Genotyping, Linkage Analysis and Haplotype Phasing in Autopolyploids

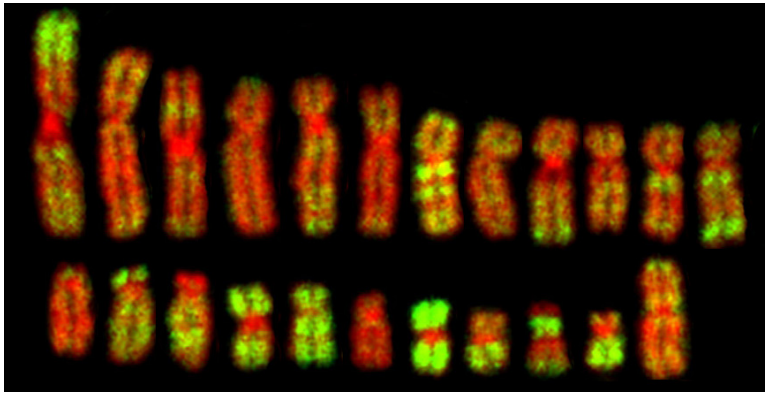
Marcelo Mollinari
mmollin@ncsu.edu

Bioinformatics Research Center - Hicks Hall 368
Department of Statistics
North Carolina State University

September 12 and 17, 2019 - GN 757 001 - North Carolina State University

Ploidy level

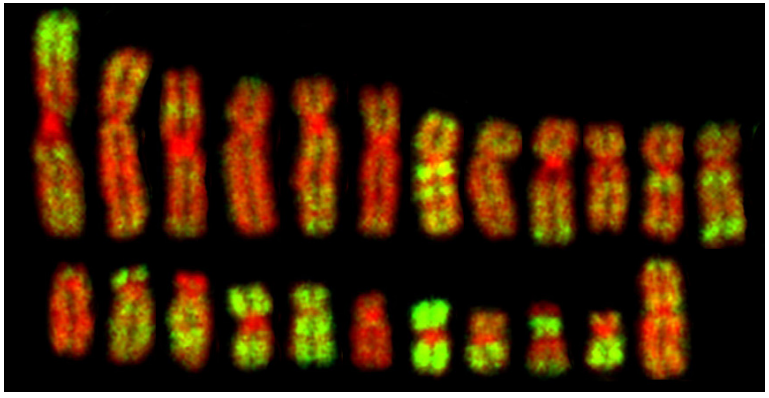
- **Basic chromosome number**: the number of different chromosomes that make up a single complete set. Humans, $x = 23$



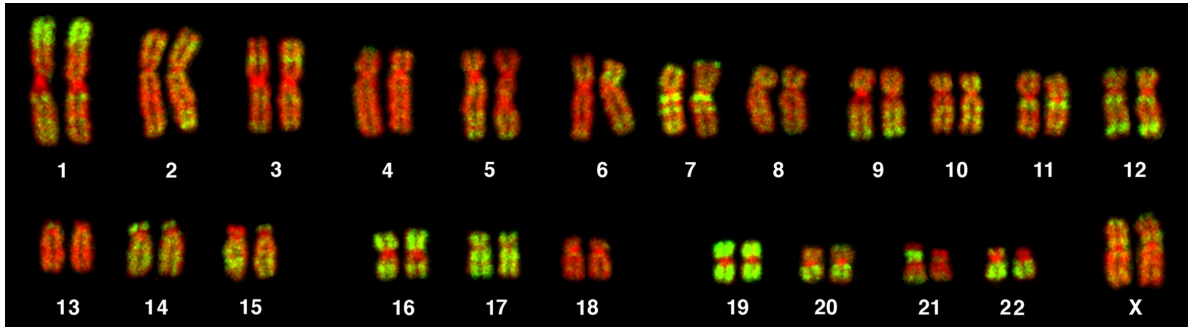
- **Ploidy level**: Number of basic chromosome sets. Humans $p = 2$

Ploidy level

- **Basic chromosome number**: the number of different chromosomes that make up a single complete set. Humans, $x = 23$

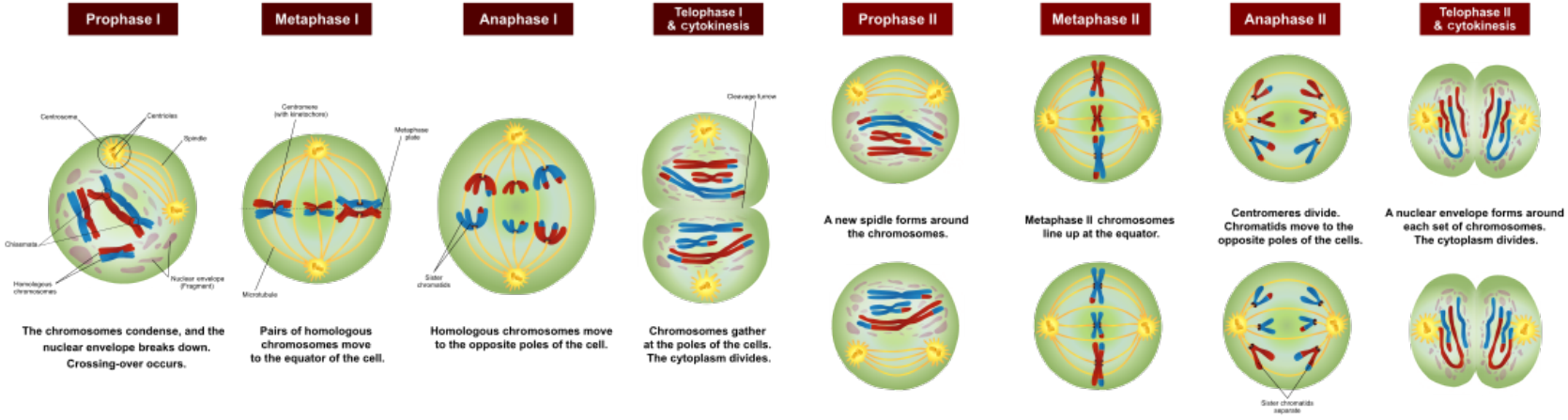


- **Ploidy level**: Number of basic chromosome sets. Humans $p = 2$



Meiotic pairing and gamete formation in diploids - Review

- Meiotic process is quite stable
- It results in four products with $n = x = basic\ number$



Meiotic pairing and gamete formation in diploids

Animation

Polyploid species

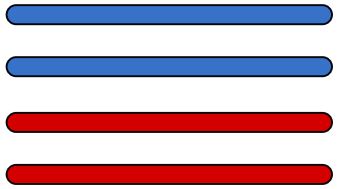
Organisms that have multiple sets of chromosomes

Polyploid species

Organisms that have multiple sets of chromosomes

Autotetraploid

Alloteraploid

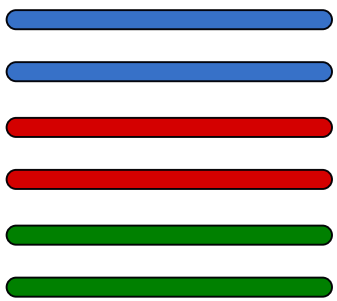


Polyploid species

Organisms that have multiple sets of chromosomes

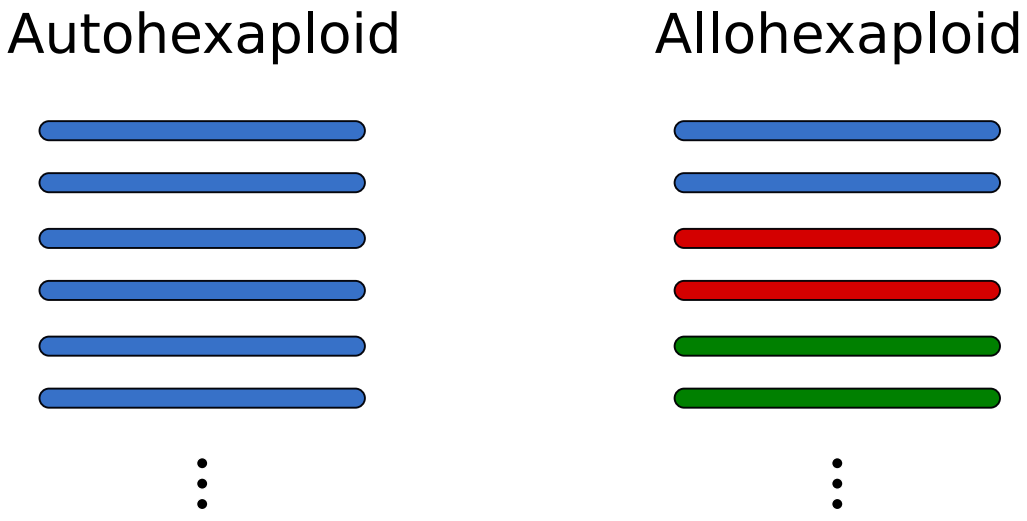
Autohexaploid

Allohexaploid

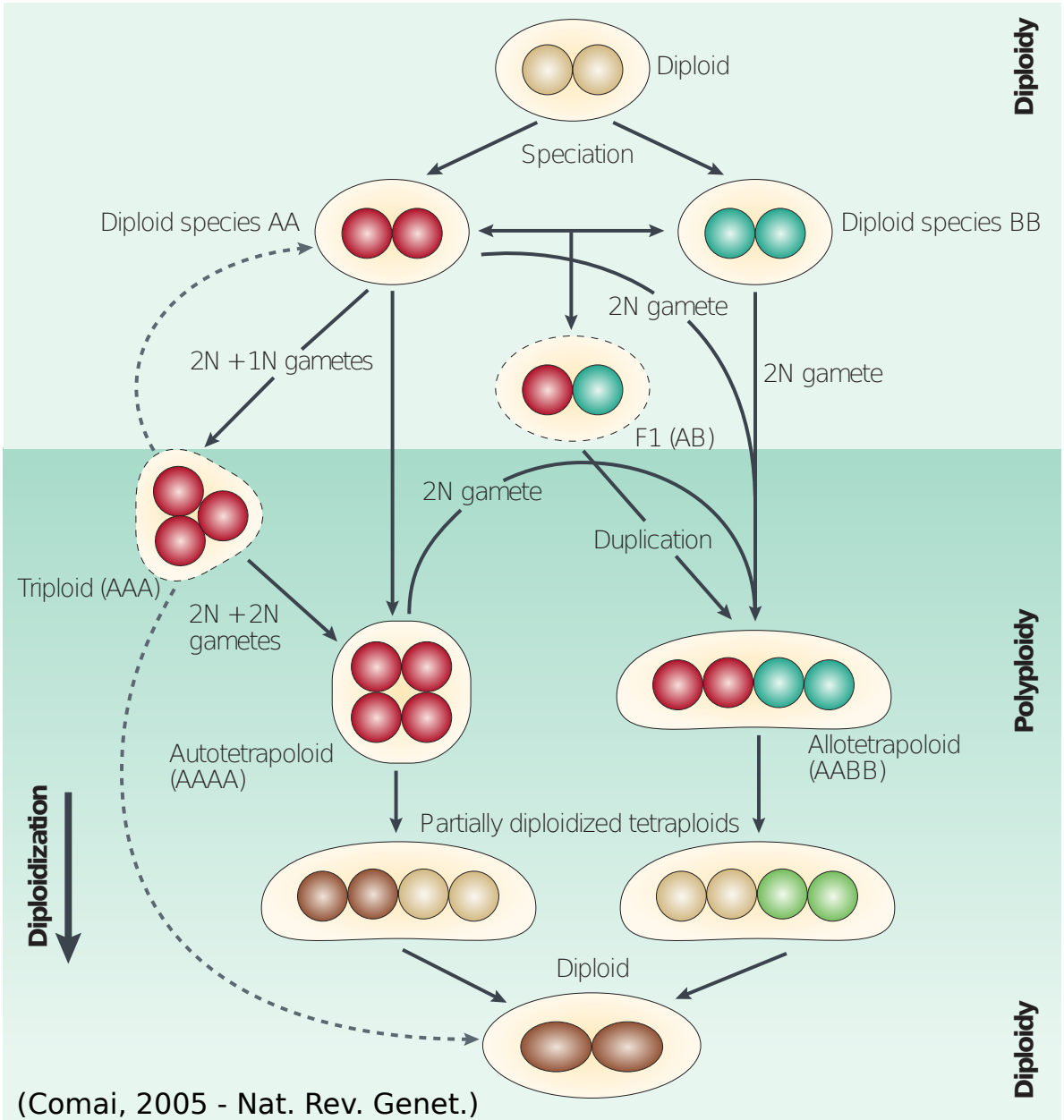


Polyploid species

Organisms that have multiple sets of chromosomes



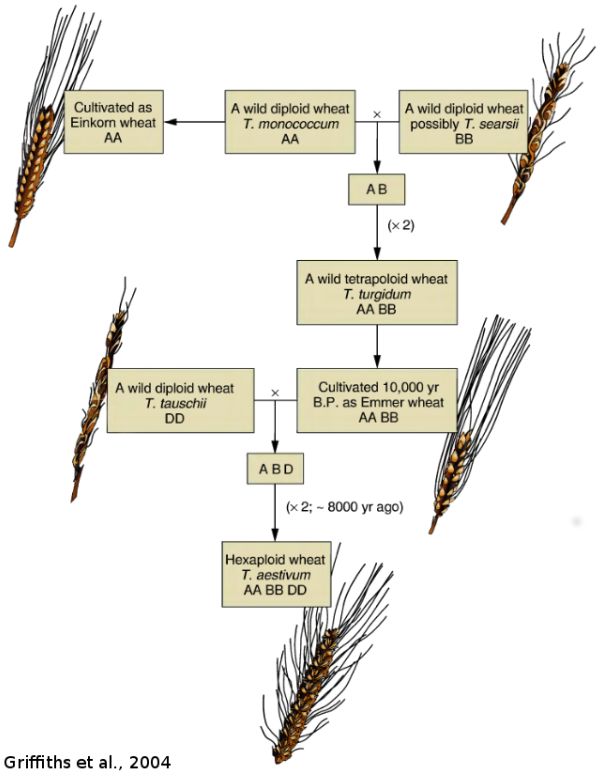
How are polyploids formed?



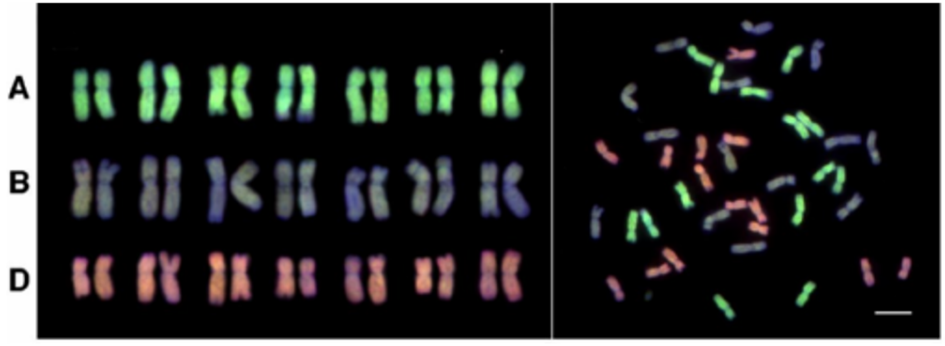
(Comai, 2005 - Nat. Rev. Genet.)

Allopolyploid example - Wheat

Wheat - *Tritucum aestivum*



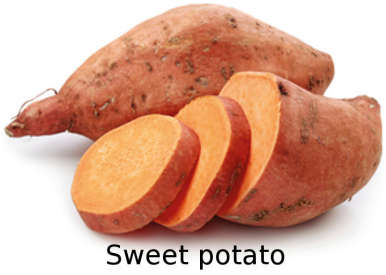
Griffiths et al., 2004

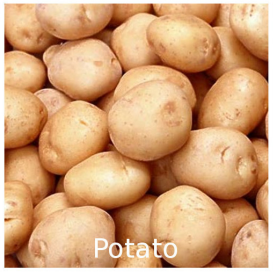


Zhang H et al. PNAS 2013;110:3447-3452

Autopolyploid example


Chrysanthemum


Sweet potato

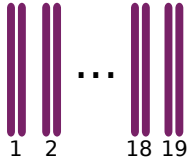

Potato


Rose

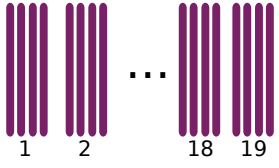

Forage crops


Sugarcane

Diploid

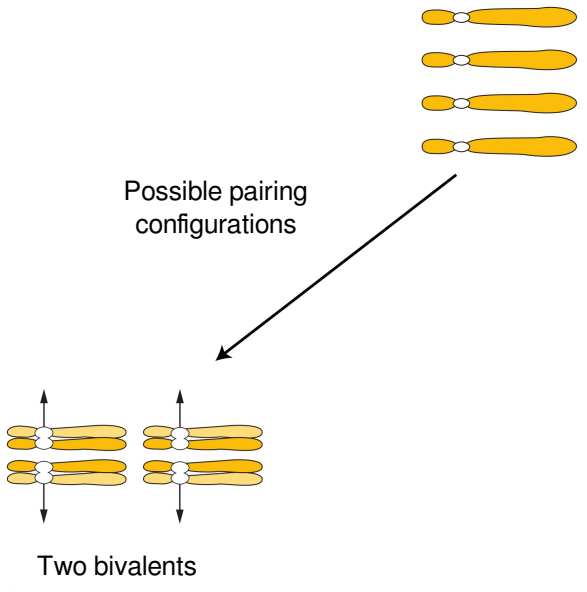


Autotetraploid



Meiotic pairing and gamete formation in autopolyploids

Tetraploid example

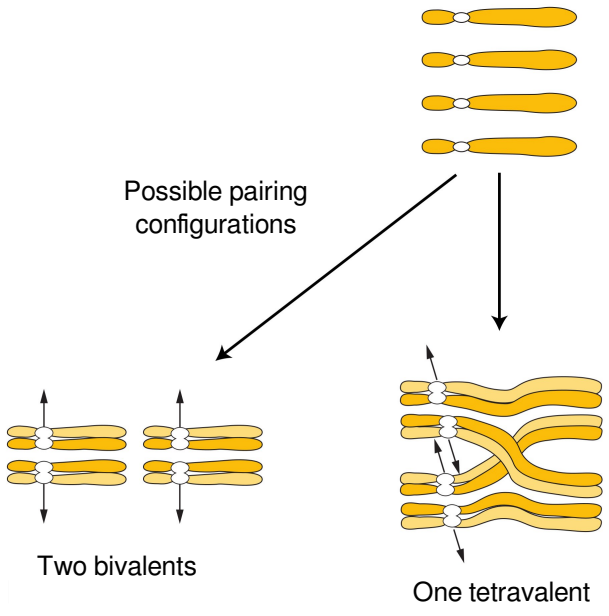


Griffiths et al. (2004)

Usually, for high ploidy levels, we assume bivalent pairing

Meiotic pairing and gamete formation in autopolyploids

Tetraploid example

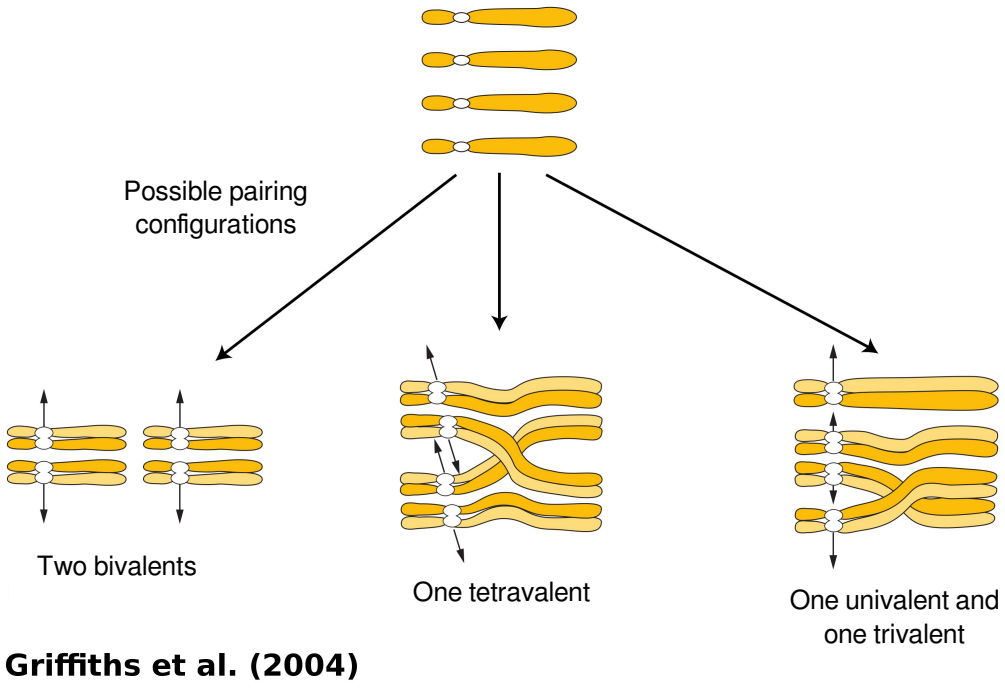


Griffiths et al. (2004)

Usually, for high ploidy levels, we assume bivalent pairing

Meiotic pairing and gamete formation in autopolyploids

Tetraploid example

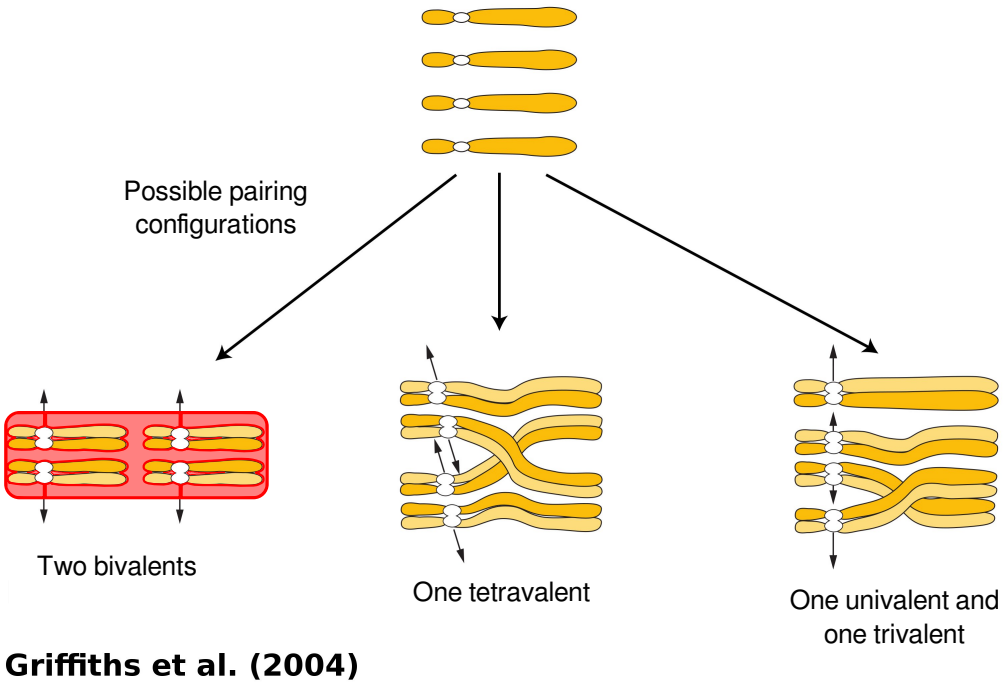


Griffiths et al. (2004)

Usually, for high ploidy levels, we assume bivalent pairing

Meiotic pairing and gamete formation in autopolyploids

Tetraploid example

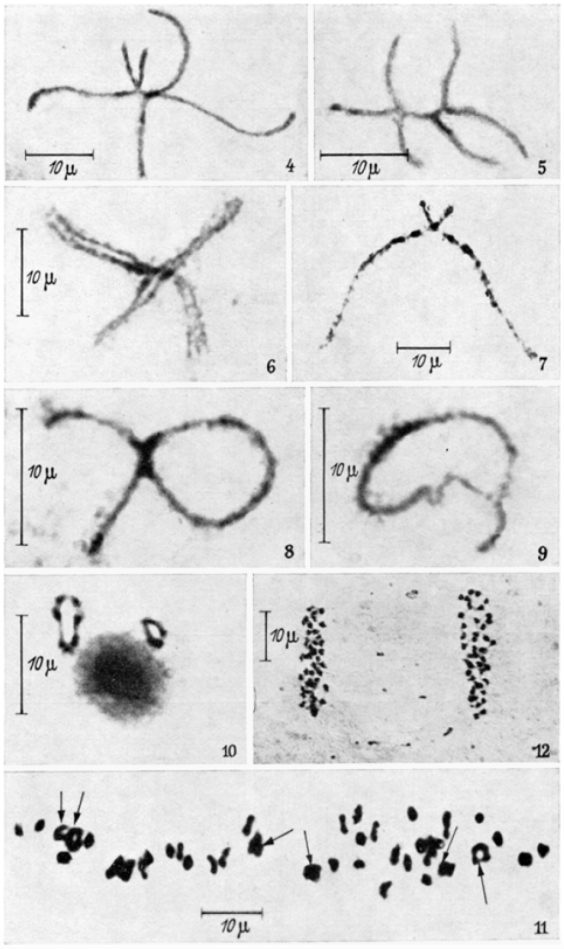


Griffiths et al. (2004)

Usually, for high ploidy levels, we assume bivalent pairing

Meiotic pairing in autopolyploids

Sweetpotato



Hexavalents, quadrivalents and bivalents in sweetpotato (Magoon *et al.* 1970)

Sugarcane

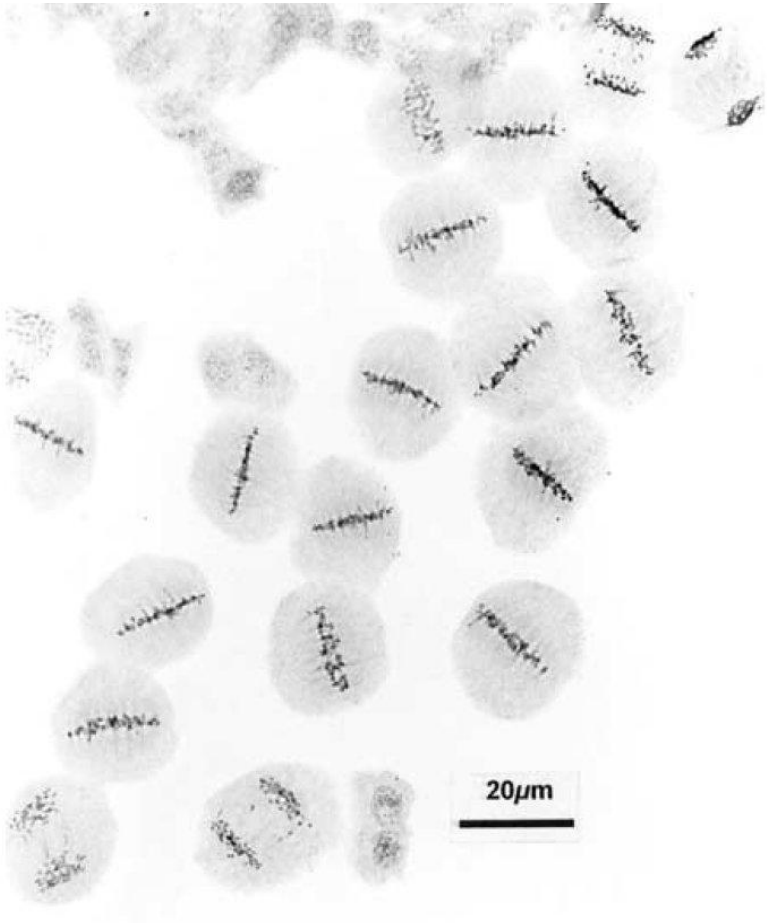
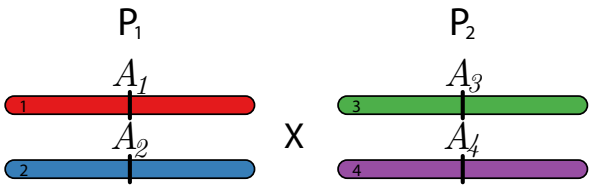


Figure 1. PMCs of *Saccharum* spp. hybrid clone 79N9059 at meiosis. As was the case in other clones, pairing was regular, bivalents generally formed. (Bielig *et al.* 2003)

Accessing the allelic variation - Diploids

Multiallelic



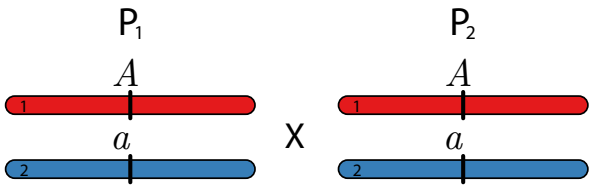
Gametes

P2 \ P1	A_1	A_2
A_3		
A_4		

4 possible genotypes

1:1:1:1

Biallelic



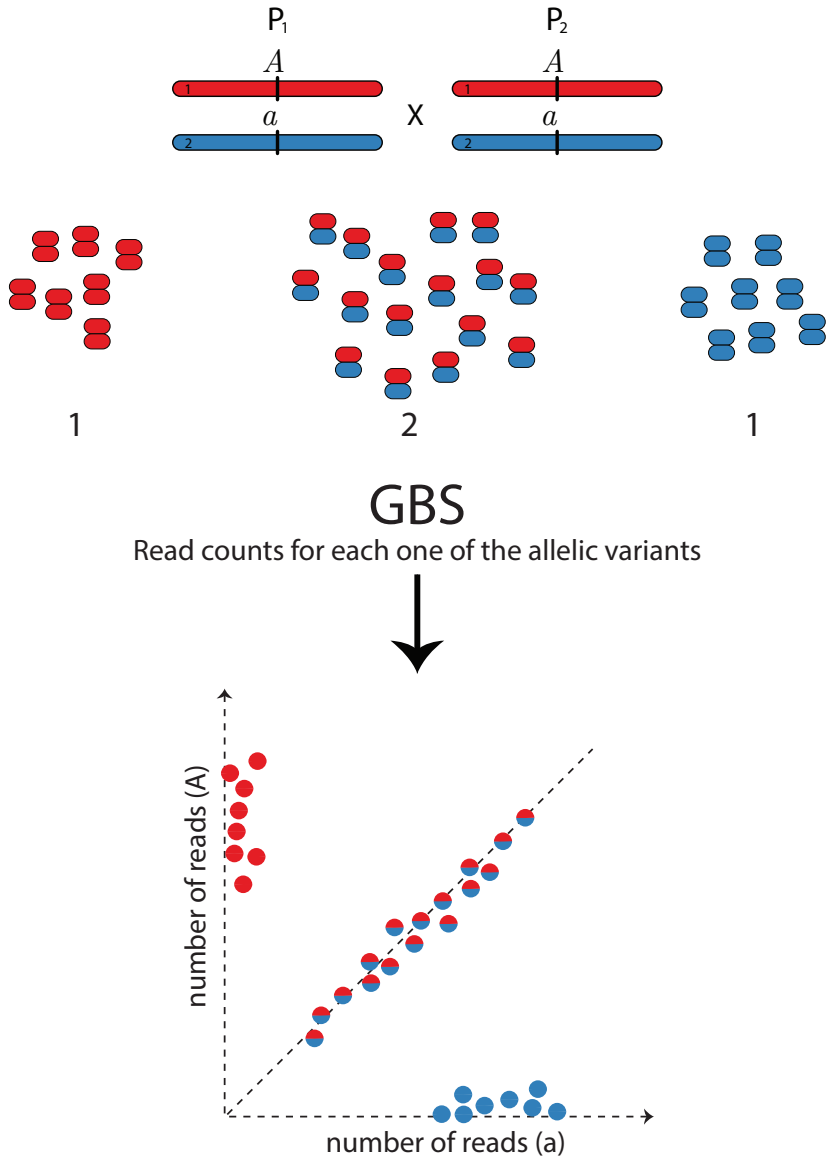
Gametes

P2 \ P1	A	a
A		
a		

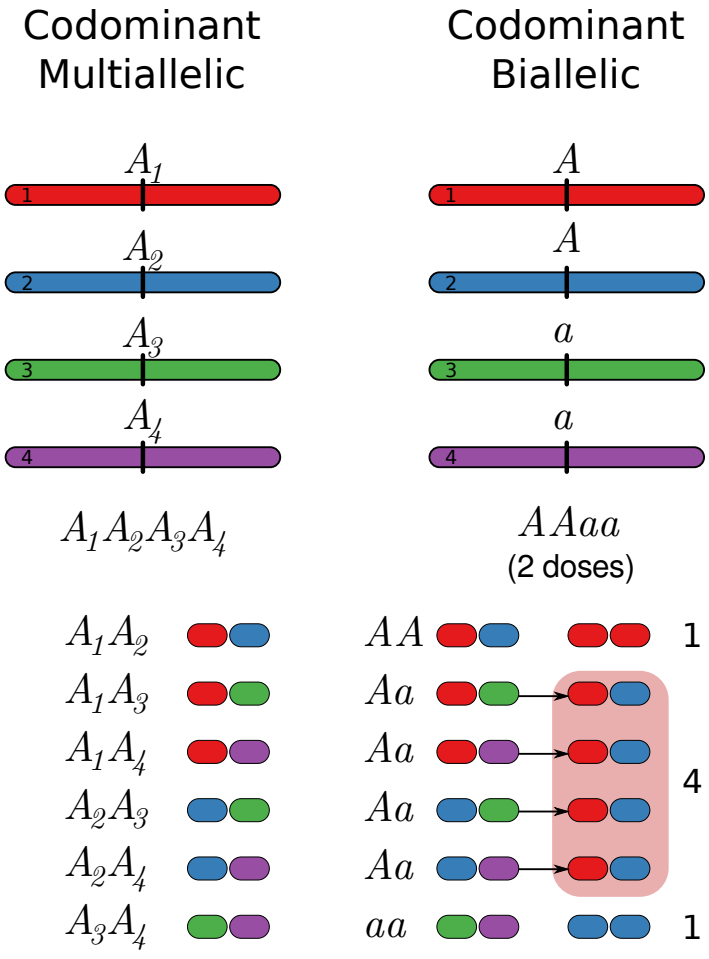
3 possible genotypes

1:2:1

Accessing the allelic variation - Diploids



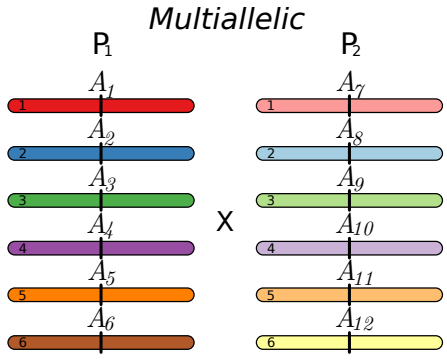
Biallelic vs. Multiallelic - Polyploids

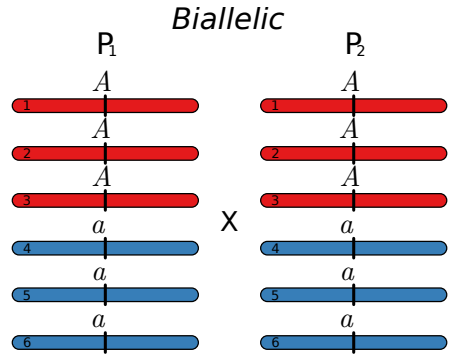


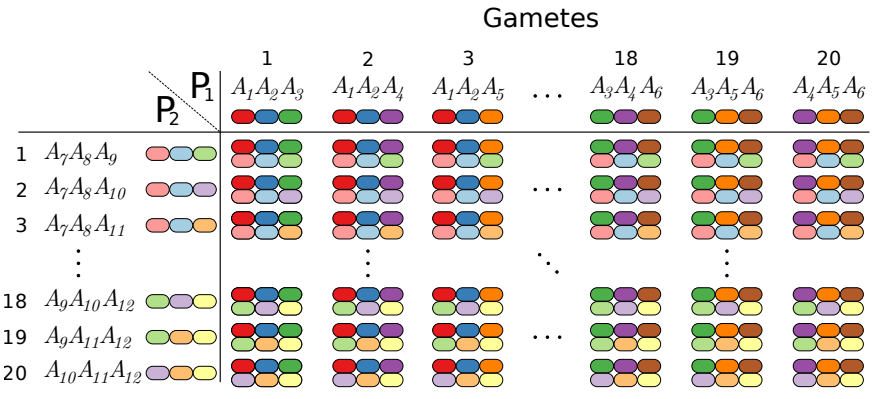
Number of possible gametes considering one locus in one and two parents

Ploidy	$\binom{p}{\frac{p}{2}}$	$\binom{p}{\frac{p}{2}}^2$
4	6	36
6	20	400
8	70	4900
10	252	63504
12	924	853776
14	3432	11778624
16	12870	165636900

Accessing the allelic variation - Hexaploid Example

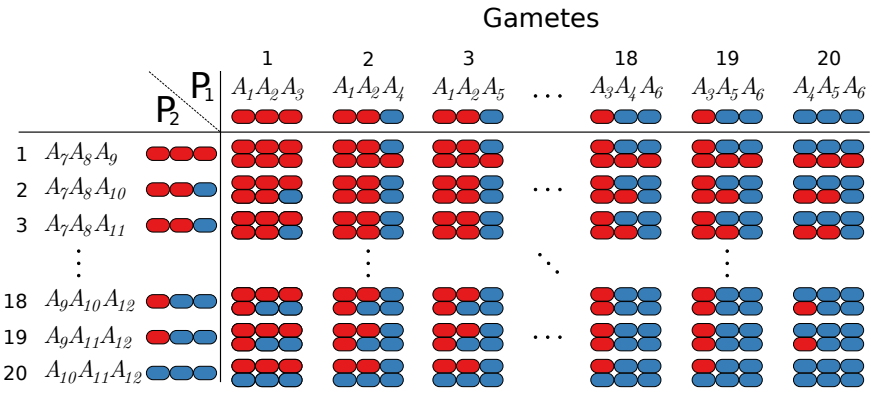






400 possible genotypes

1 : 1 : ... : 1 : 1

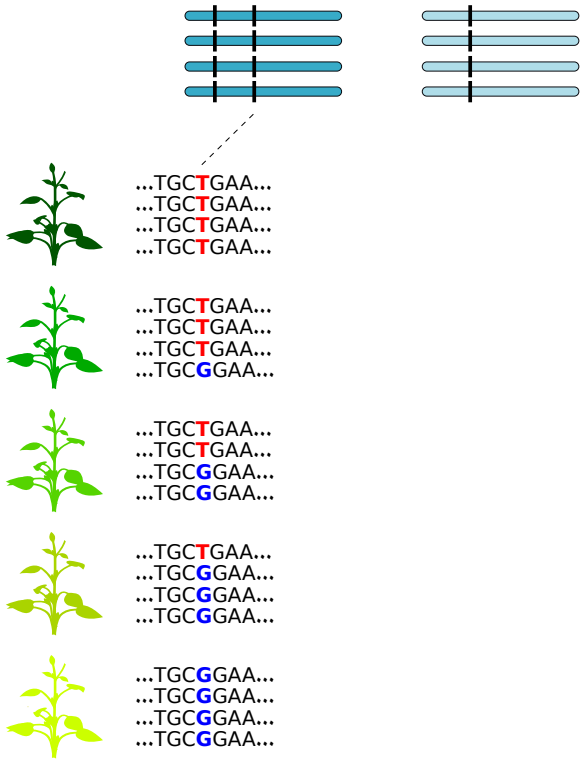


7 possible genotypes

1 : 18 : 99 : 164 : 99 : 18 : 1

Genotyping in polyploids

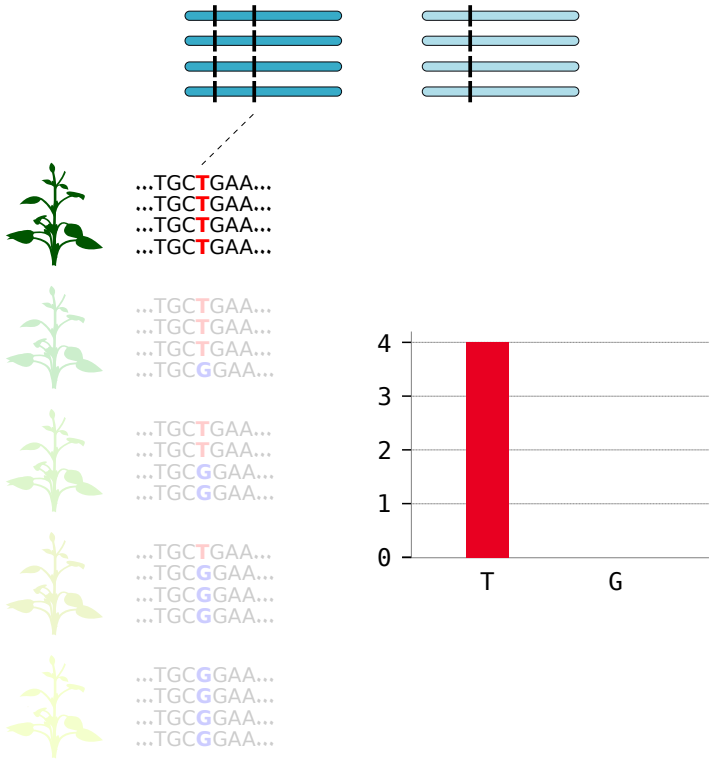
Biallelic co-dominant marker



Quantitative genotyping: access the abundance of each one of the allelic forms (in a biallelic marker)

