Quality control for SNP data

M.Timothy.Rabanus-Wallace @mtrw85

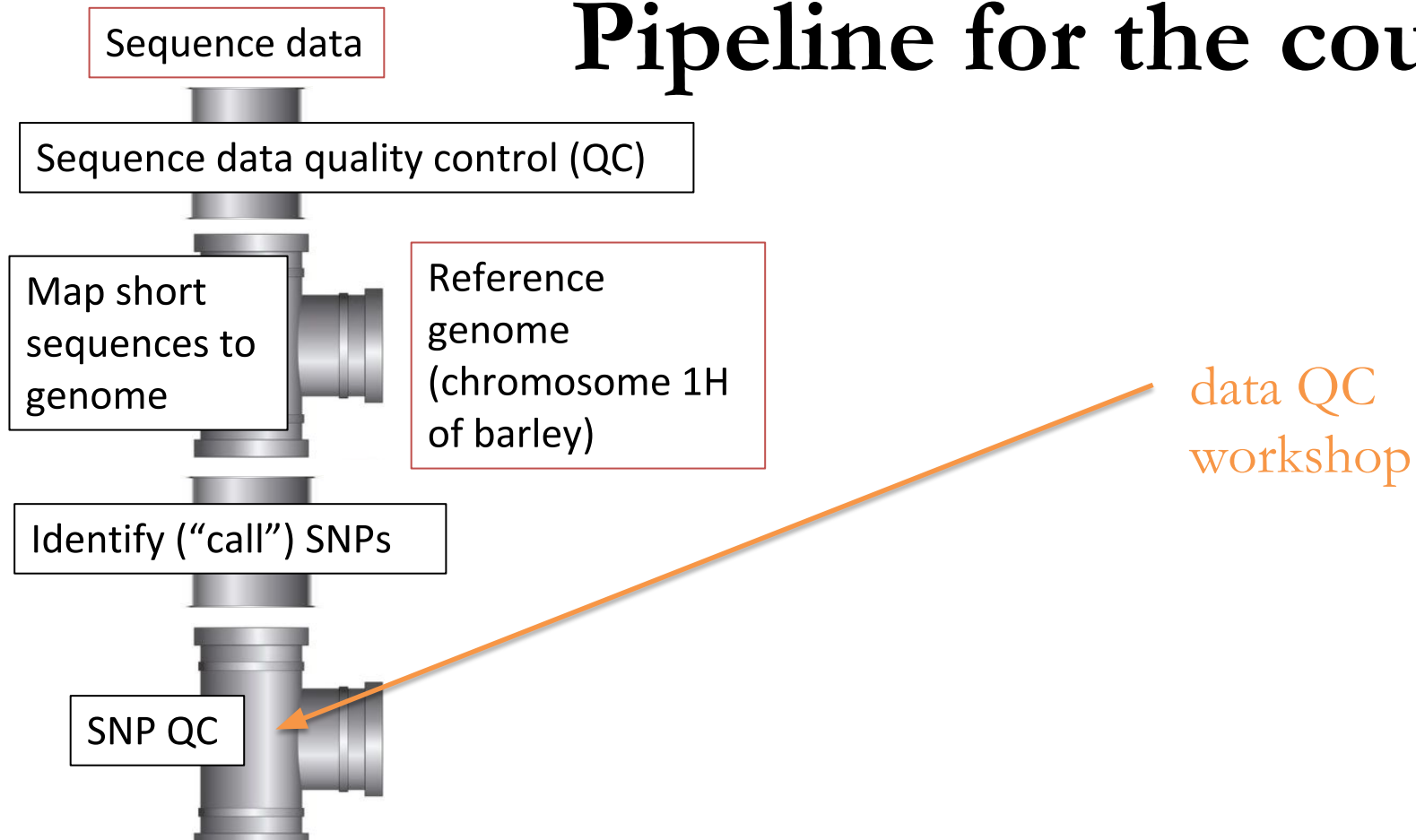# Pipeline for the course

Sequence data

Sequence data quality control (QC)

Map short sequences to genome

Reference genome (chromosome 1H of barley)

Identify ("call") SNPs

SNP QC

Population Structure Analysis

Phenotype association study

Phenotype data

Gene annotation data

Final result

primary data worksop

data QC workshop

data analysis workshop

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Pipeline for the course

Sequence data

Sequence data quality control (QC)

Map short sequences to genome

Reference genome (chromosome 1H of barley)

Identify ("call") SNPs

SNP QC

data QC workshop

Pipeline for the course

Sequence data

Sequence data quality control (QC)

Map short sequences to genome

Reference genome (chromosome 1H of barley)

