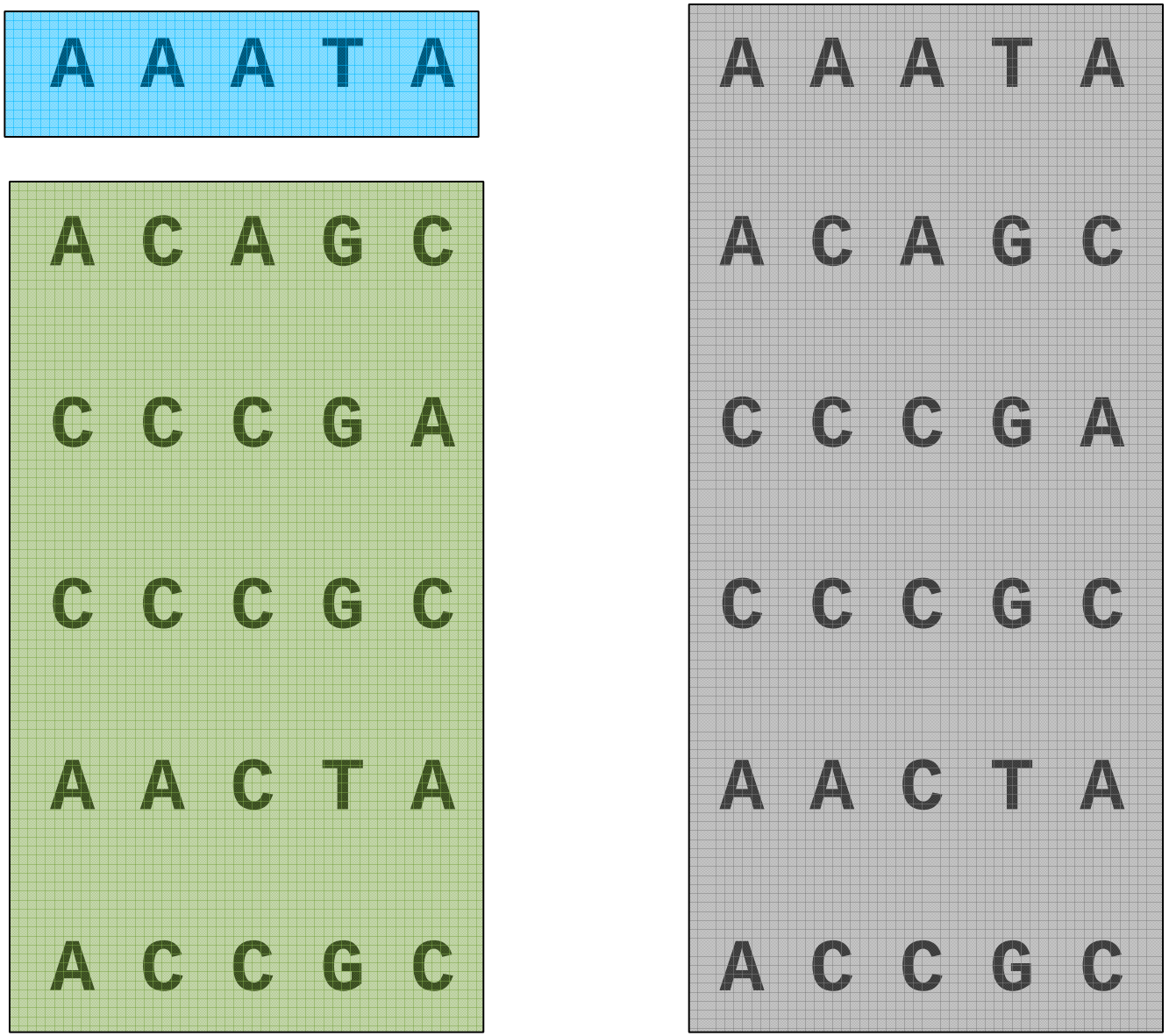
```
--alt 1,2 --null 2
```

# Conditional tests

- Does *X* have any effect after controlling for everything else?
  - *X* independently contributes (if signif.)
    - X could be a SNP or set of SNPs
    - `--alt 1,2,3,4,5  --null 2,3,4,5`
    - “independent effect test”
- Does everything else *still* have any effect after controlling for *X* ?
  - is necessary and sufficient (if test n.signif.)
    - X could be a SNP, set of SNPs, haplotype or set of haplotypes
    - `--alt 1,2,3,4,5  --null 1`
    - `--constrain 1,2,3,4,5,6/1,2,1,1,1,1`
    - “sole variant test”