Genotyping in polyploids

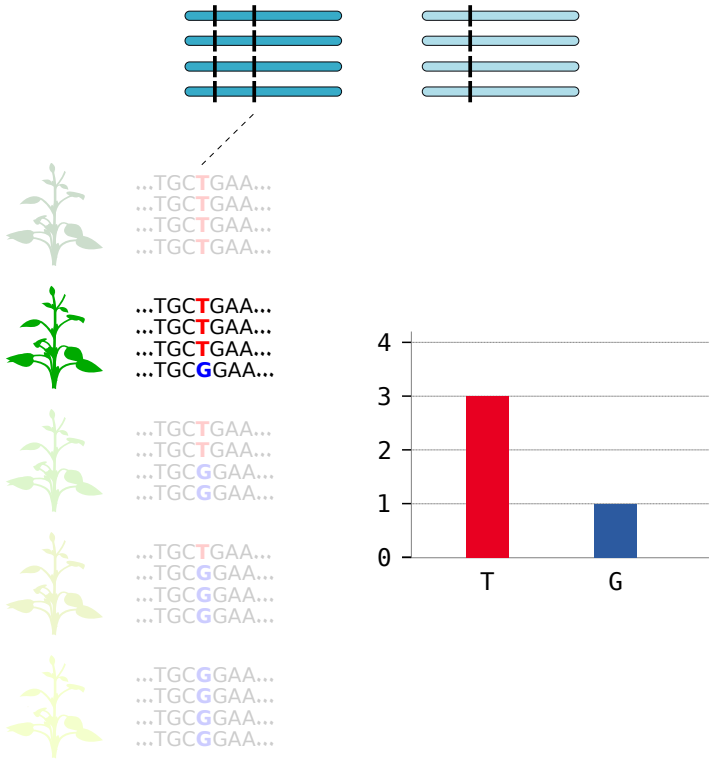
Biallelic co-dominant marker



Quantitative genotyping: access the abundance of each one of the allelic forms (in a biallelic marker)

Genotyping in polyploids

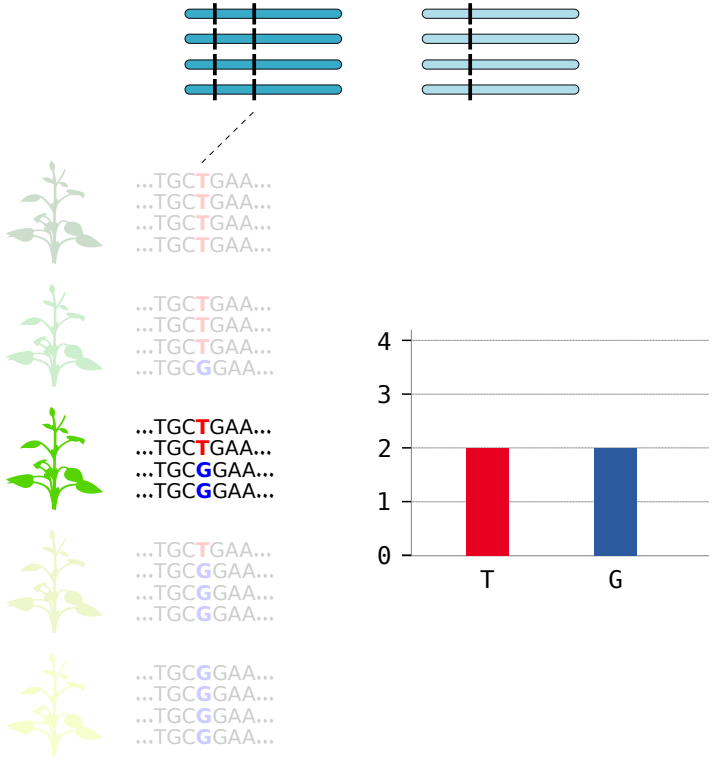
Biallelic co-dominant marker



Quantitative genotyping: access the abundance of each one of the allelic forms (in a biallelic marker)

Genotyping in polyploids

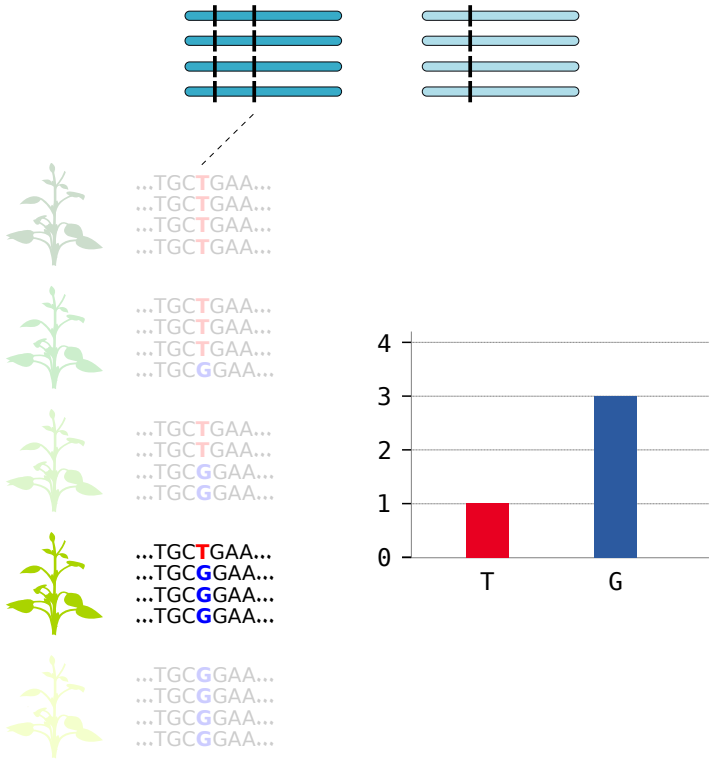
Biallelic co-dominant marker



Quantitative genotyping: access the abundance of each one of the allelic forms (in a biallelic marker)

Genotyping in polyploids

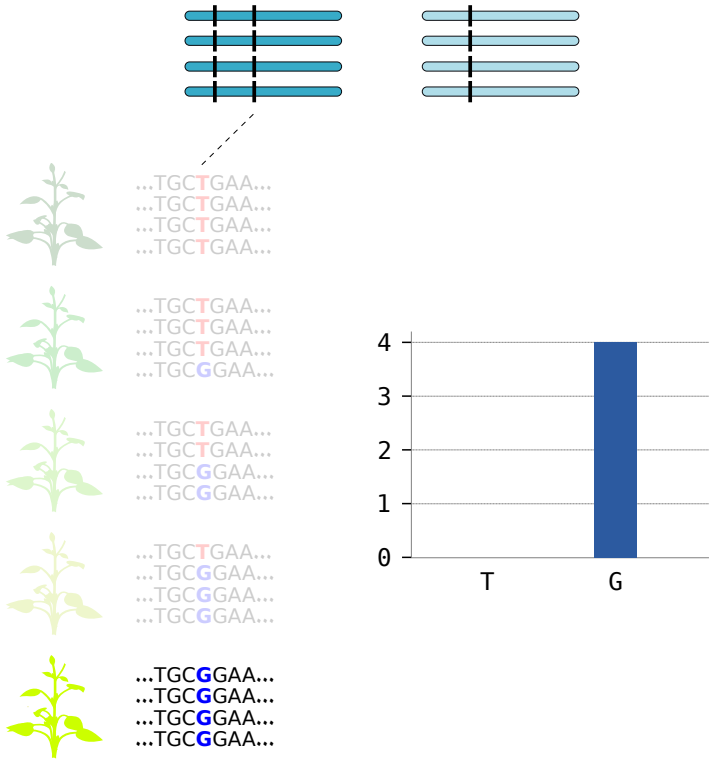
Biallelic co-dominant marker



Quantitative genotyping: access the abundance of each one of the allelic forms (in a biallelic marker)

Genotyping in polyploids

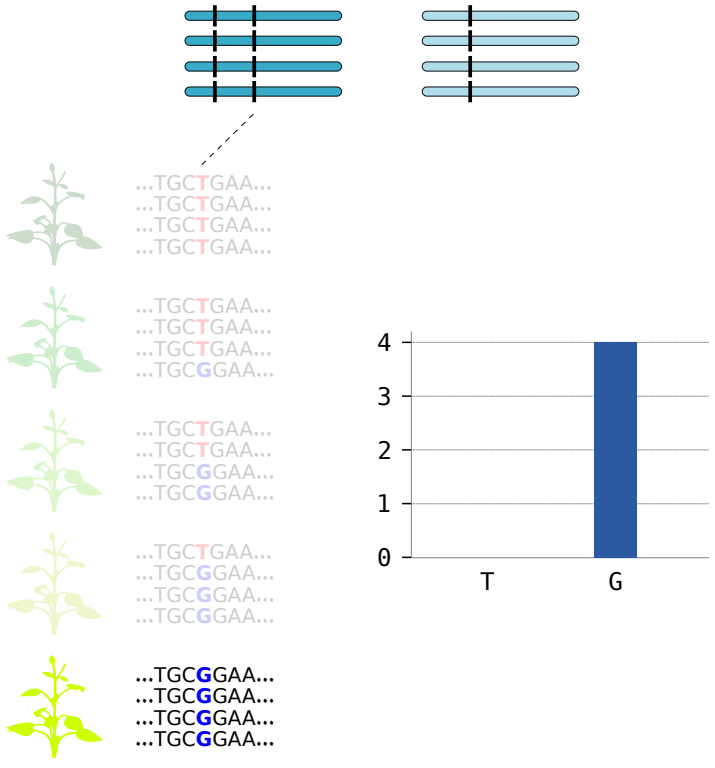
Biallelic co-dominant marker



Quantitative genotyping: access the abundance of each one of the allelic forms (in a biallelic marker)

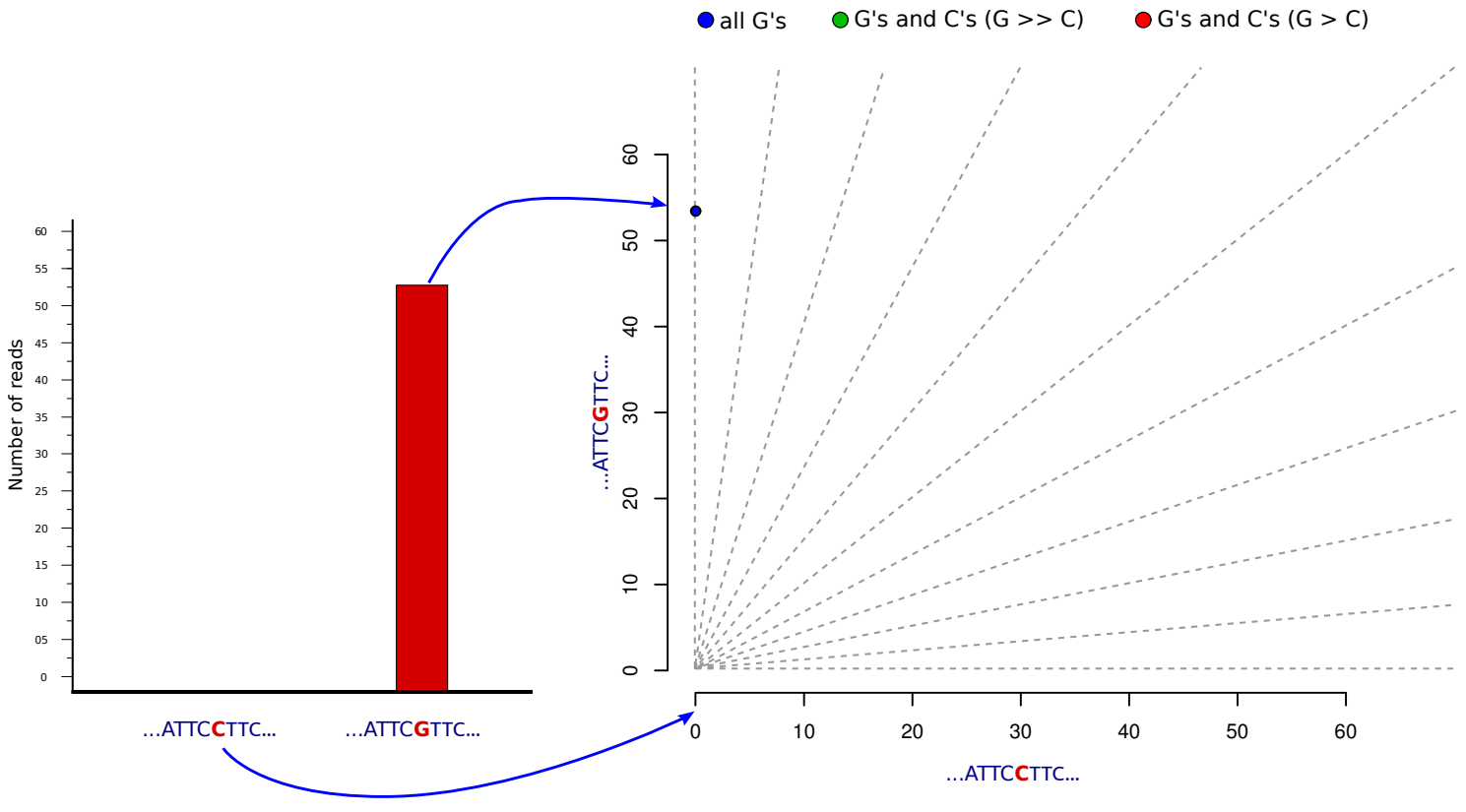
Genotyping in polyploids

Biallelic co-dominant marker



Quantitative genotyping: access the abundance of each one of the allelic forms (in a biallelic marker)

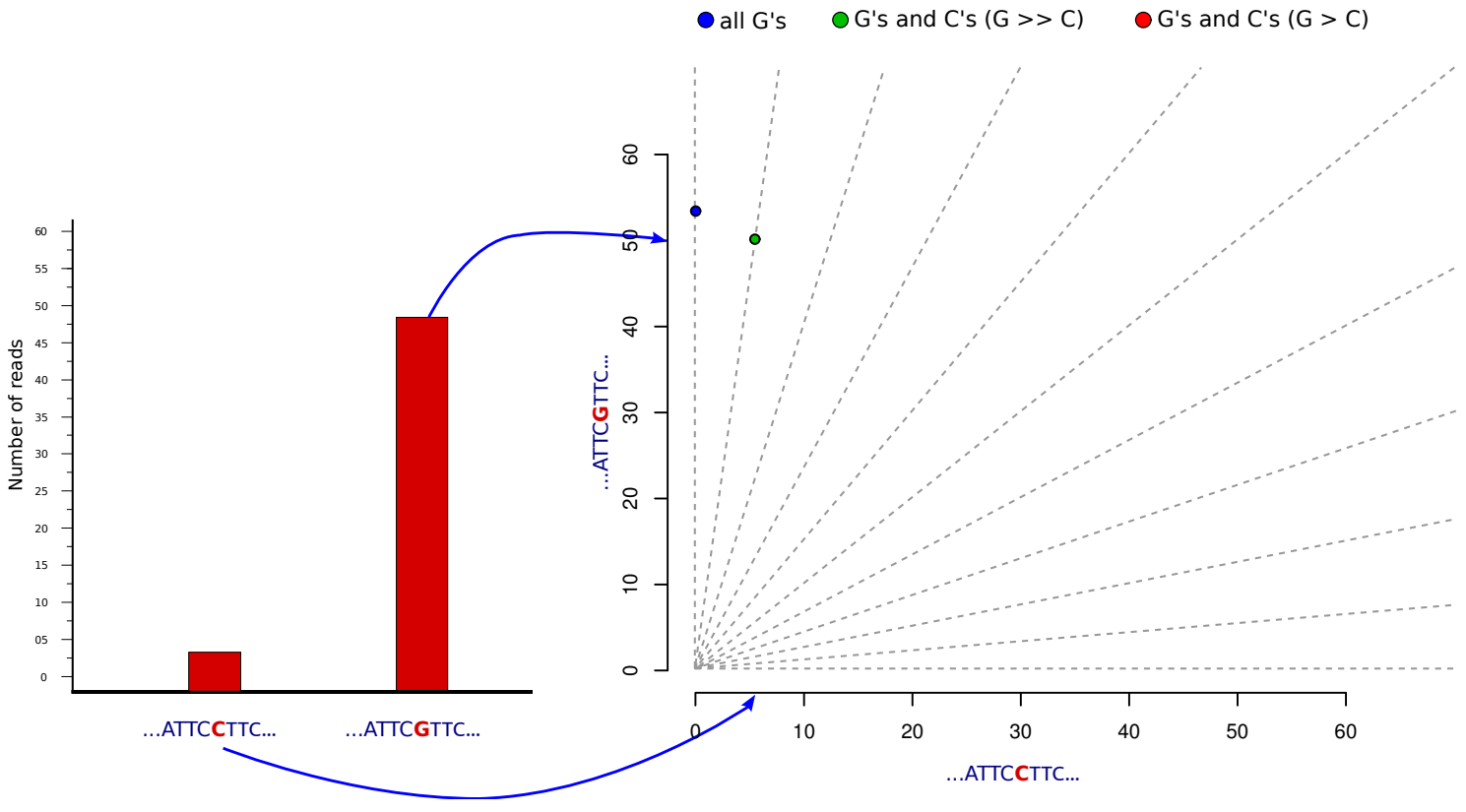
GBS scatter plot



Important

- Ratios of peaks or areas (angle): dosage

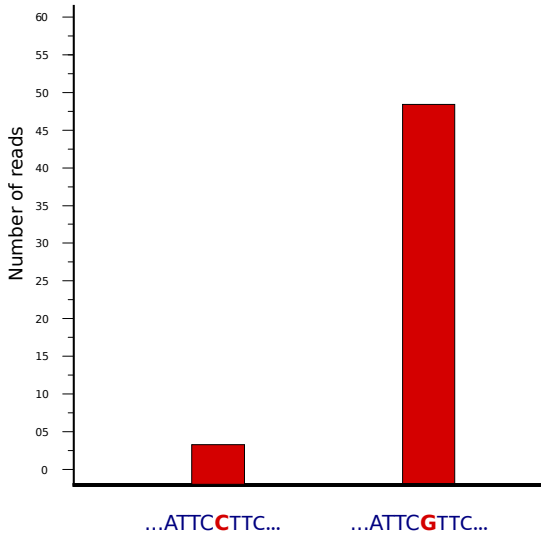
GBS scatter plot

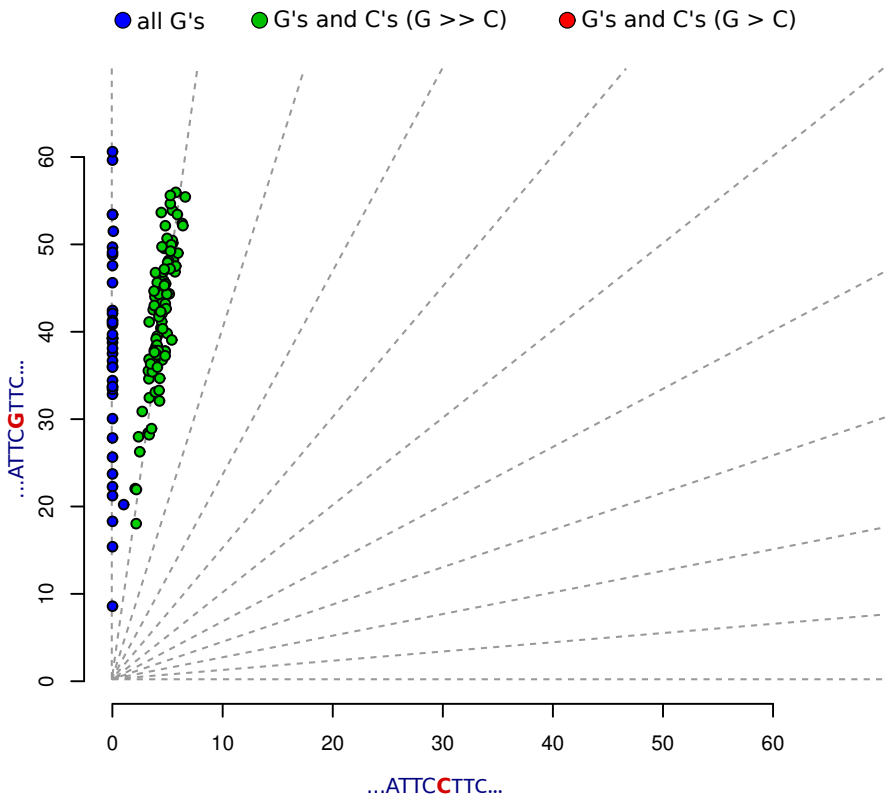


Important

- Ratios of peaks or areas (angle): dosage

GBS scatter plot

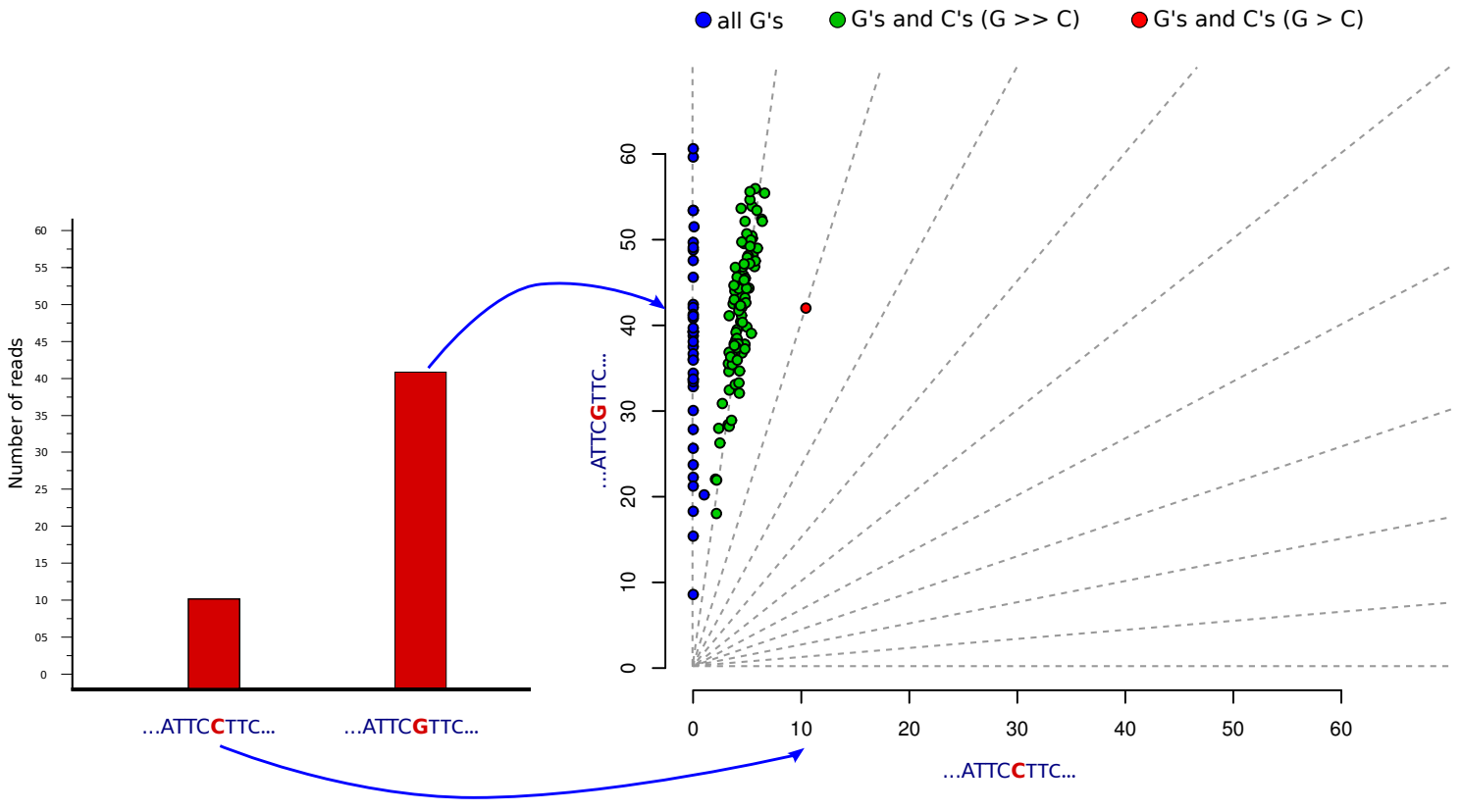




Important

- Ratios of peaks or areas (angle): dosage

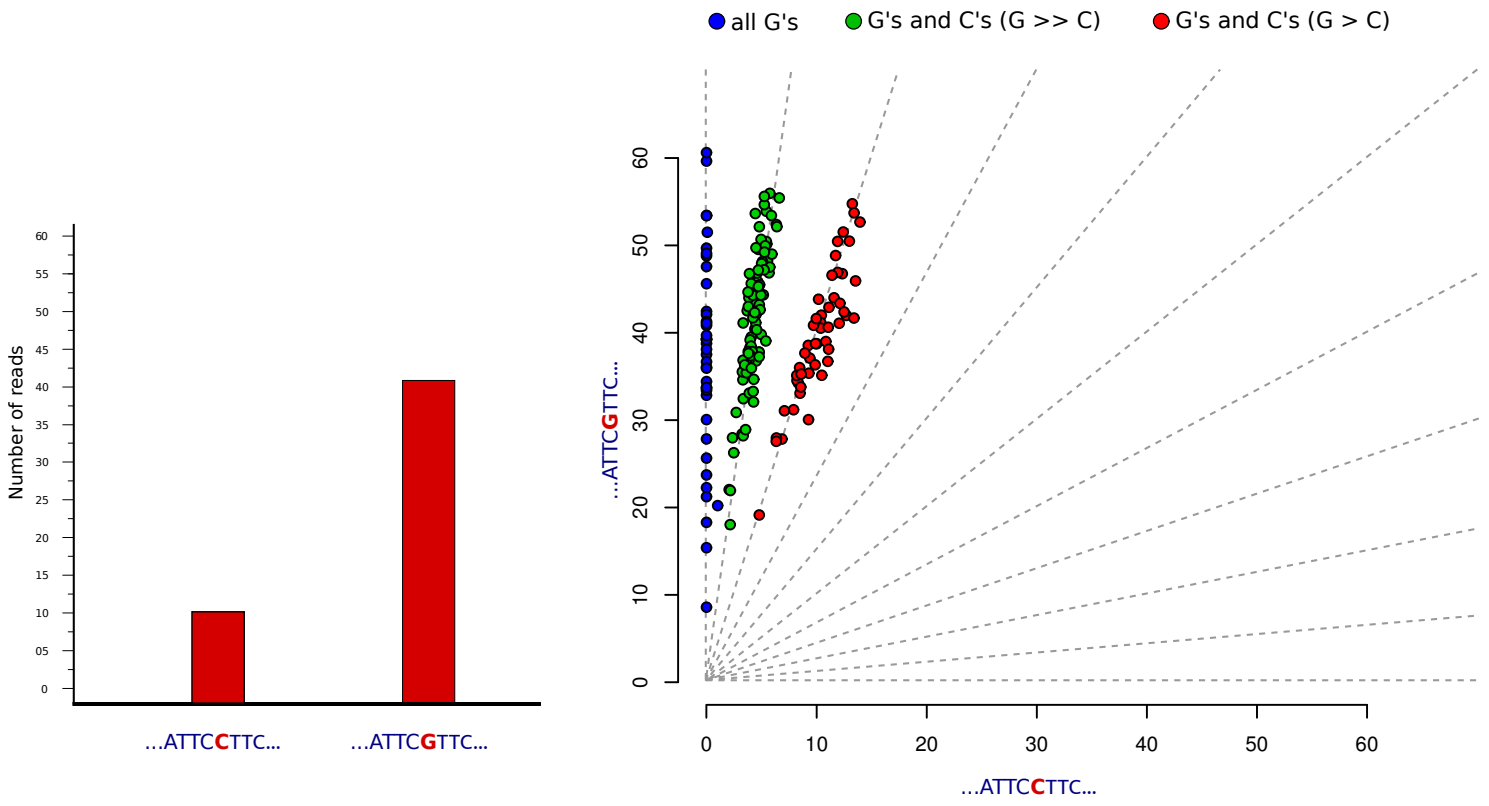
GBS scatter plot



Important

- Ratios of peaks or areas (angle): dosage

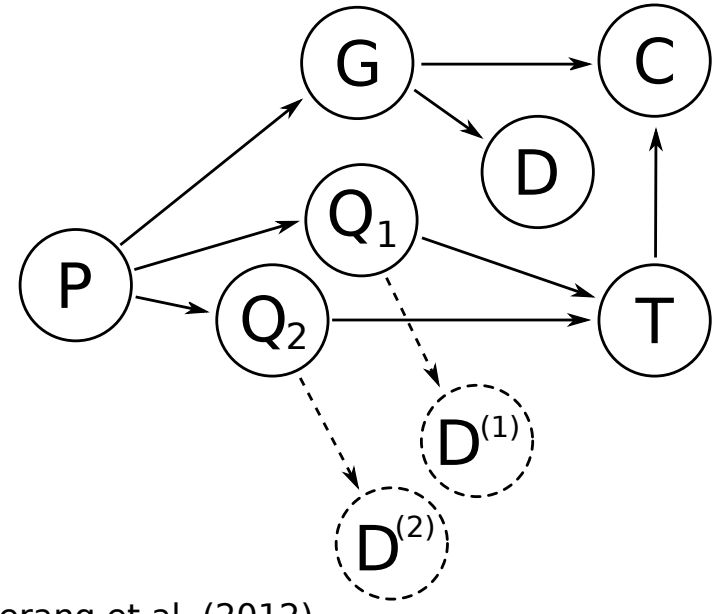
GBS scatter plot



Important

- Ratios of peaks or areas (angle): dosage

Probabilistic Graphical Model - F_1



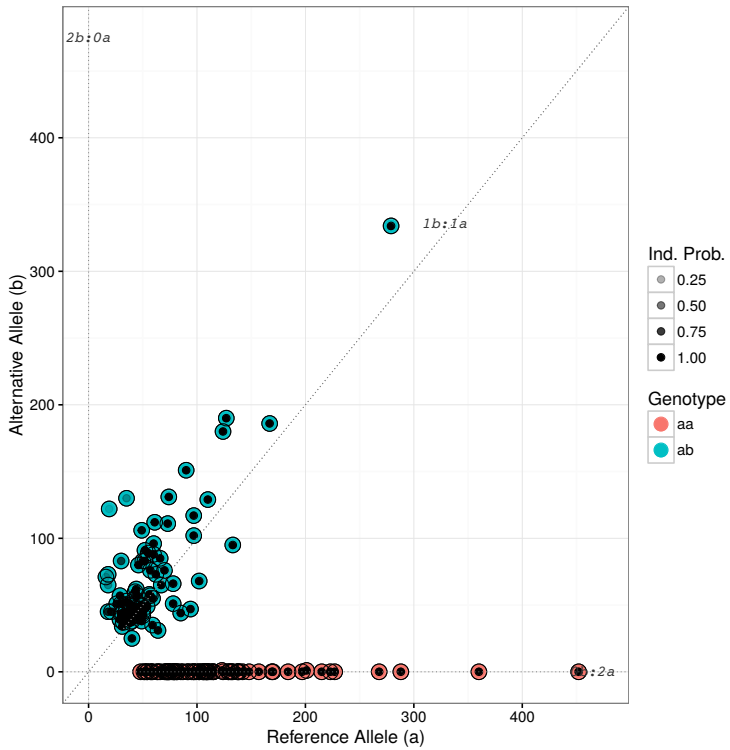
Serang et al. (2012)
Mollinari and Serang (2015)

$$\Pr(P, G, D, T, C, Q_1, Q_2, D_1, D_2)$$

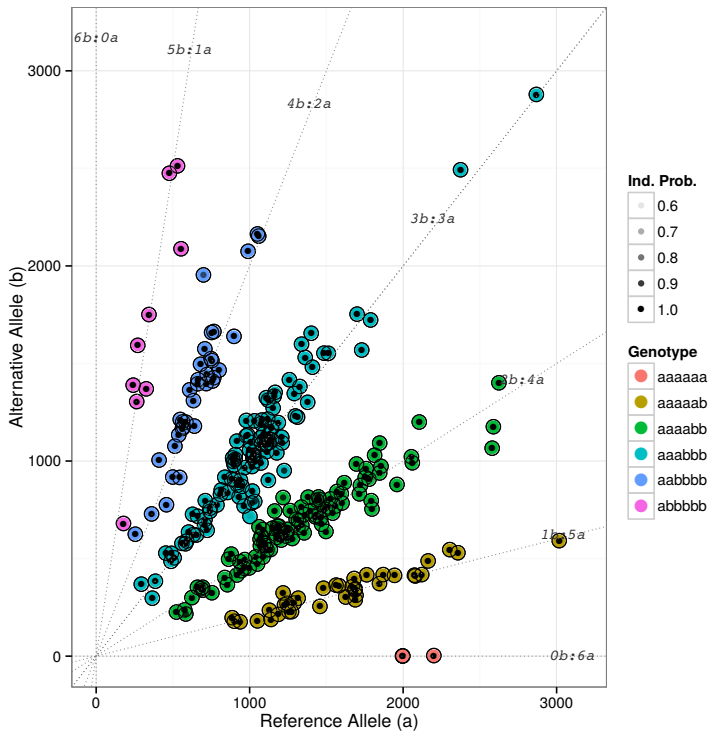
- P: ploidy
- G: genotype of all individuals
- D: observed data
- T: theoretical distribution of genotypes
- C: histogram of genotypes
- Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)

Example: SNPs in Sweet potato

Diploid



Hexaploid

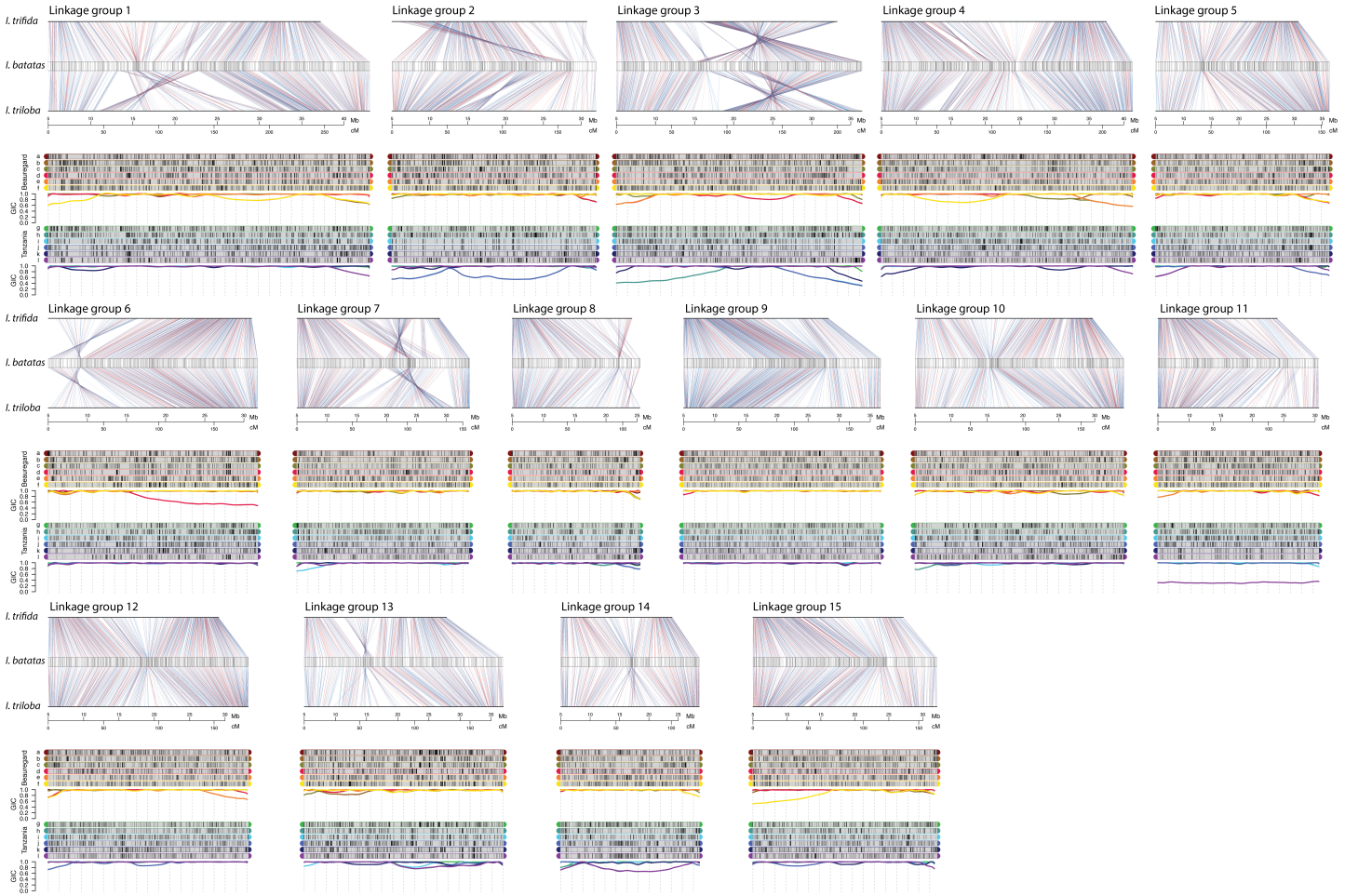


Genotype calling in polyploids

- fitTetra (tetraploids):
  `https://www.wur.nl/en/show/Software-fitTetra.htm`
- SuperMASSA (any ploidy level):
  `https://bitbucket.org/orserang/supermassa`
- updog (any ploidy level, allows preferential pairing):
  `https://github.com/dcgerard/updog`
- polyRAD (any ploidy level, reads VCF, BAM, etc):
  `https://github.com/lvclark/polyRAD`

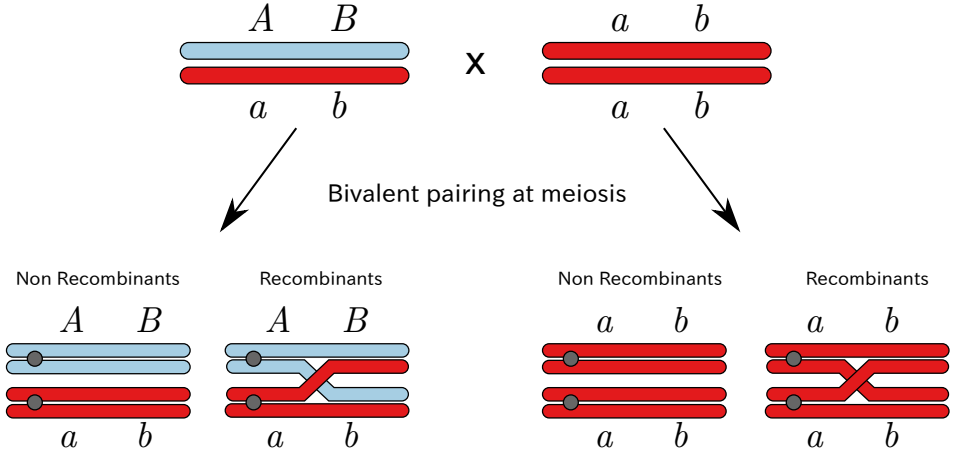
Linkage Analysis

Statement about the patterns of co-segregation of alleles or chromosomal segments. In outcrossing species, we do not know how the alleles are positioned in the parental chromosomes, thus we also need to infer the linkage phase of these alleles.

