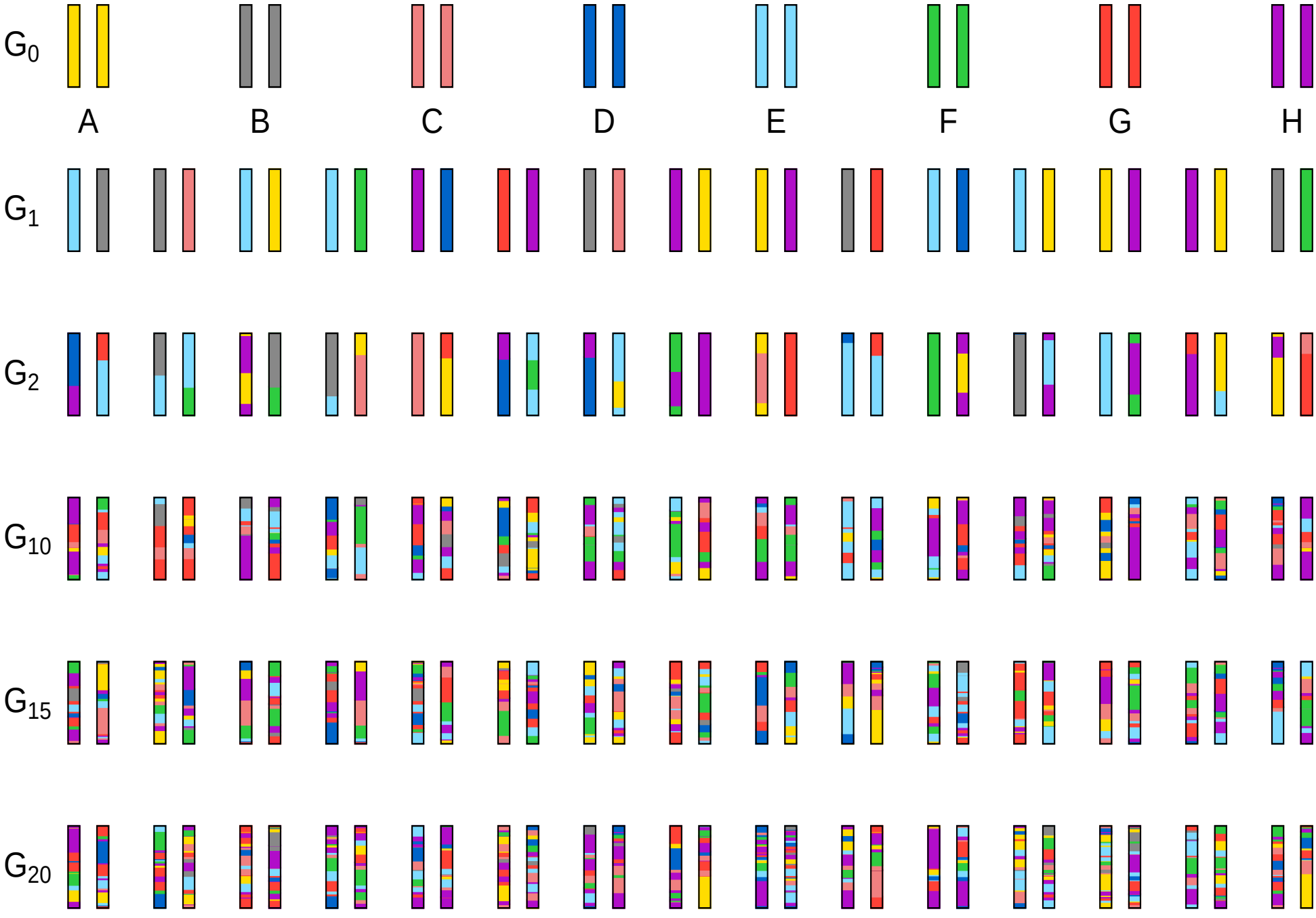
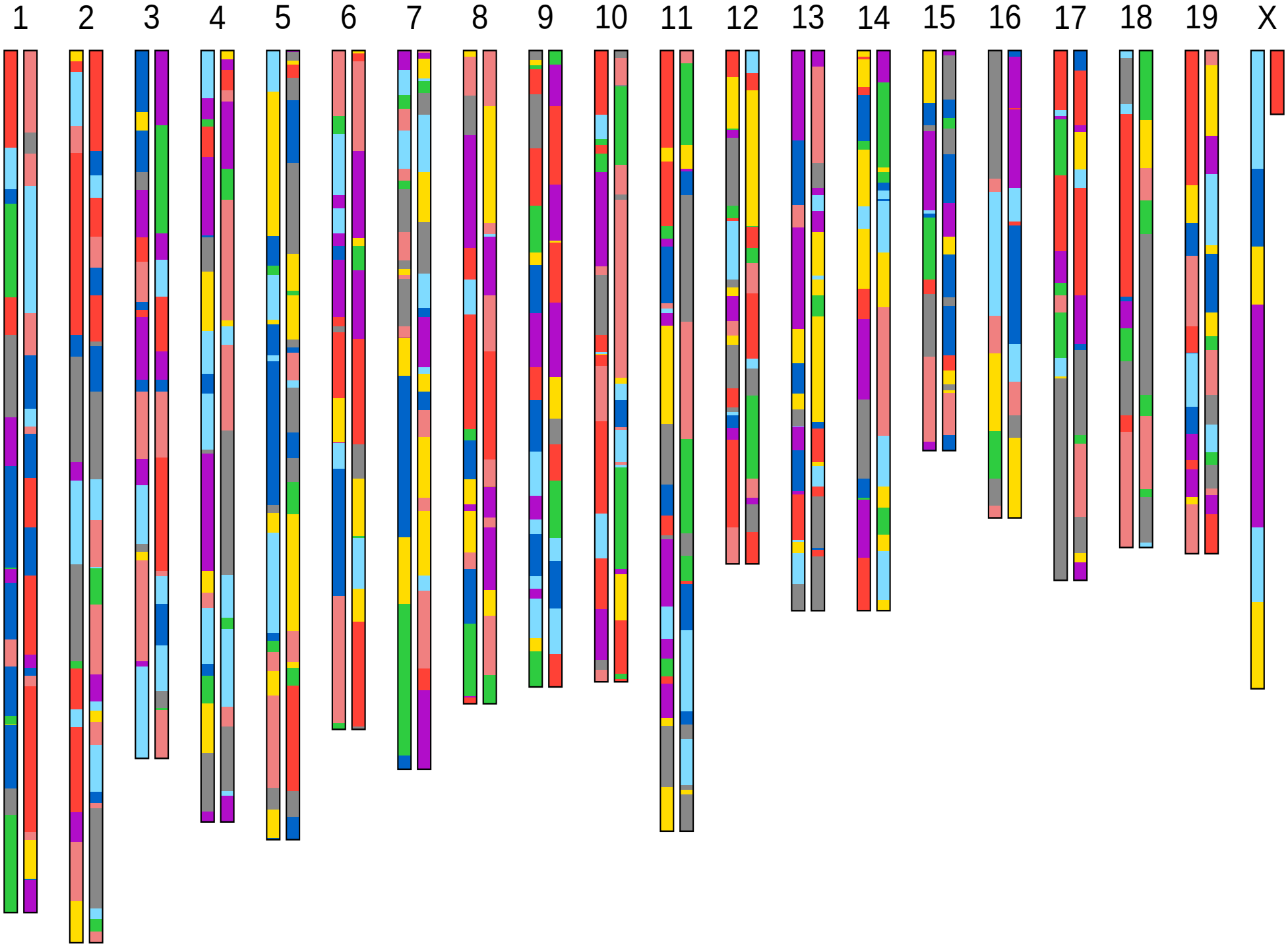
Genotype Reconstruction for Diversity Outbred Mice