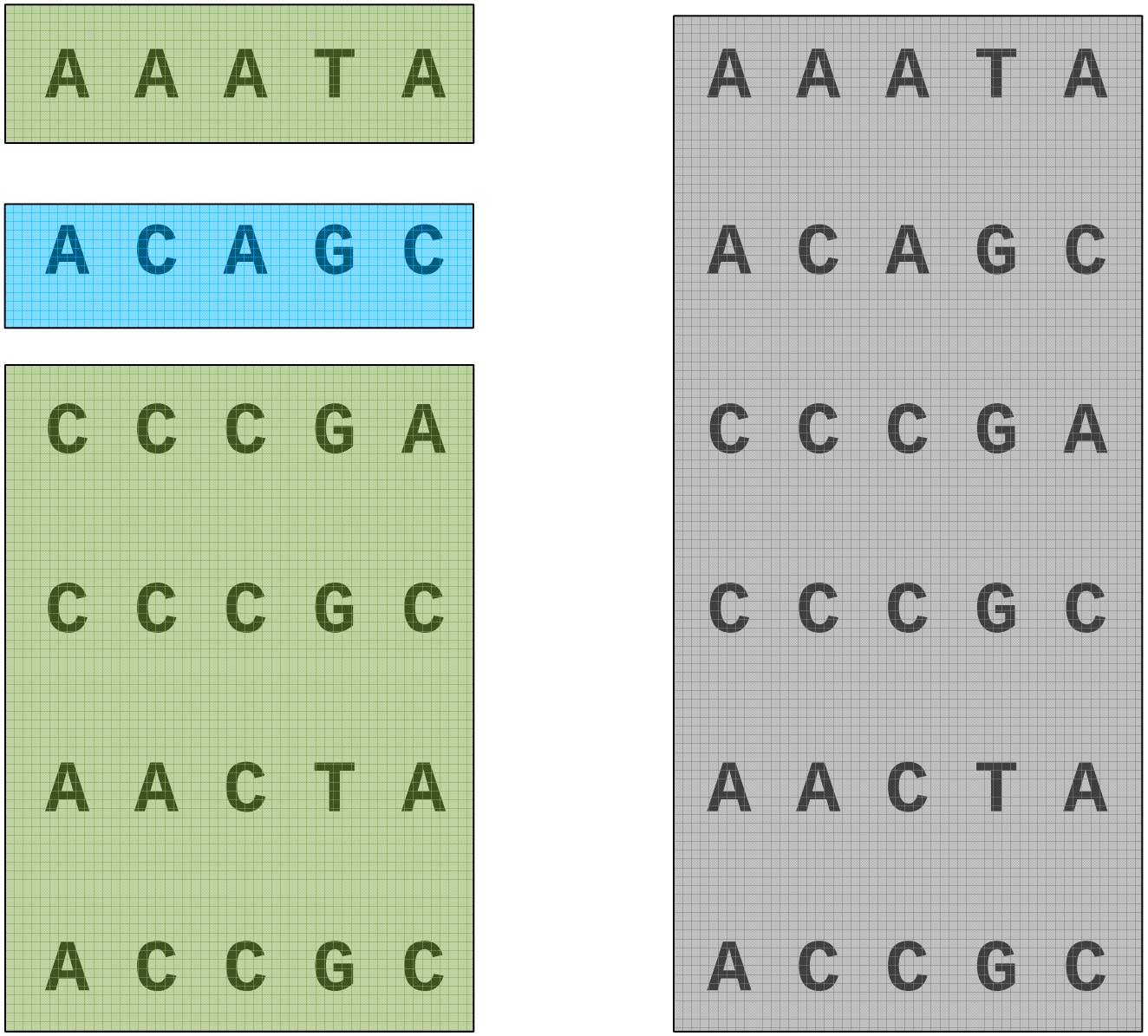
# Haplotype-specific test (H1)