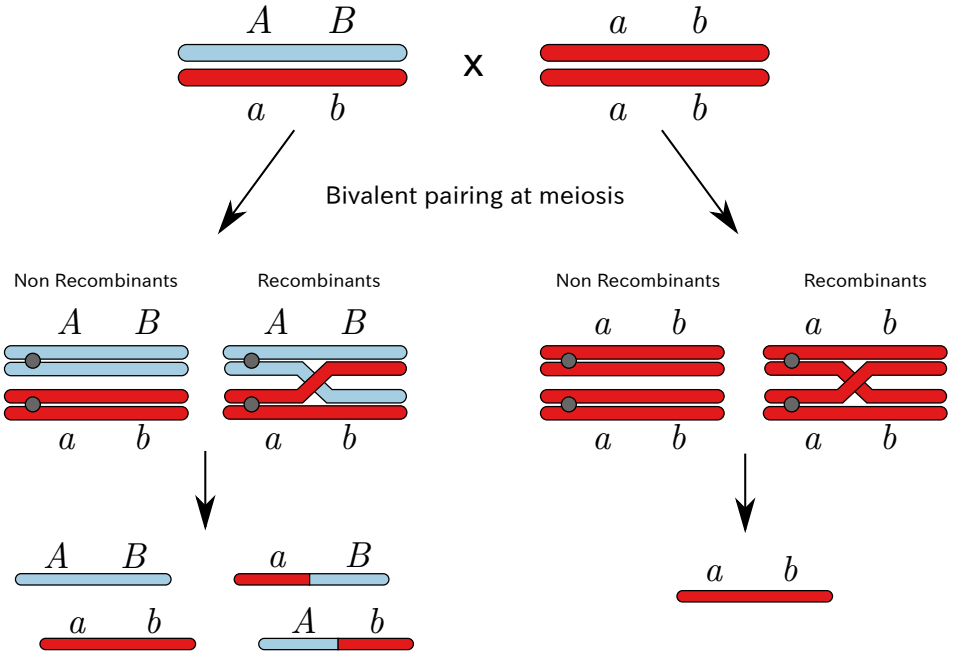


Recombination fraction in diploids

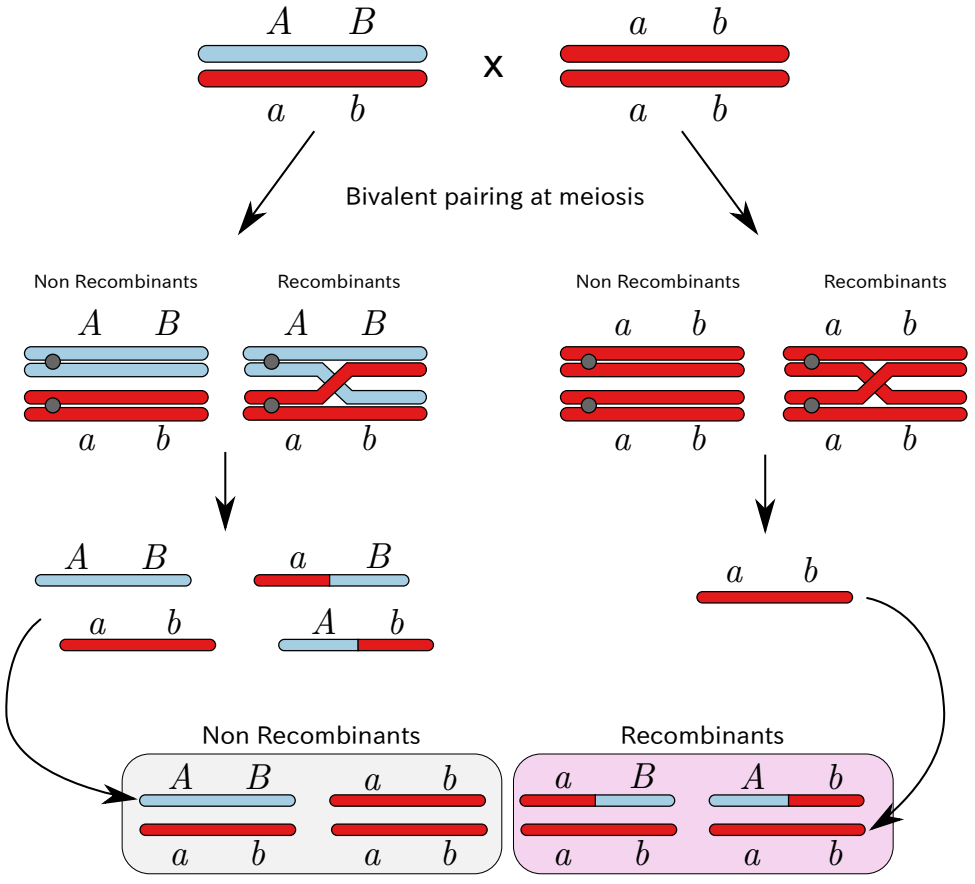
Recombination fraction in diploids



Recombination fraction in diploids



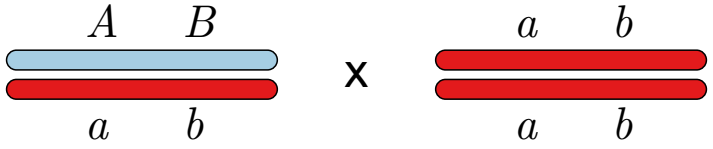
Recombination fraction in diploids

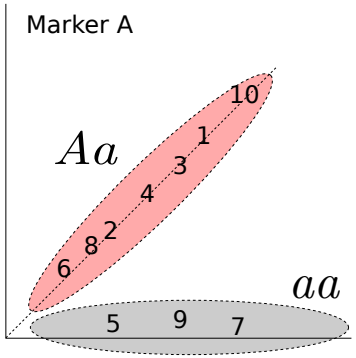


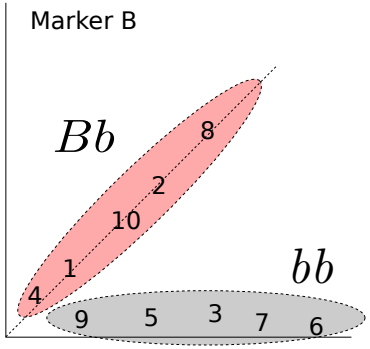
$$\widehat{r} = \frac{\#\text{recombinants}}{\#\text{total}}$$

Recombination fraction in diploids

Example



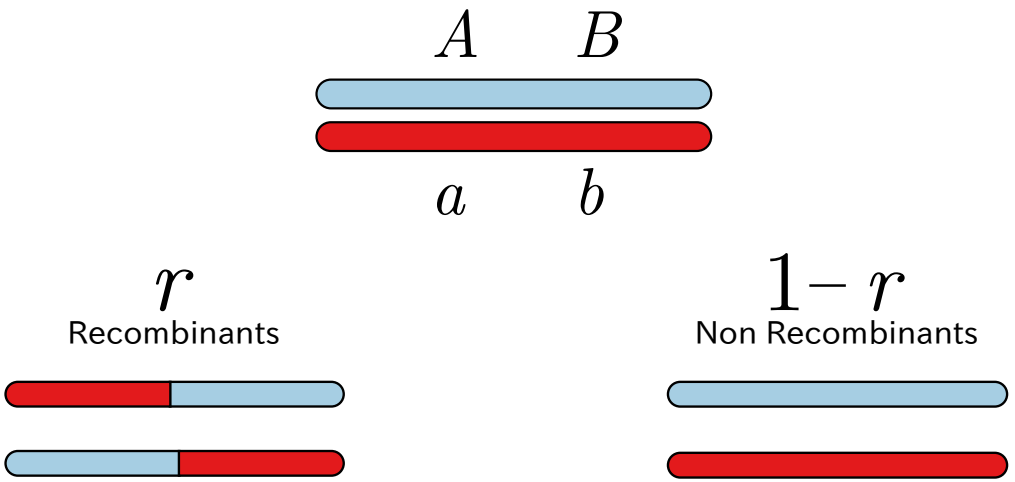




Individual	Obs. Gen.
1	$(A\ ,B)$
2	$(A\ ,B)$
3	$(A\ ,b\ )$
4	$(A\ ,B)$
5	$(a\ ,b\ )$
6	$(A\ ,b\ )$
7	$(a\ ,b\ )$
8	$(A\ ,B)$
9	$(a\ ,b\ )$
10	$(A\ ,B)$

$$\hat{r} = 2/10 = 0.2$$

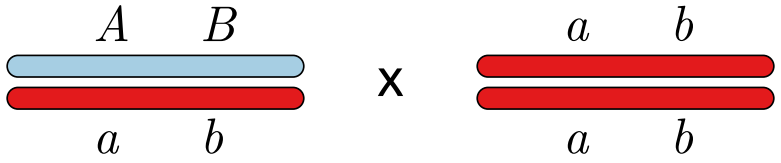
Expected gametic frequency



$$\mathbf{V} = \begin{bmatrix} \Pr(A\,,B) & \Pr(a\,,B) \\ \Pr(A\,,b) & \Pr(a\,,b) \end{bmatrix} = \begin{bmatrix} \frac{1-r}{2} & \frac{r}{2} \\ \frac{r}{2} & \frac{1-r}{2} \end{bmatrix}$$

Recombination fraction in diploids

Example (Cont.)

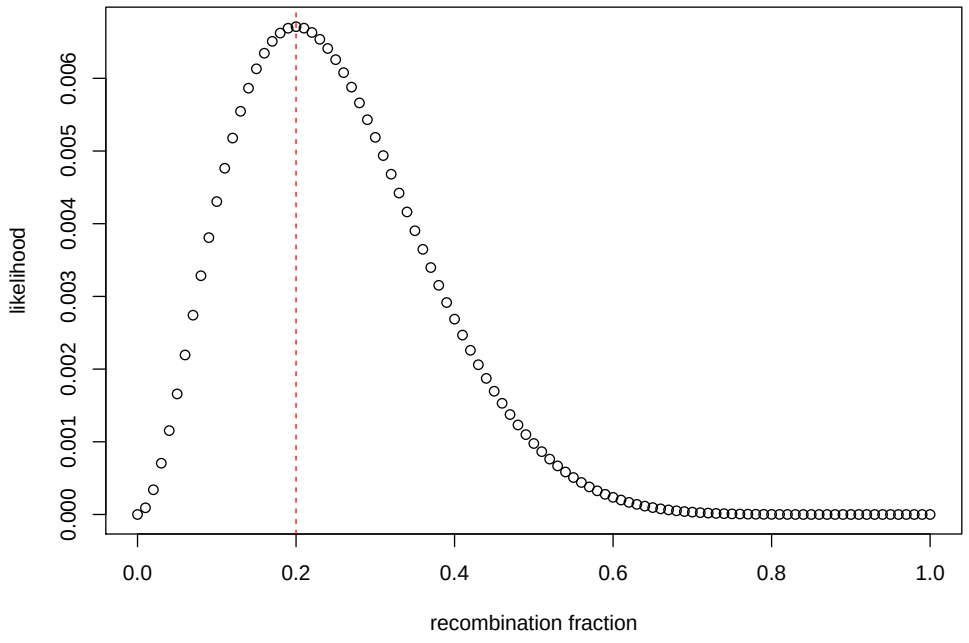


$$L = \prod_{n} \Pr(\text{loc}_B \text{ , } \text{loc}_A \mid Data)$$

Individual	Obs. Gen.	$\Pr(\text{loc}_B,\text{loc}_A)$
1	$(A\,,B)$	$\frac{1}{2}(1-r)$
2	$(A\,,B)$	$\frac{1}{2}(1-r)$
3	$(A\,,b\,)$	$\frac{1}{2}\ (r)$
4	$(A\,,B)$	$\frac{1}{2}(1-r)$
5	$(a\ \,,b\,)$	$\frac{1}{2}(1-r)$
6	$(A\,,b\,)$	$\frac{1}{2}\ (r)$
7	$(a\ \,,b\,)$	$\frac{1}{2}(1-r)$
8	$(A\,,B)$	$\frac{1}{2}(1-r)$
9	$(a\ \,,b\,)$	$\frac{1}{2}(1-r)$
10	$(A\,,B)$	$\frac{1}{2}(1-r)$

$$L = \left(\frac{r}{2}\right)^2 \left(\frac{1-r}{2}\right)^8$$

Likelihood Profile

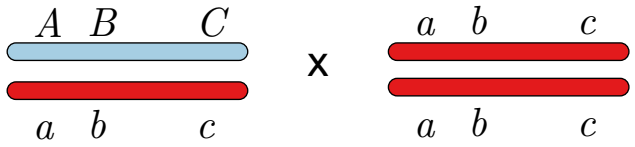


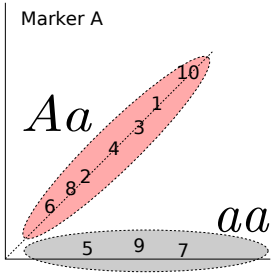
$$L = \prod_n \Pr(G_A, G_B) = \left(\frac{r}{2}\right)^2 \left(\frac{1-r}{2}\right)^8$$

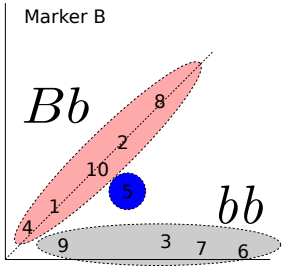
The MLE (maximum likelihood estimate) of r is $\hat{r} = 0.2$

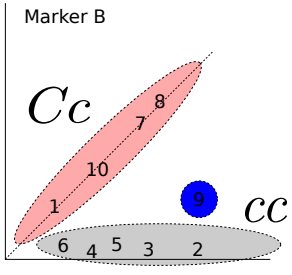
Recombination fraction in diploids

Three markers, eliminating noisy data points









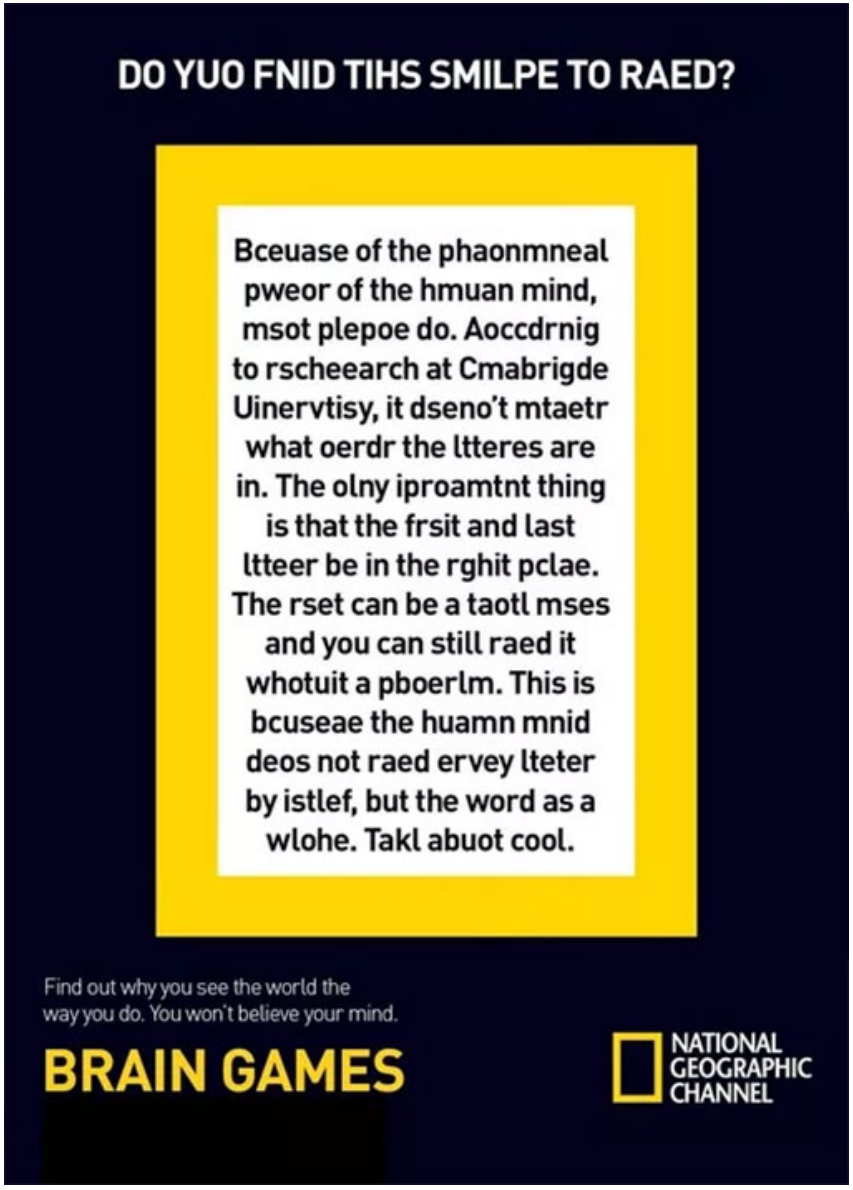
Individual	Obs. Gen.
1	$(A\ ,B,C\ )$
2	$(A\ ,B,c\ )$
3	$(A\ ,b\ ,c\ )$
4	$(A\ ,B,c\ )$
5	~~$(a\ ,?\ ,c\ )$~~
6	$(A\ ,b\ ,c\ )$
7	$(a\ ,b\ ,C\ )$
8	$(A\ ,B,C\ )$
9	~~$(a\ ,b\ ,?\ )$~~
10	$(A\ ,B,C\ )$

$$\hat{r}_{AB} = 2/9 = 0.222$$

$$\hat{r}_{BC} = 3/8 = 0.375$$

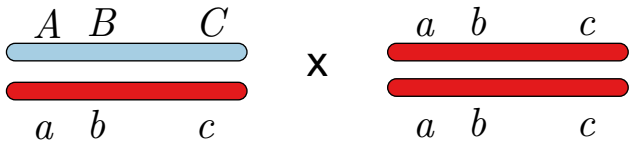
Hidden Markov Model

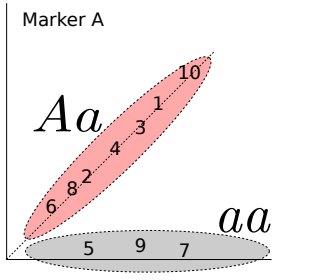
Intuition

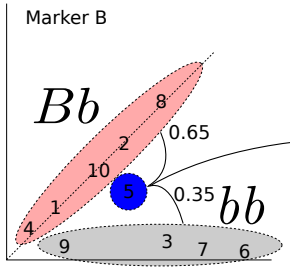


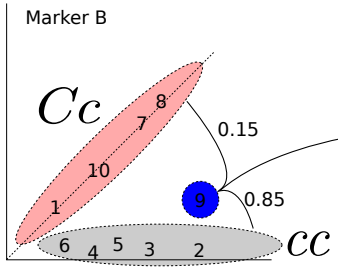
Recombination fraction in diploids

Three markers, HMM to handle noisy data points





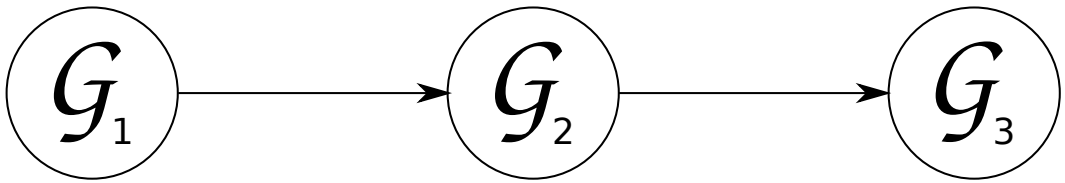




Individual	Obs. Gen.
1	$(A\ ,B\ ,C)$
2	$(A\ ,B\ ,c)$
3	$(A\ ,b\ ,c)$
4	$(A\ ,B\ ,c)$
5	$(a\ ,\ ?\ ,c)$
6	$(A\ ,b\ ,c)$
7	$(a\ ,b\ ,C)$
8	$(A\ ,B\ ,C)$
9	$(a\ ,b\ ,\ ?)$
10	$(A\ ,B\ ,C)$

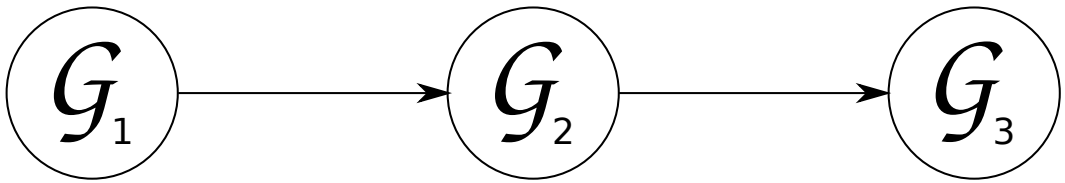
Hidden Markov Models (HMM)

Transition Probabilities



Hidden Markov Models (HMM)

Transition Probabilities

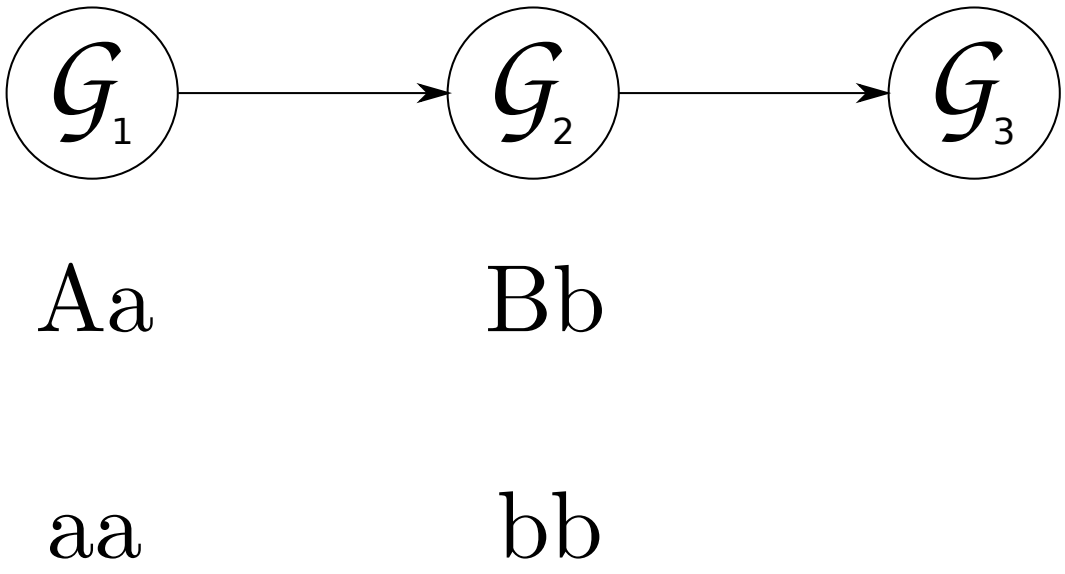


Aa

aa

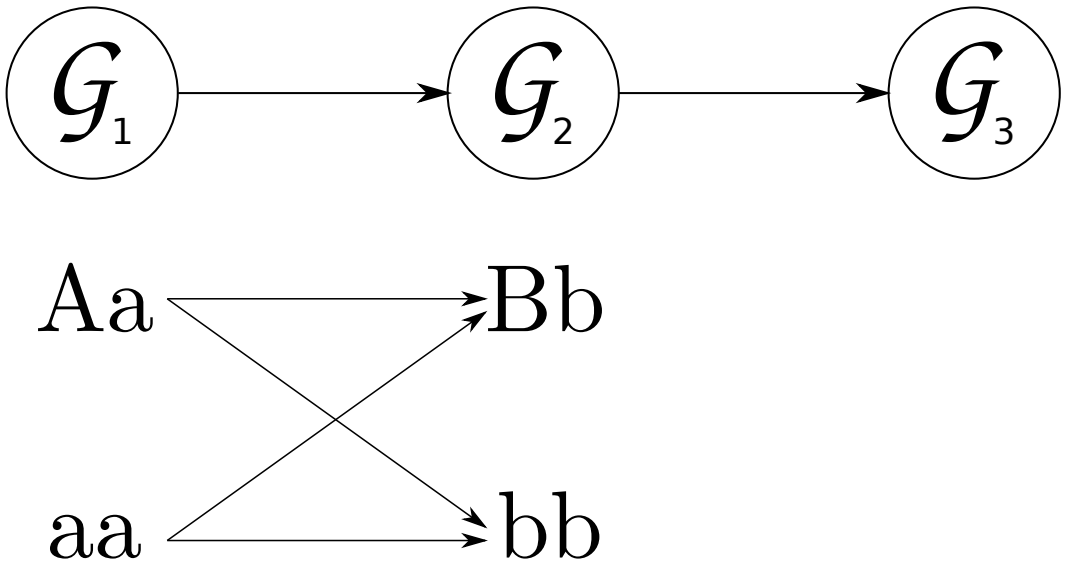
Hidden Markov Models (HMM)

Transition Probabilities



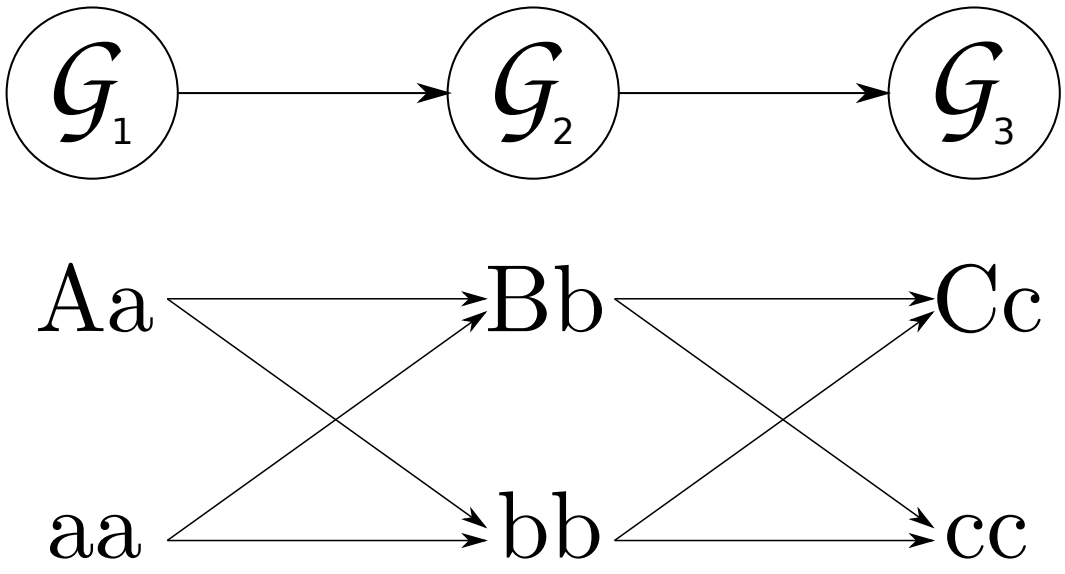
Hidden Markov Models (HMM)

Transition Probabilities



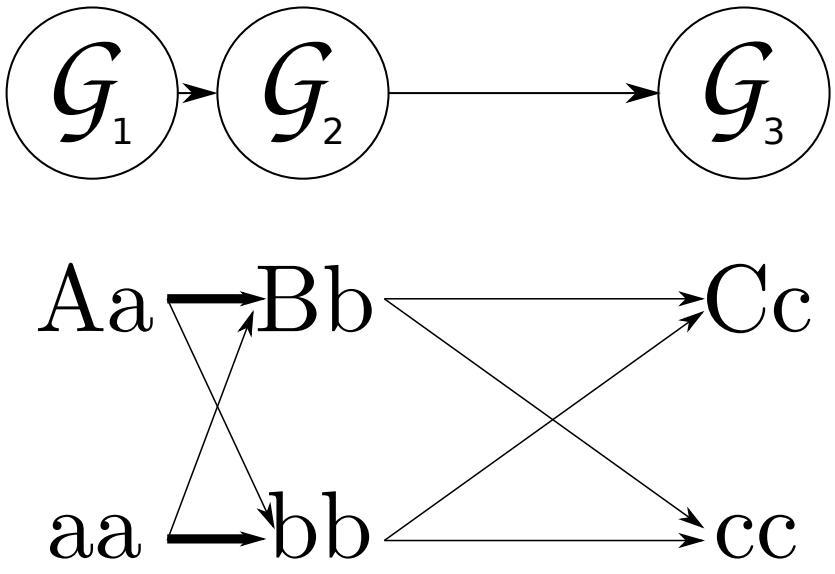
Hidden Markov Models (HMM)

Transition Probabilities



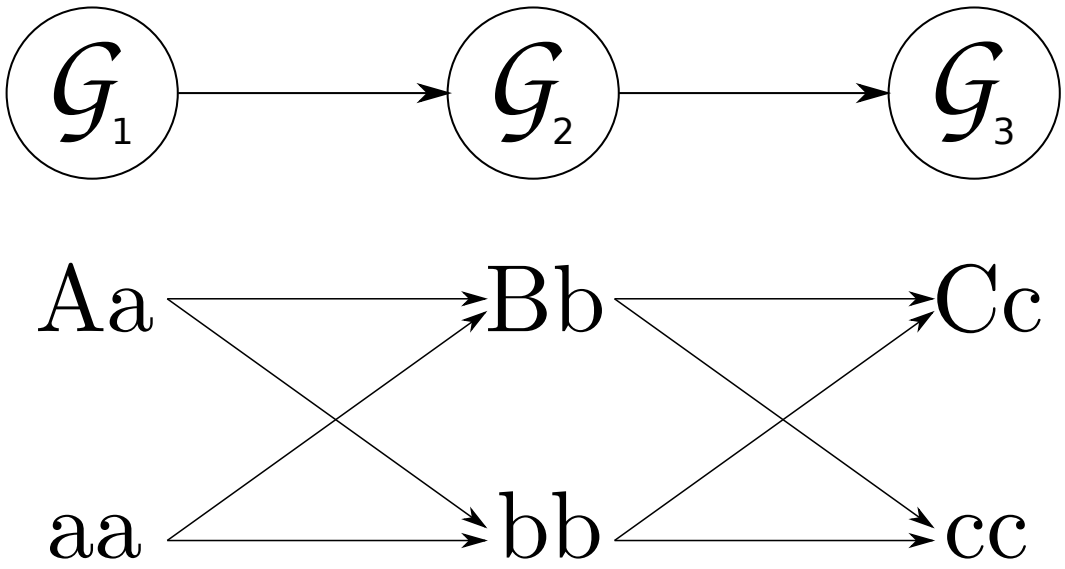
Hidden Markov Models (HMM)

Transition Probabilities



Hidden Markov Models (HMM)

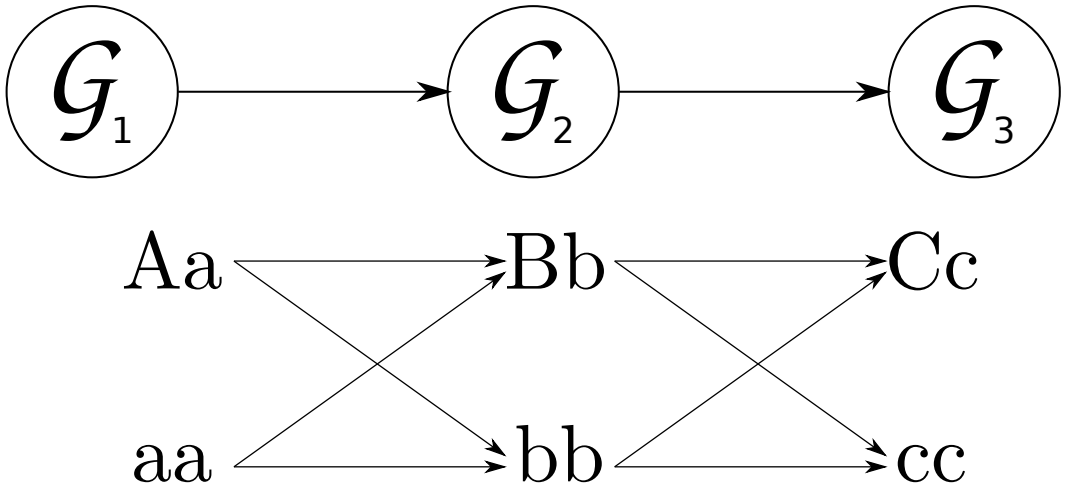
Transition Probabilities



$$\mathbf{V} = \begin{bmatrix} \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) \\ \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{aa}) \end{bmatrix} = \begin{bmatrix} 1-r & r \\ r & 1-r \end{bmatrix}$$

Hidden Markov Models (HMM)

Transition Probabilities

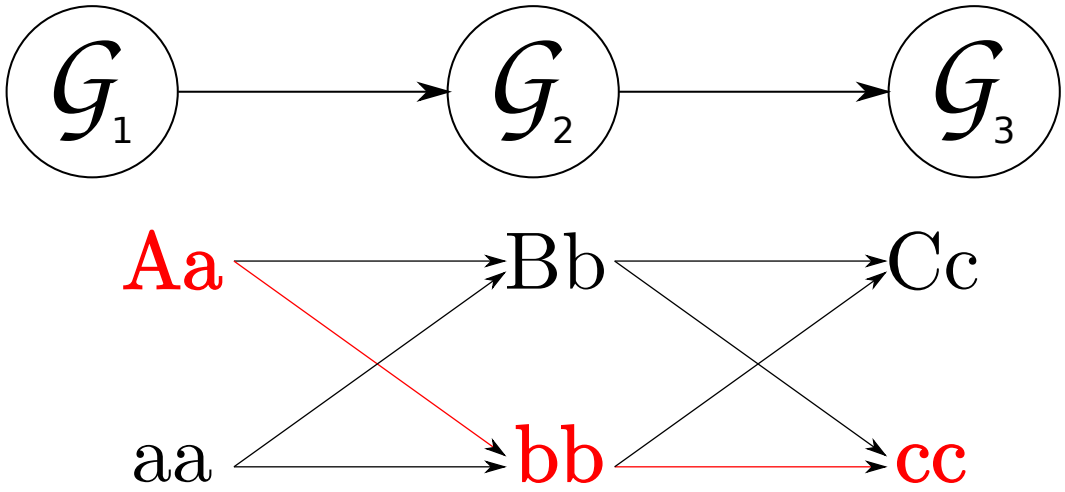


$$\begin{bmatrix} \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) \\ \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{aa}) \end{bmatrix} = \begin{bmatrix} 1-r_{AB} & r_{AB} \\ r_{AB} & 1-r_{AB} \end{bmatrix}$$

$$\begin{bmatrix} \Pr(\mathcal{G}_3{=}\mathrm{Cc}\,|\mathcal{G}_2{=}\mathrm{Bb}) & \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{Bb}) \\ \Pr(\mathcal{G}_3{=}\mathrm{Cc}\,|\mathcal{G}_2{=}\mathrm{bb}) & \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{bb}) \end{bmatrix} = \begin{bmatrix} 1-r_{BC} & r_{BC} \\ r_{BC} & 1-r_{BC} \end{bmatrix}$$

Hidden Markov Models (HMM)

Transition Probabilities



$$\begin{bmatrix} \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) \\ \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{aa}) \end{bmatrix} = \begin{bmatrix} 1-r_{AB} & r_{AB} \\ r_{AB} & 1-r_{AB} \end{bmatrix}$$

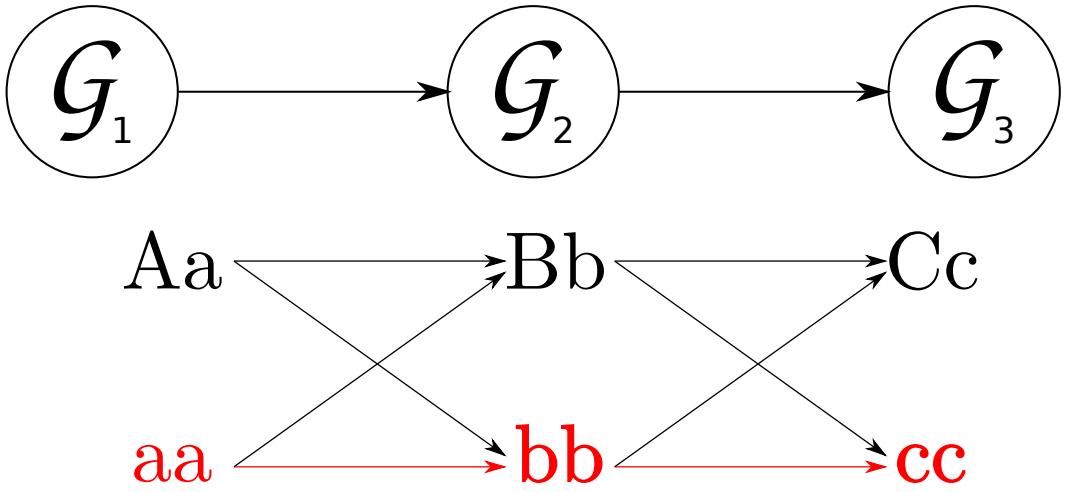
$$\begin{bmatrix} \Pr(\mathcal{G}_3{=}\mathrm{Cc}\,|\mathcal{G}_2{=}\mathrm{Bb}) & \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{Bb}) \\ \Pr(\mathcal{G}_3{=}\mathrm{Cc}\,|\mathcal{G}_2{=}\mathrm{bb}) & \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{bb}) \end{bmatrix} = \begin{bmatrix} 1-r_{BC} & r_{BC} \\ r_{BC} & 1-r_{BC} \end{bmatrix}$$