Identify ("call") SNPs

SNP QC

data QC workshop

Where we left off: Genotype calls

AGAGATTAGAGATAGTAGAT ⎫
AGAGATAGTAGATA ⎭ **Sample A**

TTAGAGATTATAGATAGTAGATA ⎫
AGAGATAGTAGA ⎬ **Sample B**
ATAGTAGATATTAGGAGAG ⎭

GATAGTAGATATTAGG ⎫
GAGATAGTAGAT ⎭ **Sample C**

AGGATTTAGTTATTAGAGATTAGAGATAGTAGATATTAGGAGAGTATAGATATAGTAA

Reference Genome

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

A more realistic scenario

M.Timothy.Rabanus-Wallace 🐦 @mtrw85
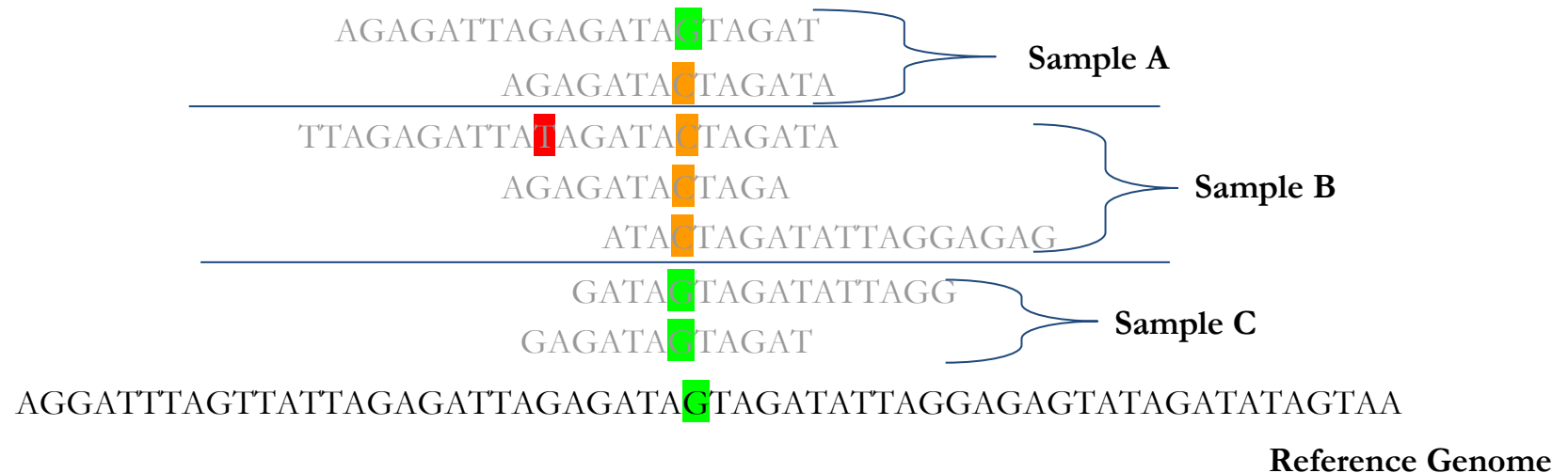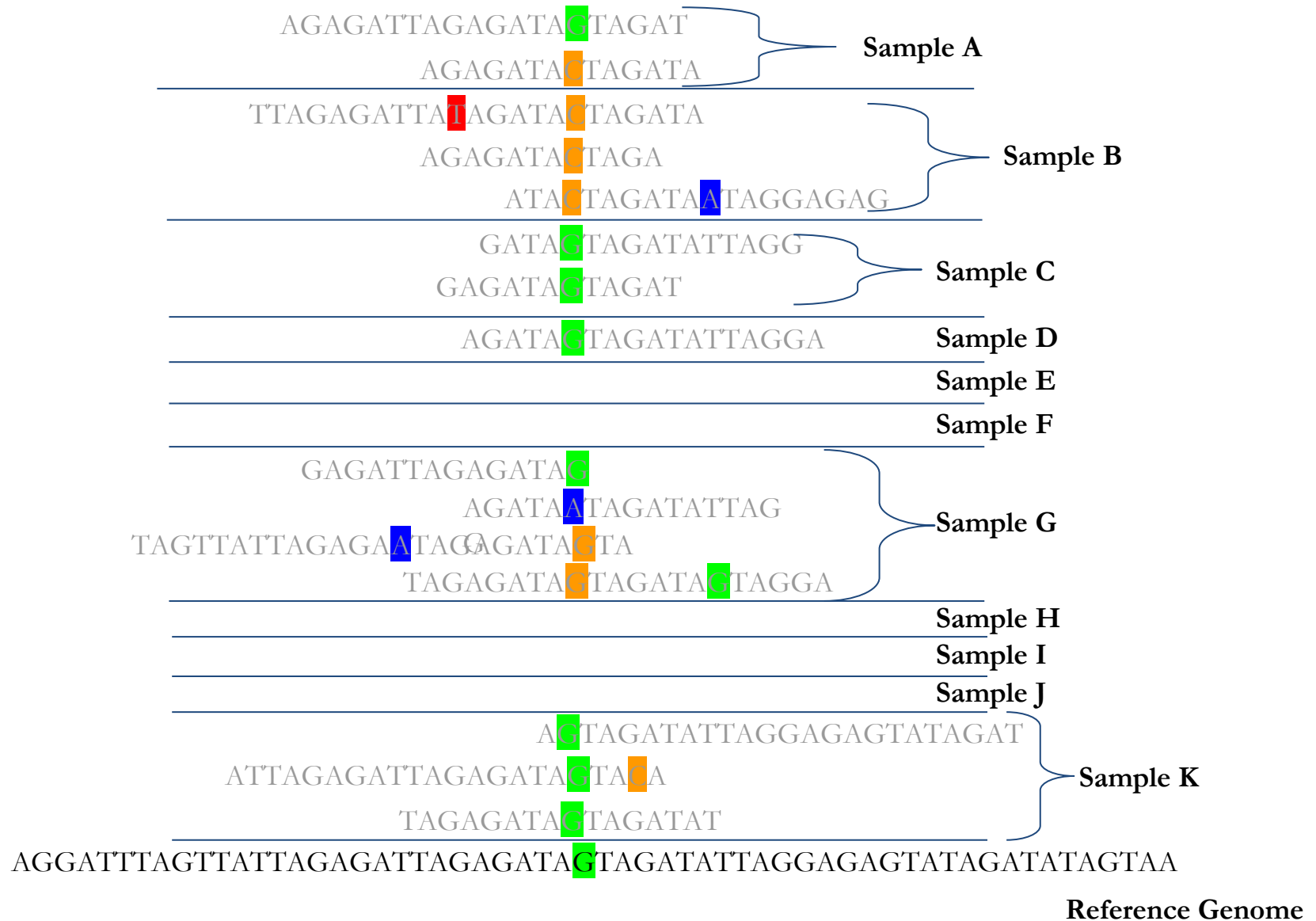
A more realistic scenario: calls
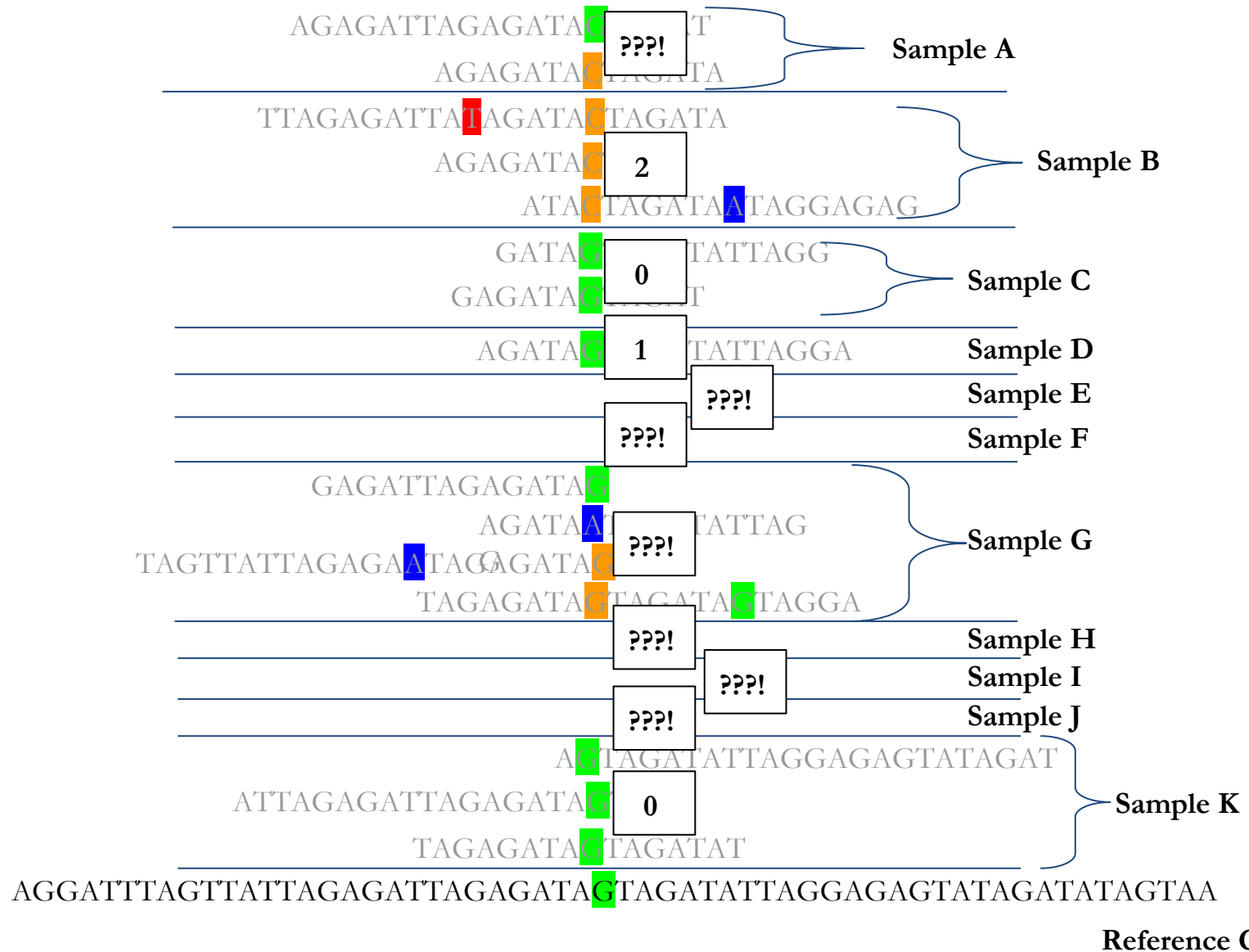
Graphical Genotypes

???!	Sample A
2	Sample B
0	Sample C
1	Sample D
???!	Sample E
???!	Sample F
???!	Sample G
???!	Sample H
???!	Sample I
???!	Sample J
0	Sample K

Reference Genome

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

???!	Sample A
2	Sample B
0	Sample C
1	Sample D
???!	Sample E
???!	Sample F
???!	Sample G
???!	Sample H
???!	Sample I
???!	Sample J
0	Sample K

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

???!	Sample A
2	Sample B
0	Sample C
1	Sample D
???!	Sample E
???!	Sample F
???!	Sample G
???!	Sample H
???!	Sample I
???!	Sample J
0	Sample K

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
Sample A	2	0	1	???!	2	0
Sample B	2	0	1	2	0	2
Sample C	0	0	1	0	2	0
Sample D	1	2	1	1	2	0
Sample E	2	0	2	???!	2	0
Sample F	0	0	1	???!	2	0
Sample G	1	0	1	???!	2	0
Sample H	0	0	1	???!	0	2
Sample I	2	0	???!	???!	0	2
Sample J	0	0	1	???!	2	0
Sample K	2	0	1	0	0	2

M.Timothy.Rabanus-Wallace 🐦 @mtrw85
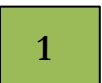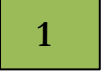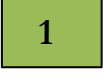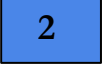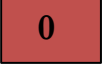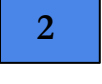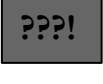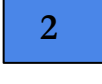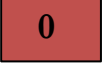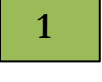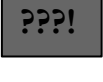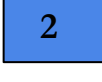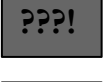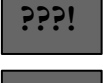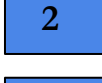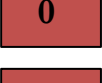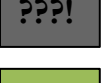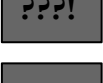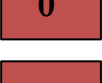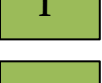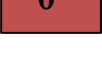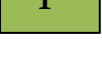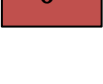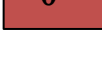
Graphical Genotypes: A real example
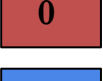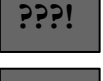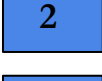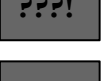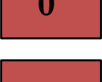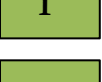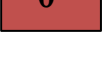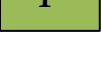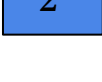
Graphical Genotypes

Each genotype call also has a genotype quality score, e.g.:

Genotype Quality=11

Genotype Quality=70

Genotype Quality=65

Genotype Quality=12

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	
	2	0	1	???!	2	0	Sample A
	2	0	1	2	0	2	Sample B
	0	0	1	0	2	0	Sample C
	1	2	1	1	2	0	Sample D
	2	0	2	???!	2	0	Sample E
	0	0	1	???!	2	0	Sample F
	1	0	1	???!	2	0	Sample G
	0	0	1	???!	0	2	Sample H
	2	0	???!	???!	0	2	Sample I
	0	0	1	???!	2	0	Sample J
	2	0	1	0	0	2	Sample K

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

Each genotype call also has a genotype quality score, e.g.:

Genotype Quality=11

Genotype Quality=70

Genotype Quality=65

Genotype Quality=12

The genotype quality scores reflect the likelihood the genotype call is correct.

We can declare SNPs with low genotype qualities "unknown" to help eliminate bad data.

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	
	2	0	1	???!	2	0	Sample A
	2	0	1	2	0	2	Sample B
	0	0	1	0	2	0	Sample C
	1	2	1	1	2	0	Sample D
	2	0	2	???!	2	0	Sample E
	0	0	1	???!	2	0	Sample F
	1	0	1	???!	2	0	Sample G
	0	0	1	???!	0	2	Sample H
	2	0	???!	???!	0	2	Sample I
	0	0	1	???!	2	0	Sample J
	2	0	1	0	0	2	Sample K

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

EXAMPLE:

Before
genotype
quality filter:

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	
	2	0	1	???!	2	0	**Sample A**
	2	0	1	2	0	2	**Sample B**
	0	0	1	0	2	0	**Sample C**
	1	2	1	1	2	0	**Sample D**
	2	0	2	???!	2	0	**Sample E**
	0	0	1	???!	2	0	**Sample F**
	1	0	1	???!	2	0	**Sample G**
	0	0	1	???!	0	2	**Sample H**
	2	0	???!	???!	0	2	**Sample I**
	0	0	1	???!	2	0	**Sample J**
	2	0	1	0	0	2	**Sample K**

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

EXAMPLE:

After genotype quality filter:

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	
	2	0	1	???	2	0	**Sample A**
	???	0	1	2	0	2	**Sample B**
	0	???	1	0	2	???	**Sample C**
	1	2	1	1	2	???	**Sample D**
	2	0	2	???	???	0	**Sample E**
	0	???	1	???	???	???	**Sample F**
	1	???	???	???	2	0	**Sample G**
	???	0	1	???	0	2	**Sample H**
	2	0	???	???	???	2	**Sample I**
	???	???	1	???	2	0	**Sample J**
	2	0	1	0	0	2	**Sample K**

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Breaking down graphical genotypes

| 2 |
| 2 |
| 0 |
| 1 |
| 2 |
| 0 |
| 1 |
| 0 |
| 2 |
| 0 |
| 2 |
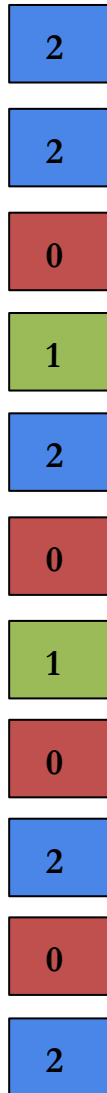
- A fairly "normal" SNP!
- Enough of each allele to detect associations
- A "reasonable" number of heterozygotes
  - How can we decide what is "reasonable"? We can use population genetic assumptions such as Hardy-Weinberg Equilibrium.
  - However … no natural populations are actually in HWE.