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



# Haplotype-specific test (H2)

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



# Omnibus test (df=5)

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

A A A T A

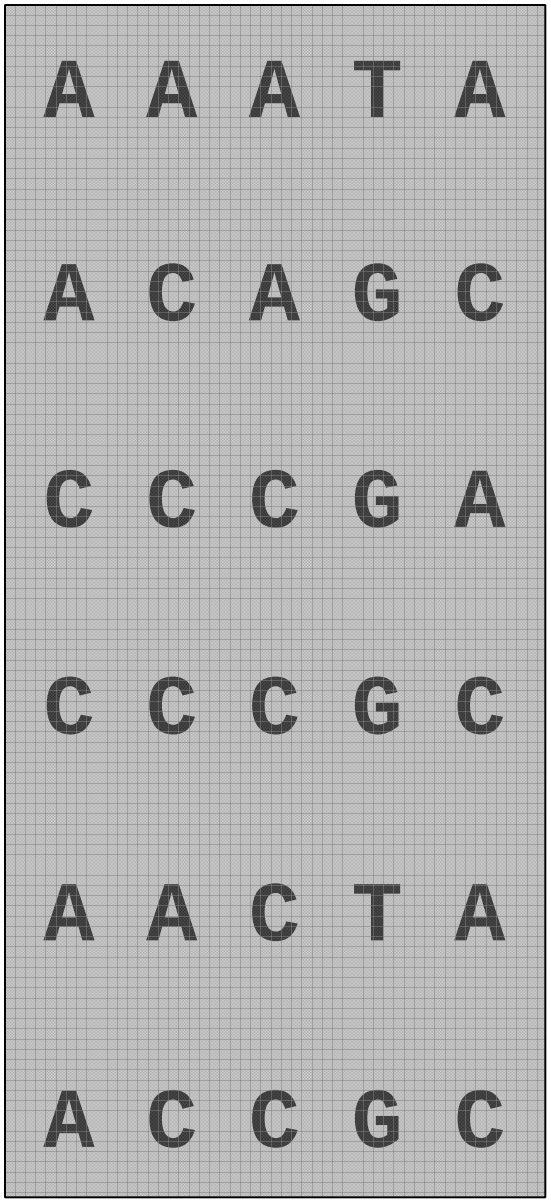
A C A G C

C C C G A

C C C G C

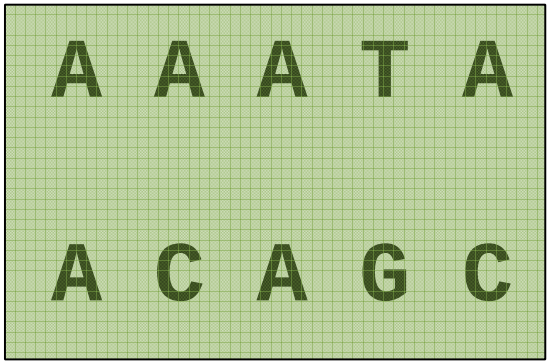
A A C T A

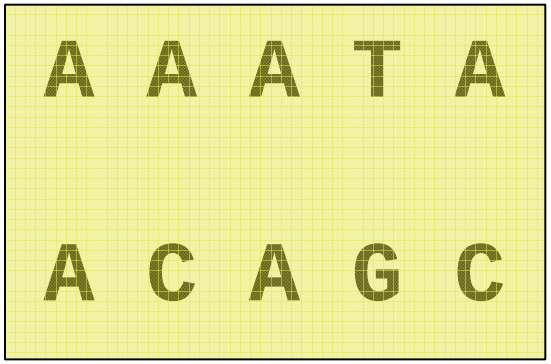
A C C G C

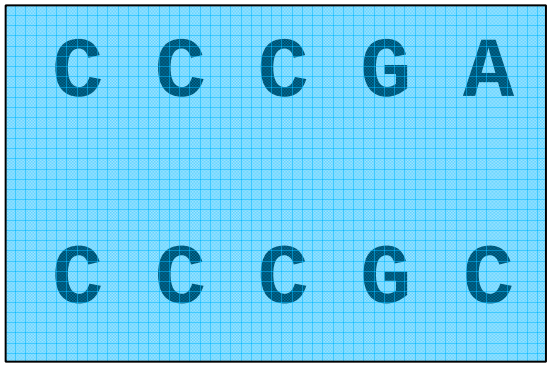


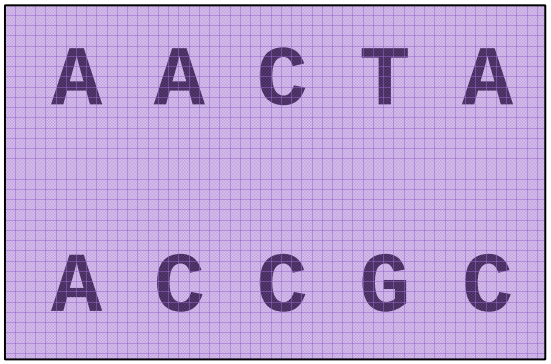
# Clade-based homogeneity test (1df)