Observed Genotype: Aa bb cc

$$\Pr(\mathrm{Aa},\ \mathrm{bb},\ \mathrm{cc}) = \Pr(\mathcal{G}_1{=}\mathrm{Aa})\,\Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{Aa})\,\Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{bb})$$

Hidden Markov Models (HMM)

Transition Probabilities



$$\begin{bmatrix} \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) \\ \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{aa}) \end{bmatrix} = \begin{bmatrix} 1-r_{AB} & r_{AB} \\ r_{AB} & 1-r_{AB} \end{bmatrix}$$

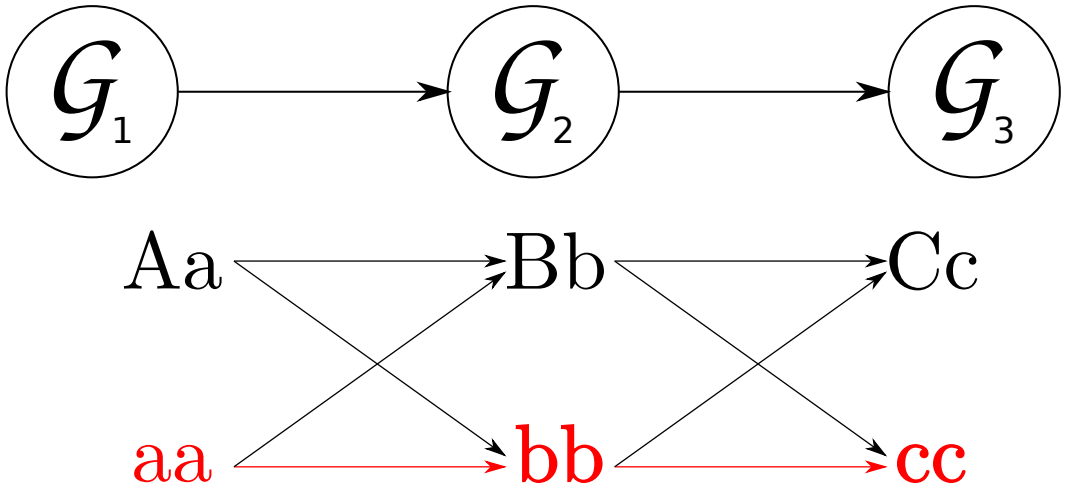
$$\begin{bmatrix} \Pr(\mathcal{G}_3{=}\mathrm{Cc}\,|\mathcal{G}_2{=}\mathrm{Bb}) & \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{Bb}) \\ \Pr(\mathcal{G}_3{=}\mathrm{Cc}\,|\mathcal{G}_2{=}\mathrm{bb}) & \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{bb}) \end{bmatrix} = \begin{bmatrix} 1-r_{BC} & r_{BC} \\ r_{BC} & 1-r_{BC} \end{bmatrix}$$

Observed Genotype: aa bb cc

$$\Pr(\mathrm{aa},\,\mathrm{bb},\,\mathrm{cc}) = \Pr(\mathcal{G}_1{=}\mathrm{aa})\ \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{aa})\ \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{bb})$$

Hidden Markov Models (HMM)

Transition Probabilities



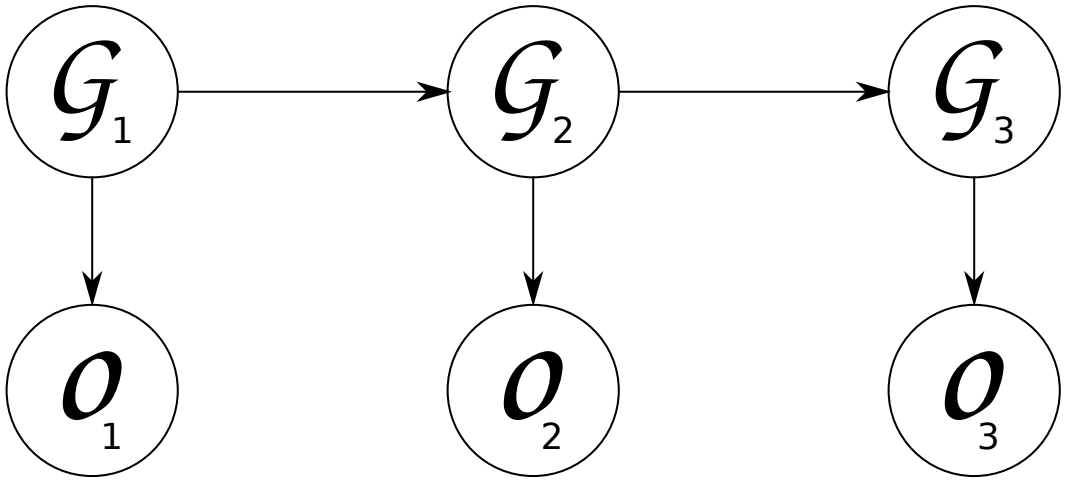
$$\begin{bmatrix} \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{Aa}) \\ \Pr(\mathcal{G}_2{=}\mathrm{Bb}\,|\mathcal{G}_1{=}\mathrm{aa}) & \Pr(\mathcal{G}_2{=}\mathrm{bb}\,|\mathcal{G}_1{=}\mathrm{aa}) \end{bmatrix} = \begin{bmatrix} 1-r_{AB} & r_{AB} \\ r_{AB} & 1-r_{AB} \end{bmatrix}$$

$$\begin{bmatrix} \Pr(\mathcal{G}_3{=}\mathrm{Cc}\,|\mathcal{G}_2{=}\mathrm{Bb}) & \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{Bb}) \\ \Pr(\mathcal{G}_3{=}\mathrm{Cc}\,|\mathcal{G}_2{=}\mathrm{bb}) & \Pr(\mathcal{G}_3{=}\mathrm{cc}\,|\mathcal{G}_2{=}\mathrm{bb}) \end{bmatrix} = \begin{bmatrix} 1-r_{BC} & r_{BC} \\ r_{BC} & 1-r_{BC} \end{bmatrix}$$

$$L = \prod_n \Pr(\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3 |\ Data)$$

Hidden Markov Models (HMM)

Emission Probabilities



$\Pr(\mathcal{O}_1|\mathcal{G}_1)$

$\mathcal{O}_1$ \ $\mathcal{G}_1$	Aa	aa
Aa	1	0
aa	0	1
—	1/2	1/2
or	or	
—	p_1	1-p_1

$\Pr(\mathcal{O}_2|\mathcal{G}_2)$

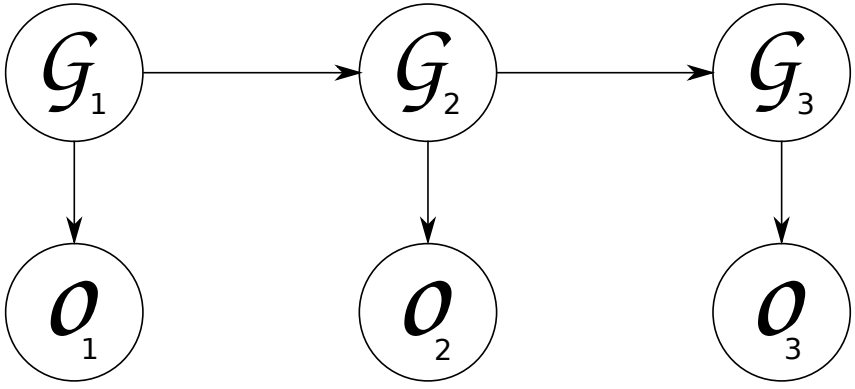
$\mathcal{O}_2$ \ $\mathcal{G}_2$	Bb	bb
Bb	1	0
bb	0	1
—	1/2	1/2
or	or	
—	p_2	1-p_2

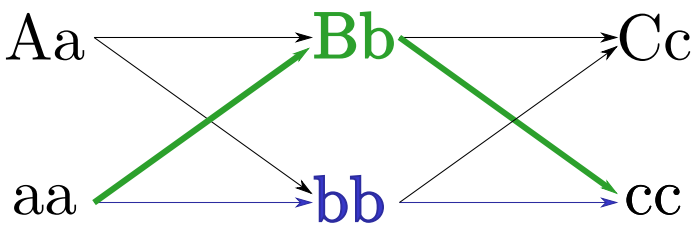
$\Pr(\mathcal{O}_3|\mathcal{G}_3)$

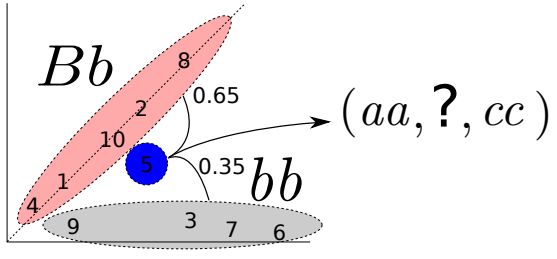
$\mathcal{O}_3$ \ $\mathcal{G}_3$	Cc	cc
Cc	1	0
cc	0	1
—	1/2	1/2
or	or	
—	p_3	1-p_3

Hidden Markov Models (HMM)

Emission Probabilities





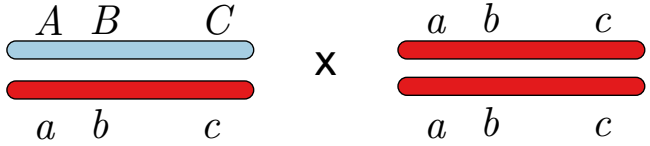


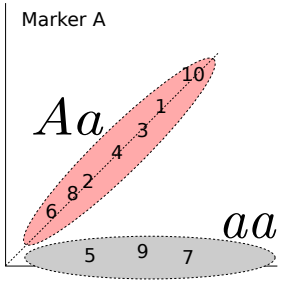
$\Pr(aa)\Pr(Bb|aa)\Pr(O|Bb)\Pr(cc|Bb) = 1/2 \cdot (r_{ab}).(1\text{-}r_{bc}).0.65 \quad \times$

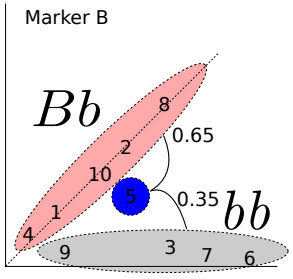
$\Pr(aa)\Pr(Bb|aa)\Pr(O|bb)\Pr(cc|bb) = 1/2 \cdot (1\text{-}r_{ab}).(1\text{-}r_{bc}).0.35$

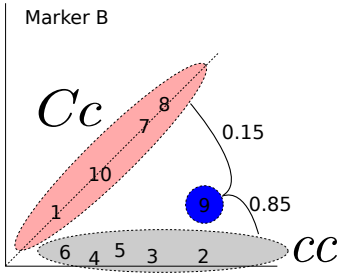
Hidden Markov Models (HMM)

Likelihood









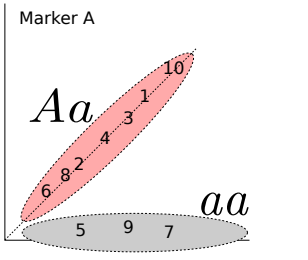
Ind.	Obs. Gen.	$Pr(G_1,G_2,G_3\|Data)$
1	(A,B,C)	$Pr(Aa)Pr(Bb\|Aa)Pr(Cc\|Bb) = 1/2 \cdot (1-r_{ab}) \cdot (1-r_{bc})$
2	(A,B,c)	$Pr(Aa)Pr(Bb\|Aa)Pr(cc\|Bb) = 1/2 \cdot (1-r_{ab}) \cdot (r_{bc})$
3	(A,b,c)	$Pr(Aa)Pr(bb\|Aa)Pr(cc\|bb) = 1/2 \cdot (r_{ab}) \cdot (1-r_{bc})$
4	(A,B,c)	$Pr(Aa)Pr(Bb\|Aa)Pr(cc\|Bb) = 1/2 \cdot (1-r_{ab}) \cdot (r_{bc})$
5	$(a,?,c)$	$Pr(aa)Pr(Bb\|aa).Pr(O\|Bb).Pr(cc\|bb) = 1/2 \cdot (r_{ab}) \cdot (1-r_{bc}).0.65$ $Pr(aa)Pr(bb\|aa).Pr(O\|bb).Pr(cc\|bb) = 1/2 \cdot (1-r_{ab}) \cdot (1-r_{bc}).0.35$
6	(A,b,c)	$Pr(Aa)Pr(bb\|Aa)Pr(cc\|bb) = 1/2 \cdot (r_{ab}) \cdot (1-r_{bc})$
7	(a,b,C)	$Pr(aa)Pr(bb\|aa)Pr(Cc\|bb) = 1/2 \cdot (1-r_{ab}) \cdot (r_{bc})$
8	(A,B,C)	$Pr(Aa)Pr(Bb\|Aa)Pr(Cc\|Bb) = 1/2 \cdot (1-r_{ab}) \cdot (1-r_{bc})$
9	$(a,b,?)$	$Pr(aa)Pr(bb\|aa)Pr(cc\|bb).Pr(O\|cc) = 1/2 \cdot (1-r_{ab}) \cdot (1-r_{bc}).0.85$ $Pr(aa)Pr(bb\|aa)Pr(Cc\|bb).Pr(O\|Cc) = 1/2 \cdot (1-r_{ab}) \cdot (r_{bc}).0.15$
10	(A,B,C)	$Pr(Aa)Pr(Bb\|Aa)Pr(Cc\|Bb) = 1/2 \cdot (1-r_{ab}) \cdot (1-r_{bc})$

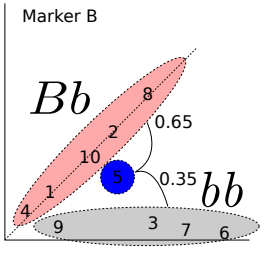
$$L = \prod_n Pr(G_1, G_2, G_3 \mid Data) = 7.1 \cdot 10^{-6} (1-r_{ab})^9 \; r_{ab}^3 \; (1-r_{bc})^8 \; r_{bc}^4$$

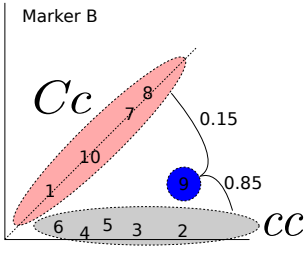
Hidden Markov Models (HMM)

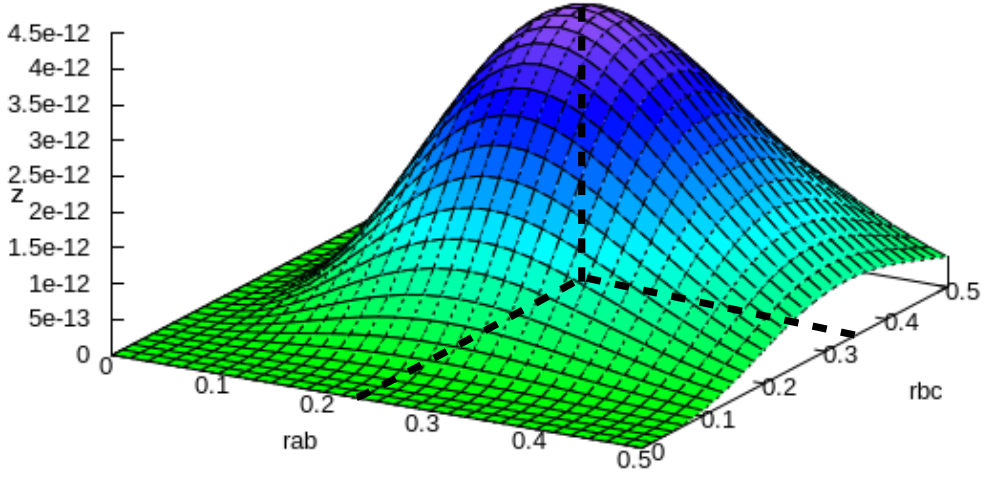
Maximum Likelihood Estimate

$$L = \prod_{n} \Pr(\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3 | \ Data) = 7.1 \cdot 10^{-6} (1\text{-}r_{ab})^{9} \ r_{ab}^{3} \ (1\text{-}r_{bc})^{8} \ r_{bc}^{4}$$









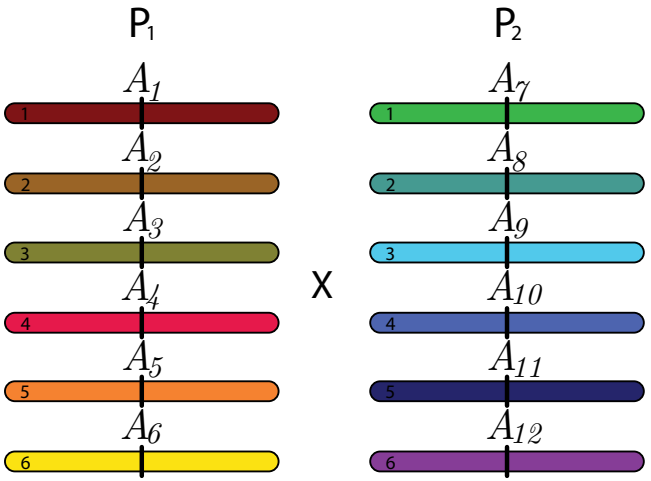
HMM	two-point
$\hat{r}_{AB} = 1/4 = 0.250$	$\hat{r}_{AB} = 2/9 = 0.222$
$\hat{r}_{BC} = 1/3 = 0.333$	$\hat{r}_{BC} = 3/8 = 0.375$

Linkage analysis in autopolyploids

- Linkage Analysis and Haplotype Phasing in Experimental Autopolyploid Populations with High Ploidy Level Using Hidden Markov Models (`https://doi.org/10.1534/g3.119.400378`)

- Unraveling the hexaploid sweetpotato inheritance using ultra-dense multilocus mapping (`https://doi.org/10.1101/689638`)

Hexaploid bi-parental cross



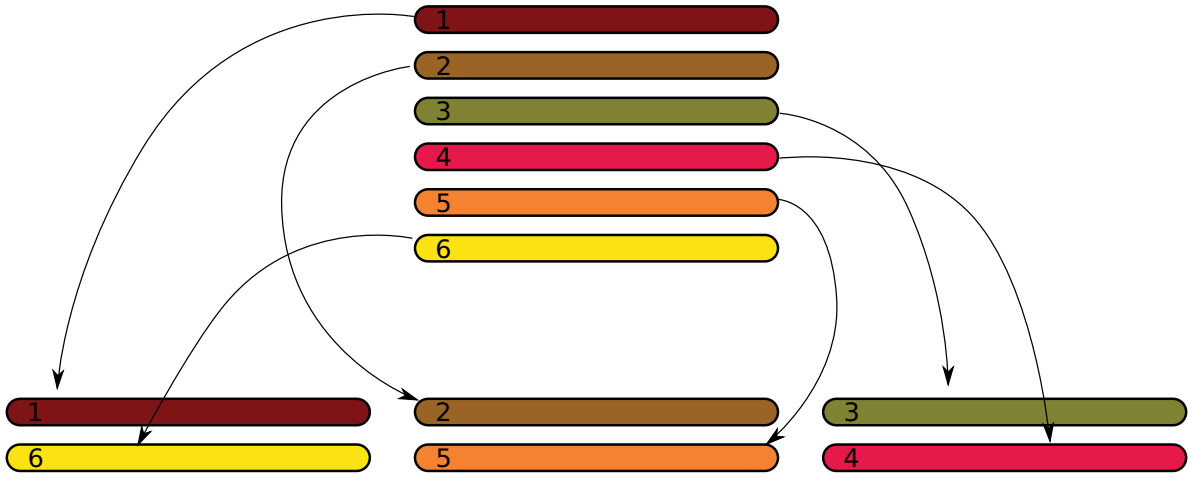
Gametes

P2 \ P1	1 $A_1A_2A_3$	2 $A_1A_2A_4$	3 $A_1A_2A_5$	...	18 $A_3A_4A_6$	19 $A_3A_5A_6$	20 $A_4A_5A_6$
1 $A_7A_8A_9$							
2 $A_7A_8A_{10}$				...			
3 $A_7A_8A_{11}$							
⋮	⋮	⋮		⋱		⋮	
18 $A_9A_{10}A_{12}$							
19 $A_9A_{11}A_{12}$				...			
20 $A_{10}A_{11}A_{12}$							

400 possible genotypes

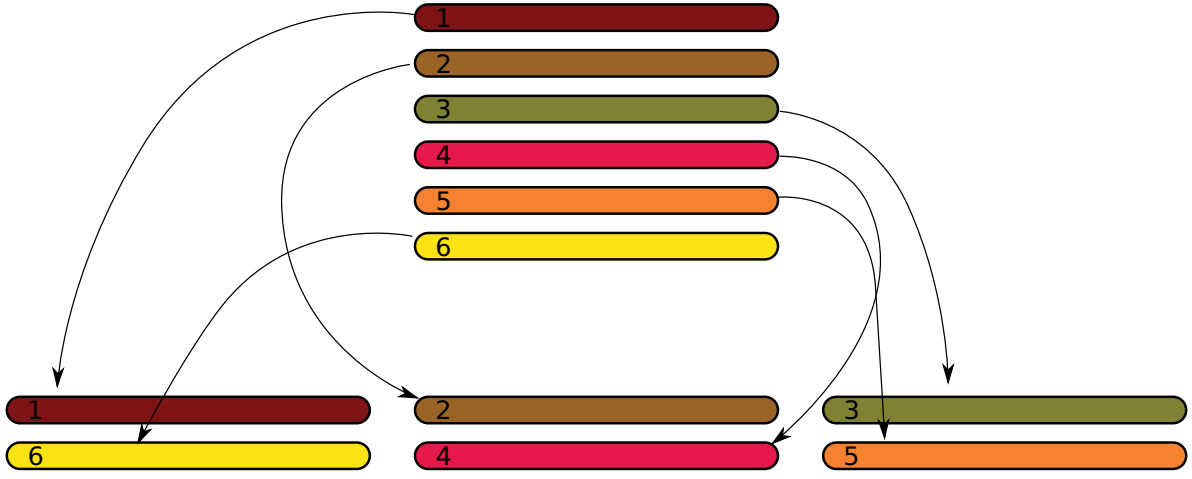
Gamete formation - random pairing

Gamete formation - random pairing



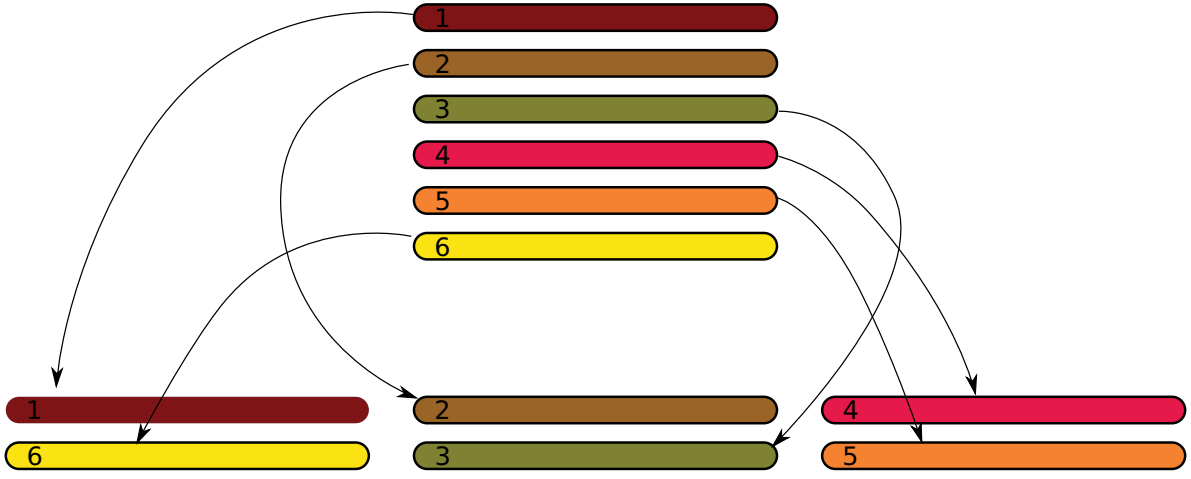
$$\psi_1$$

Gamete formation - random pairing



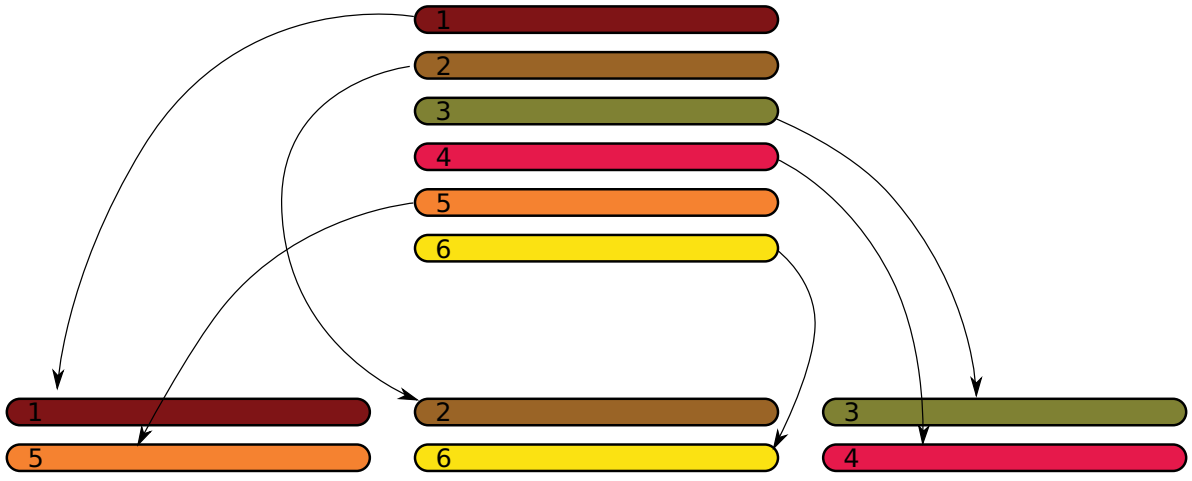
$$\psi_2$$

Gamete formation - random pairing



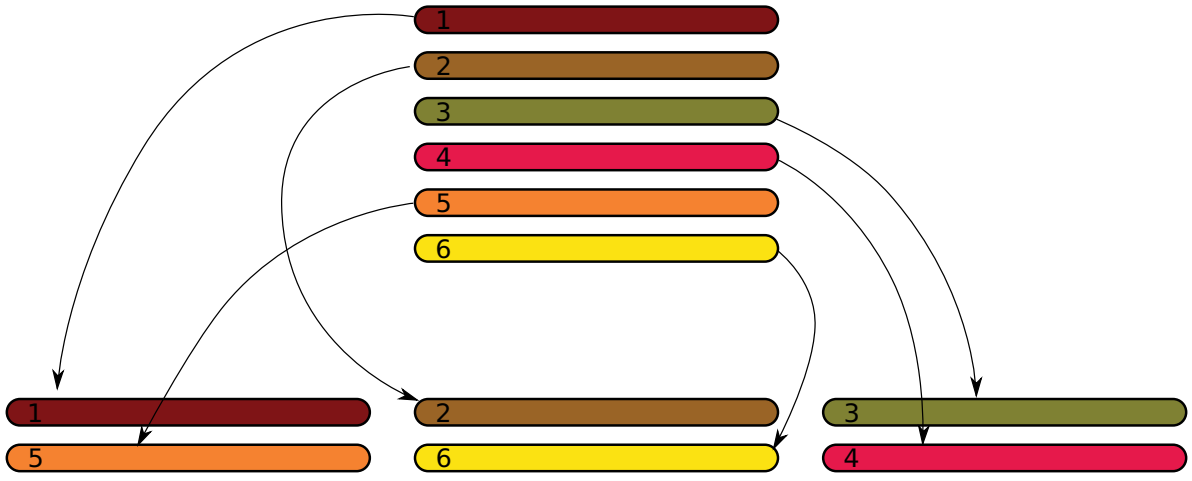
$$\psi_3$$

Gamete formation - random pairing



$$\psi_4$$

Gamete formation - random pairing

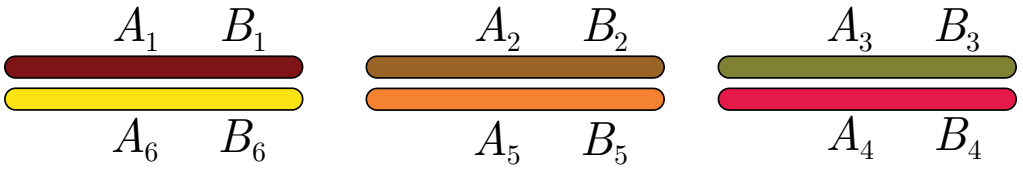


$$\psi_4$$

In this case: 15 possible configurations. For any ploidy level p

$$\frac{1}{\frac{p}{2}!}\prod_{i=1}^{\frac{p}{2}}\binom{2i}{2}$$

Expected gametic frequency for a bivalent configuration



$$\begin{bmatrix} \frac{1-r}{2} & \frac{r}{2} \\ \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} \otimes \begin{bmatrix} \frac{1-r}{2} & \frac{r}{2} \\ \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} \otimes \begin{bmatrix} \frac{1-r}{2} & \frac{r}{2} \\ \frac{r}{2} & \frac{1-r}{2} \end{bmatrix}$$

- In general:

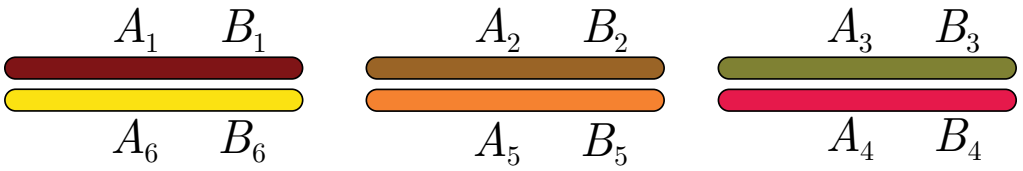
$$\mathbf{V}_1 \otimes \mathbf{V}_2 \otimes \cdots \otimes \mathbf{V}_{\frac{p}{2}}$$

- All elements of this product are of the form

$$\Pr(G_A, G_B \mid \psi_j) = \frac{(1-r)^{\frac{p}{2}-i}(r)^i}{2^{\frac{p}{2}}}$$

i: known number of recombinat bivalents between loci A and B

Expected gametic frequency for a bivalent configuration



$$\begin{bmatrix} \frac{1-r}{2} & \frac{r}{2} \\ \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} \otimes \begin{bmatrix} \frac{1-r}{2} & \frac{r}{2} \\ \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} \otimes \begin{bmatrix} \frac{1-r}{2} & \frac{r}{2} \\ \frac{r}{2} & \frac{1-r}{2} \end{bmatrix}$$

- In general:

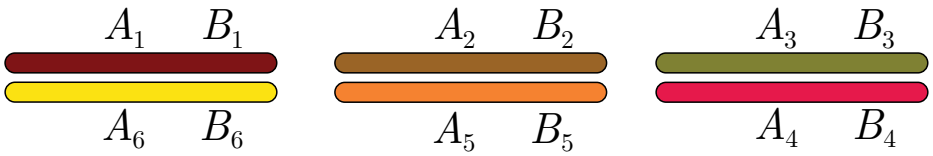
$$\mathbf{V}_1 \otimes \mathbf{V}_2 \otimes \cdots \otimes \mathbf{V}_{\frac{p}{2}}$$

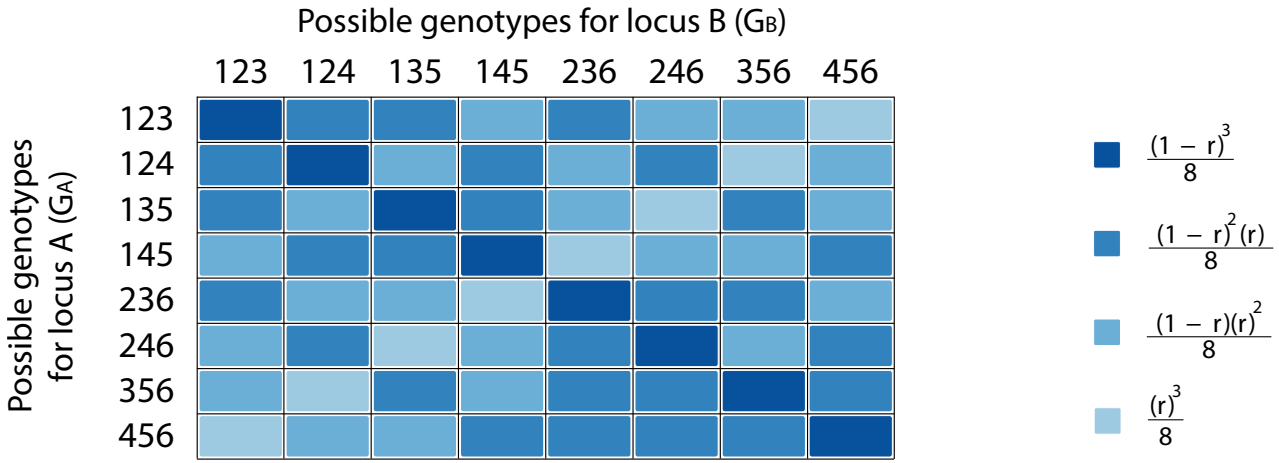
- All elements of this product are of the form

$$\Pr(G_A, G_B \mid \psi_j) = \frac{(1-r)^{\frac{p}{2}-i}(r)^i}{2^{\frac{p}{2}}}$$

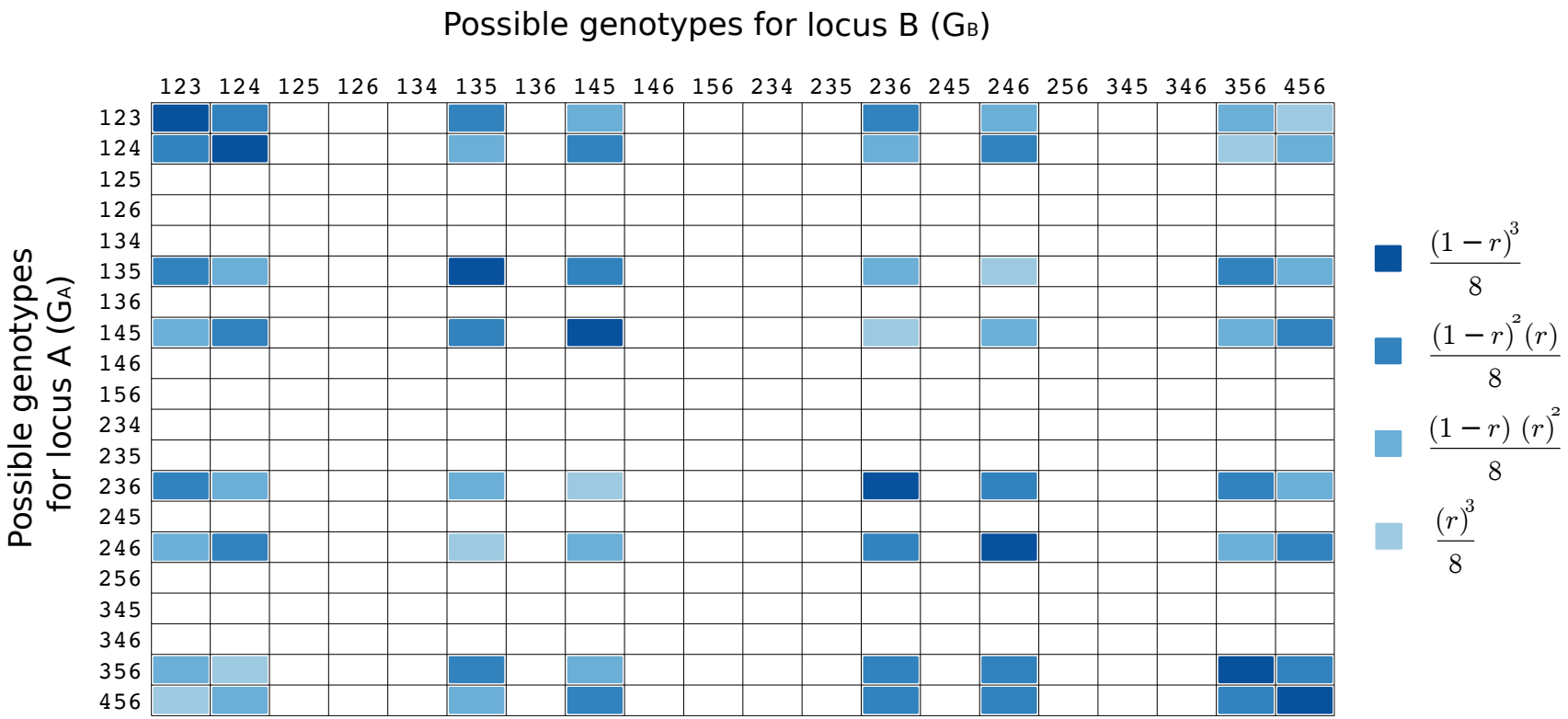
i: known number of recombinat bivalents between loci A and B