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

0

0

0

2

0

0

0

0

0

0

0
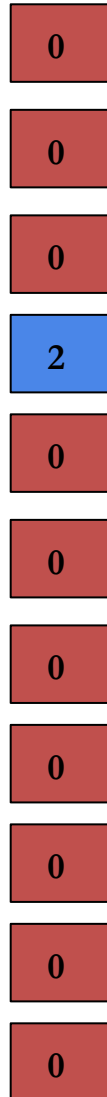
- Rare allele
- Low statistical power
- Possibly sequencing error

M.Timothy.Rabanus-Wallace @mtrw85

Graphical Genotypes

1

1

1

1

2

1

1

1

???!

1

1
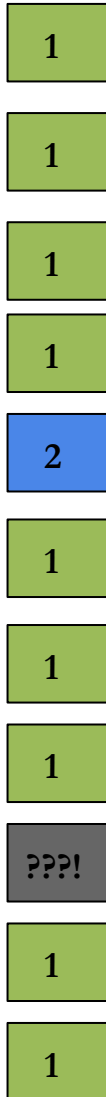
- Excess of heterozygotes
- Very likely a mapping artifact (close paralogous sequences in the real genome, but with only one representative in the reference sequence?)
- Also suffers from low association power (see previous slide)

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

???	
2	
0	
1	
???	

- Too much missing data!
- Low statistical power
- Called genotypes have high likelihood of errors
- Will rely too heavily on imputation ("guessing")

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

2	**0**
0	**2**
2	**0**
2	**0**
2	**0**
2	**0**
2	**0**
0	**2**
0	**2**
2	**0**
0	**2**