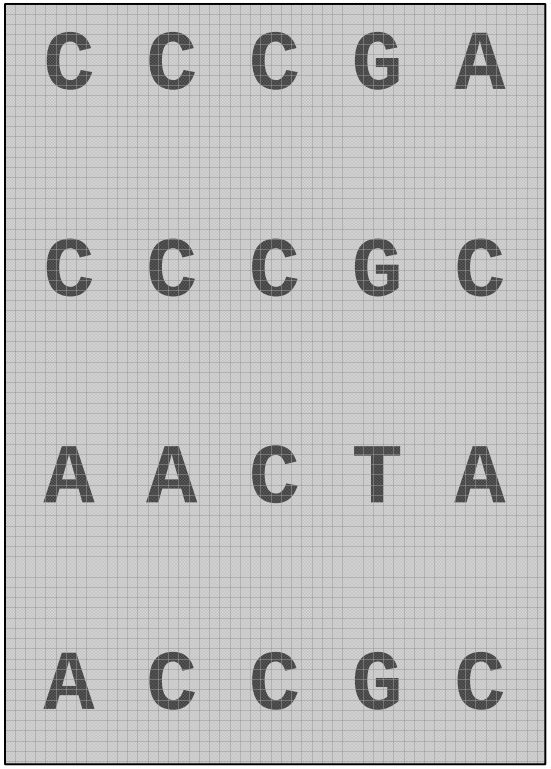
```
--constrain 1,1,2,2,3,3 / 1,1,2,2,2,2
```