A comparison of `r/qtl2` and `DOQTL`

John Spaw[1]
Karl Broman[2]

`github.com/JohnPSpaw`[1]
`github.com/kbroman`[2]


G0
A
B
C
D
E
F
G
H
G1
G2
G10
G15
G20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X

Hidden Markov Model

Hidden Markov Model

Probabilities are generated from:

Hidden Markov Model

Probabilities are generated from:

1. Transition Model

Hidden Markov Model

Probabilities are generated from:

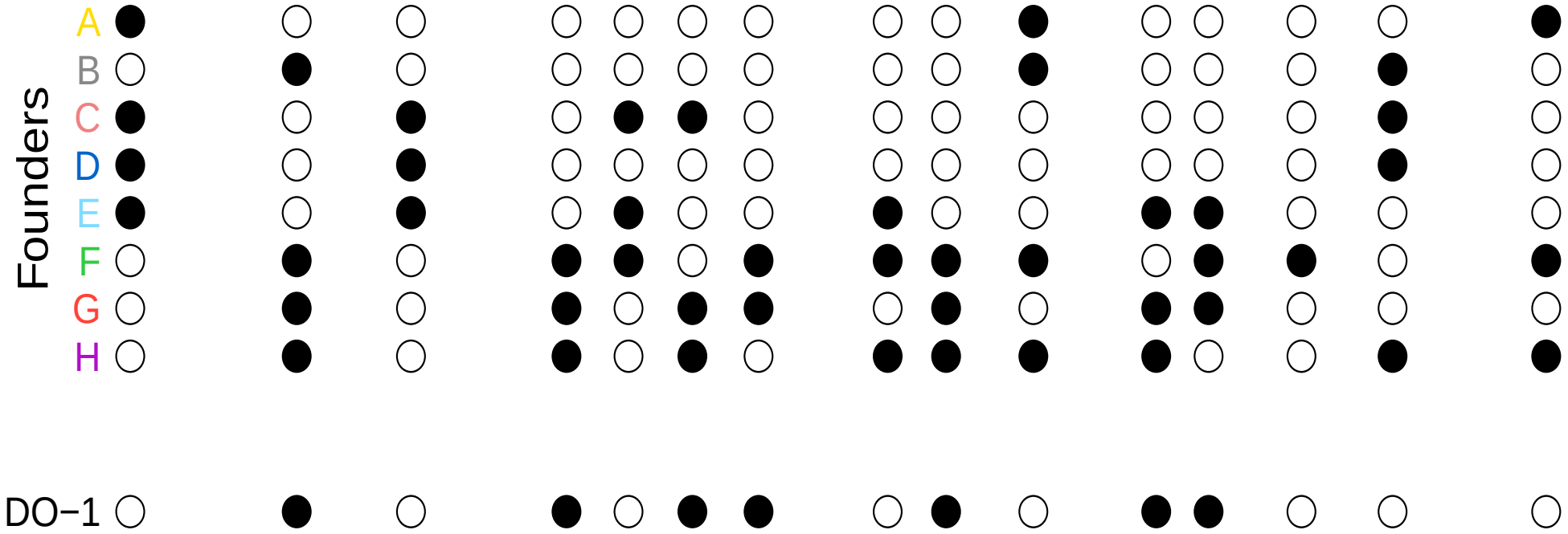
1. Transition Model
2. **Emission Model**