M.Timothy.Rabanus-Wallace @mtrw85

Linkage Disequilibrium

Ancestral haplotypes

Generations

Linkage Disequilibrium

Ancestral haplotypes

Generations

Linkage Disequilibrium

Ancestral haplotypes

Generations

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium

Linkage Disequilibrium

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium

M.Timothy.Rabanus-Wallace @mtrw85
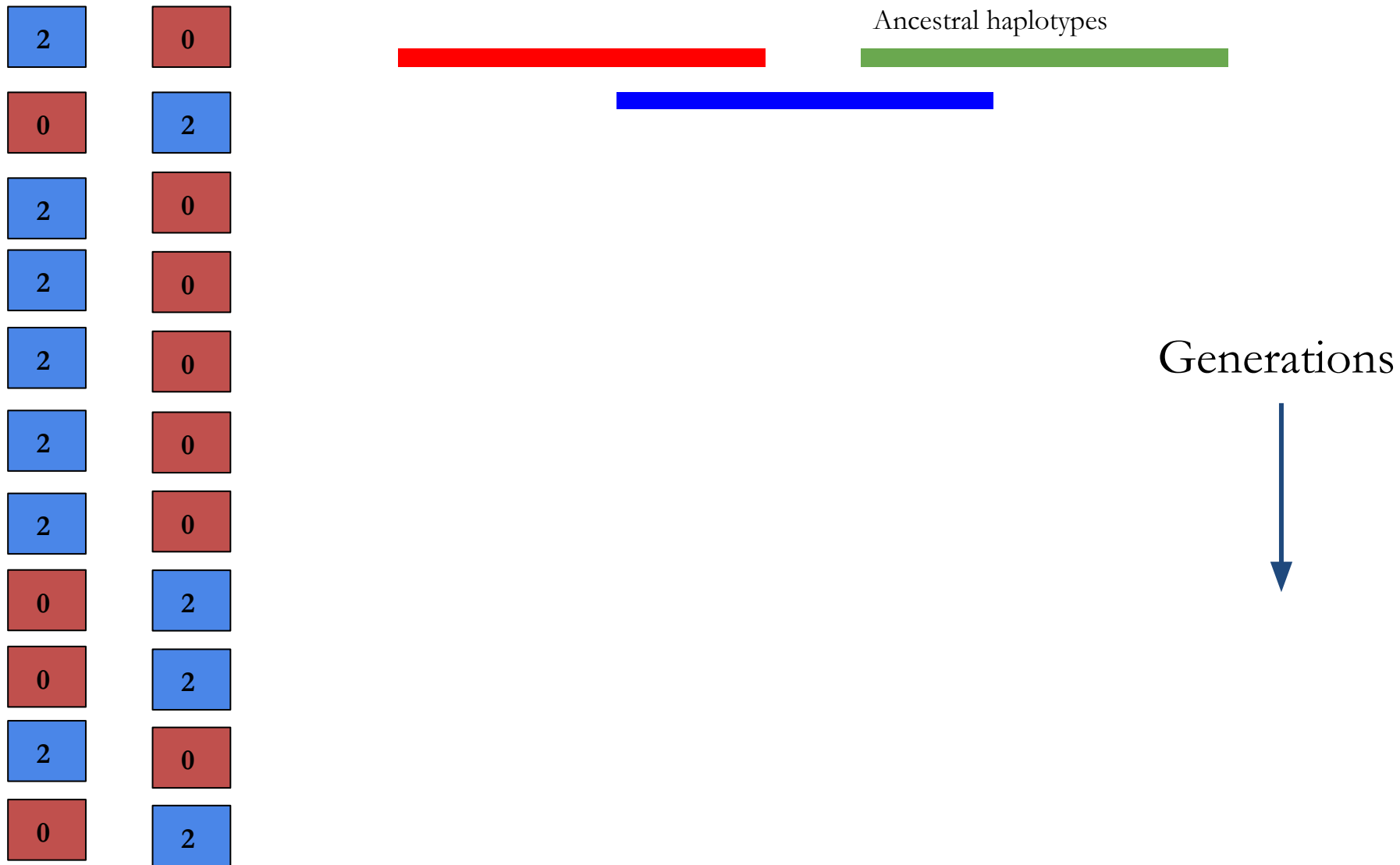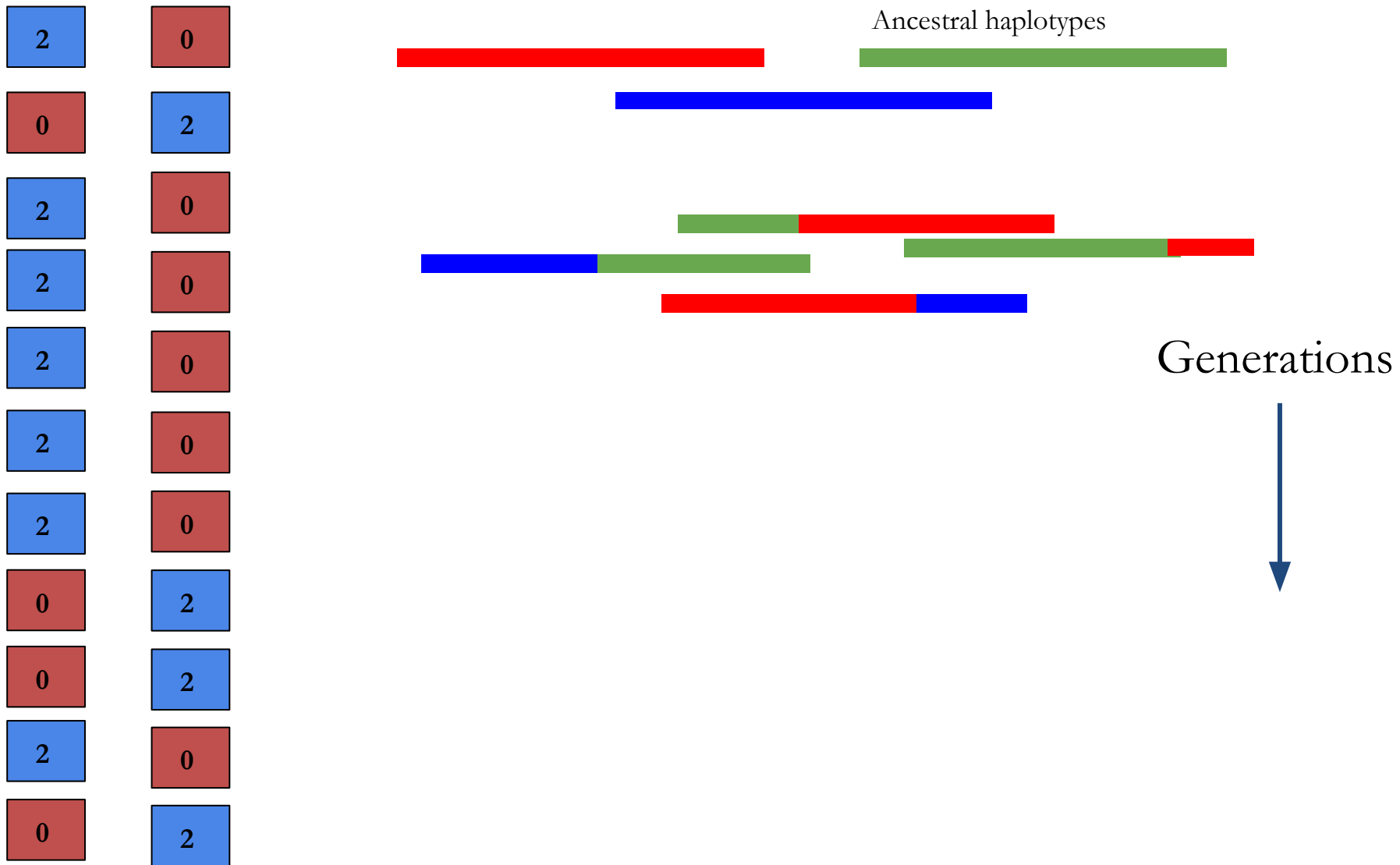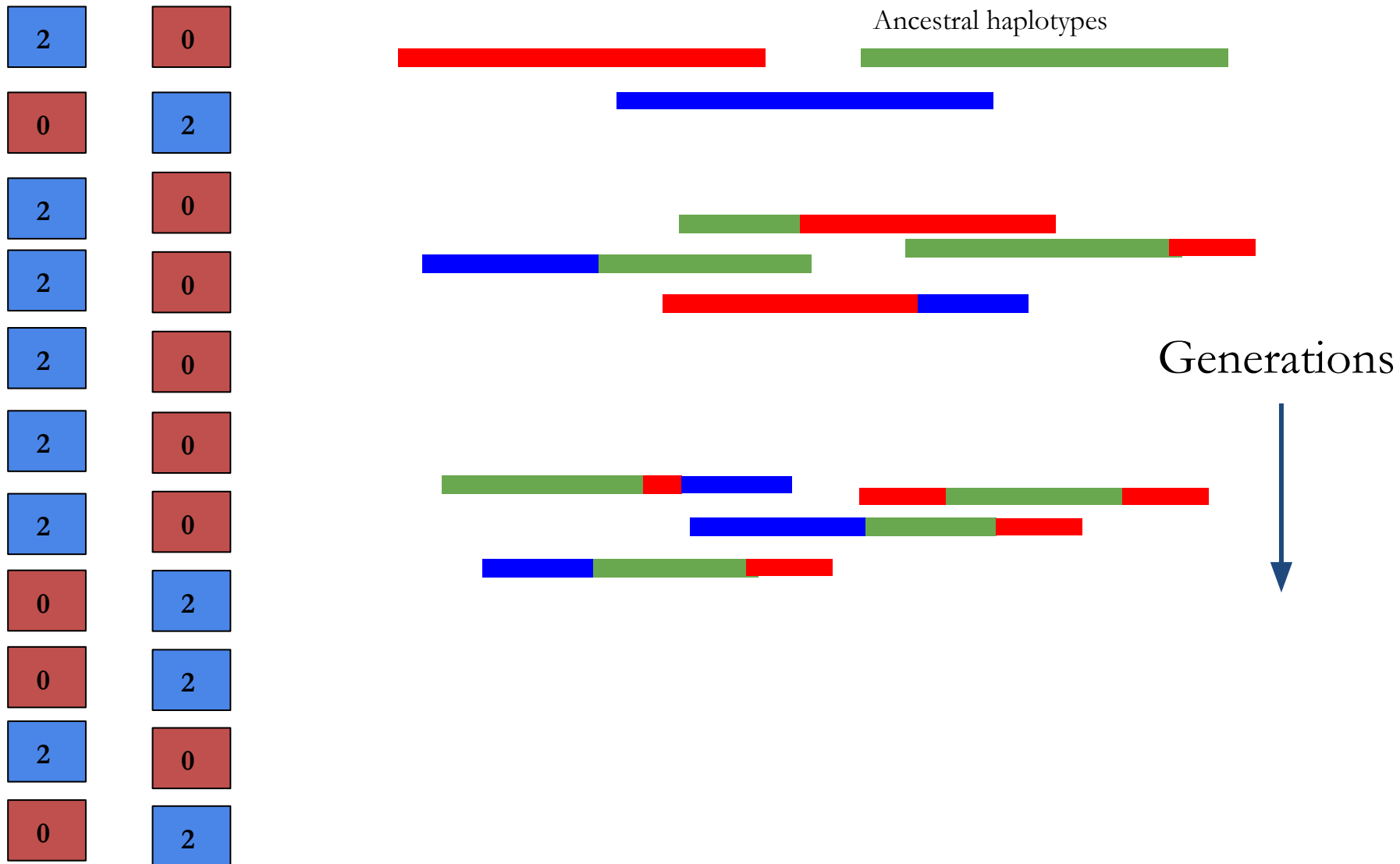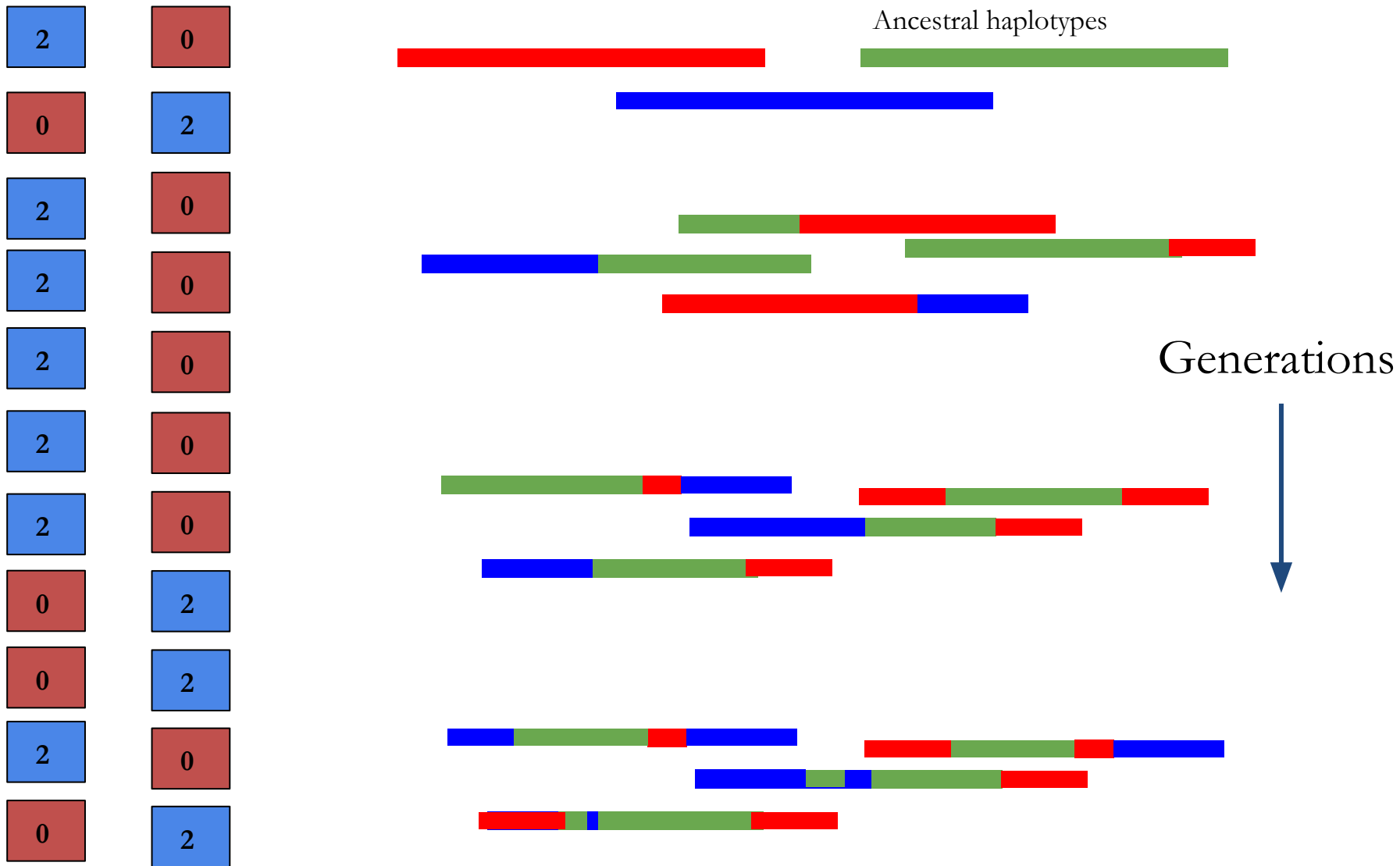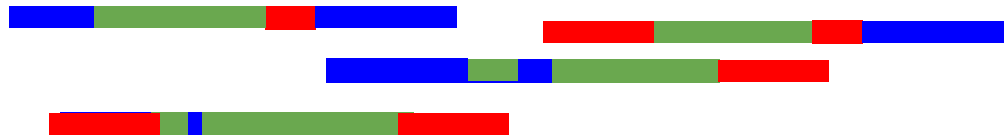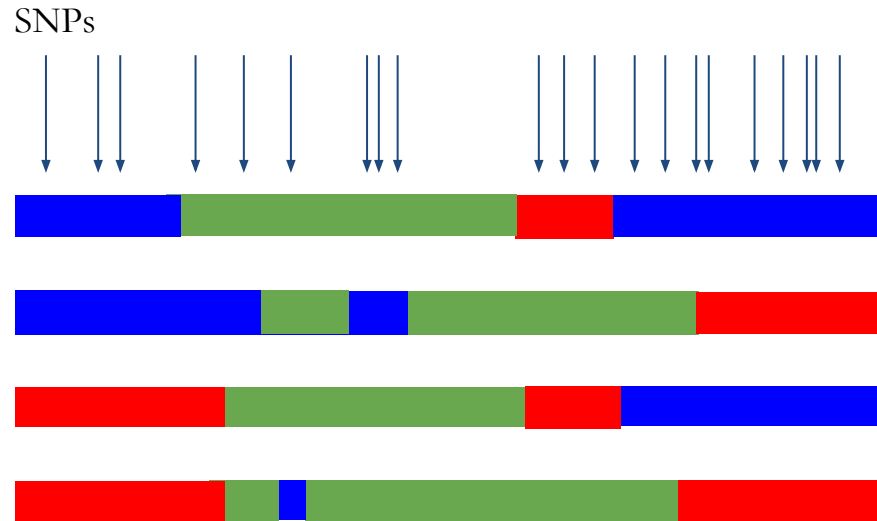
Linkage Disequilibrium

Nearby SNPs often have similar states because they come from the same ancestral haplotype blocks.