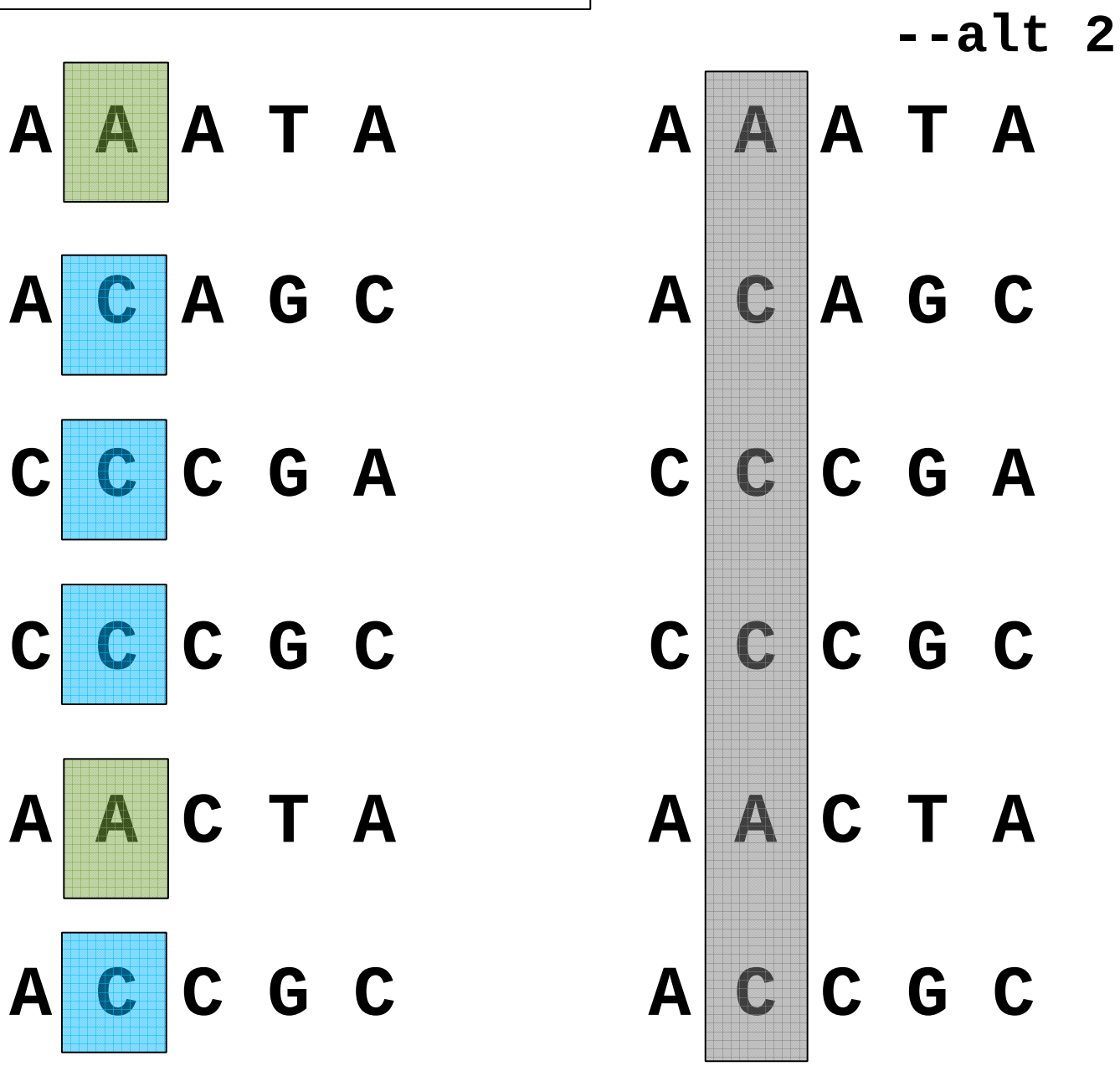



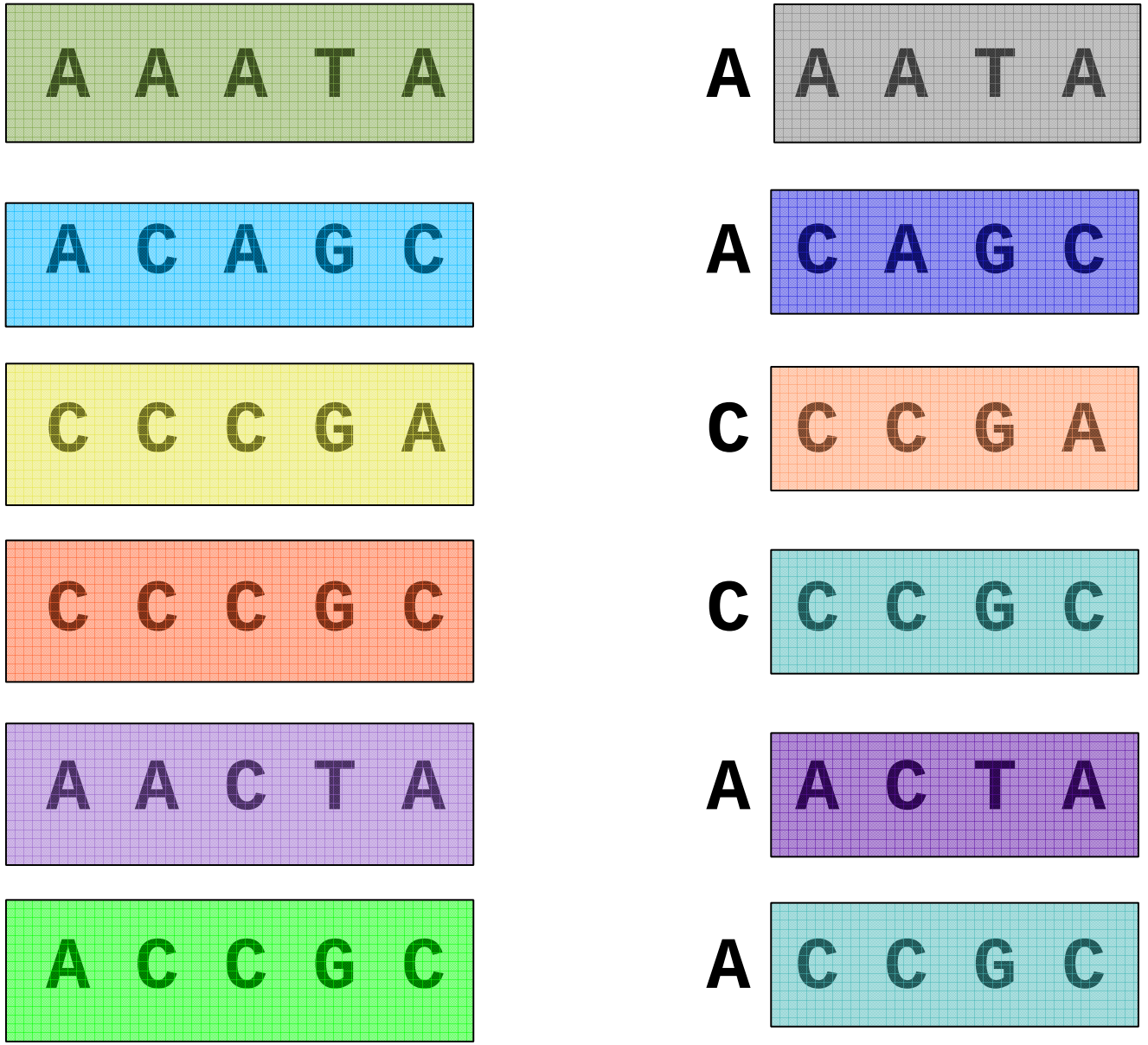



# Single SNP test (2nd marker)

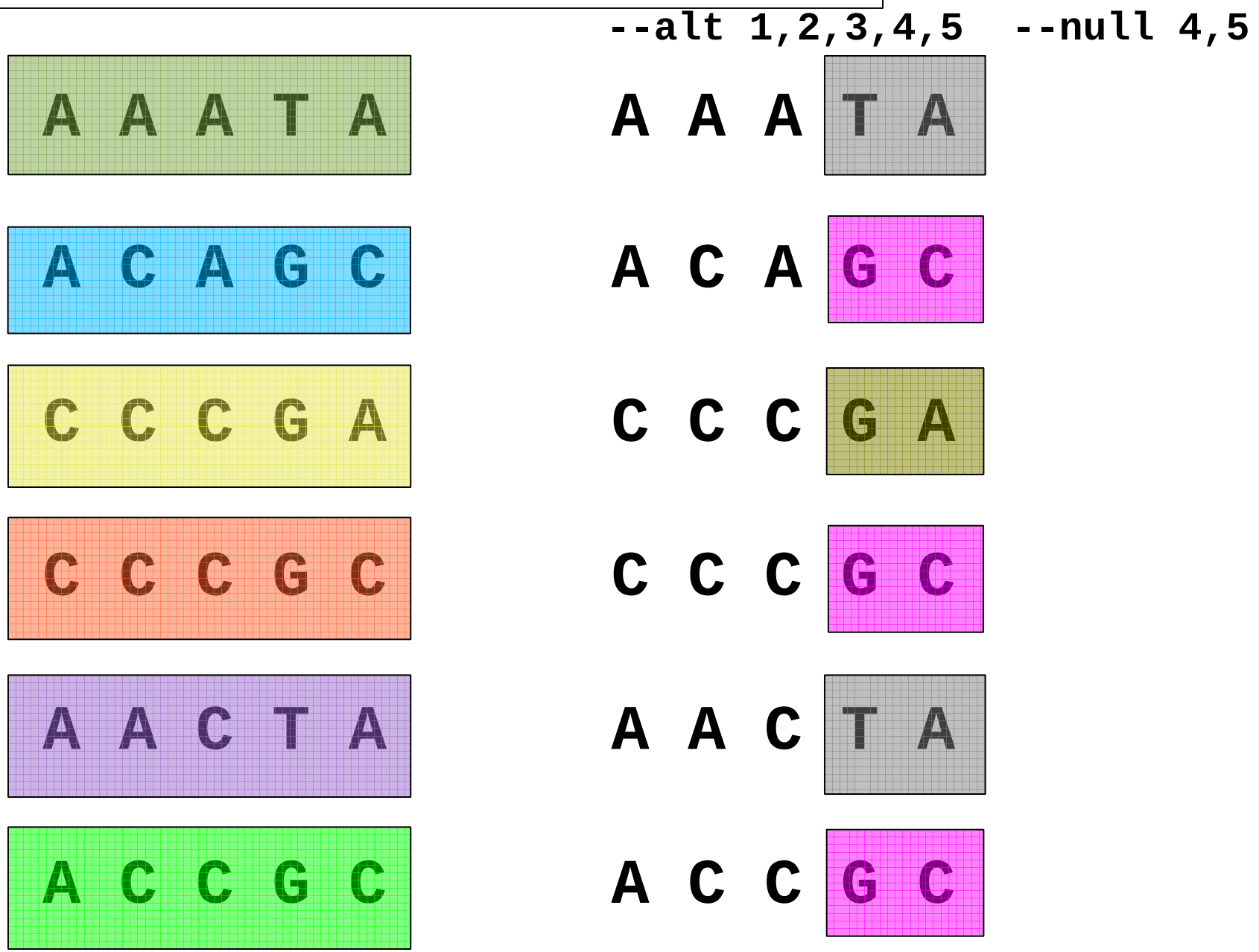


# Independent effect test for SNP 1

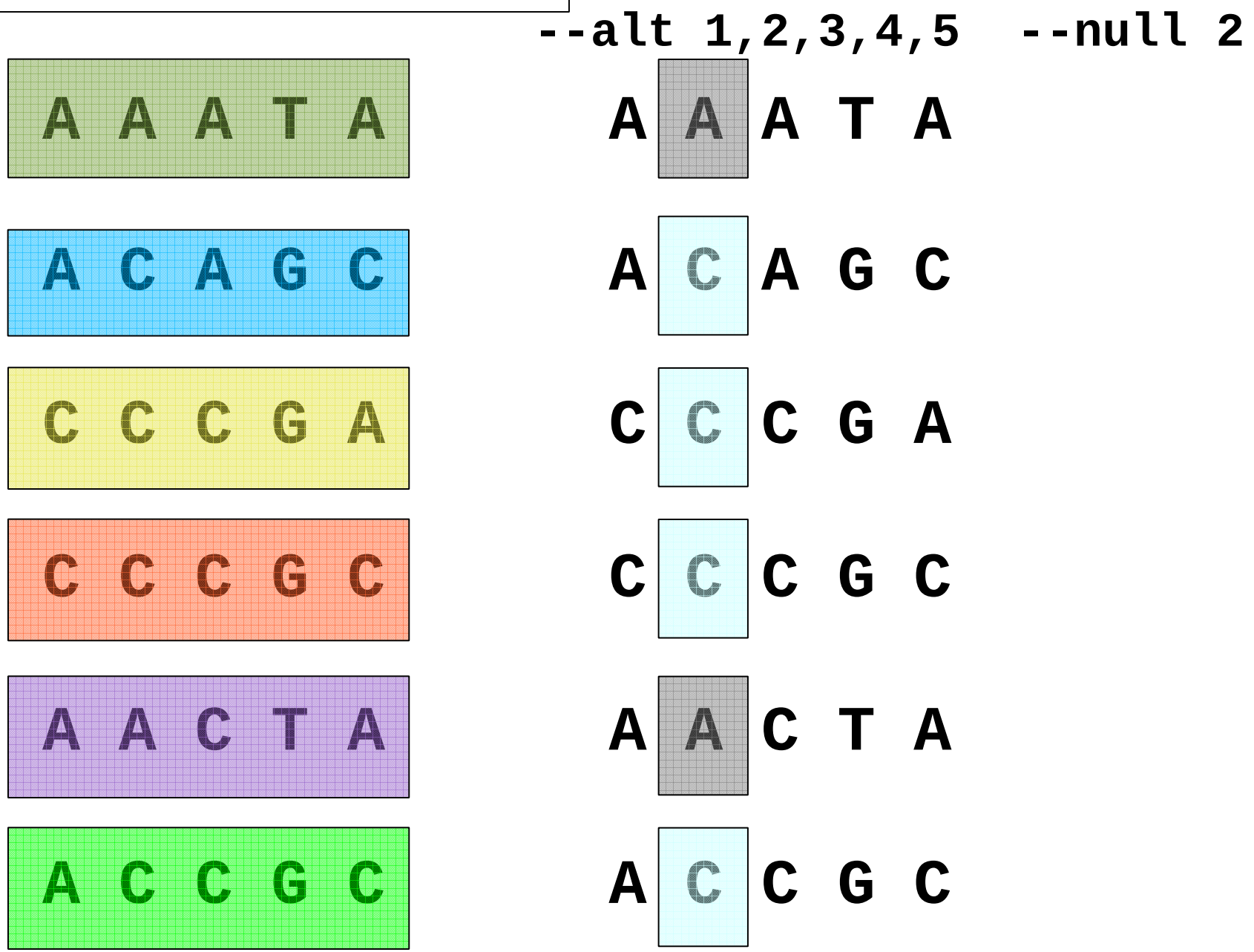
```
--alt 1,2,3,4,5 --null 2,3,4,5
```



# Independent effect test for SNPs 1, 2 and 3

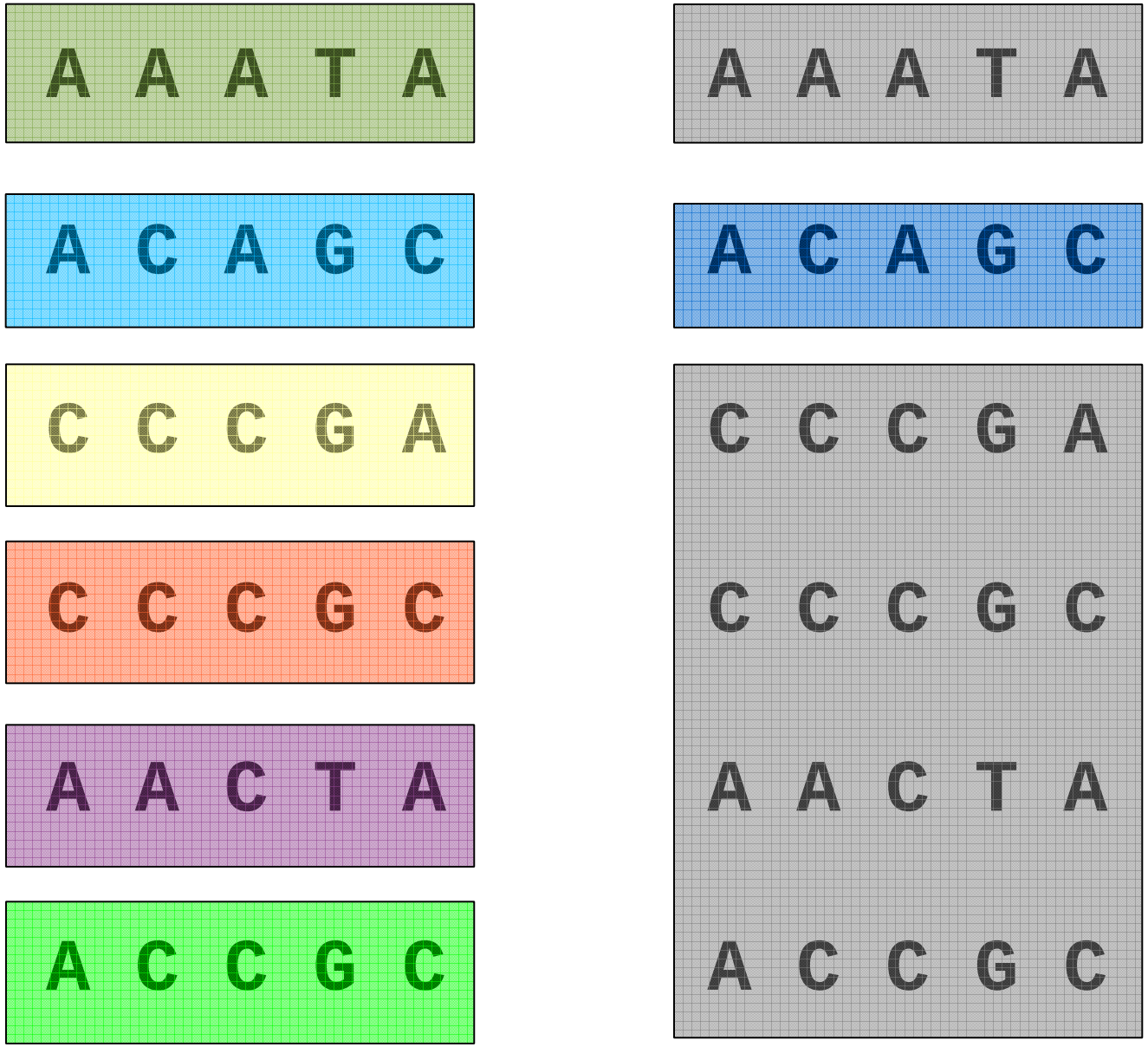