Gametic probability for ψ_1

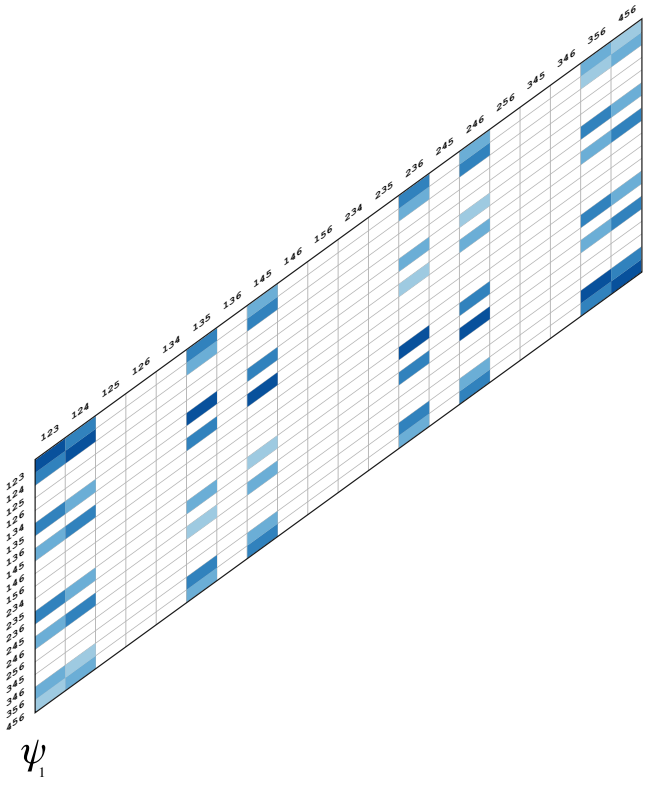




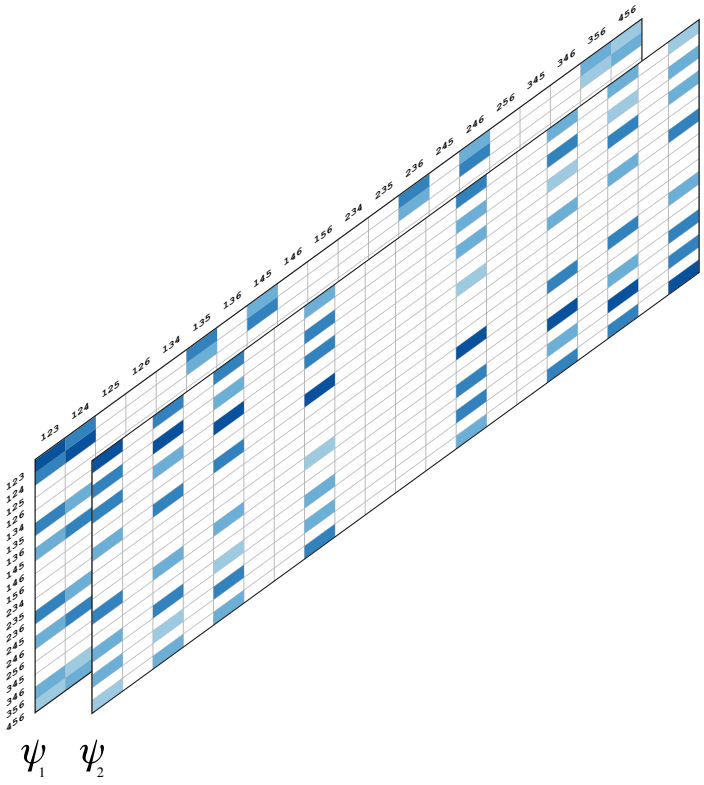
Gametic probability for ψ_1



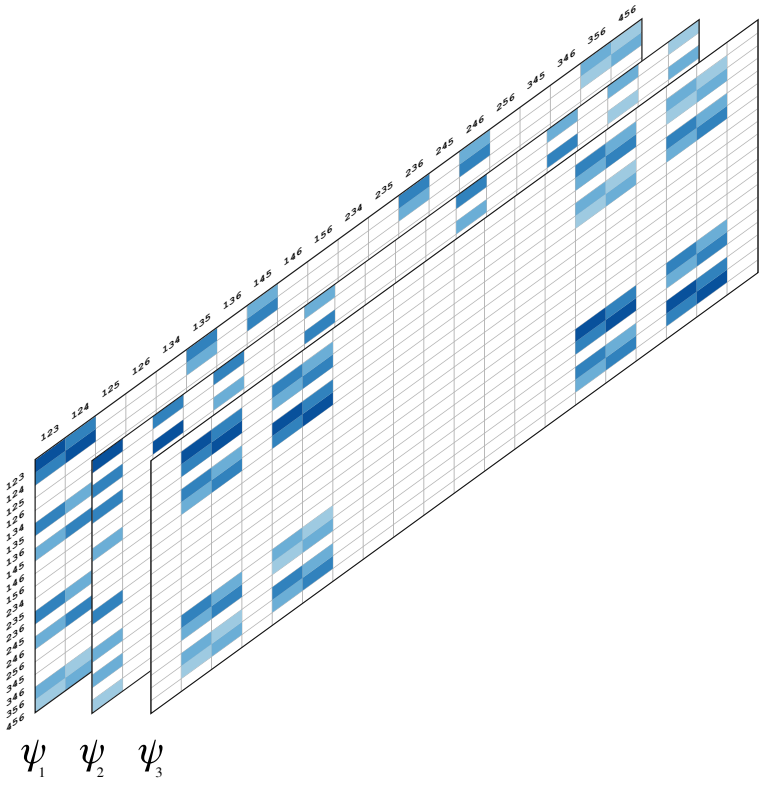
Gametic probability for all $\psi's$



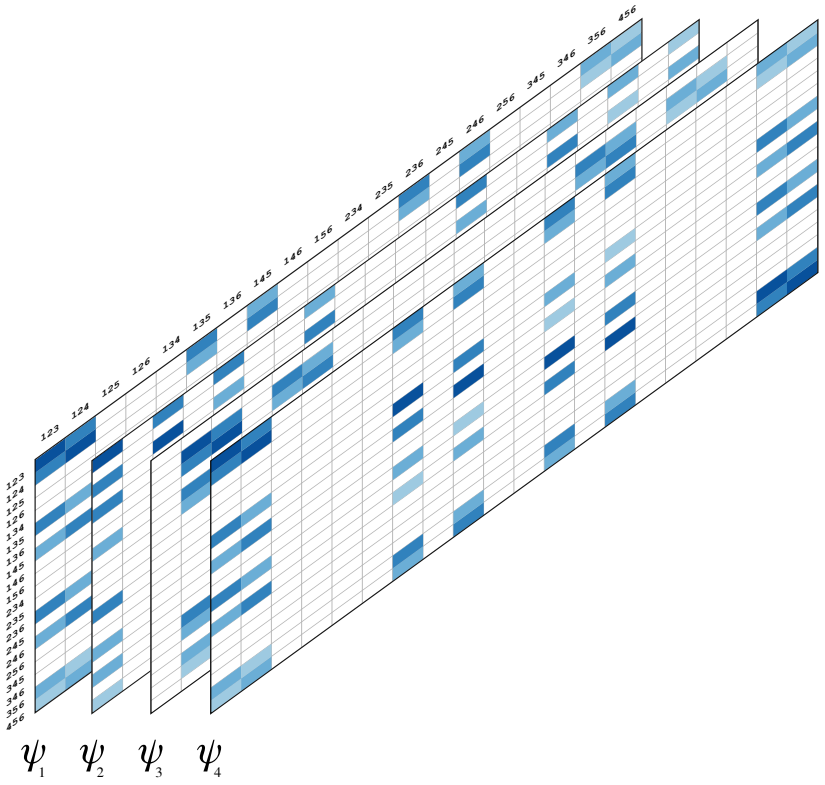
Gametic probability for all $\psi's$



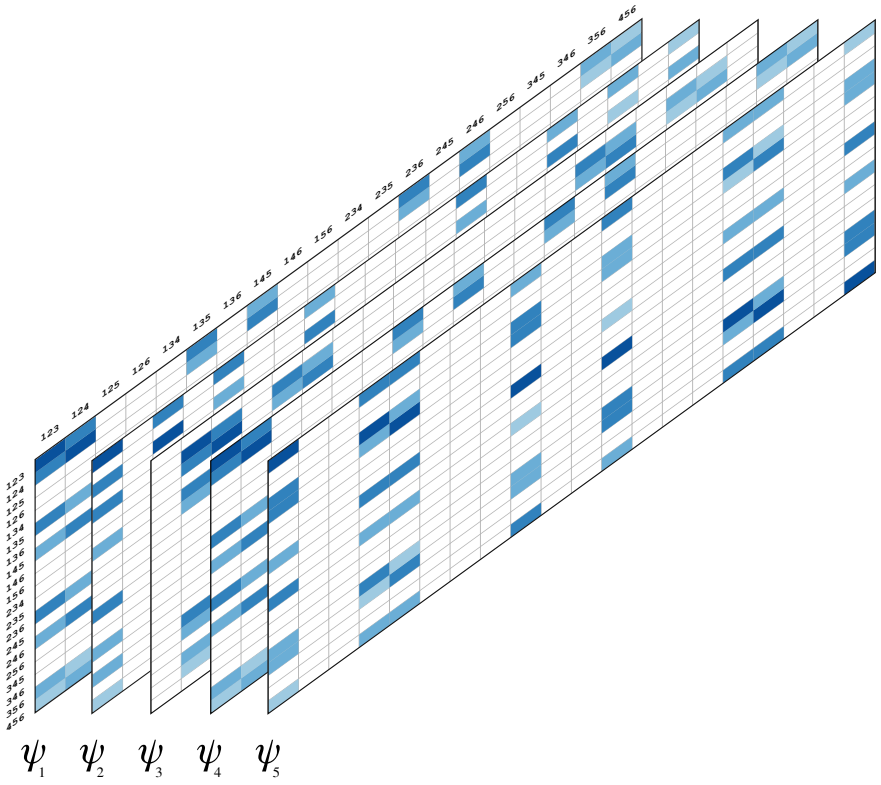
Gametic probability for all $\psi's$



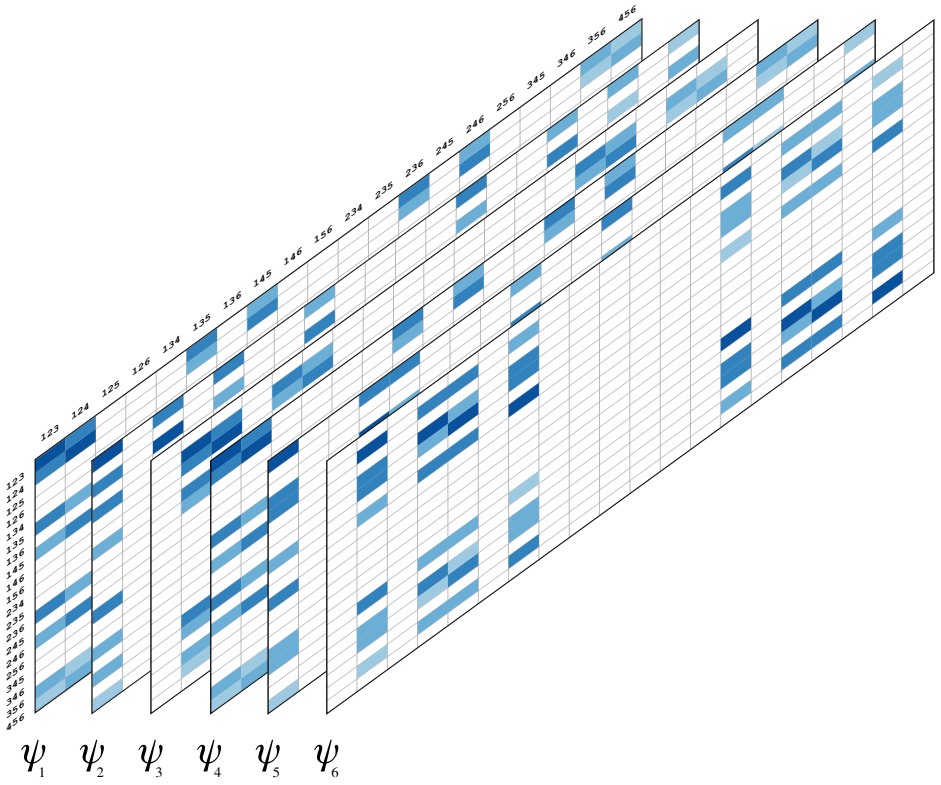
Gametic probability for all $\psi's$



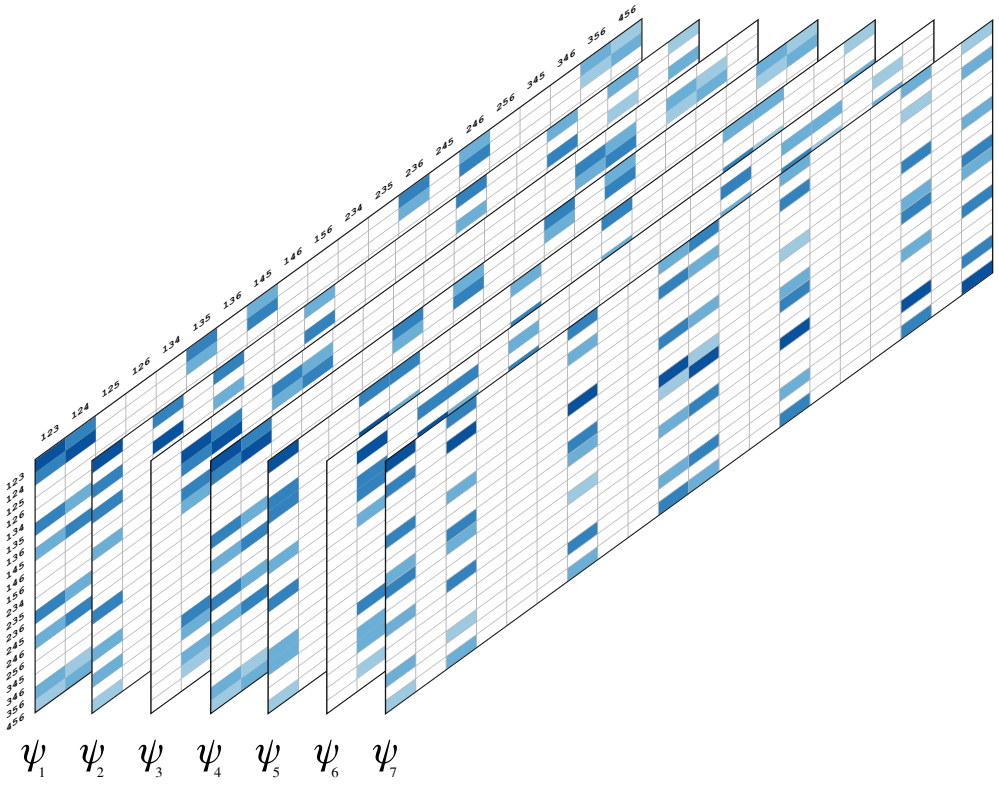
Gametic probability for all $\psi's$



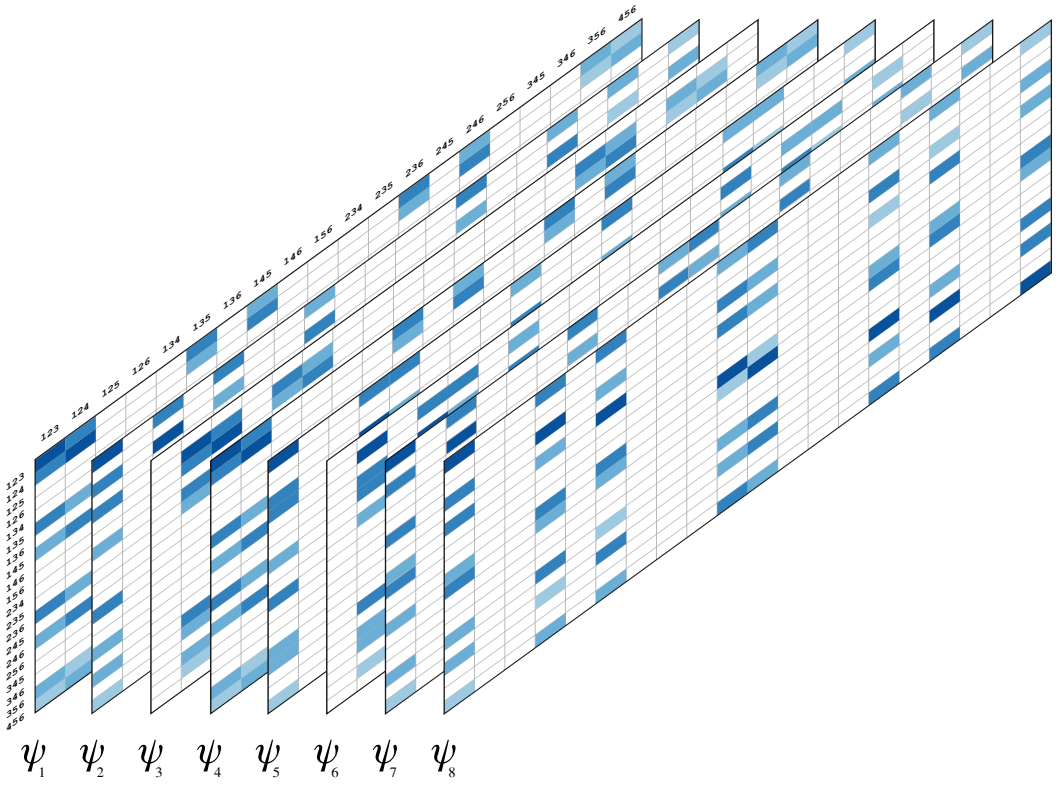
Gametic probability for all $\psi's$



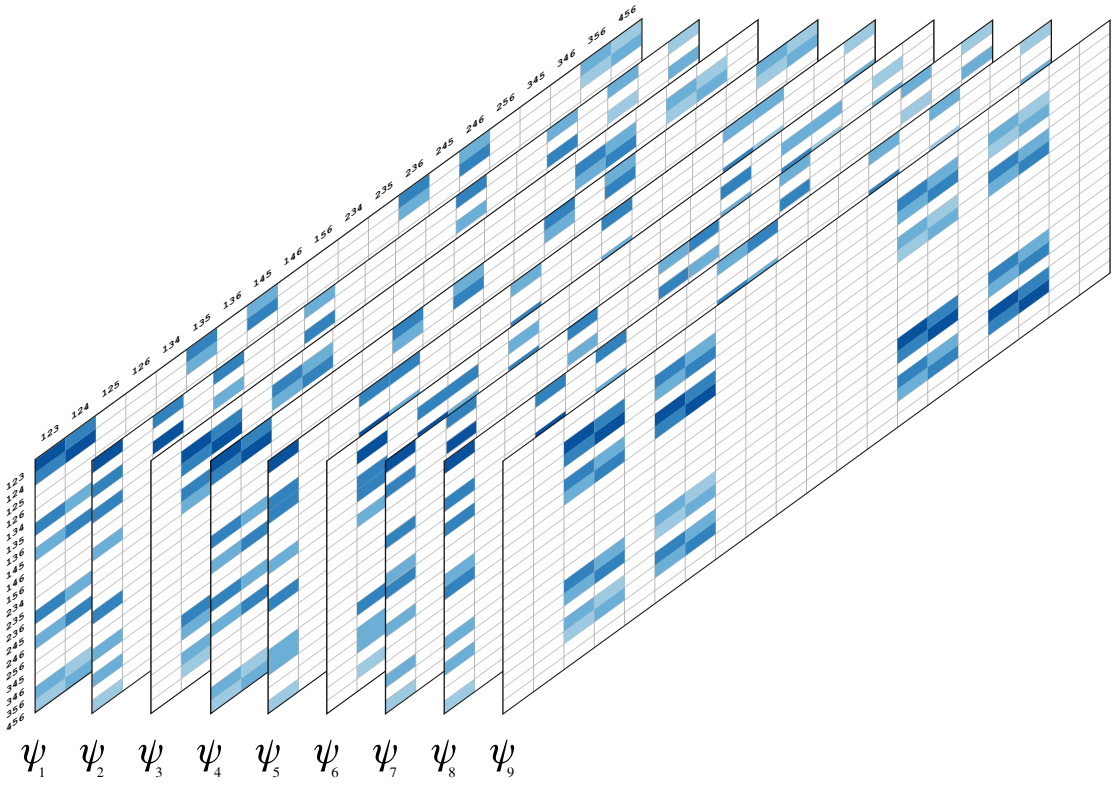
Gametic probability for all $\psi's$



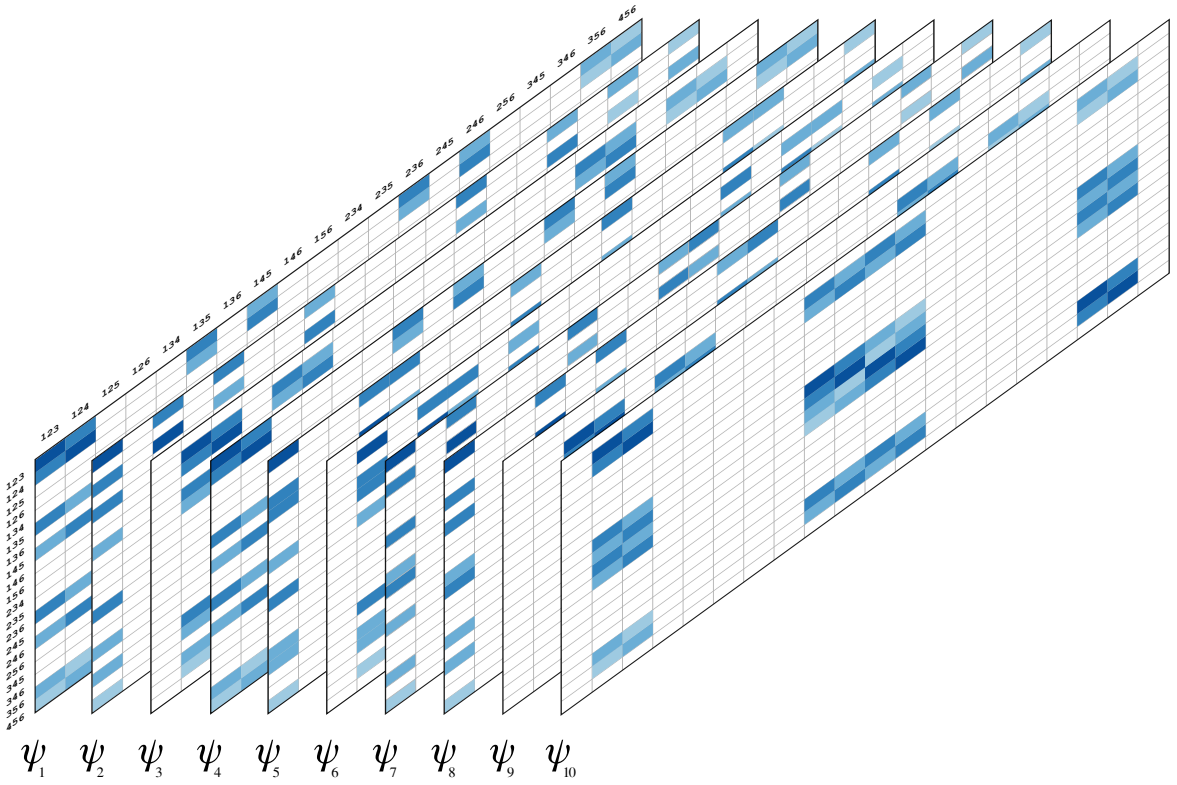
Gametic probability for all $\psi's$



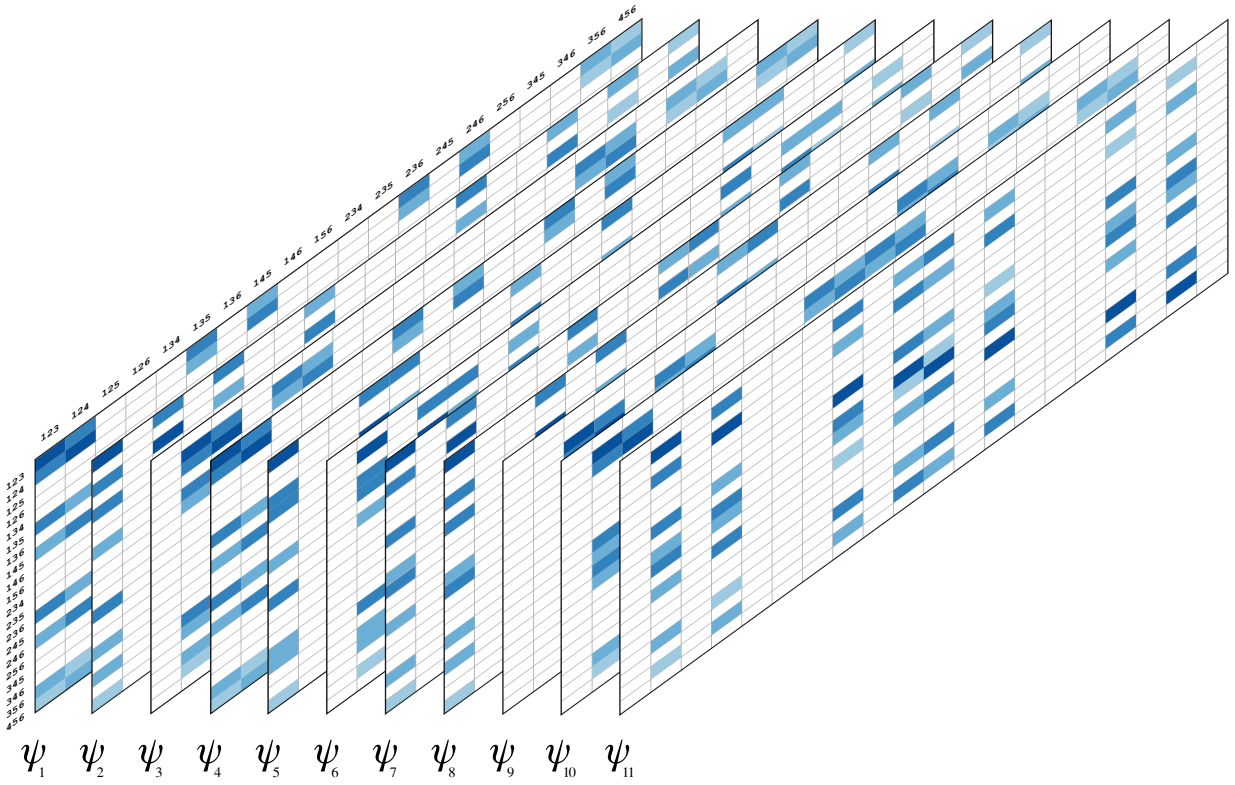
Gametic probability for all $\psi's$



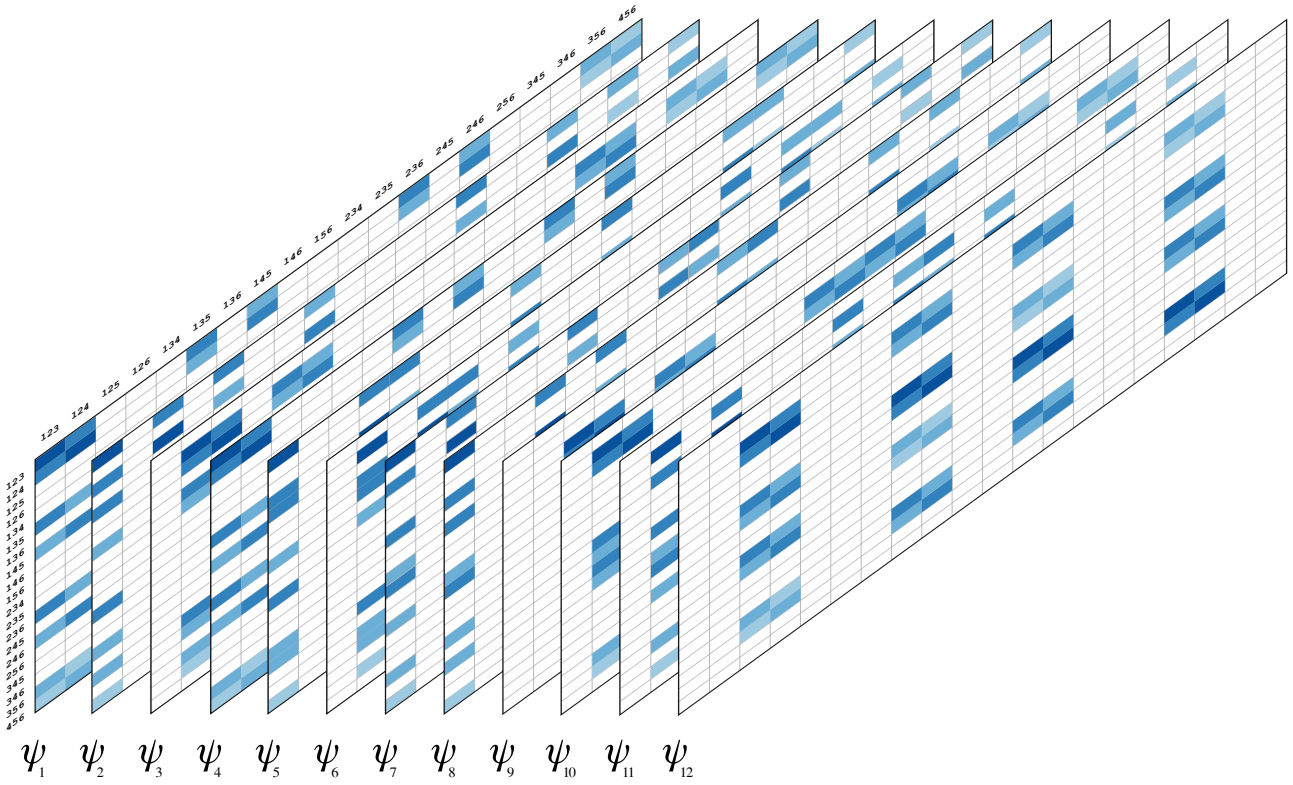
Gametic probability for all $\psi's$



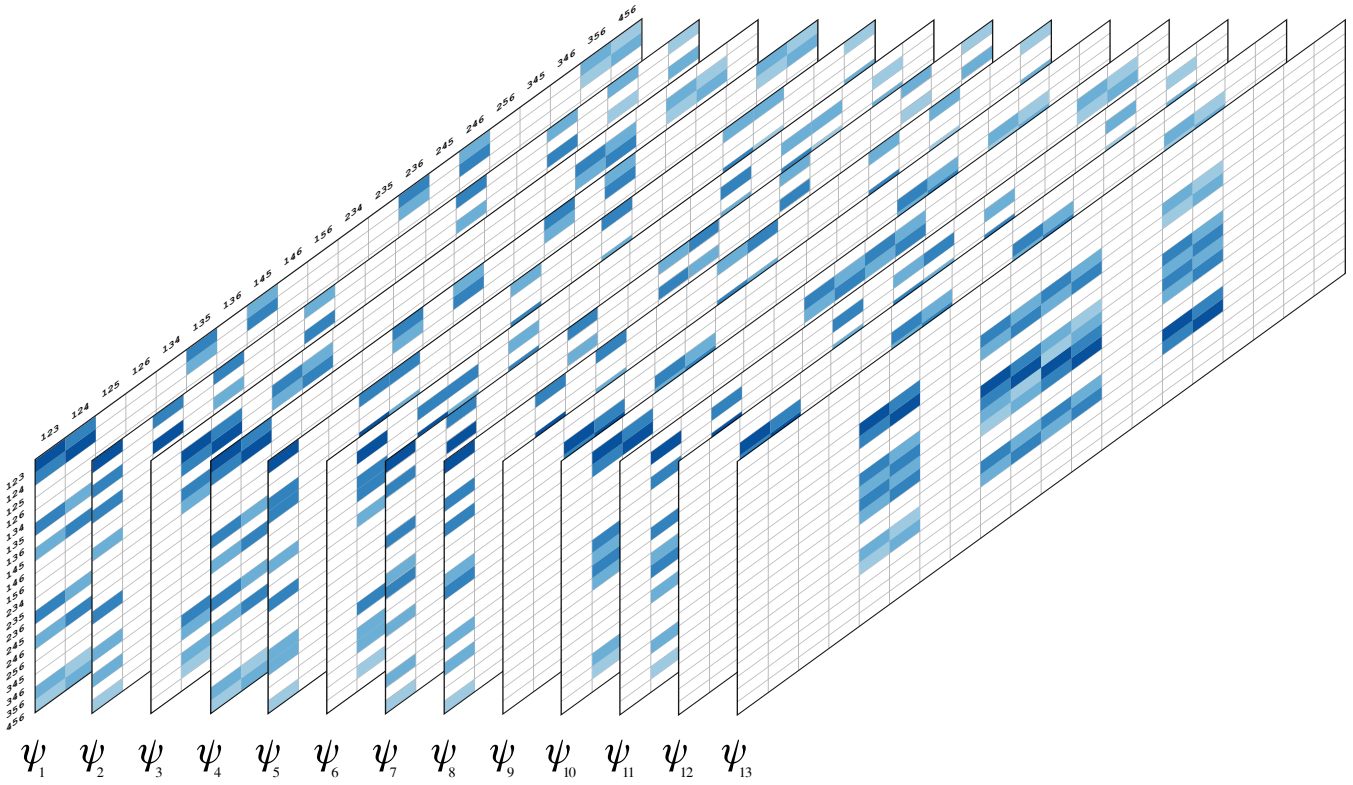
Gametic probability for all $\psi's$



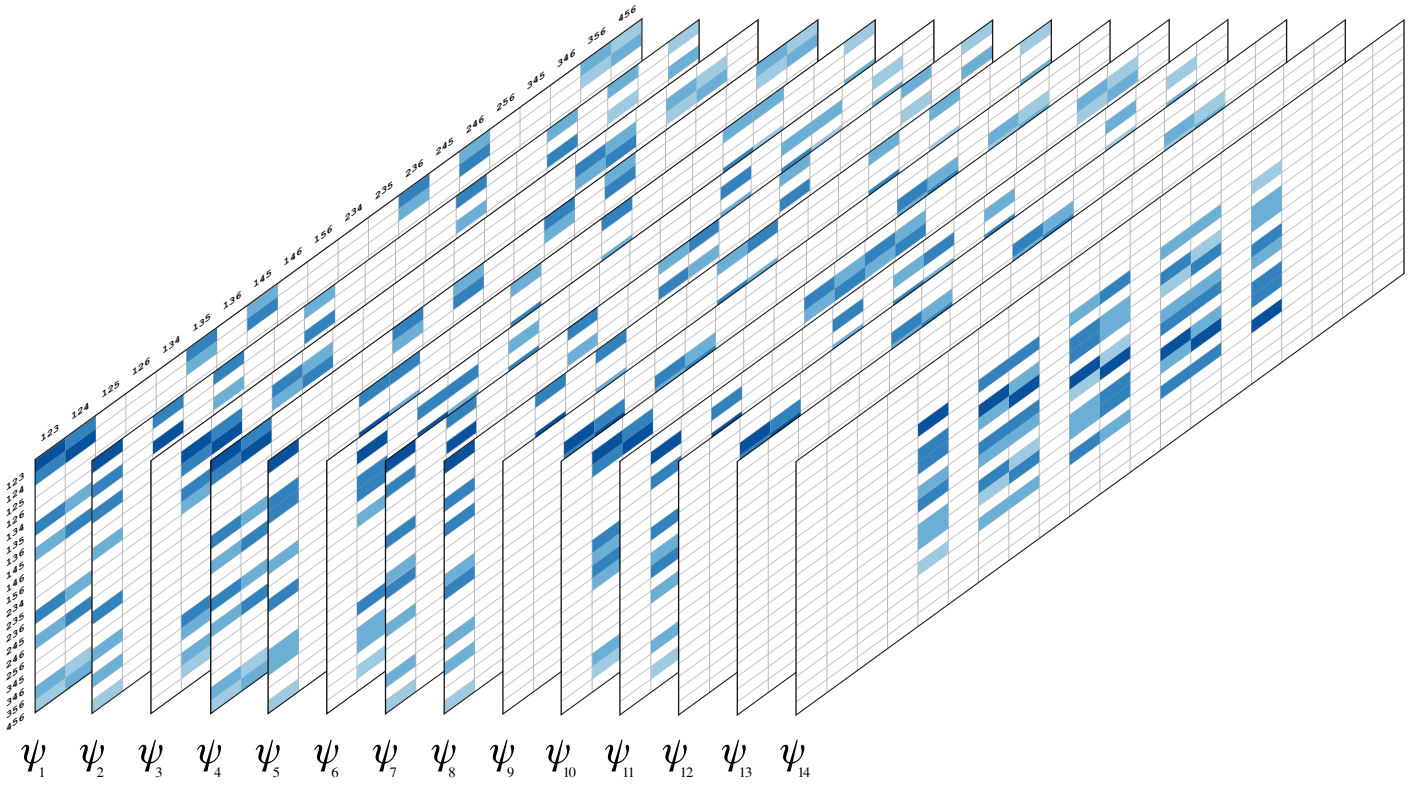
Gametic probability for all $\psi's$



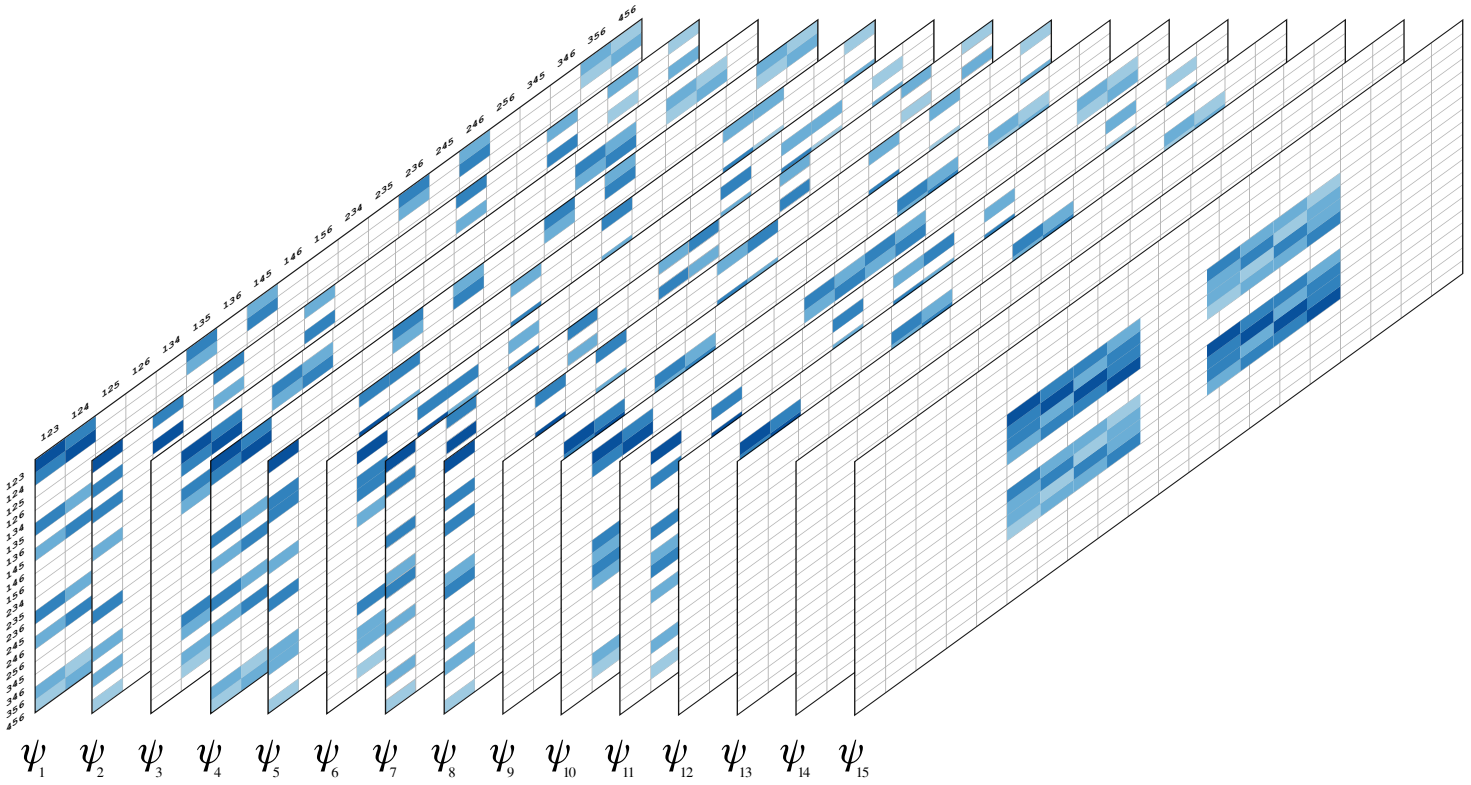
Gametic probability for all $\psi's$



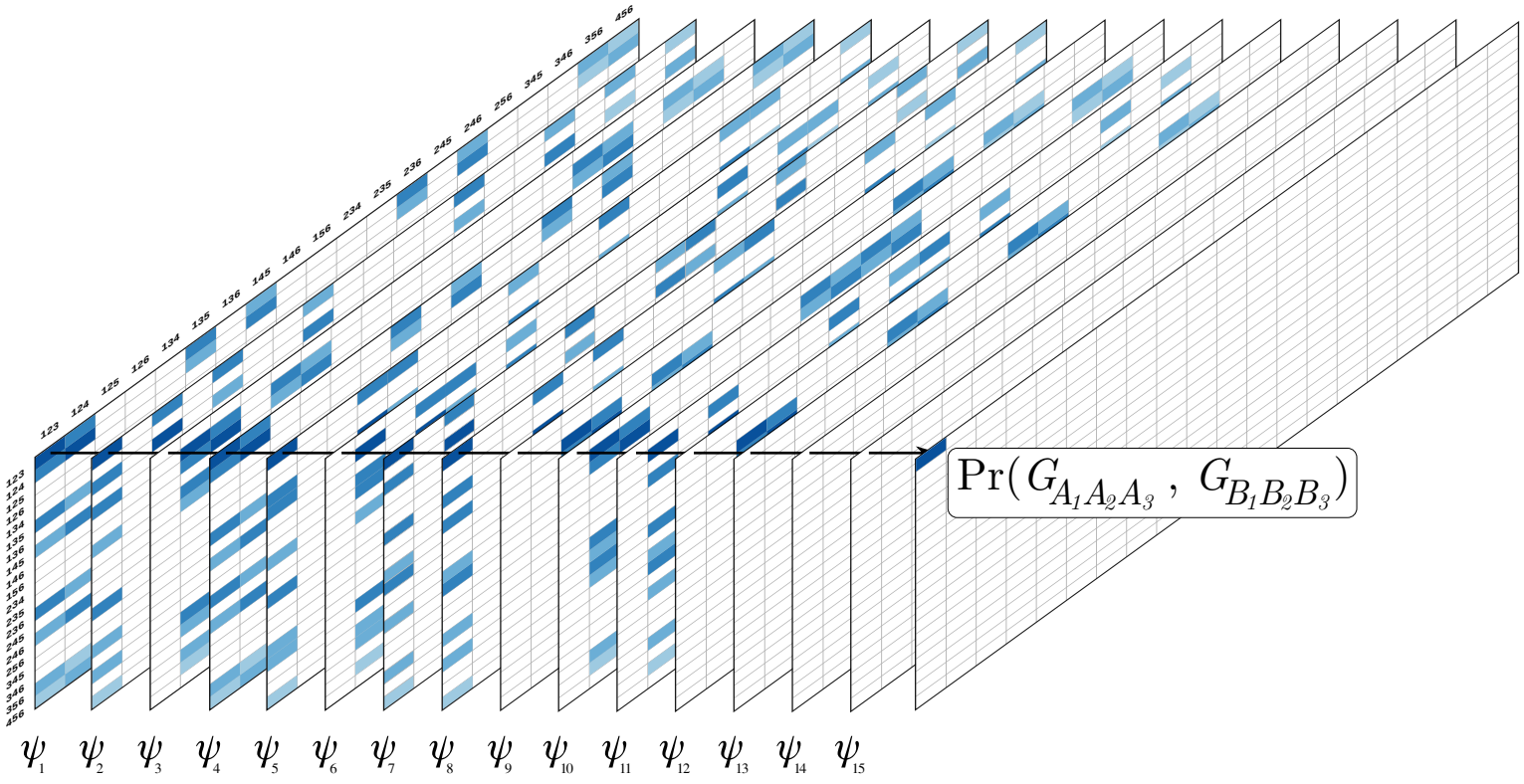
Gametic probability for all $\psi's$



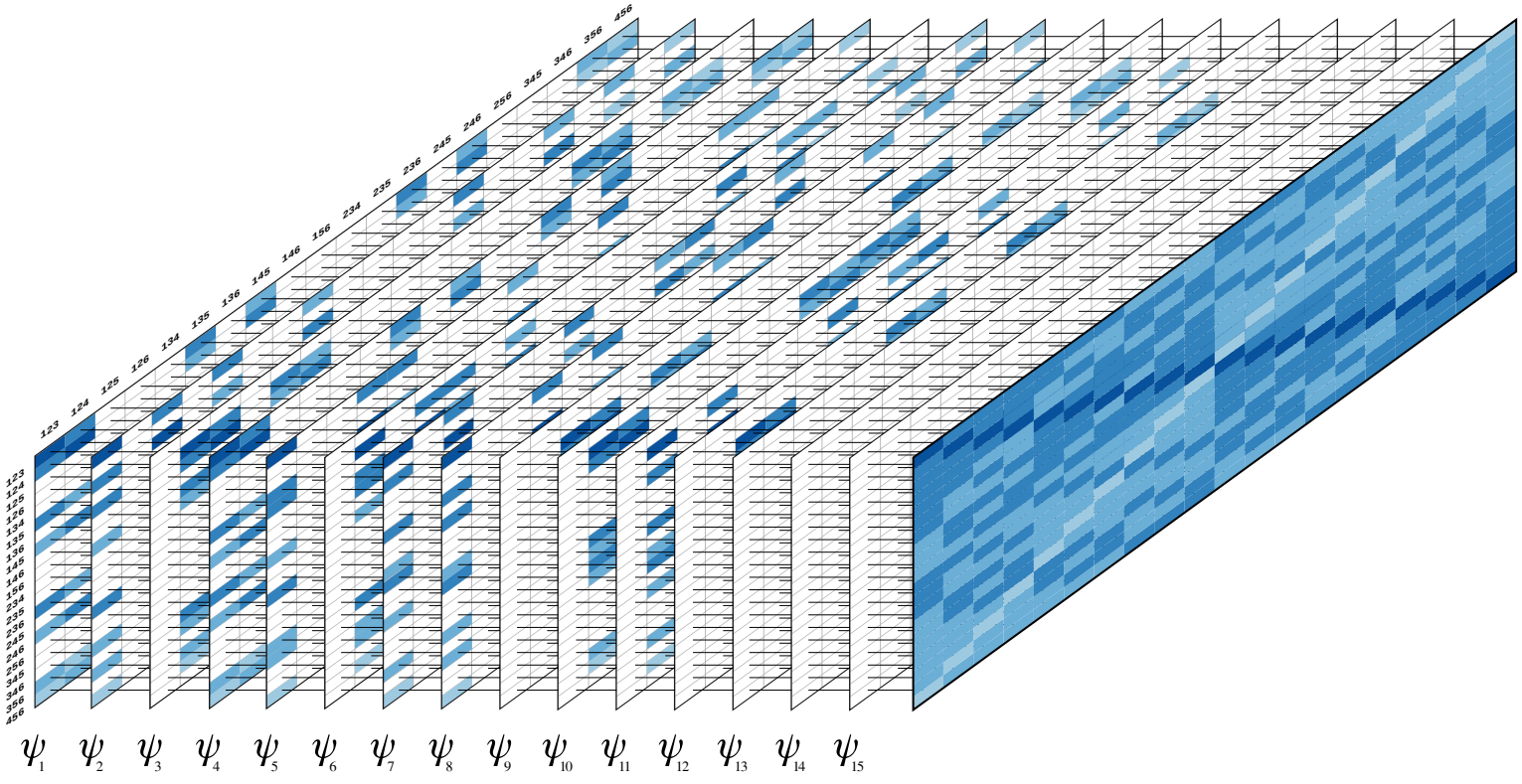
Gametic probability for all $\psi's$



Gametic probability for all $\psi's$

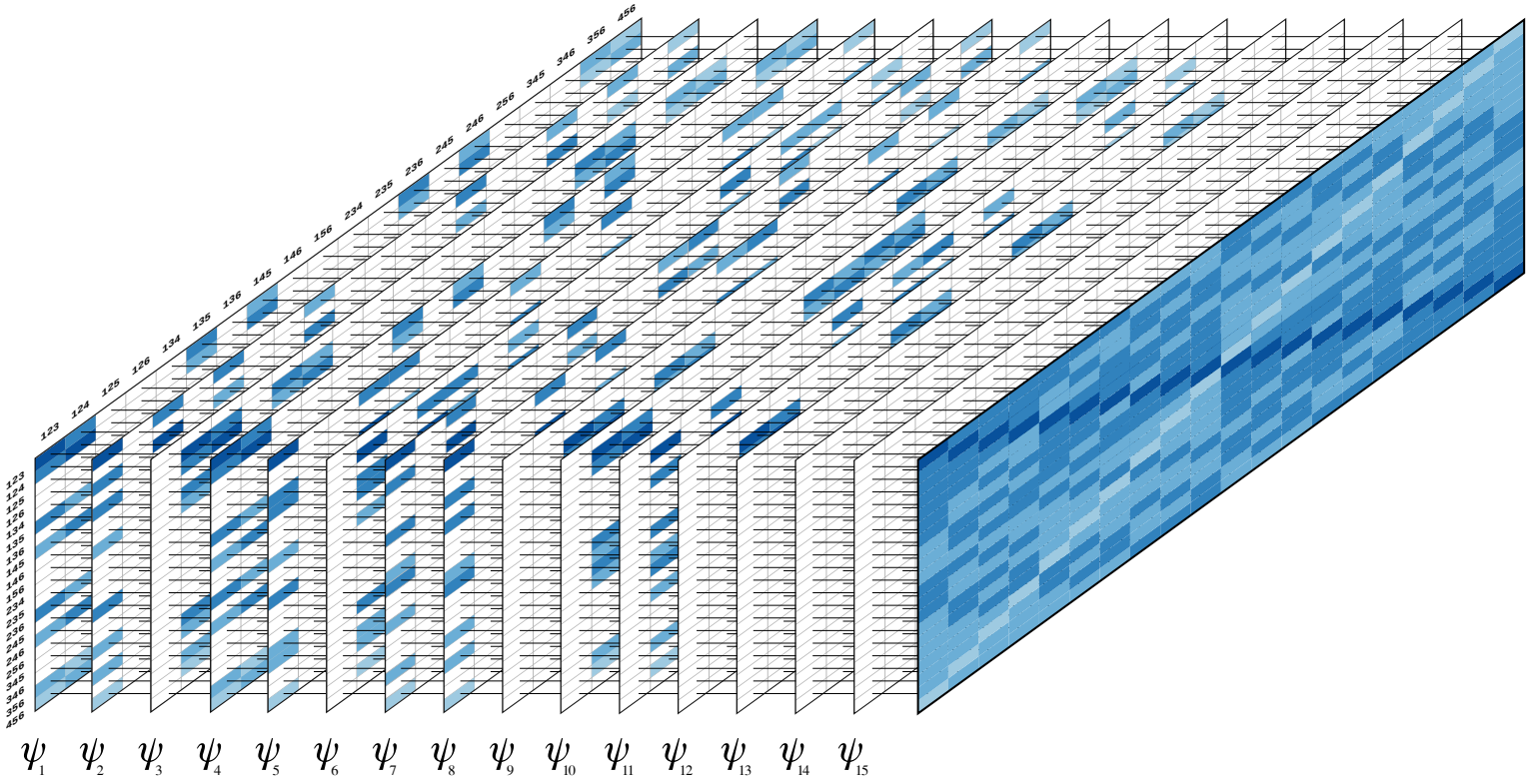


Gametic probability for all $\psi's$



$$\Pr(G_A, G_B) = \sum_j \Pr(G_A, G_B \mid \psi_j) \Pr(\psi_j) \tag{1}$$

Gametic probability for all $\psi's$