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium

SNPs

= Allele from red ancestor

= Allele from blue ancestor

= Allele from green ancestor

Nearby SNPs often have similar states because they come from the same ancestral haplotype blocks.

M.Timothy.Rabanus-Wallace @mtrw85

Linkage Disequilibrium Plots

Colour according to similarity (in this case)

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium "Pruning"

2	0
0	2
2	0
2	0
2	0
2	0
2	0
0	2
0	2
2	0
0	2

Represent linked groups of SNPs with representative markers.

SNPs are usually part of small linkage blocks

Since there are many more SNPs than linkage blocks, some are in a sense "redundant".

A pruned set ...

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

2	0
0	2
2	0
2	0
2	0
2	0
2	0
0	2
0	2
2	0
0	2

- Linked alleles
- Not a problem per se, but:
  - Increase computing time
  - Lead to excessively strict corrections for multiple hypothesis testing (e.g. Bonferroni correction)

Graphical Genotypes

2	0
0	2
2	0
2	0
2	0
2	0
2	0
0	2
0	2
2	0
0	2

- Linked alleles
- Not a problem per se, but:
  - Increase computing time
  - Lead to excessively strict corrections for multiple hypothesis testing (e.g. Bonferroni correction)
- LD Pruning is *not always performed!*
  - We do it in the course mainly to increase your familiarity with how genetic markers and linkage works
- Linkage is an important concept because it underlies:
  - Imputation
  - Association studies

Imputation

M.Timothy.Rabanus-Wallace @mtrw85

Imputation

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Imputation
(same challenge, rows rearranged)

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Imputation

- Linkage between these SNPs can help us impute ("guess") missing data
  - In this case, two distinct haplotypes exist at this linkage block
  - Imputation, therefore, needs to be done BEFORE removing linked markers!
- Missing data can be difficult to handle for some analyses
- Imputing data can increase the number of available markers and increase the statistical power of markers with missing data
- Imputation can be automated by several algorithms
- Accuracy depends on the quality of the non-missing data and the type of algorithm used
- Accuracy can be tested by "imputing" sites that are actually non-missing, and checking the error rate

Pipeline for the course

Sequence data

Sequence data quality control (QC)

Map short sequences to genome

Reference genome (chromosome 1H of barley)

Identify ("call") SNPs

SNP QC

Population Structure Analysis

Phenotype association study

Phenotype data

Gene annotation data

Final result

primary data worksop

data QC workshop

data analysis workshop

M.Timothy.Rabanus-Wallace ✈ @mtrw85
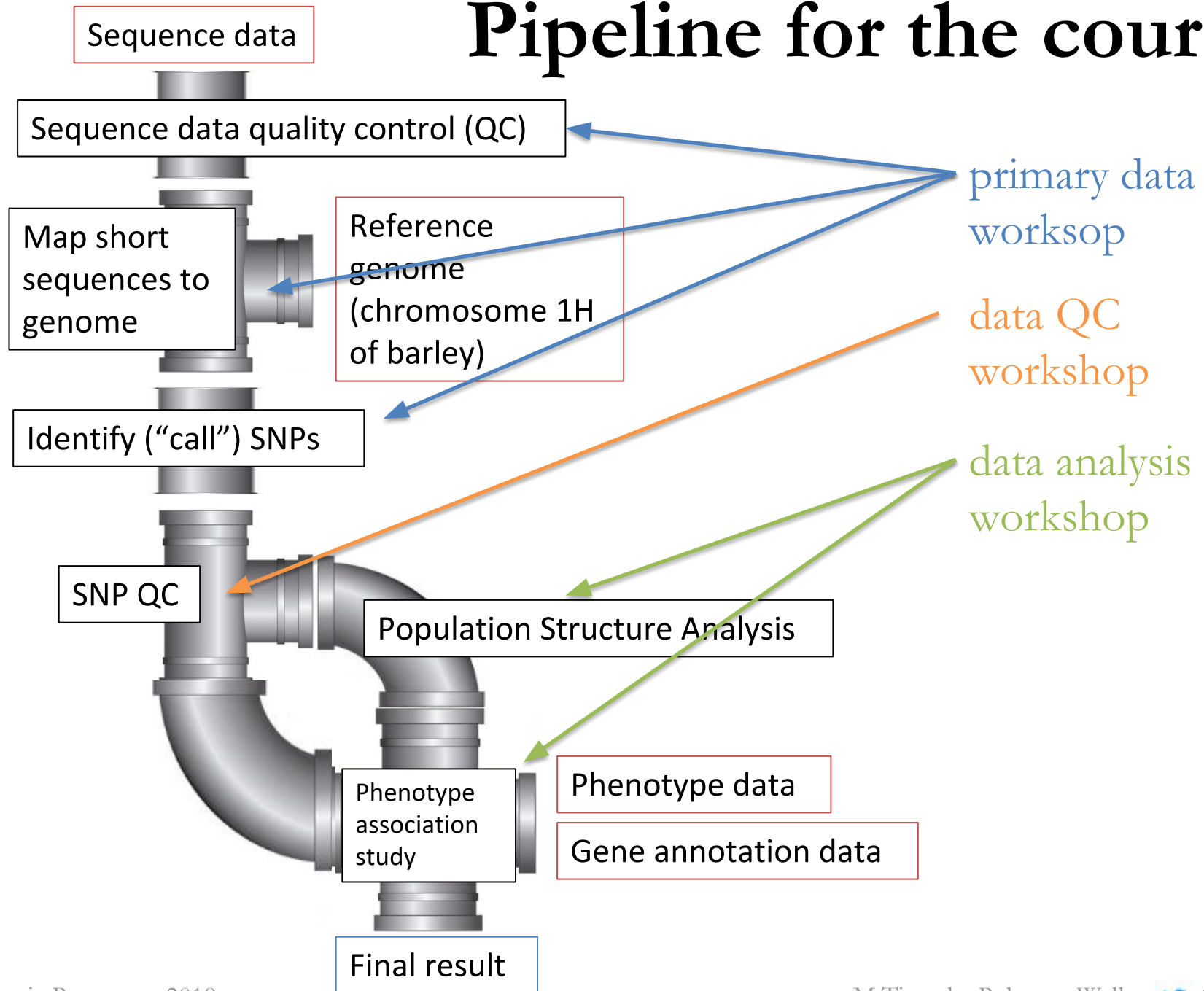
Quality control for SNP data: Recap session

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Pipeline for the course

Sequence data

Sequence data quality control (QC)

Map short sequences to genome

Reference genome (chromosome 1H of barley)

Identify ("call") SNPs

SNP QC

Population Structure Analysis

Phenotype association study

Phenotype data

Gene annotation data

Final result

primary data worksop

data QC workshop

data analysis workshop

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

Rare allele (low MAF)

M.Timothy.Rabanus-Wallace @mtrw85

Graphical Genotypes

Rare allele (low MAF)

Excessive heterozygosity

Graphical Genotypes

Rare allele (low MAF)

Excessive heterozygosity

Too much missing data

Graphical Genotypes

Rare allele (low MAF)

Excessive heterozygosity

Too much missing data

Linked alleles (can be pruned to single marker)

Sequential Filtering

Set low genotype qualities to NA

Set low genotype qualities to NA

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Remove SNPs with too much missing data

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Remove SNPs with too much missing data

Remove SNPs with low MAF (rare alleles)

Remove SNPs with low MAF (rare alleles)

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Remove SNPs with excessive heterozygosity

Remove SNPs with excessive heterozygosity

Impute missing genotypes ...

M.Timothy.Rabanus-Wallace @mtrw85

Impute missing genotypes ...

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Prune linked SNPs to give a representative set

Prune linked SNPs to give a representative set

2	… Final SNP data set …	**2**
2		**0**
0		**2**
1		**2**
2		**2**
0		**2**
1		**2**
0		**0**
2		**0**
0		**2**
2		**0**

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium

Ancestral haplotypes

Generations

Linkage Disequilibrium

Ancestral haplotypes

Generations

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium

Ancestral haplotypes

Generations

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium

Linkage Disequilibrium

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium

SNPs

Linkage Disequilibrium

SNPs

Nearby SNPs often have similar states because they come from the same ancestral haplotype blocks.

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium

SNPs

● = Allele from red ancestor

● = Allele from blue ancestor

● = Allele from green ancestor

Nearby SNPs often have similar states because they come from the same ancestral haplotype blocks.

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium Plots

Colour according to similarity (in this case)

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Linkage Disequilibrium "Pruning"

2	**0**
0	**2**
2	**0**
2	**0**
2	**0**
2	**0**
2	**0**
0	**2**
0	**2**
2	**0**
0	**2**

Represent linked groups of SNPs with representative markers.

SNPs are usually part of small linkage blocks

Since there are many more SNPs than linkage blocks, some are in a sense "redundant".

A pruned set ...