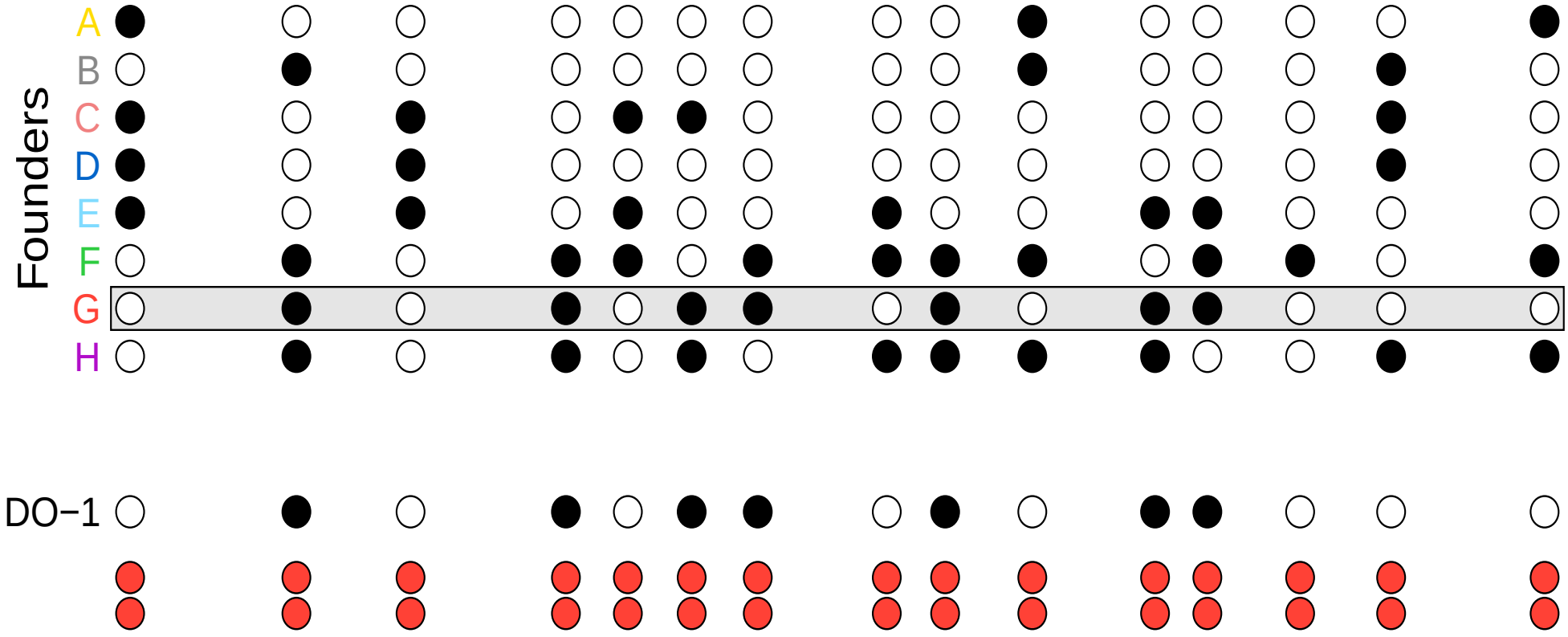
Hidden Markov Model

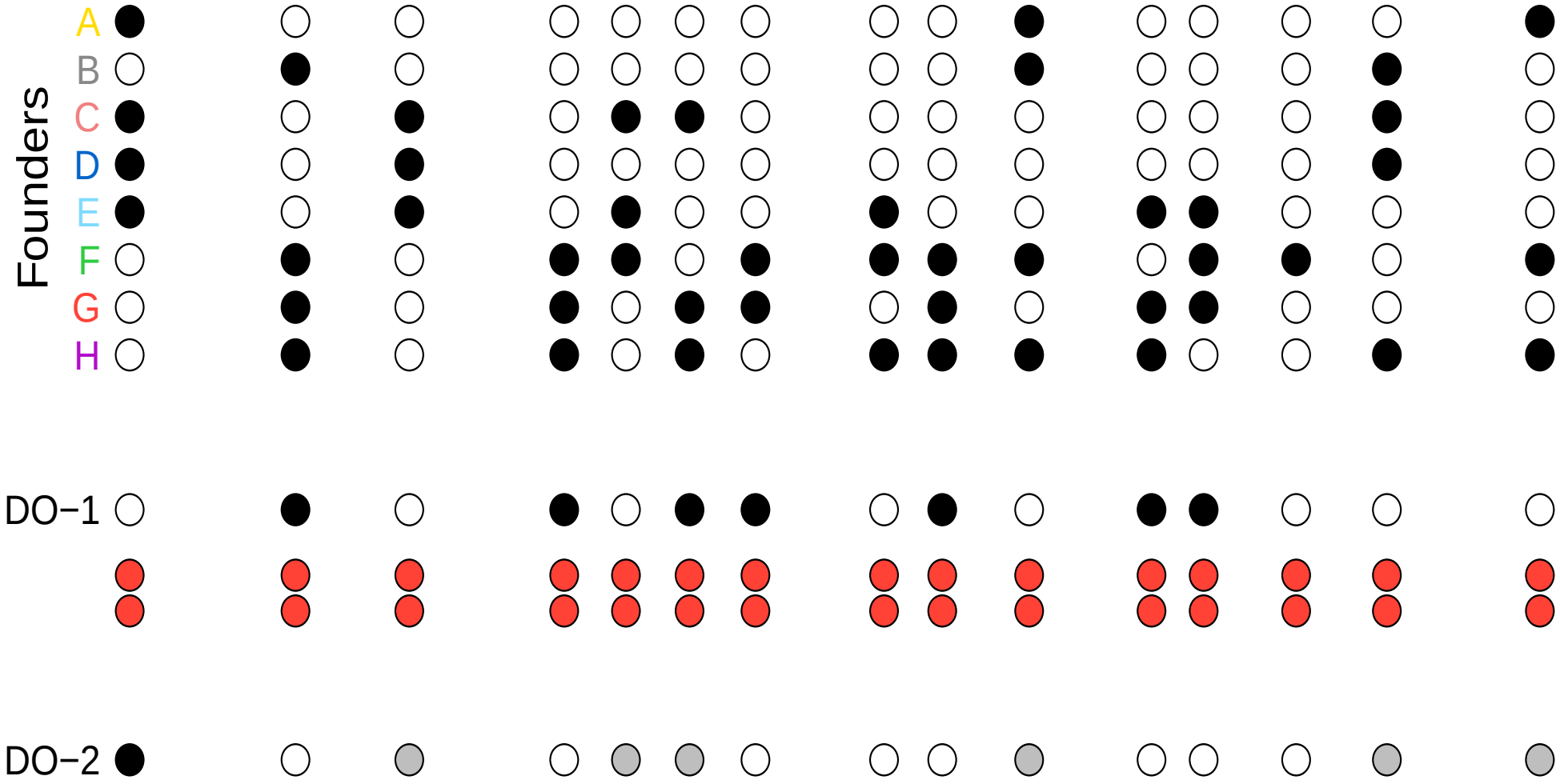
Probabilities are generated from:

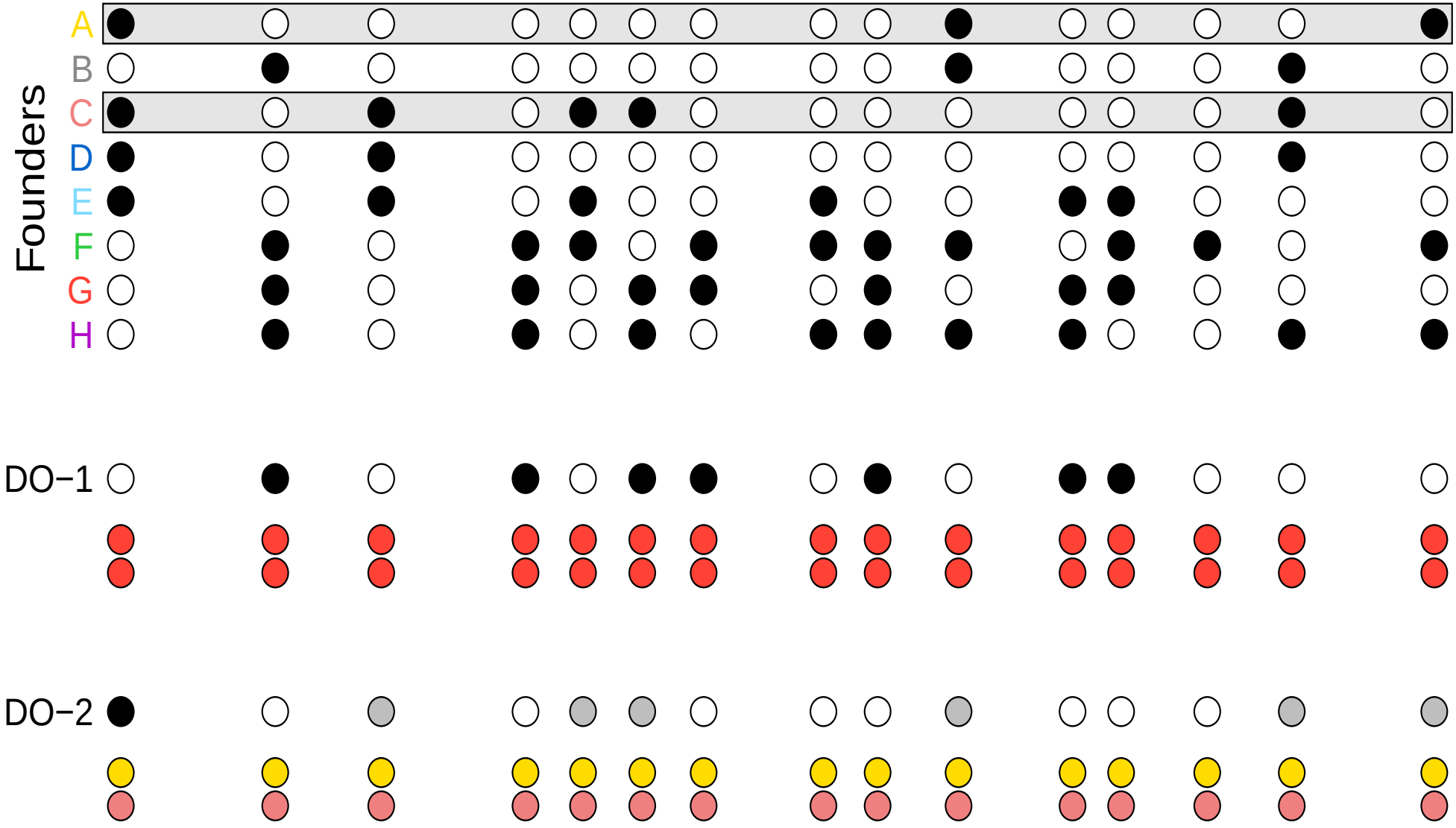
1. Transition Model
2. **Emission Model**