# Sole-variant test for 2nd SNP



# Sole-variant test for haplotype 2

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



# Practical exercise

- Now continue practical session:

  “SECOND PART: DISSECTING THE EFFECT”

- Perform conditional tests

- What do these suggest about the nature of the association?

***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 | |
|---|---|---|---|
| **ACAGC** | **8.546** | **0.00346** | 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 |
| **AACTA** | **13.929** | **0.00019** | 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 |
| **AACTA** | **5.150** | **0.272** | 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
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
```

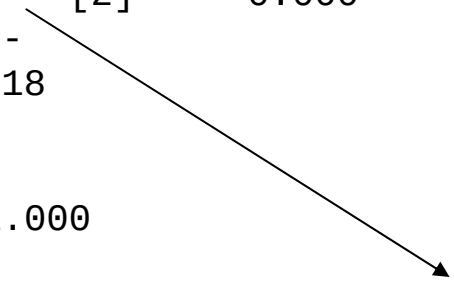
# Single locus test of the CV

```
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) | |
|---|---|---|---|---|---|---|---|
| AC**C**AGC | 0.261 | 0.000 | 0.000 | [1] | 0.000 | 0.000 | [1] |
| CC**C**CGC | 0.237 | 0.411 | 0.411 | [2] | 0.000 | 0.000 | [1] |
| CC**C**CGA | 0.212 | 0.276 | 0.276 | [3] | 0.000 | 0.000 | [1] |
| AA**C**ATA | 0.171 | 0.406 | 0.406 | [4] | 0.000 | 0.000 | [1] |
| **AA<u>T</u>CTA** | **0.065** | **1.317** | **1.317** | **[5]** | 0.000 | 0.000 | [1] |
| AC**C**CGC | 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”