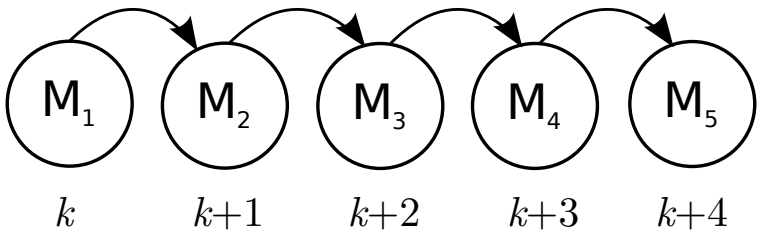


$$\Pr(G_A, G_B) = \sum_j \Pr(G_A, G_B \mid \psi_j) \Pr(\psi_j)$$

$$= \frac{l! \left(\frac{m}{2} - l\right)!}{w_m} \frac{(1 - r_k)^{\frac{m}{2} - l} (r_k)^l}{2^{\frac{m}{2}}}. \quad (1)$$

Transition Probabilities and the Hidden Markov Model (HMM)

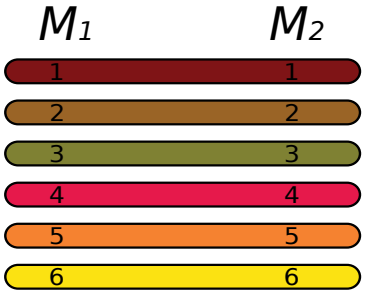
Markov Model: conditional independence



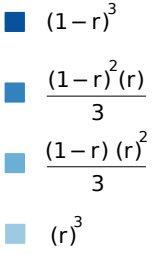
$$\Pr(G_{k+1}|G_k) = \frac{(1-r_k)^{\frac{p}{2}-l}(r_k)^l}{\binom{\frac{p}{2}}{l}}$$

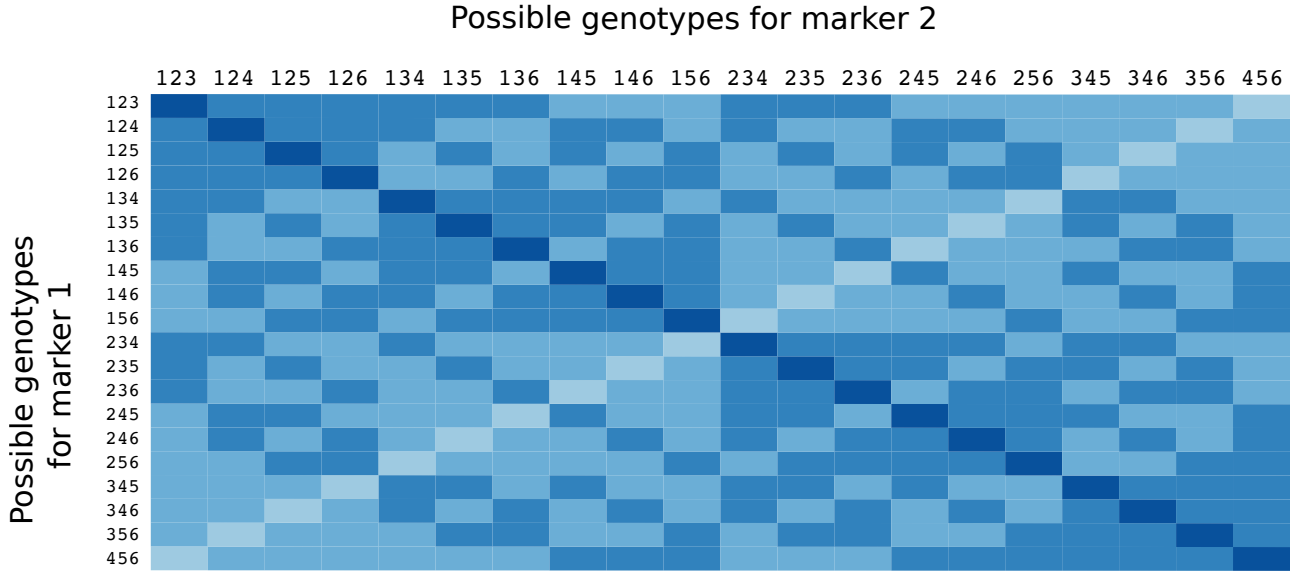
where r_k is teh recombination fraction between loci k and $k+1$, p is teh ploidy level and l is the number of recombinant events between k and $k+1$.

Final transition matrix for hexaploids (one parent)

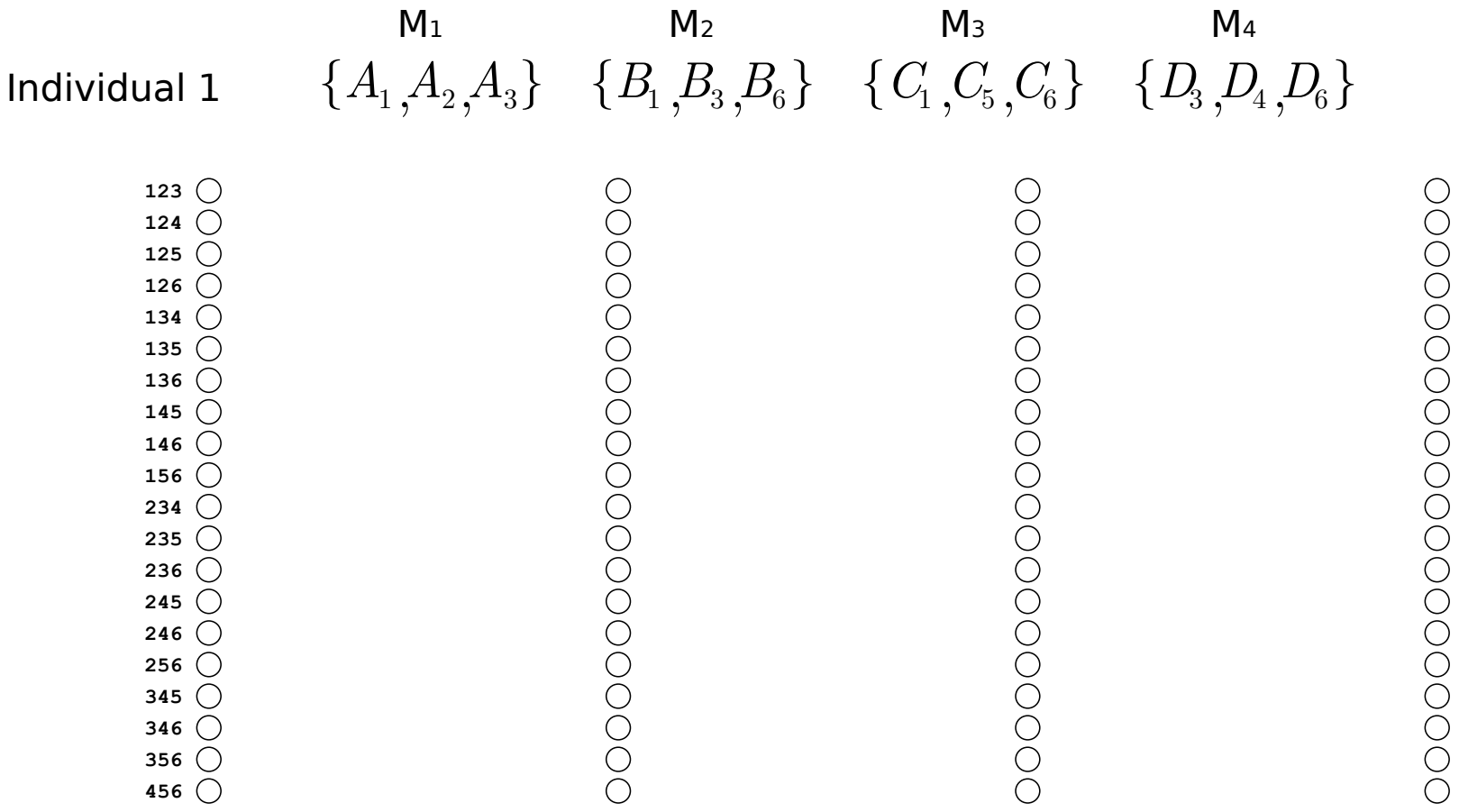


$$Pr(M_2|M_1)$$

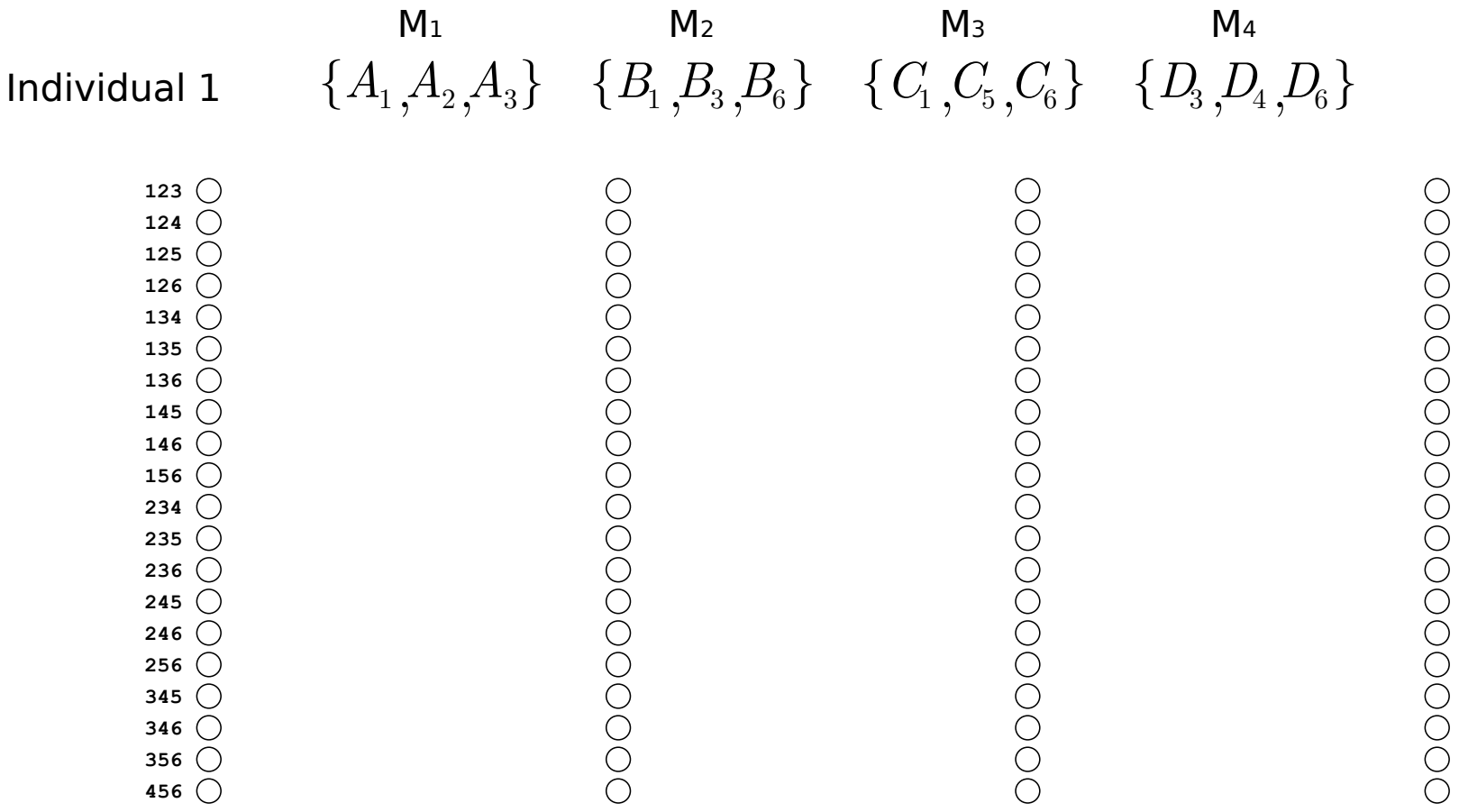




The Markov Model

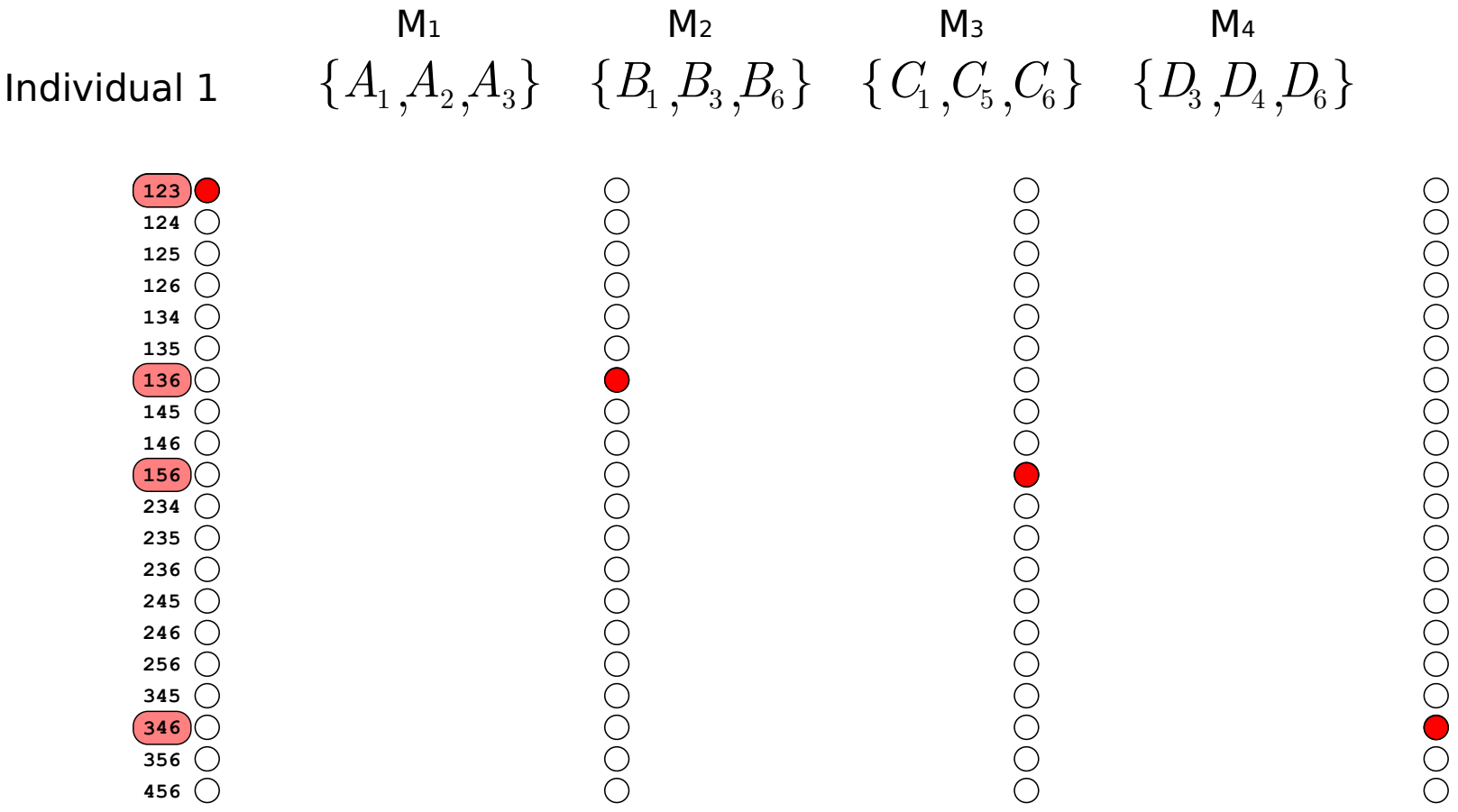


The Markov Model



$$\Pr(\text{ Sequence}_i \mid \text{Model}) =$$

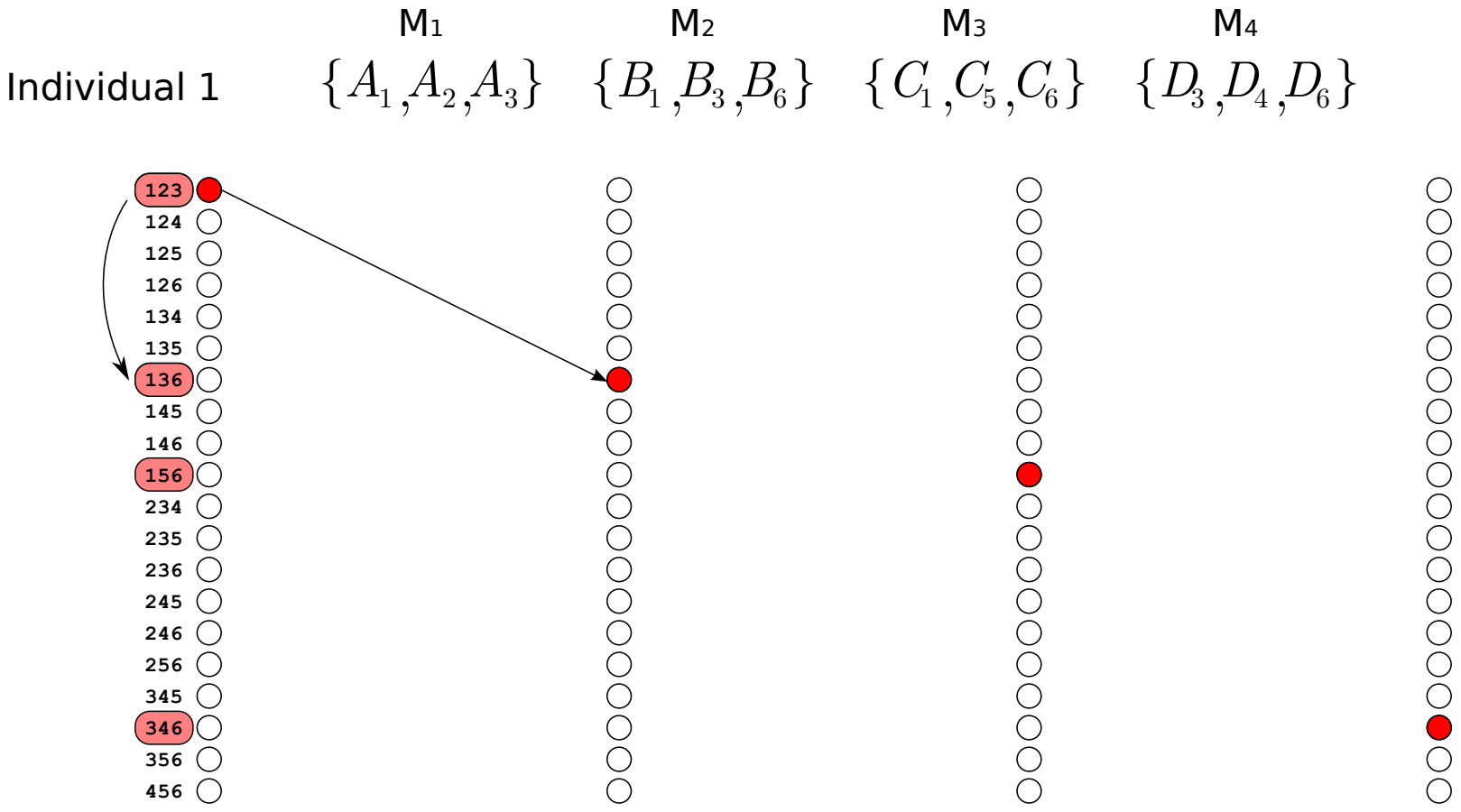
The Markov Model



$\Pr(\text{ Sequence}_1 \mid \text{Model}) =$

$\Pr(\{A_1, A_2\ A_3\})$

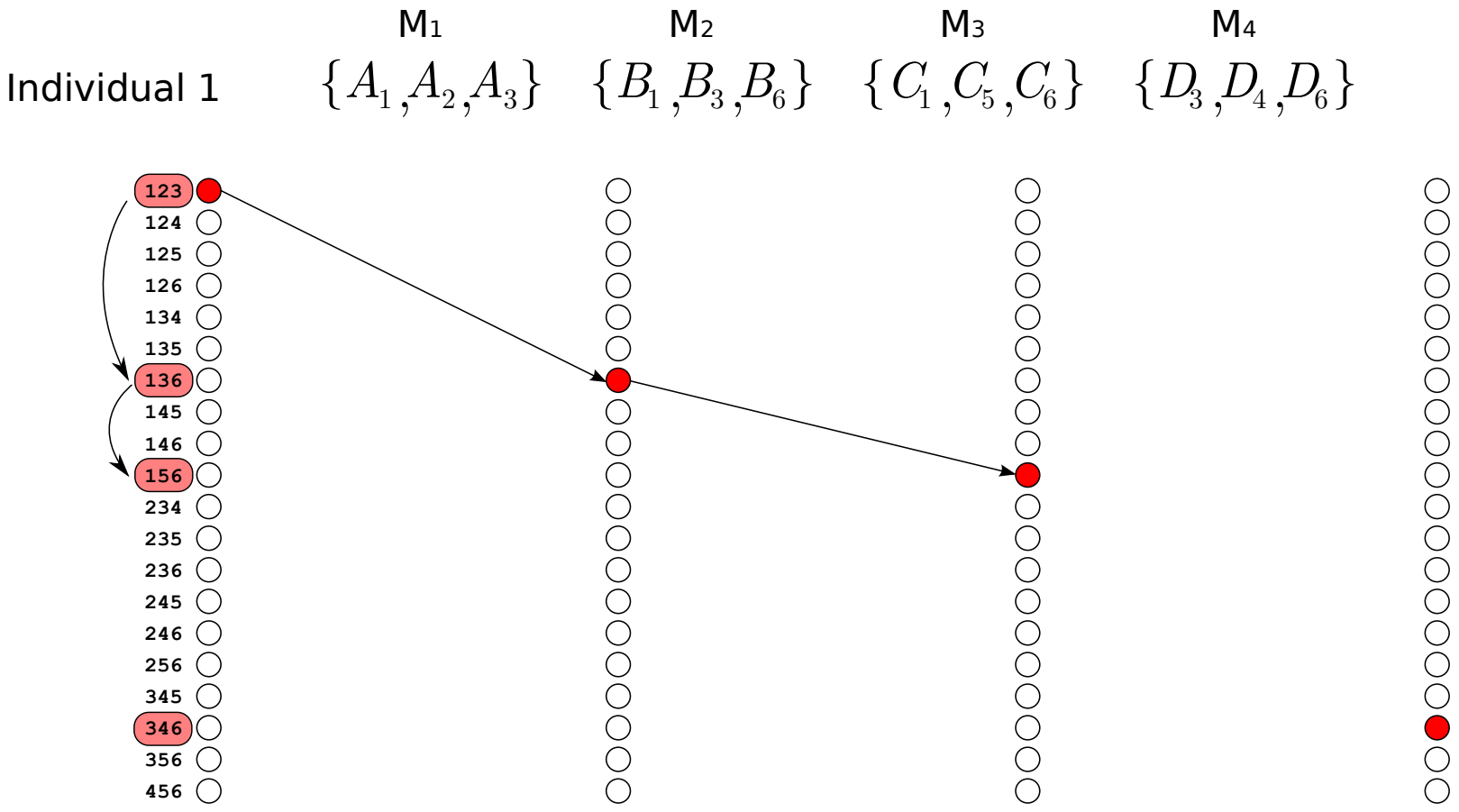
The Markov Model



$$\Pr(\text{ Sequence}_1 \mid \text{Model}) =$$

$$\Pr(\{A_1\, A_2\, A_3\}) \times \Pr(\{B_1\, B_3\, B_6\} \mid \{A_1\, A_2\, A_3\})$$

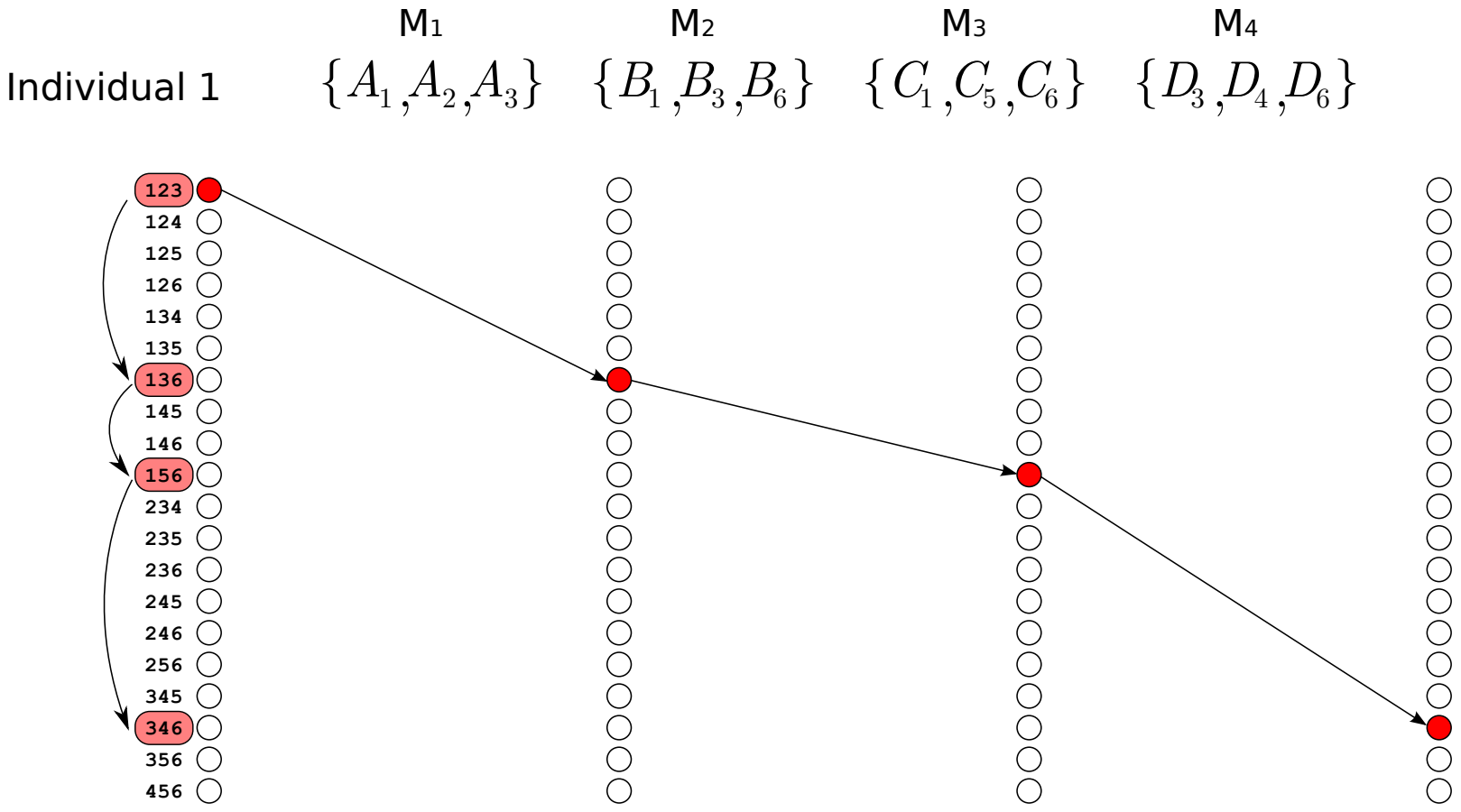
The Markov Model



$$\Pr(\text{ Sequence}_1 \,|\, \text{Model}) =$$

$$\Pr(\{A_1\, A_2\, A_3\}) \times \Pr(\{B_1\, B_3\, B_6\} | \{A_1\, A_2\, A_3\}) \times \Pr(\{C_1\, C_5\, C_6\} | \{B_1\, B_3\, B_6\})$$