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

2	0
0	2
2	0
2	0
2	0
2	0
2	0
0	2
0	2
2	0
0	2

- Linked alleles
- Not a problem per se, but:
  - Increase computing time
  - Lead to excessively strict corrections for multiple hypothesis testing (e.g. Bonferroni correction)

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Graphical Genotypes

2	0
0	2
2	0
2	0
2	0
2	0
2	0
0	2
0	2
2	0
0	2

- Linked alleles
- Not a problem per se, but:
  - Increase computing time
  - Lead to excessively strict corrections for multiple hypothesis testing (e.g. Bonferroni correction)
- LD Pruning is *not always performed!*
  - We do it in the course mainly to increase your familiarity with how genetic markers and linkage works
- Linkage is an important concept because it underlies:
  - Imputation
  - Association studies

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Imputation

M.Timothy.Rabanus-Wallace @mtrw85

Imputation

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Imputation
(same challenge, rows rearranged)

2	0	2	2	2	2	0	0	2	Sample A
2	0	2	2	2	2	0	0	2	Sample E
2	0	2	2	2	2	0	0	2	Sample I
2	0	2	2	2	2	0	0	2	Sample J
0	2	2	0	2	0	0	0	0	Sample B
0	2	2	0	2	0	0	0	0	Sample C
0	2	2	0	2	0	0	0	0	Sample D
0	2	2	0	2	0	0	0	0	Sample F
0	2	2	0	2	0	0	0	0	Sample G
0	2	2	0	2	0	0	0	0	Sample H

M.Timothy.Rabanus-Wallace @mtrw85

Imputation

2	0	2	2	2	2	0	0	2	Sample A
2	0	2	2	2	2	0	0	2	Sample E
2	0	2	2	2	2	0	0	2	Sample I
2	0	2	2	2	2	0	0	2	Sample J
0	2	2	0	2	0	0	0	0	Sample B
0	2	2	0	2	0	0	0	0	Sample C
0	2	2	0	2	0	0	0	0	Sample D
0	2	2	0	2	0	0	0	0	Sample F
0	2	2	0	2	0	0	0	0	Sample G
0	2	2	0	2	0	0	0	0	Sample H

- Linkage between these SNPs can help us impute ("guess") missing data
  - In this case, two distinct haplotypes exist at this linkage block
  - Imputation, therefore, needs to be done BEFORE removing linked markers!
- Missing data can be difficult to handle for some analyses
- Imputing data can increase the number of available markers and increase the statistical power of markers with missing data
- Imputation can be automated by several algorithms
- Accuracy depends on the quality of the non-missing data and the type of algorithm used
- Accuracy can be tested by "imputing" sites that are actually non-missing, and checking the error rate

Further Material
(NOT REQUIRED IN THIS COURSE)
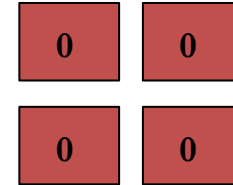
Hardy-Weinberg Equilibrium
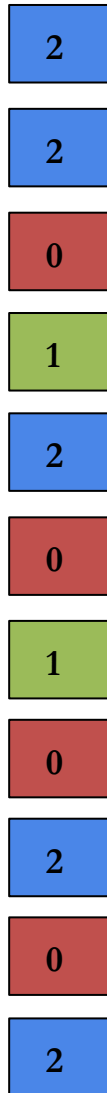
M.Timothy.Rabanus-Wallace @mtrw85

Hardy-Weinberg Equilibrium

Estimating the allele frequencies ...

| 2 | 2 |

| 0 | 0 |

| 2 | 2 |

| 0 | 0 |

| 2 | 2 |

| 1 | 1 |

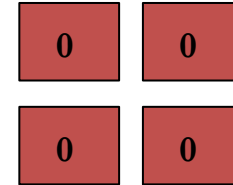# Hardy-Weinberg Equilibrium

Estimating the allele frequencies ...

2	2
2	2
2	2

0	0
0	0

M.Timothy.Rabanus-Wallace @mtrw85

Hardy-Weinberg Equilibrium

Estimating the allele frequencies ...

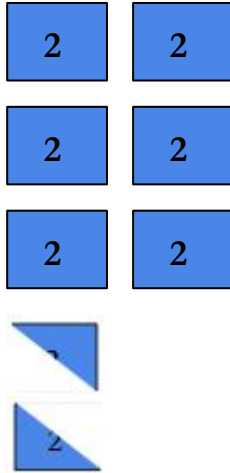M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Hardy-Weinberg Equilibrium

Predict the effects of
random mating

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Hardy-Weinberg Equilibrium

Predict the effects of random mating

A heterozygote!

M.Timothy.Rabanus-Wallace 🐦 @mtrw85
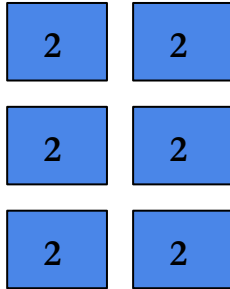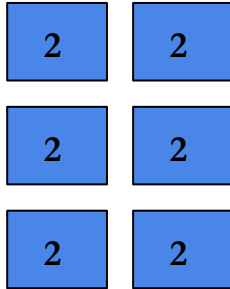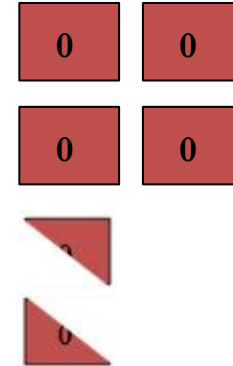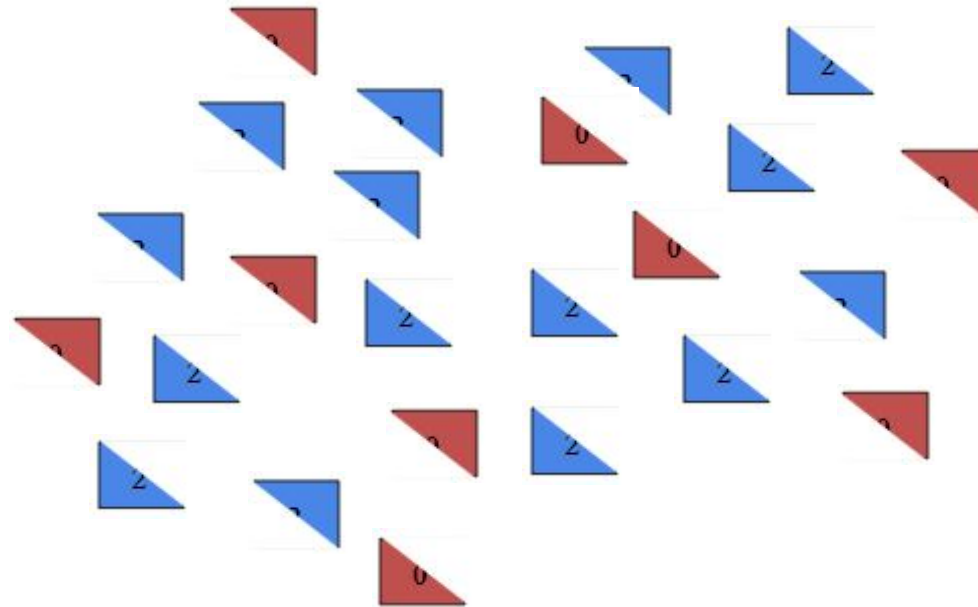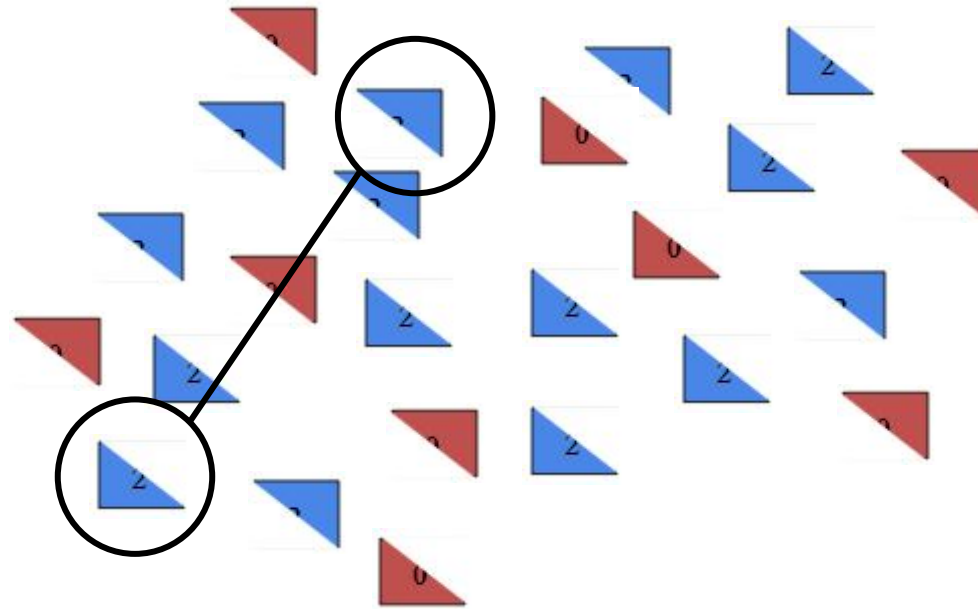
Hardy-Weinberg Equilibrium

Predict the effects of
random mating
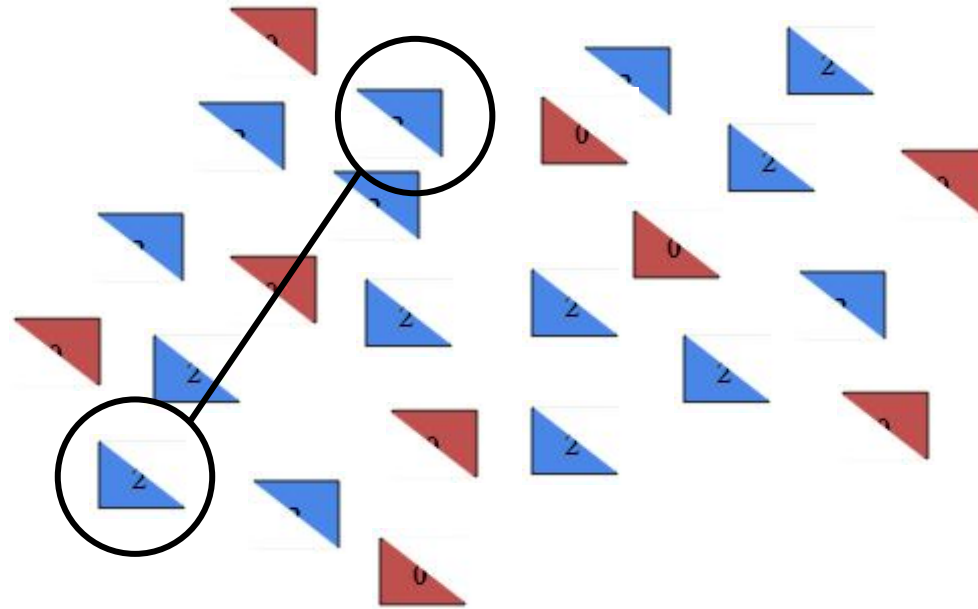
A homozygote blue!