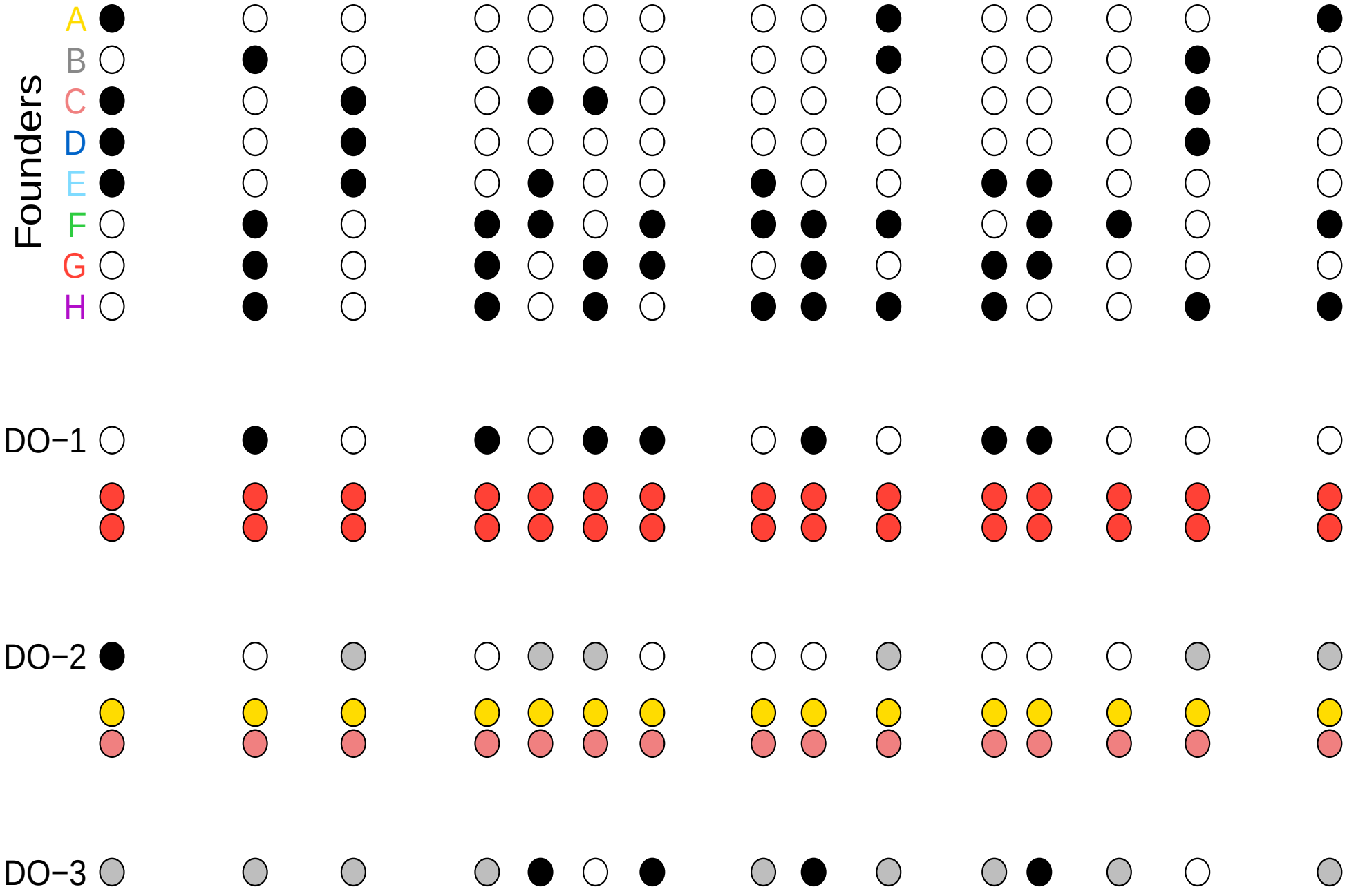
Conditional probability of observed data given underlying diplotype state