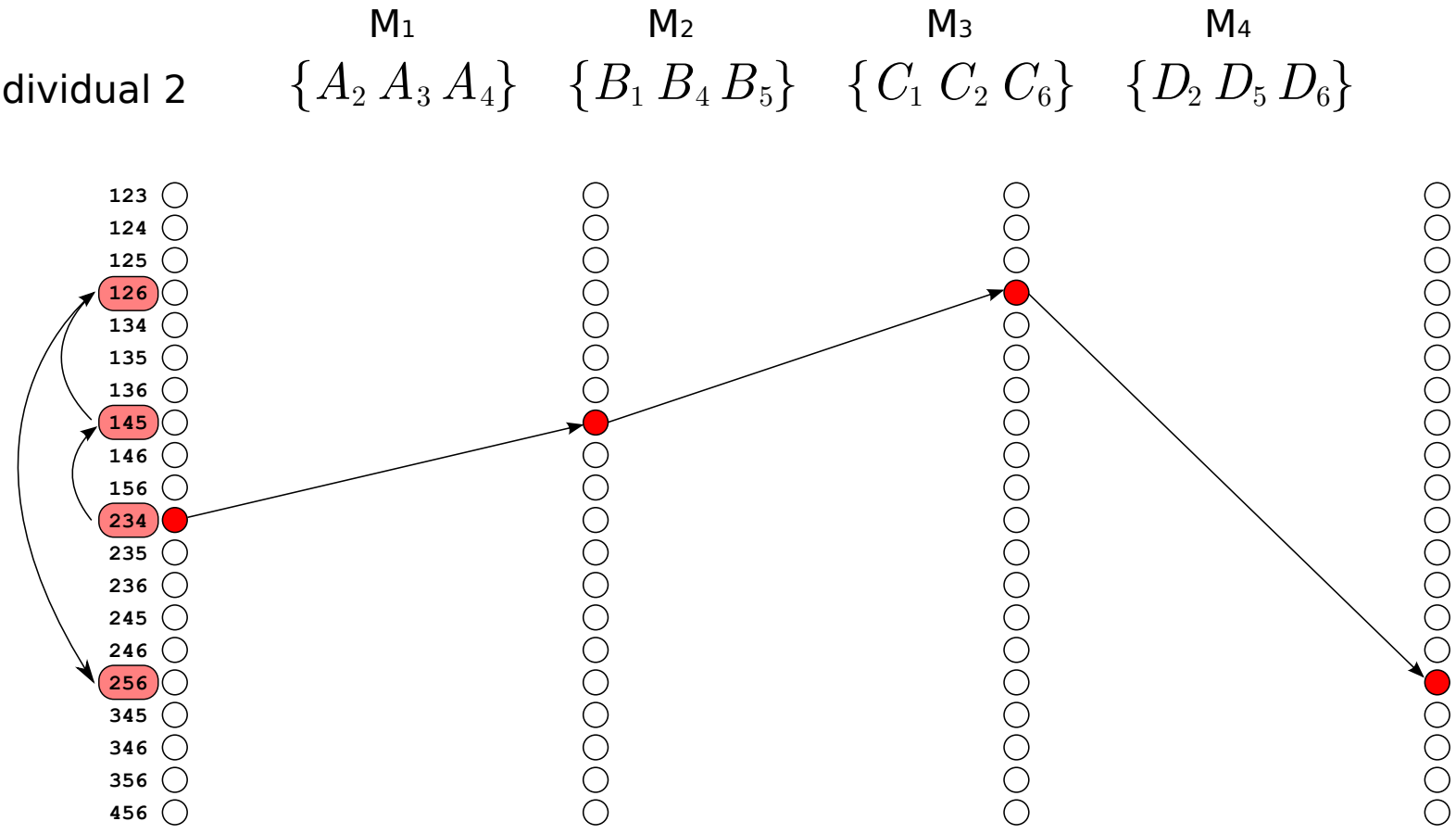
The Markov Model



$$\Pr(\text{ Sequence}_1 \mid \text{Model}) =$$

$$\Pr(\{A_1\ A_2\ A_3\}) \times \Pr(\{B_1\ B_3\ B_6\} \mid \{A_1\ A_2\ A_3\}) \times \Pr(\{C_1\ C_5\ C_6\} \mid \{B_1\ B_3\ B_6\}) \times \Pr(\{D_3\ D_4\ D_6\} \mid \{C_1\ C_5\ C_6\})$$

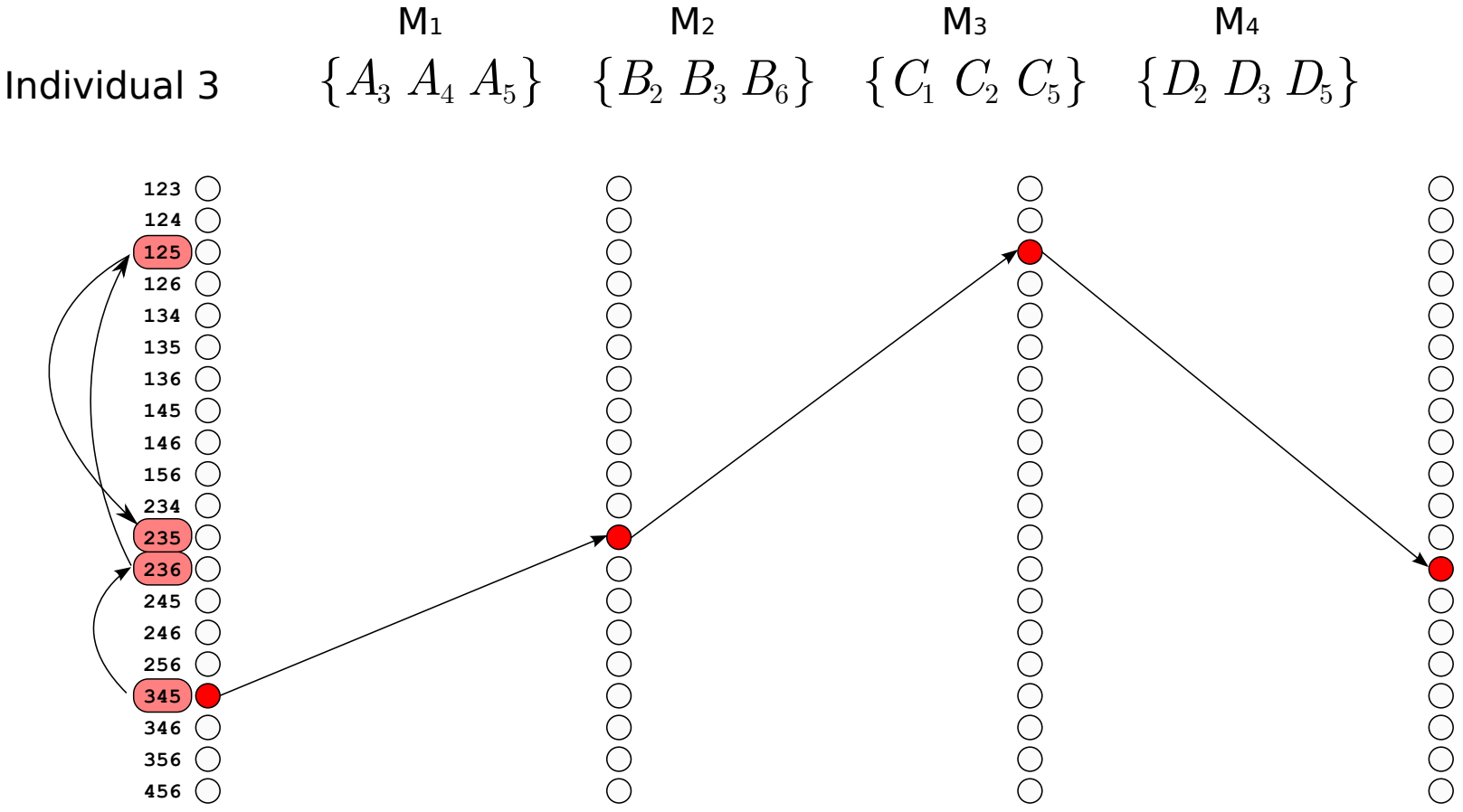
The Markov Model



$$\Pr(\text{ Sequence}_2 \big| \text{ Model}) =$$

$$\Pr(\{A_2\ A_3\ A_4\}) \times \Pr(\{B_1\ B_4\ B_5\} \big| \{A_2\ A_3\ A_4\}) \times \Pr(\{C_1\ C_2\ C_6\} \big| \{B_1\ B_4\ B_5\}) \times \Pr(\{D_2\ D_5\ D_6\} \big| \{C_1\ C_2\ C_6\})$$

The Markov Model



$$\Pr(\text{ Sequence}_3 \big| \text{ Model}) =$$

$$\Pr(\{A_3\,A_4\,A_5\}) \times \Pr(\{B_2\,B_3\,B_6\} \big| \{A_3\,A_4\,A_5\}) \times \Pr(\{C_1\,C_2\,C_5\} \big| \{B_2\,B_3\,B_6\}) \times \Pr(\{D_2\,D_3\,D_5\} \big| \{C_1\,C_2\,C_5\})$$

Likelihood

$$L = \prod_{\text{all individuals}} \Pr(\text{Sequence}_i | \text{Model})$$

Maximizing the likelihood function, it is possible to obtain the estimates of

$$[r_{AB}, r_{BC}, r_{CD}]$$

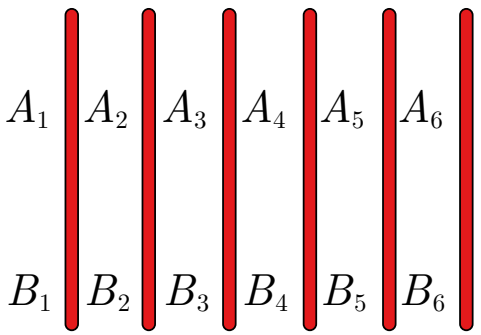
Likelihood

$$L = \prod_{\text{all individuals}} \Pr(\text{Sequence}_i | \text{Model})$$

Maximizing the likelihood function, it is possible to obtain the estimates of

$$[r_{AB}, r_{BC}, r_{CD}]$$

However, we access the genotypes using biallelic SPNs.



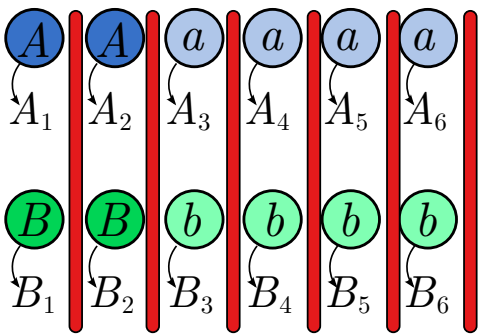
Likelihood

$$L = \prod_{\text{all individuals}} \Pr(\text{Sequence}_i | \text{Model})$$

Maximizing the likelihood function, it is possible to obtain the estimates of

$$[r_{AB}, r_{BC}, r_{CD}]$$

However, we access the genotypes using biallelic SPNs.



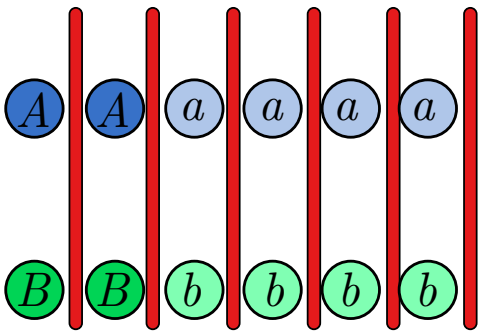
Likelihood

$$L = \prod_{\text{all individuals}} \Pr(\text{Sequence}_i|\text{Model})$$

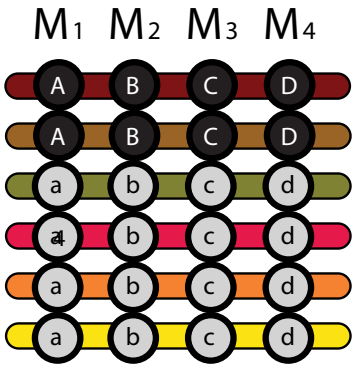
Maximizing the likelihood function, it is possible to obtain the estimates of

$$[r_{AB}, r_{BC}, r_{CD}]$$

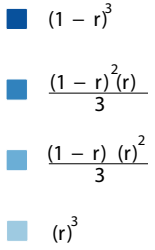
However, we access the genotypes using biallelic SPNs.

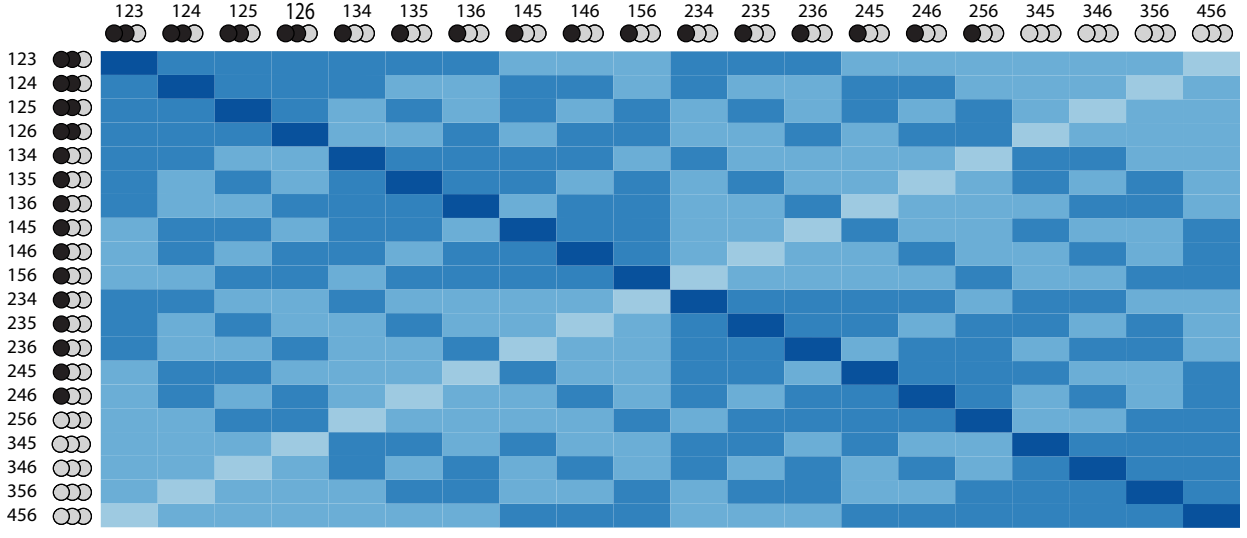


Final transition matrix for hexaploids (one parent)



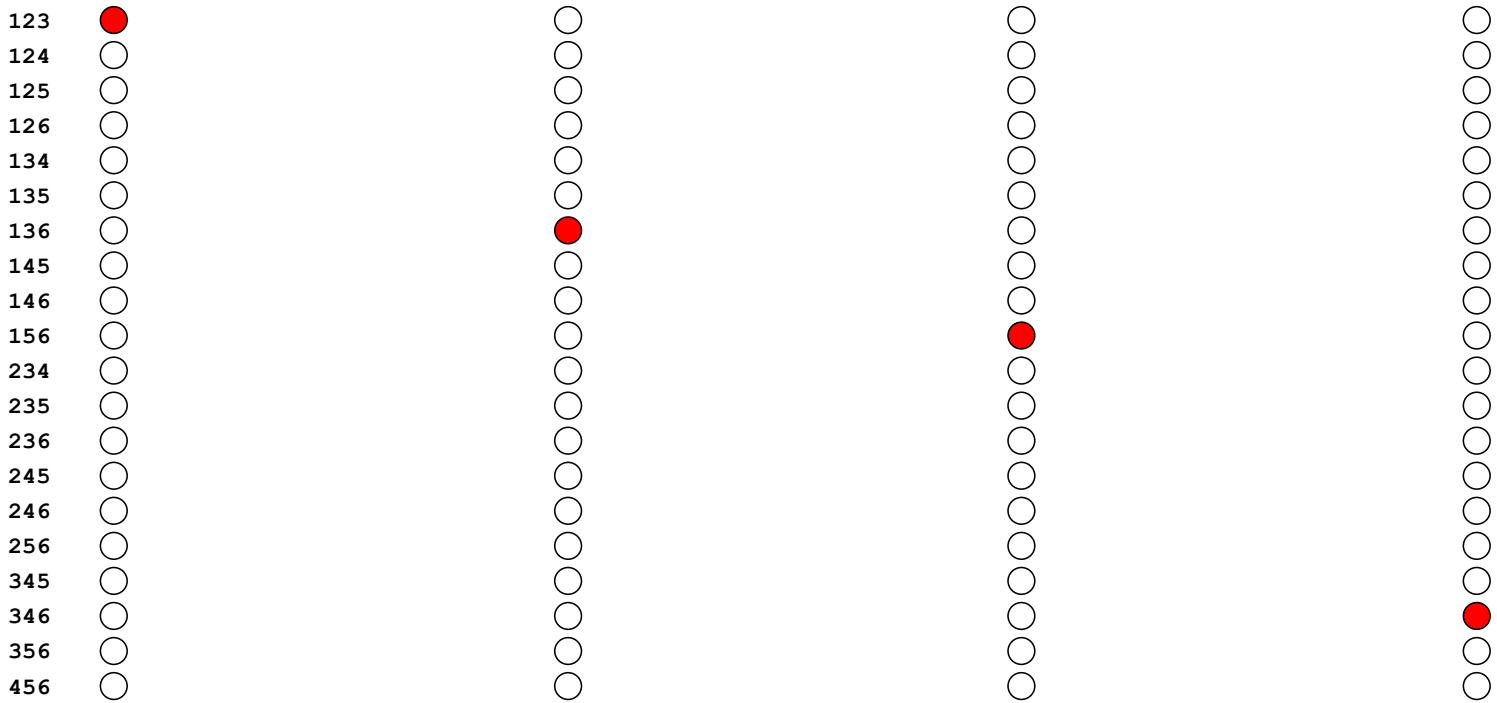
$$Pr(M_2|M_1)$$



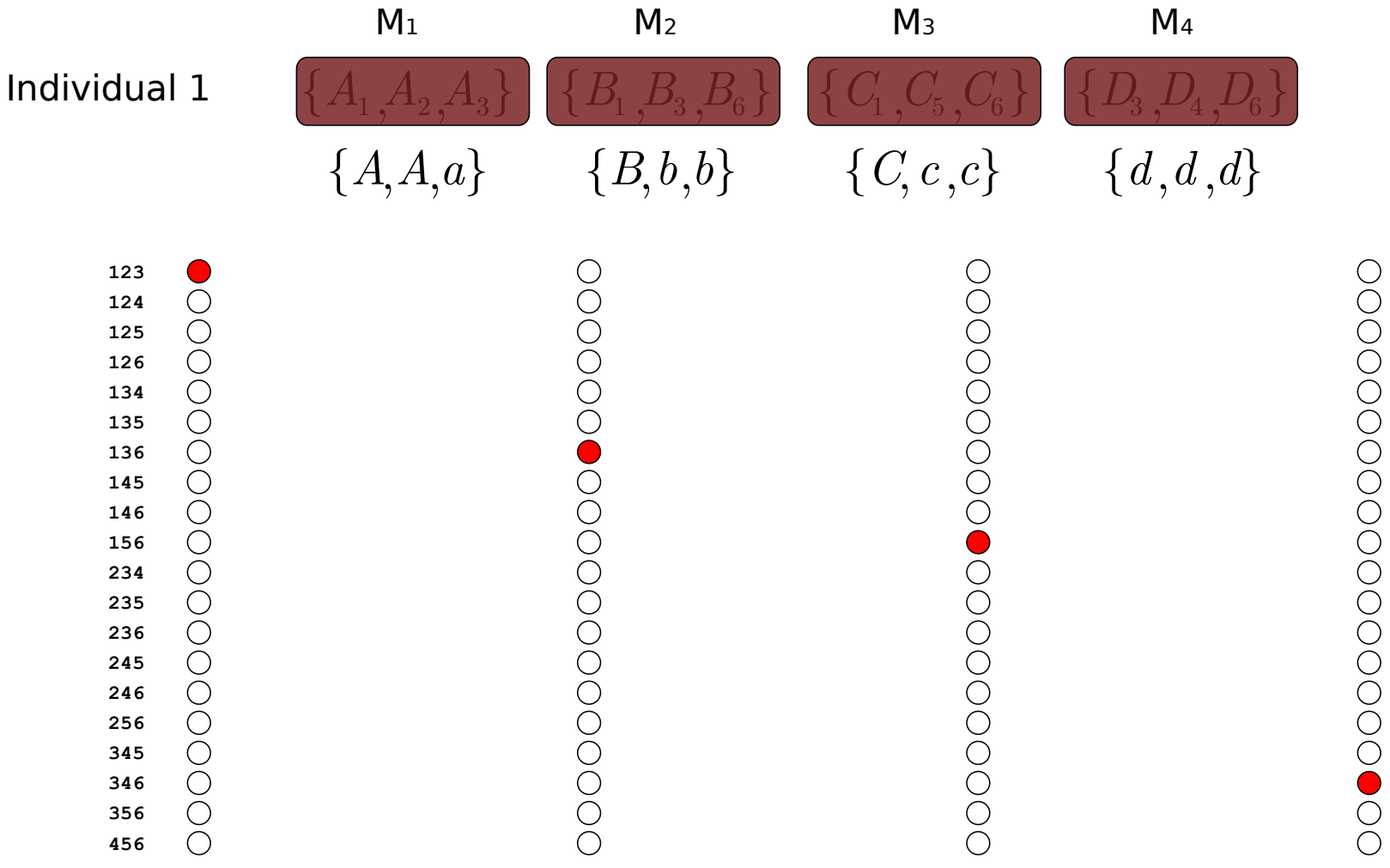


Hidden Markov model (HMM)

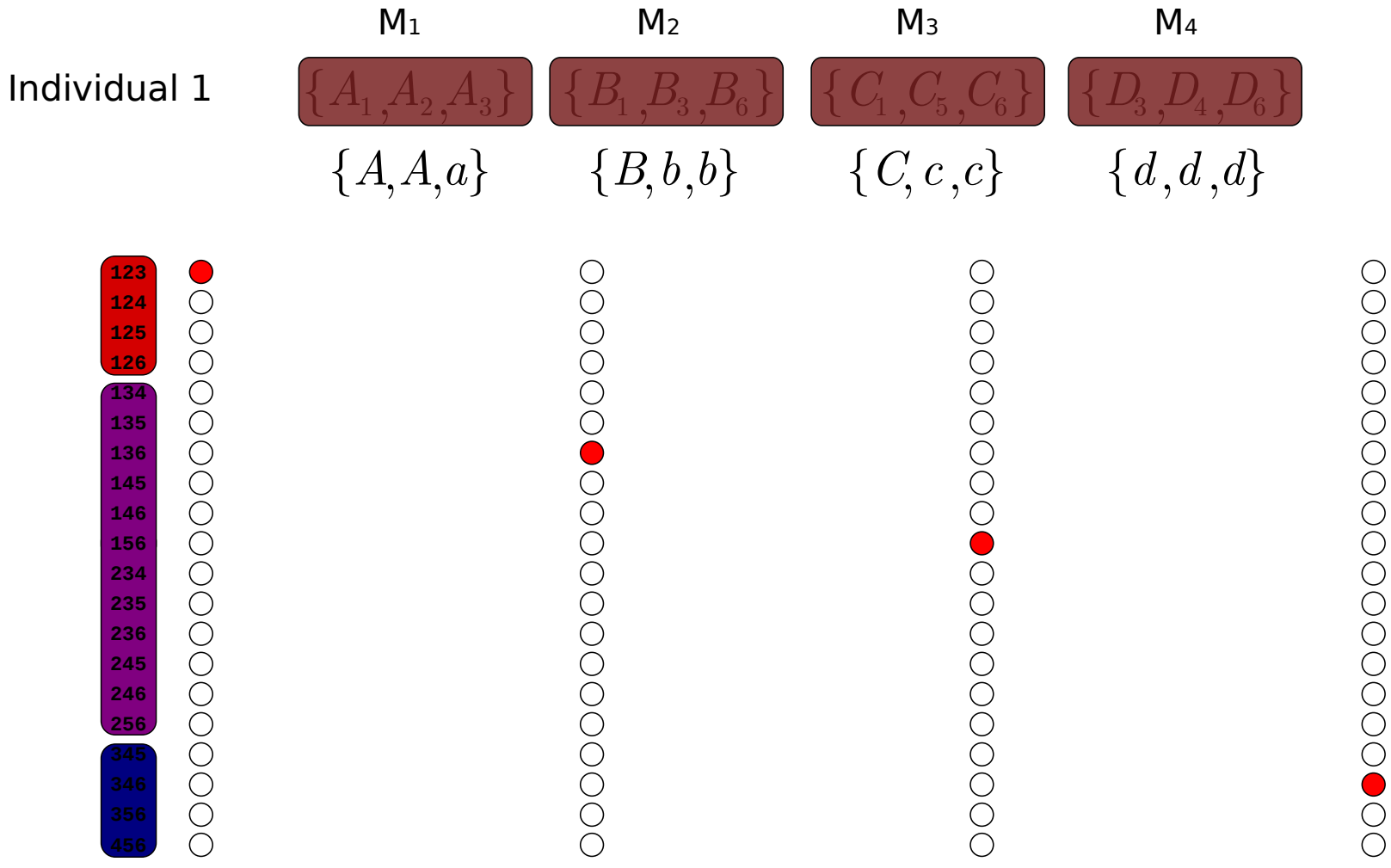
	M_1	M_2	M_3	M_4
Individual 1	$\{P_1^1, P_1^2, P_1^3\}$	$\{P_2^1, P_2^3, P_2^6\}$	$\{P_3^1, P_3^5, P_3^6\}$	$\{P_4^3, P_4^4, P_4^6\}$



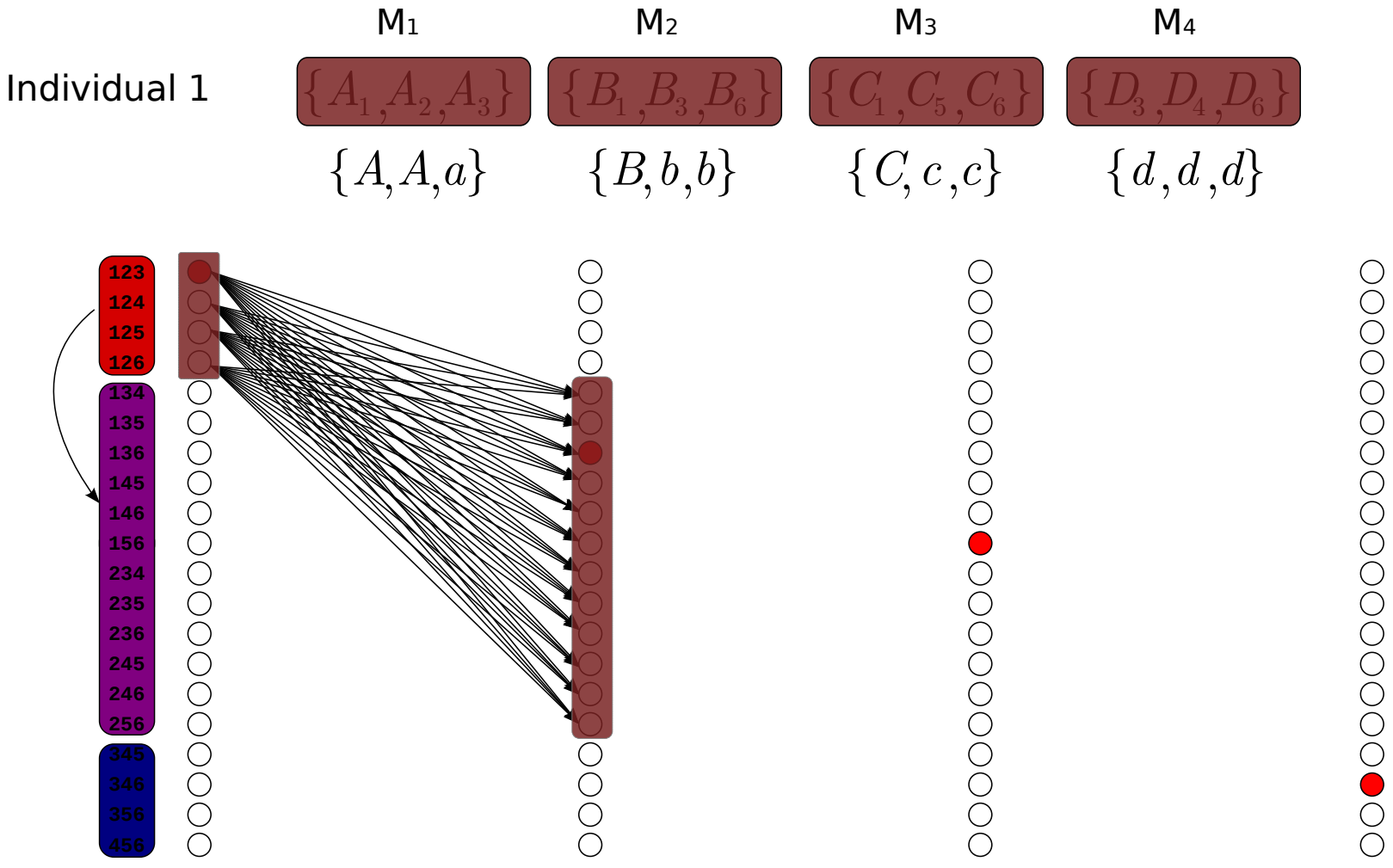
Hidden Markov model (HMM)



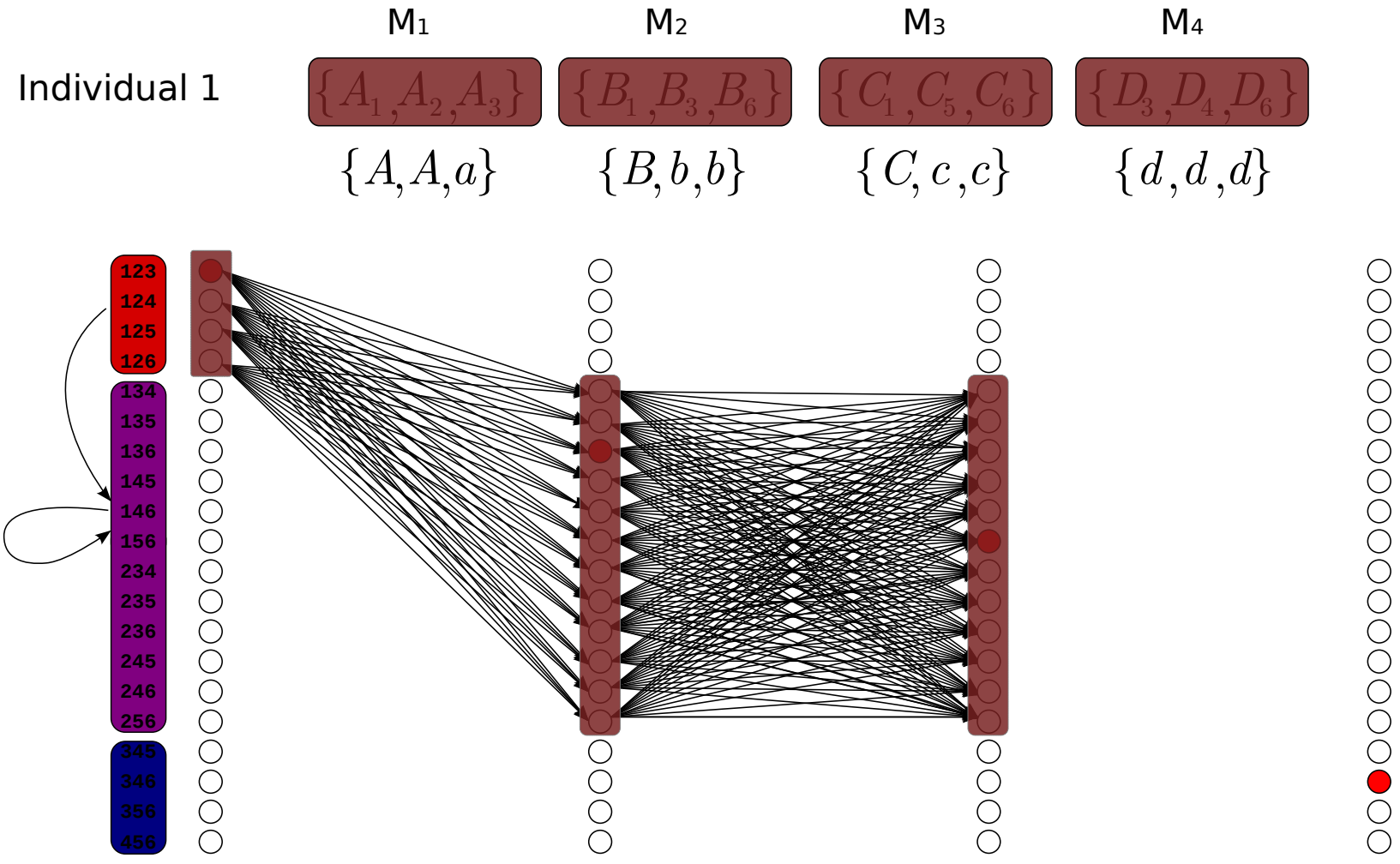
Hidden Markov model (HMM)



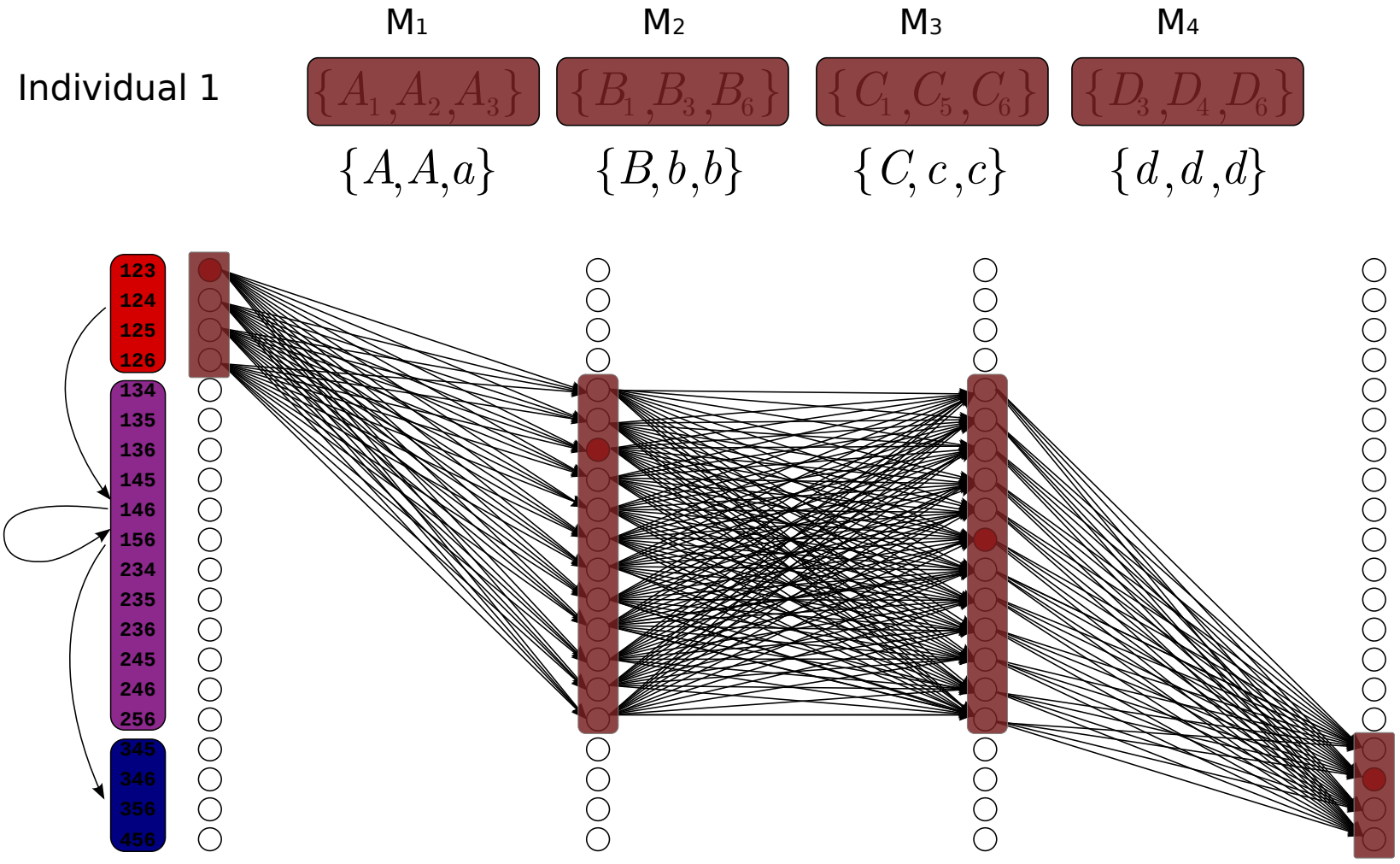
Hidden Markov model (HMM)



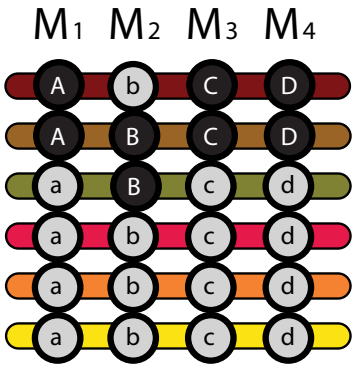
Hidden Markov model (HMM)



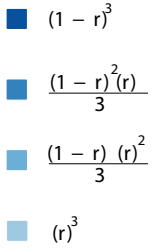
Hidden Markov model (HMM)

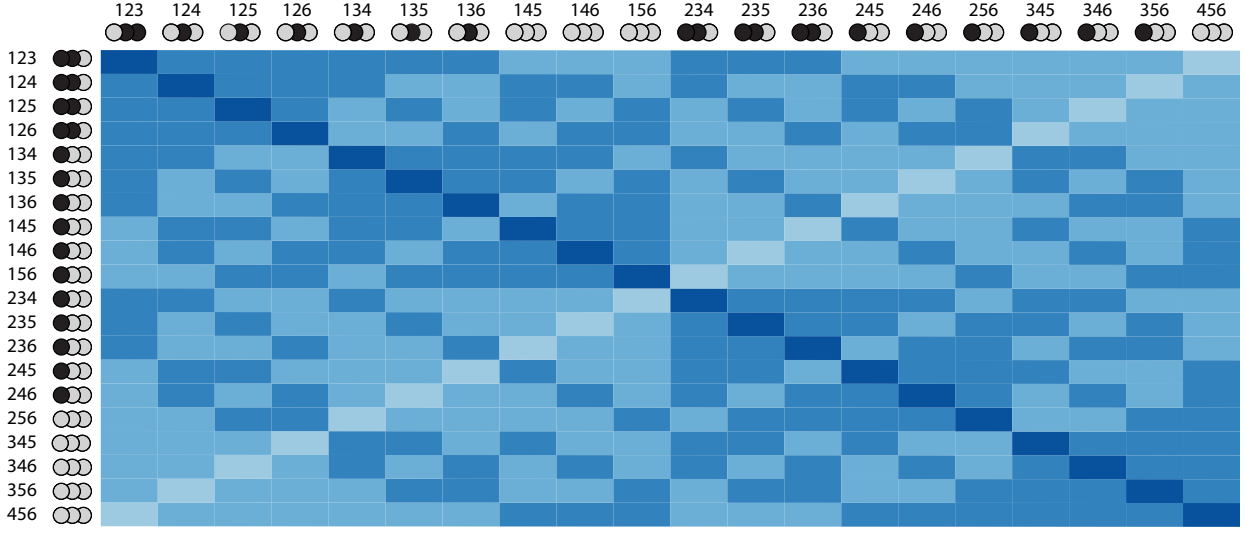


Final transition matrix for hexaploids (one parent)