Red : 0
Het : 0
Blue : 1

Hardy-Weinberg Equilibrium

Predict the effects of random mating

A heterozygote!

Red : 0
Het : 1
Blue : 1

Hardy-Weinberg Equilibrium

Predict the effects of random mating

Red : 0
Het : 1
Blue : 2

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Hardy-Weinberg Equilibrium

Predict the effects of
random mating

Red : 1
Het : 1
Blue : 2

M.Timothy.Rabanus-Wallace 🐦 @mtrw85
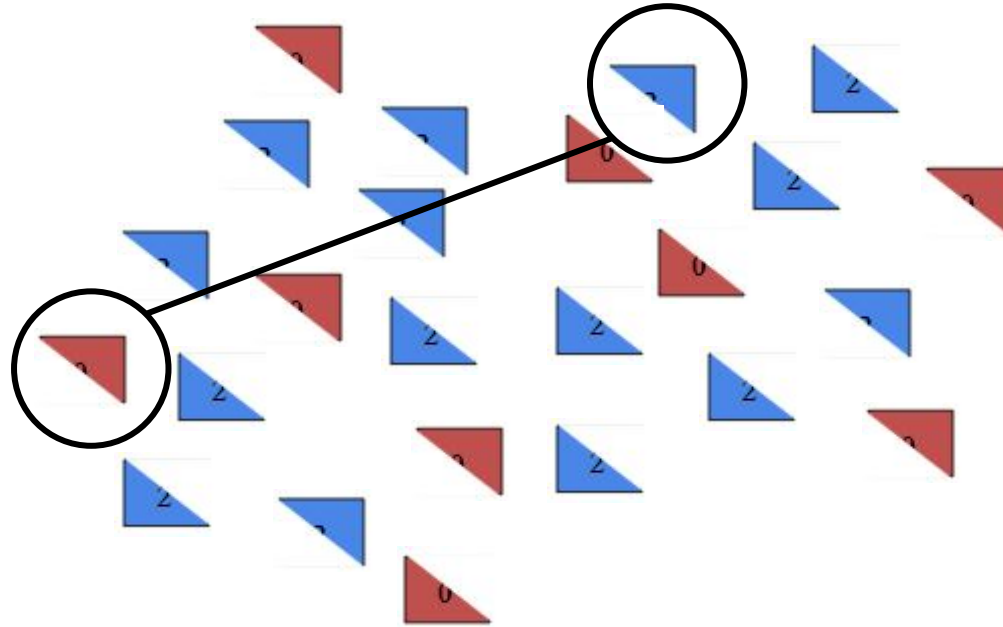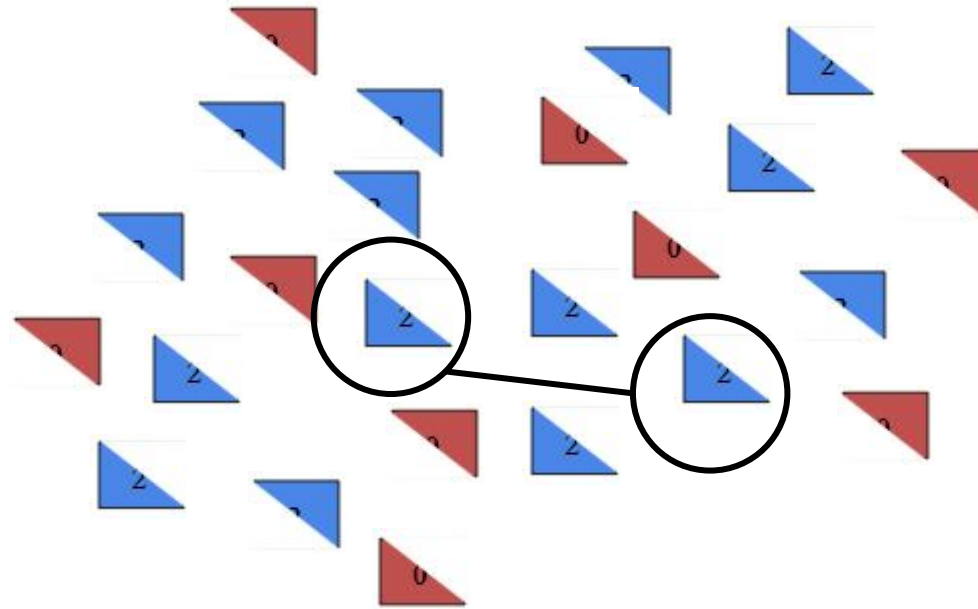
Hardy-Weinberg Equilibrium

Predict the effects of random mating

Red : 1
Het : 1
Blue : 3

Hardy-Weinberg Equilibrium

Predict the effects of random mating
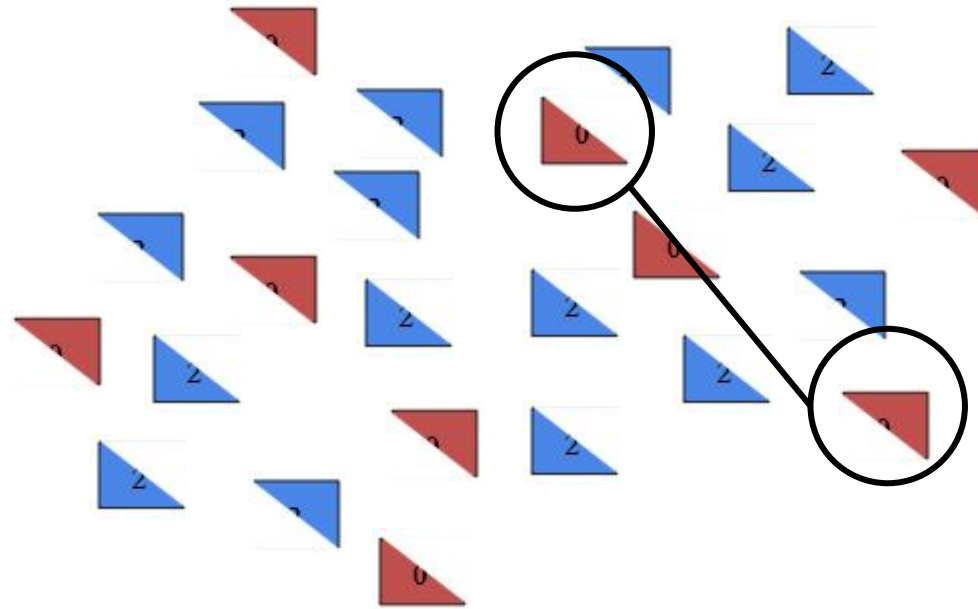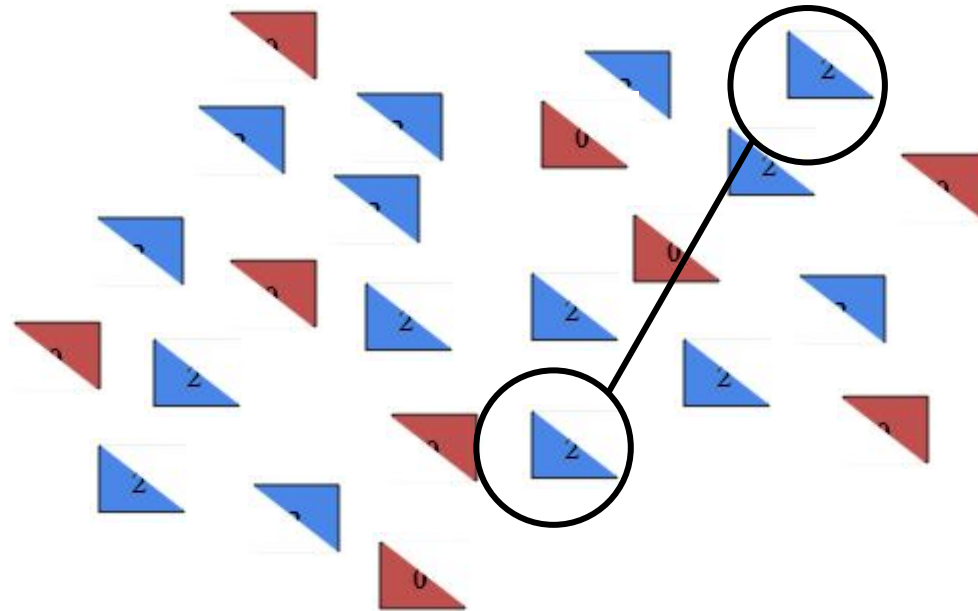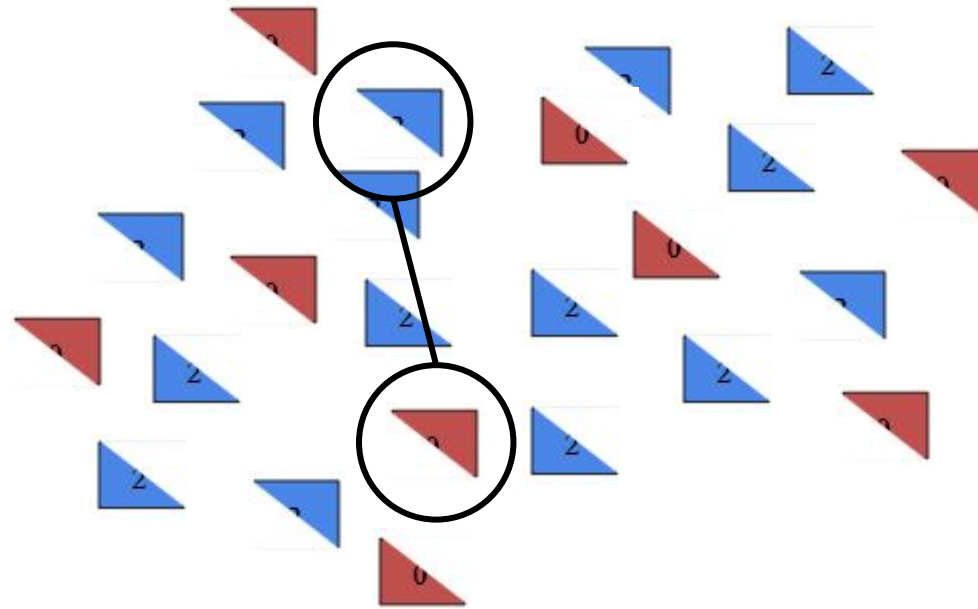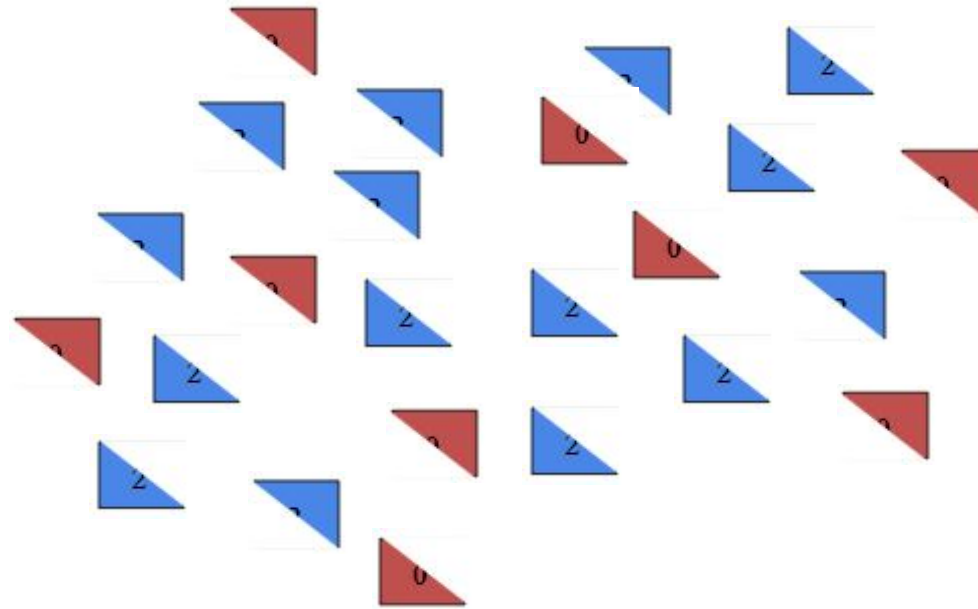