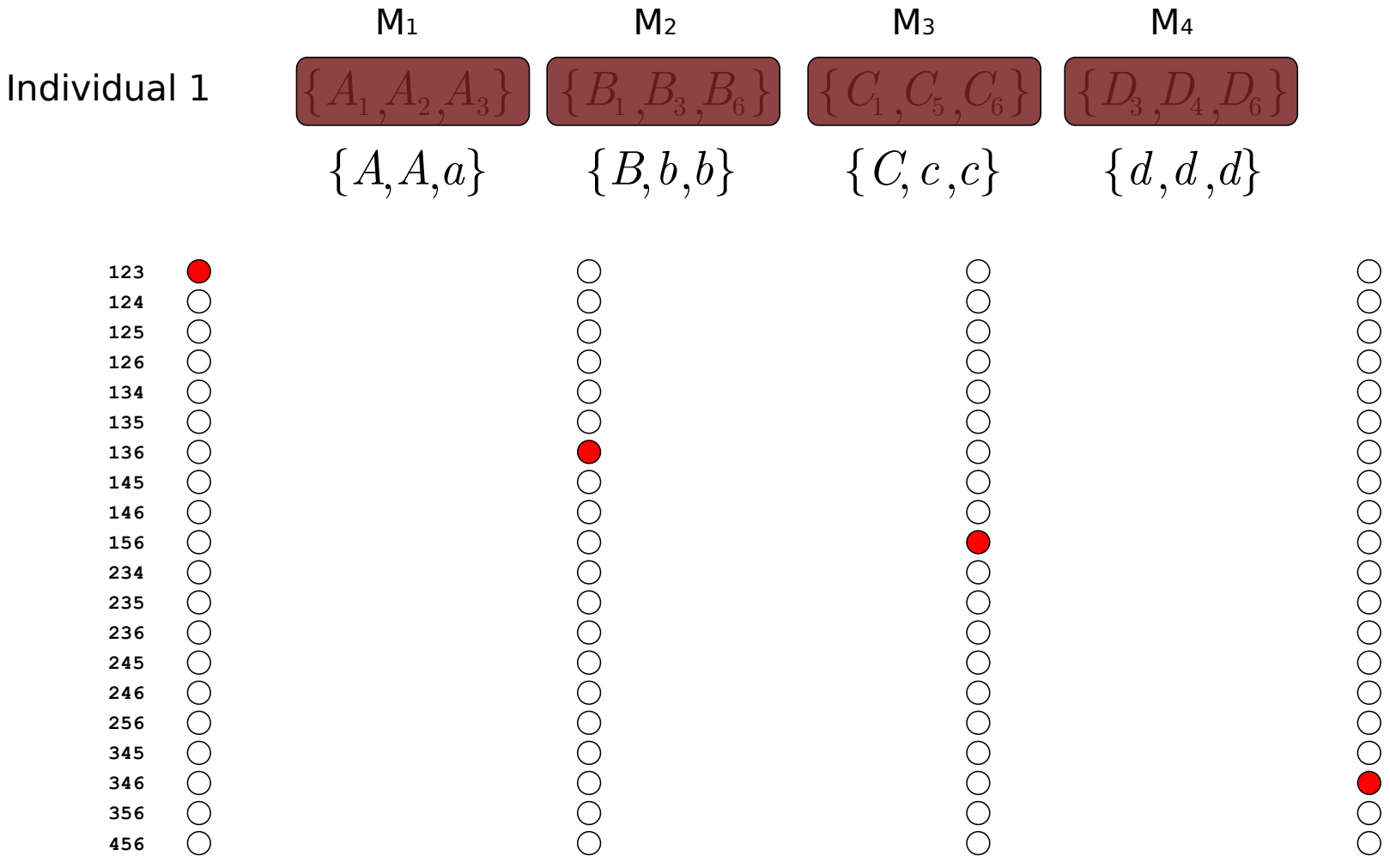


$$Pr(M_2|M_1)$$

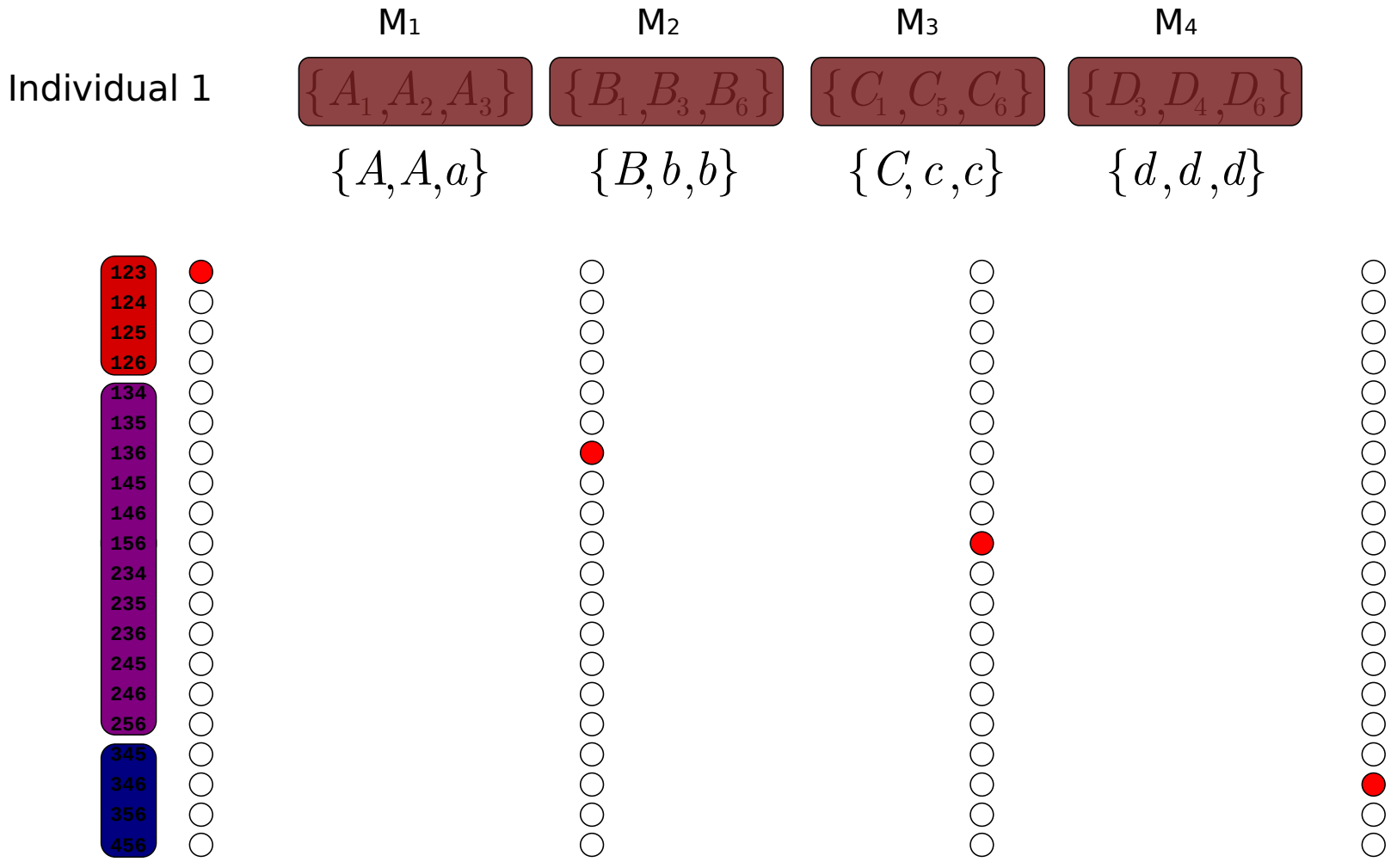




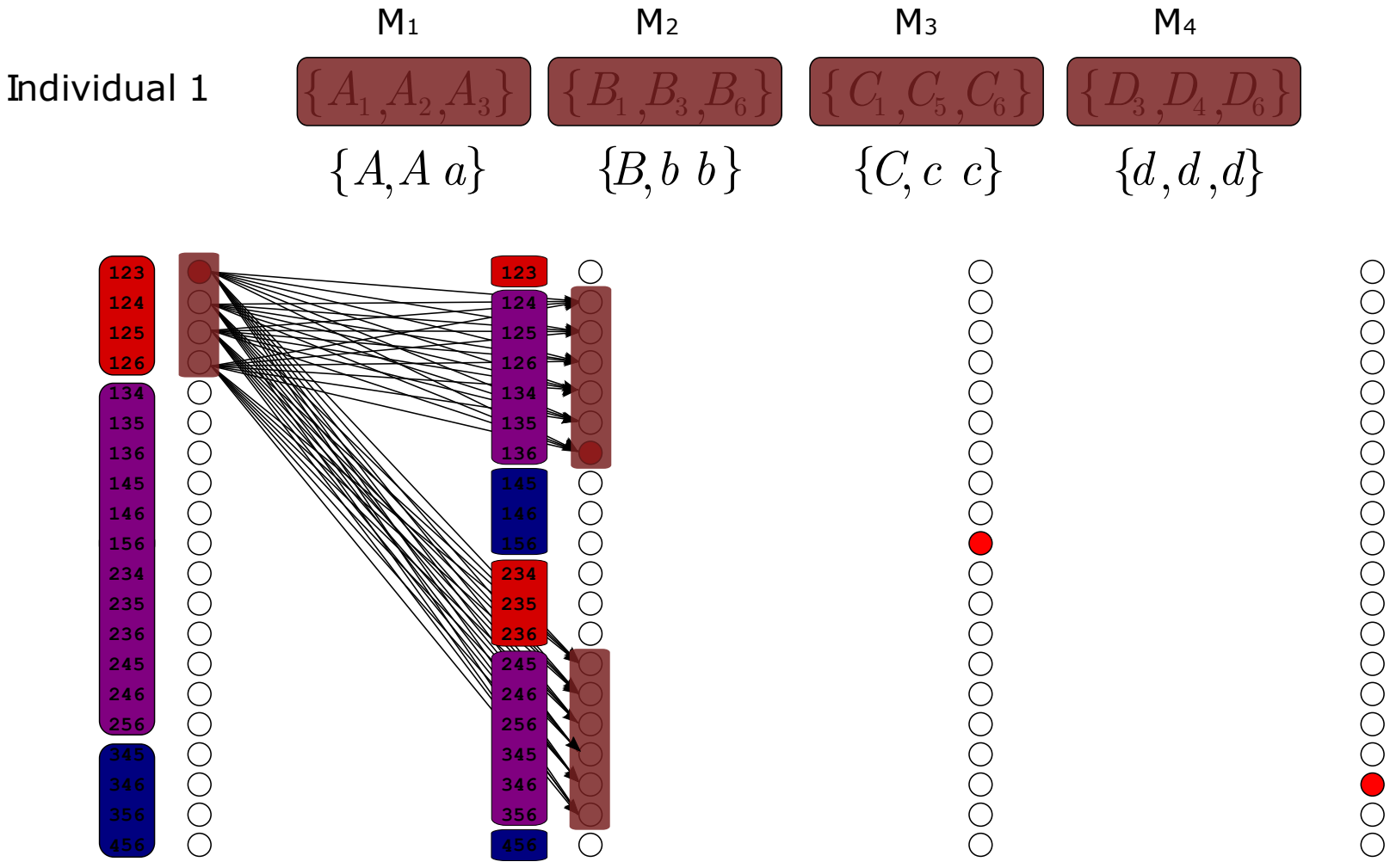
Hidden Markov model (HMM)



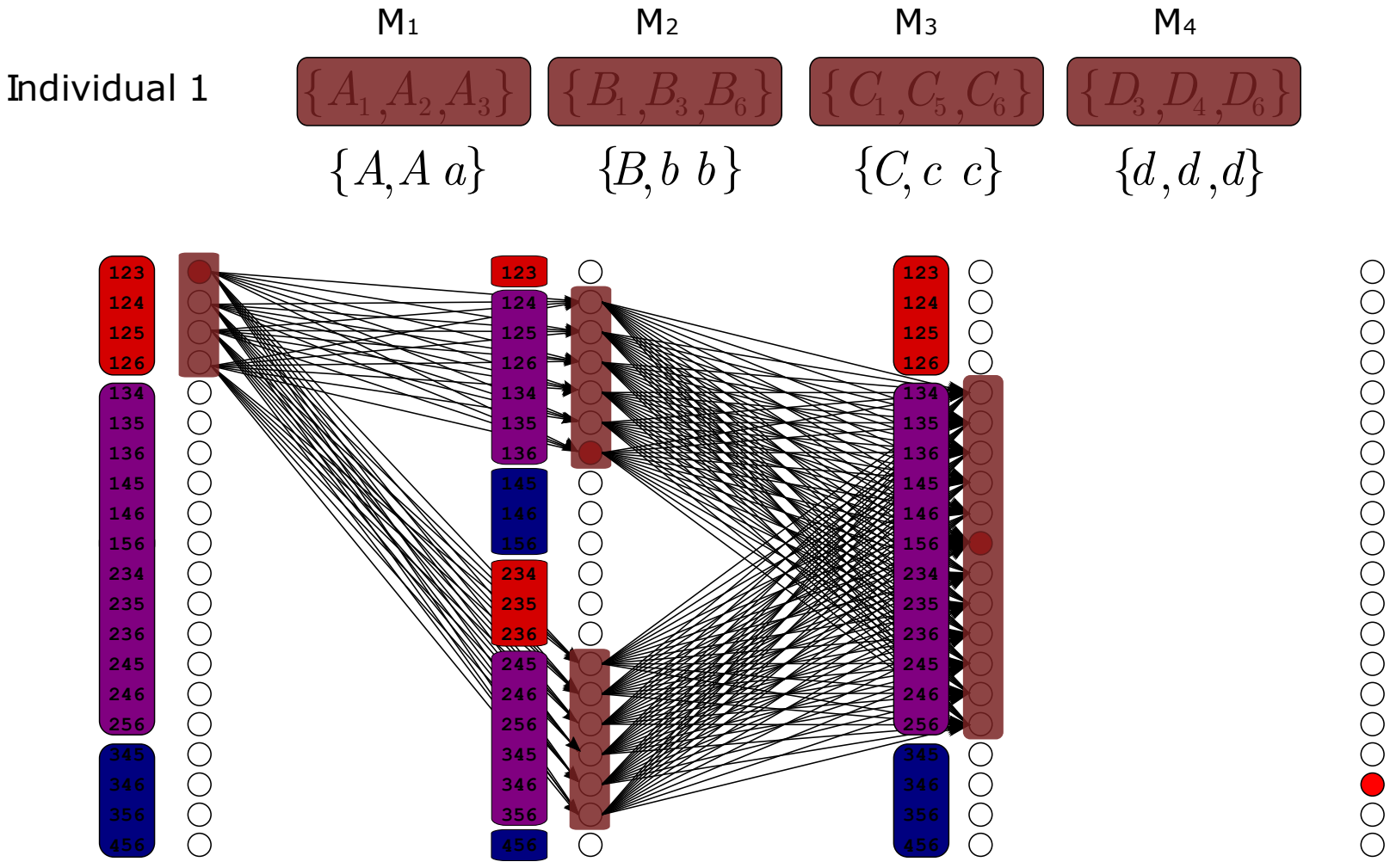
Hidden Markov model (HMM)



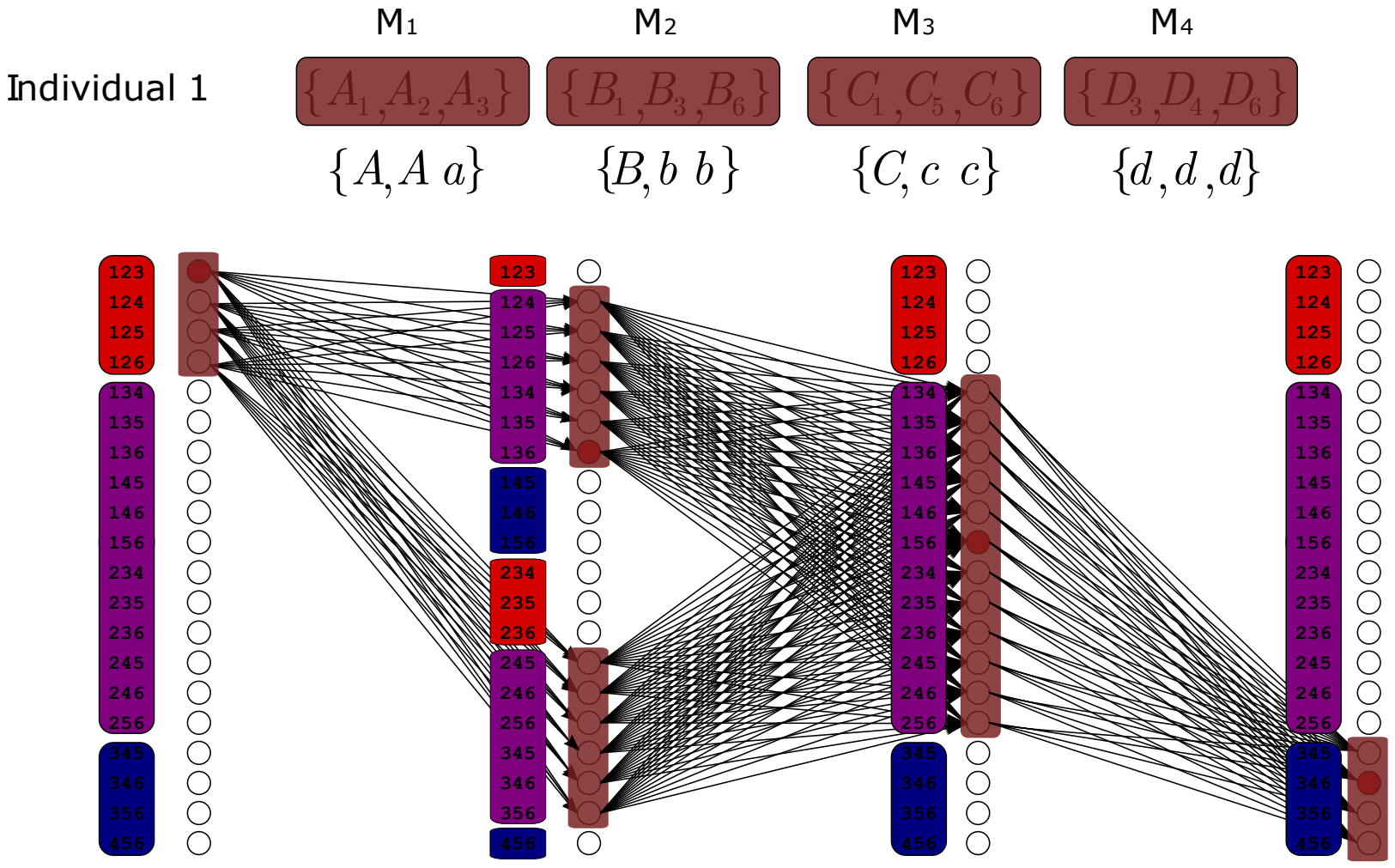
Hidden Markov model (HMM)



Hidden Markov model (HMM)



Hidden Markov model (HMM)



Likelihood - *Forward* procedure

1. Initialization:
$$\alpha_1(j) = \pi_j b_j(o_1)), \; j = 1, \cdots, g_m$$
2. Induction:
$$\alpha_{k+1}\left(j'\right) = \left[\sum_{j}^{g_m} \alpha_k\left(j\right) t_k(j, j')\right] b_{j'}(o_{k+1})$$
where $k = 1, \cdots, z-1$ and $j = 1, \cdots, g_m$
3. Termination:
$$\Pr(o_1, \cdots, o_z | \mathbf{r}, \Omega_P, \Omega_Q) = \sum_{j=1}^{g_m} \alpha_z(j)$$

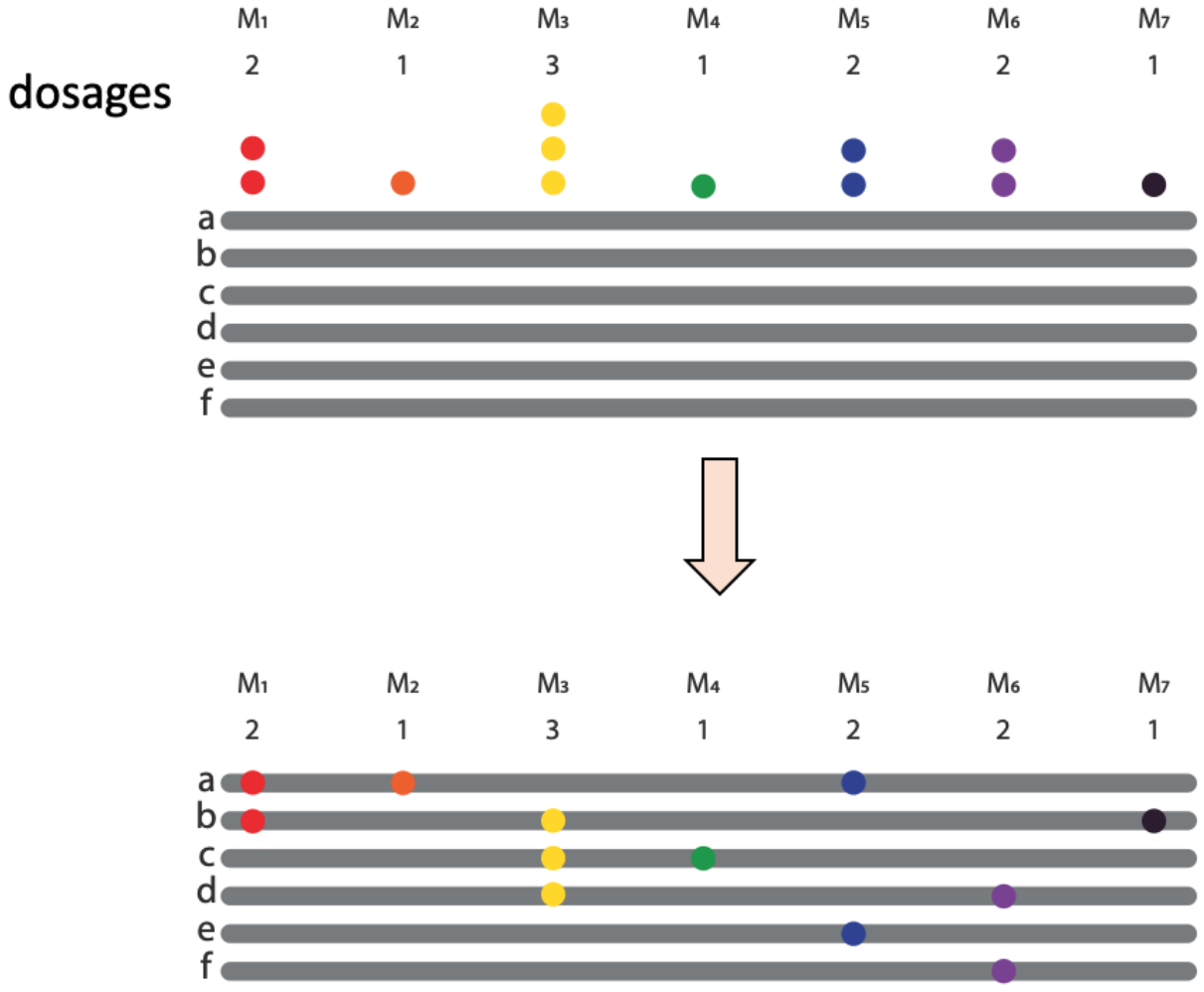
Then, the likelihood of the model can be defined as

$$L = \prod_{i=1}^{n} \Pr(o_1, \cdots, o_z | \mathbf{r}, \Omega_P, \Omega_Q)$$

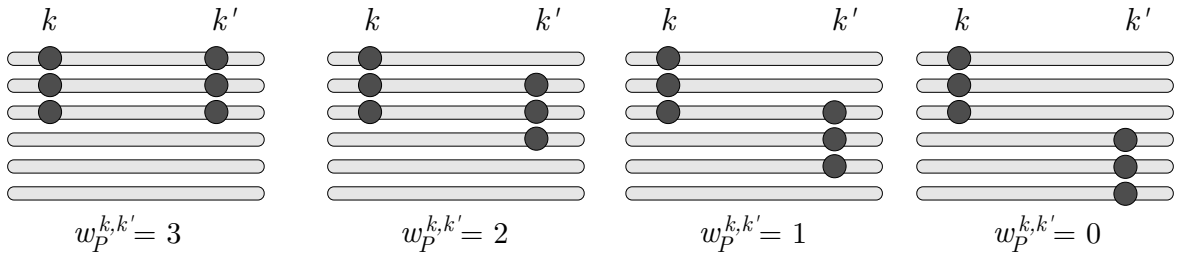
where n is the number of individuals on the full sib family.

Haplotyping of phasing

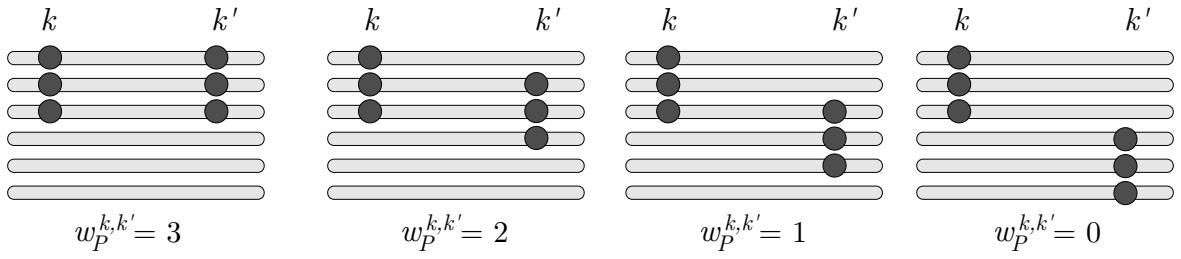
- Disposition of allelic variants in the homologs in a homology group

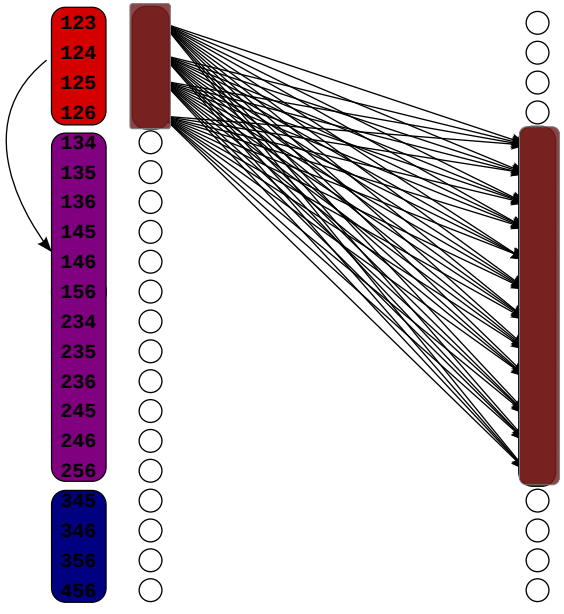


Reduction of complexity - the two-point approach

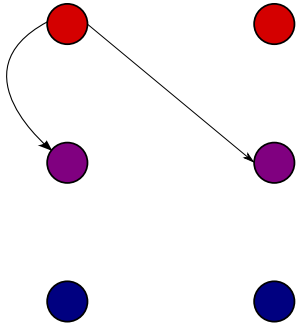


Reduction of complexity - the two-point approach

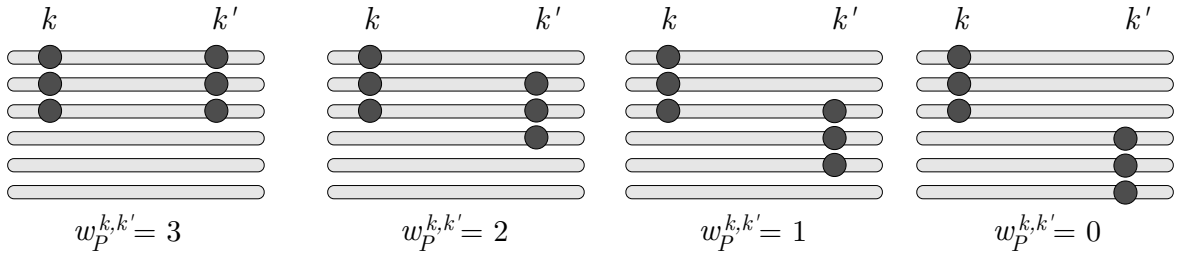




vs.

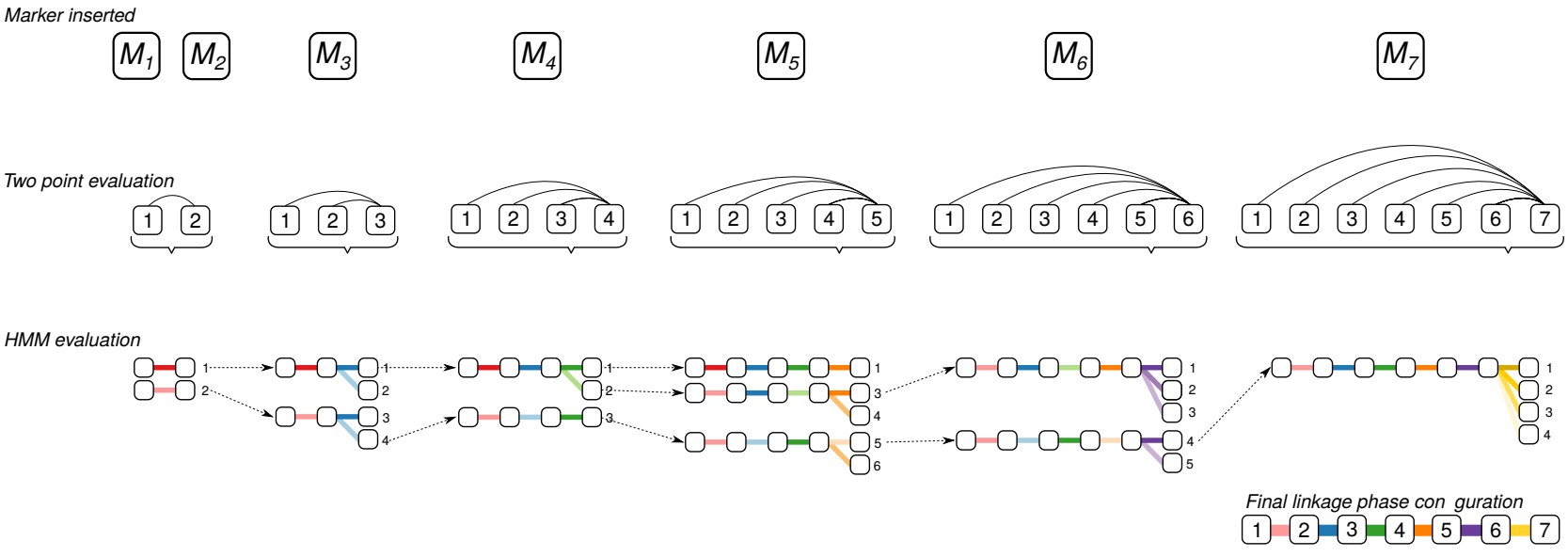


Reduction of complexity - the two-point approach

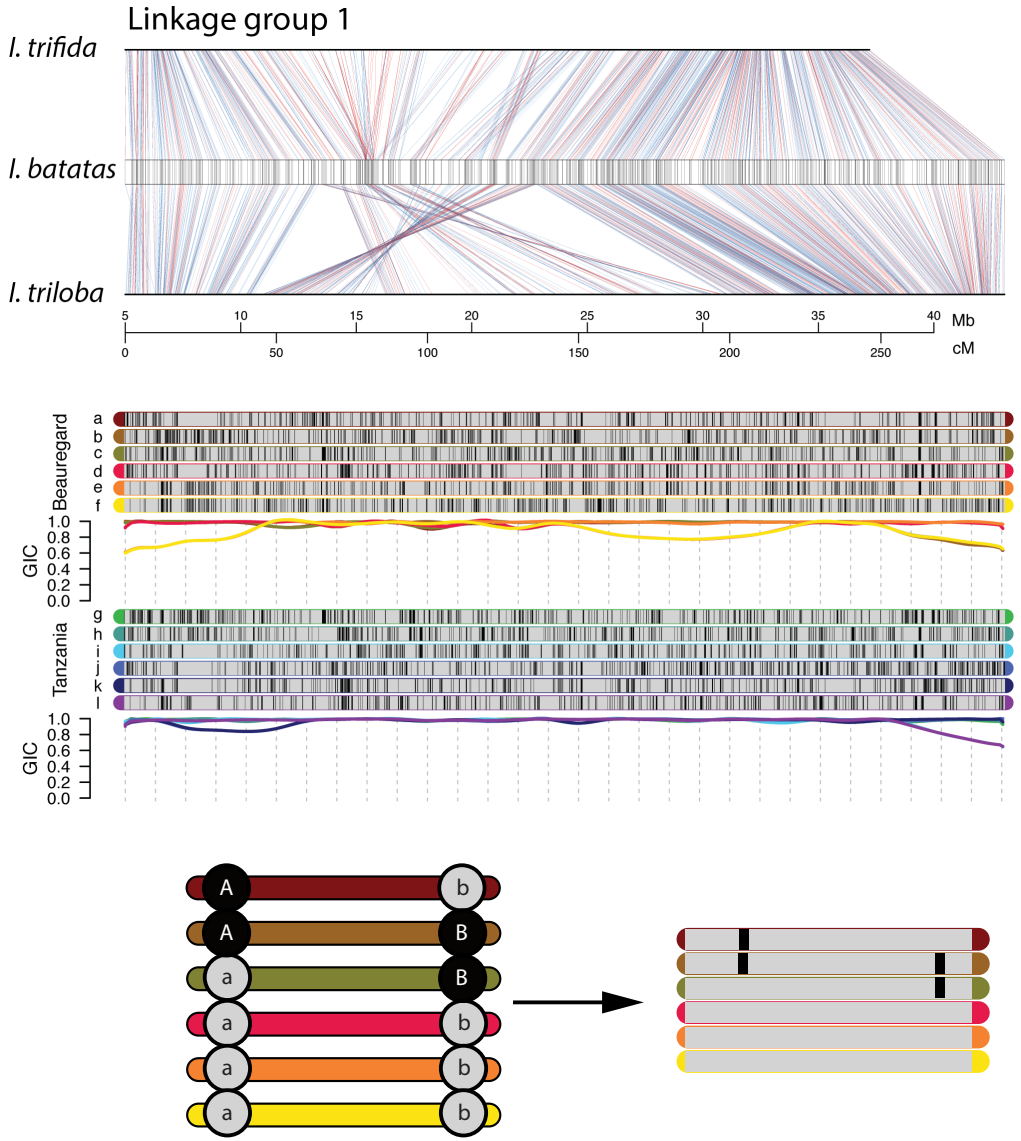


	LOD_ph	rf	LOD_rf
1	0.000000	0.07239439	6.7773070858
2	2.302351	0.09233424	4.8678664544
3	2.855213	0.36392079	4.0093145291
0	6.864653	0.49994804	0.0001254893

Hybrid phasing - two-point and HMM approaches



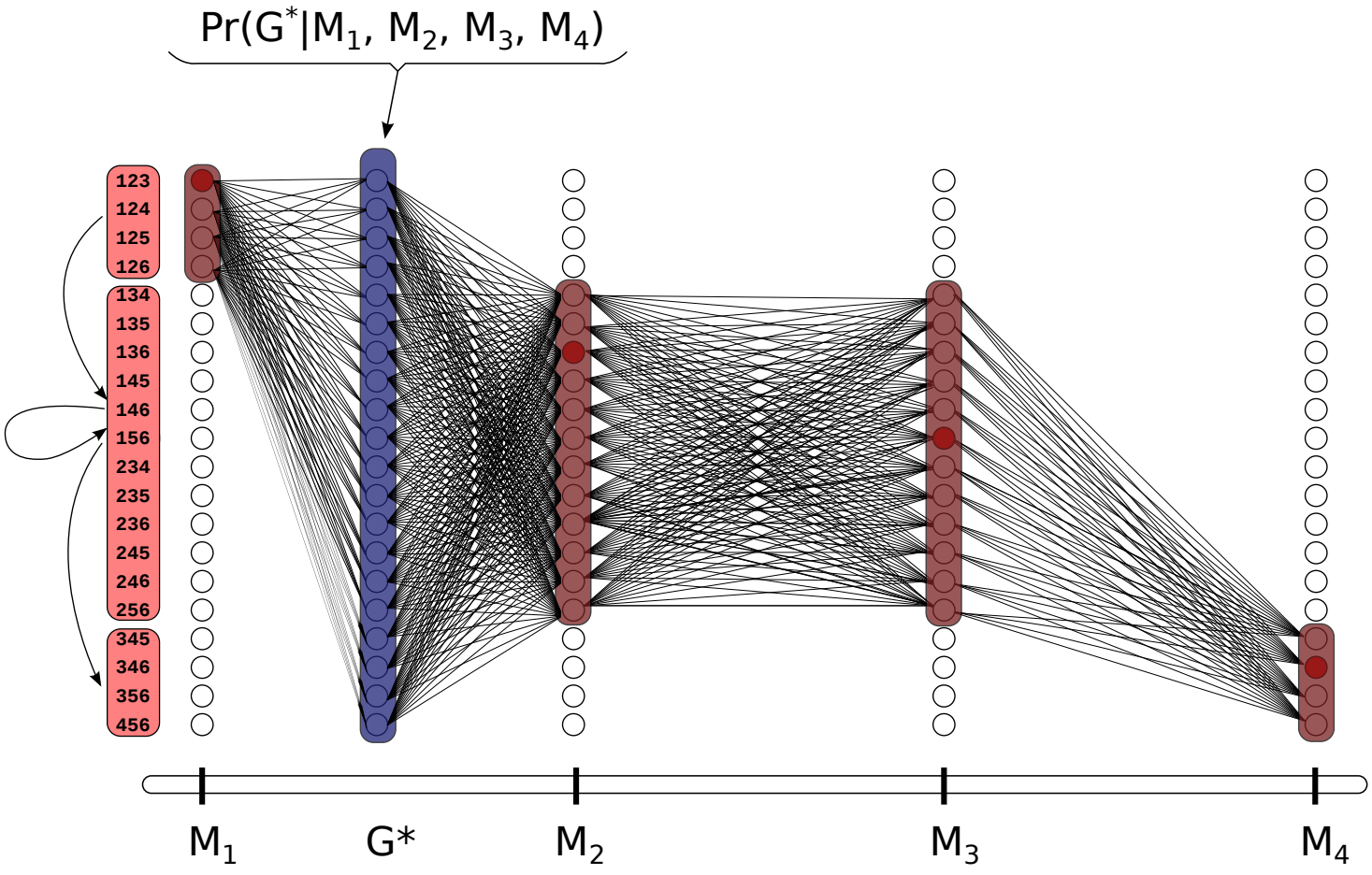
Homologous group 1 - Sweetpotato



HMM - Estimation procedure and Implementation

- Maximum likelihood estimation: Baum-Welch algorithm
- Implemented in `C++` and `R`
- Available under GPL.
- GitHub: https://github.com/mmollina/
- R package - MAPpoly

QTL conditional Probabilities and HMM



QTL conditional Probabilities and HMM