Red : 1
Het : 2
Blue : 3

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Hardy-Weinberg Equilibrium

Predict the effects of random mating

Red : 34%
Het : 49%
Blue : 17%

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Hardy-Weinberg Equilibrium

Calculate probabilities of certain numbers of heterozygotes in the population …
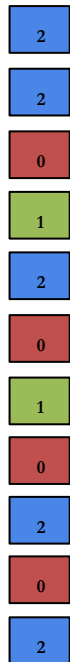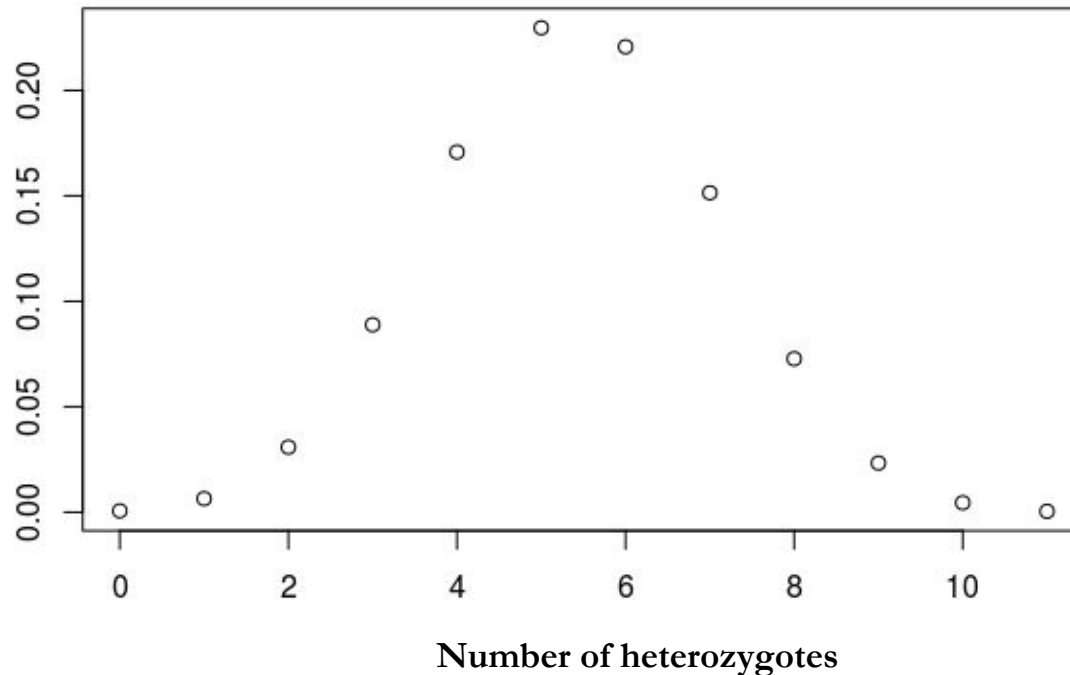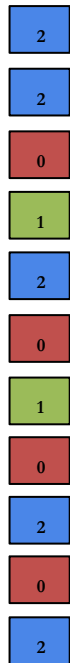
Red : 34%
Het : 49%
Blue : 17%

Hardy-Weinberg Equilibrium

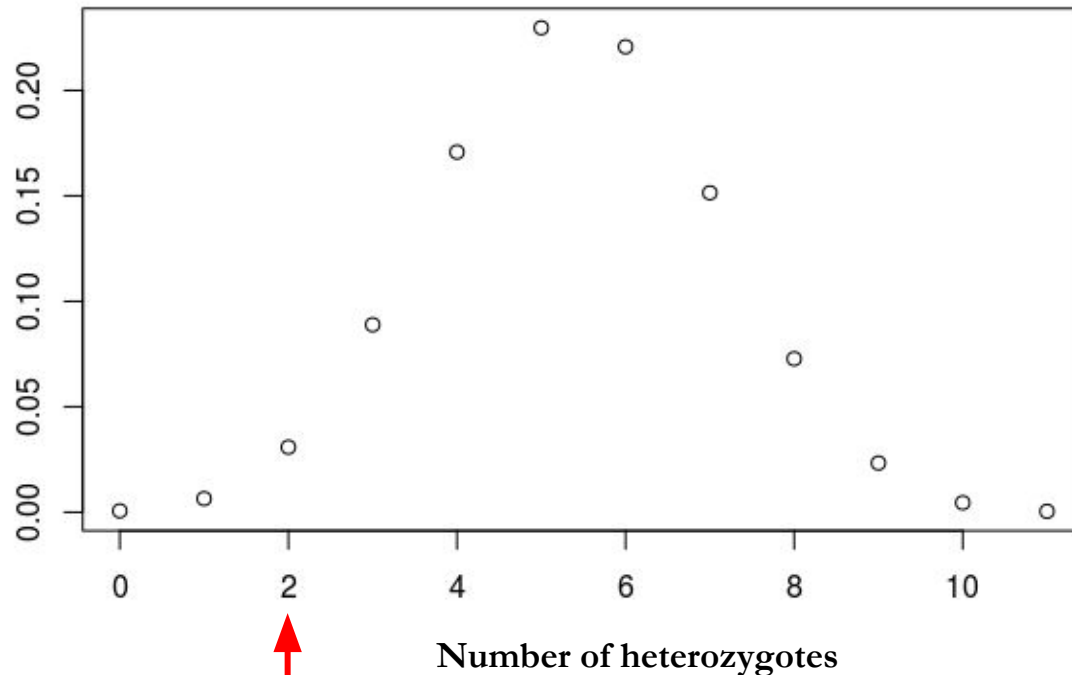Calculate probabilities of certain numbers of heterozygotes in the population ...

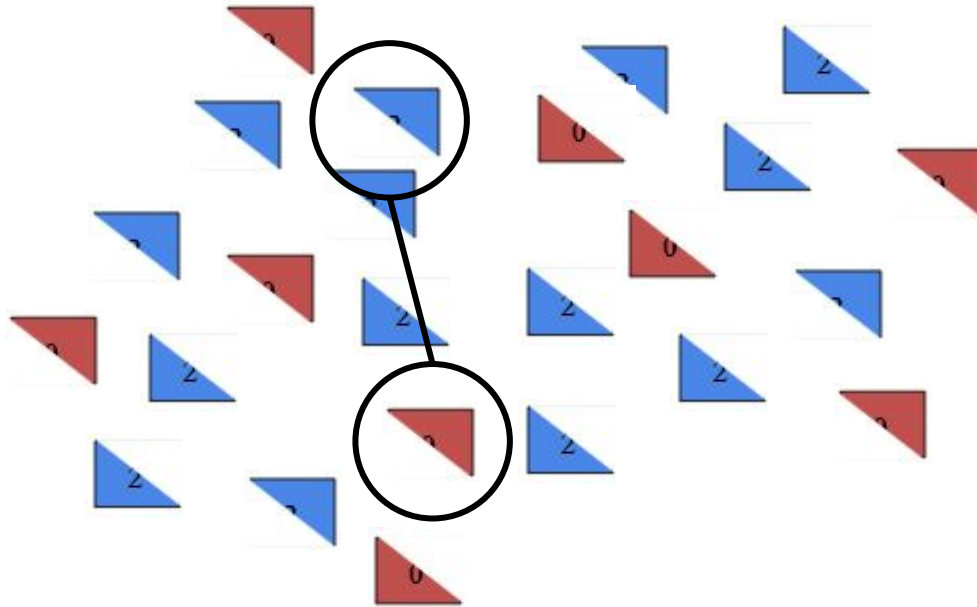**Red : 34%**
Het : 49%
Blue : 17%

Probability

Number of heterozygotes

Hardy-Weinberg Equilibrium

So, based on HWE assumptions, not very likely to be a correctly-genotyped, representative SNP … BUT

M.Timothy.Rabanus-Wallace 🐦 @mtrw85

Hardy-Weinberg Equilibrium

So, based on HWE assumptions, not very likely to be a correctly-genotyped, representative SNP … BUT

Does anyone actually believe this is how mating works in a natural population?