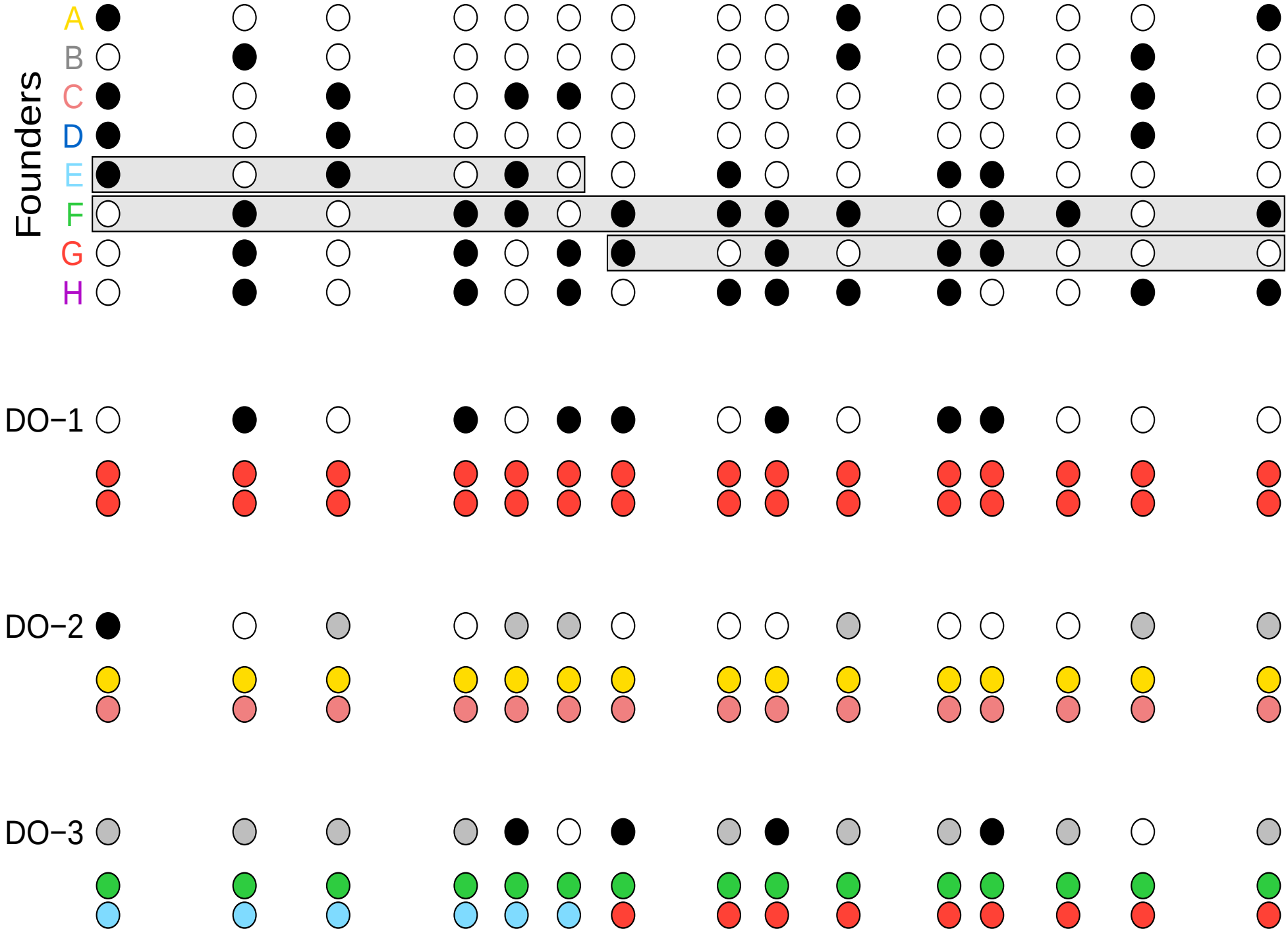

Founders
A
B
C
D
E
F
G
H
DO−1


Founders
A
B
C
D
E
F
G
H
DO−1

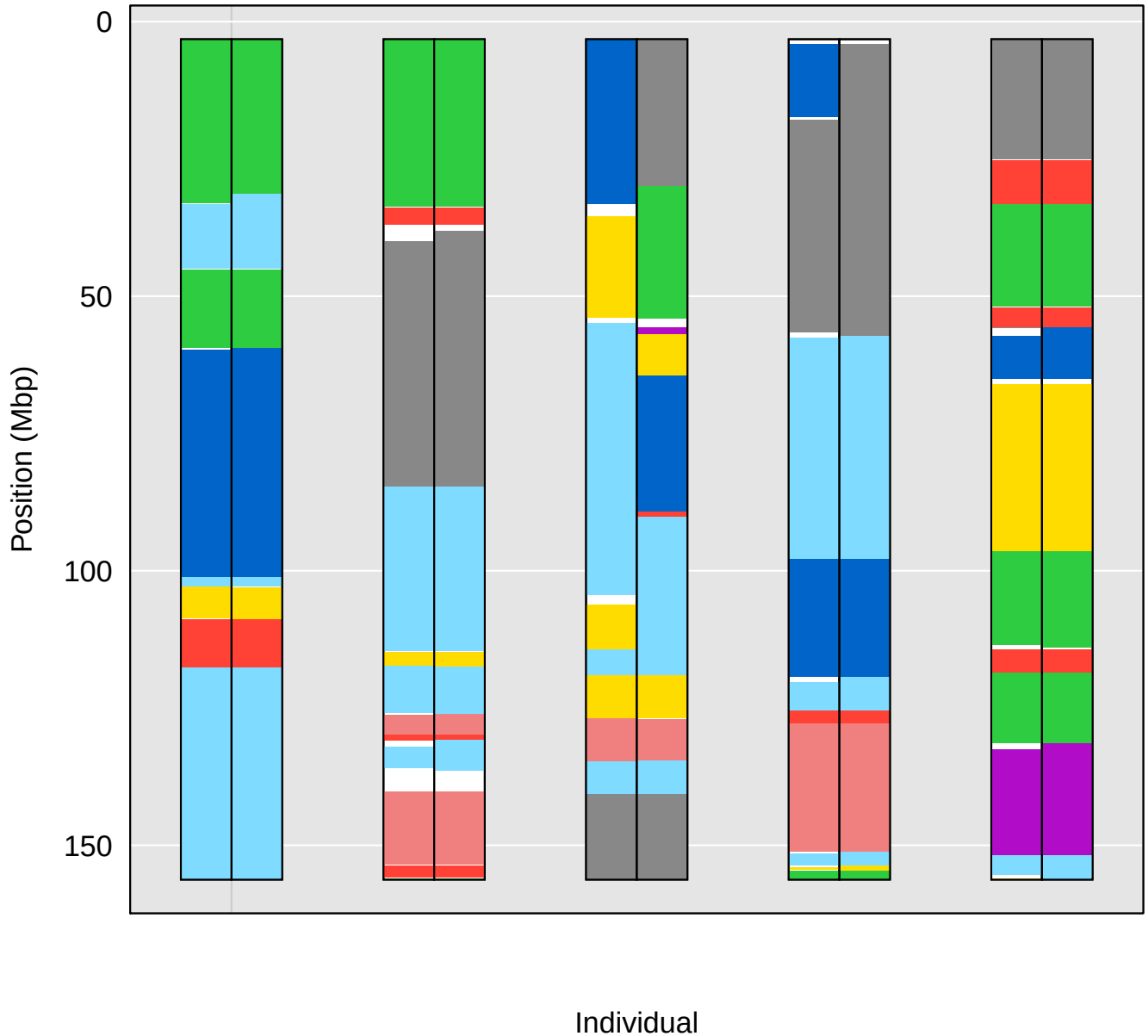

Founders
A
B
C
D
E
F
G
H
DO−1
DO−2


Founders
A
B
C
D
E
F
G
H
DO−1
DO−2


Founders
A
B
C
D
E
F
G
H
DO−1
DO−2
DO−3


Founders
A
B
C
D
E
F
G
H
DO−1
DO−2
DO−3

Inferred Haplotypes



Data

Two *large* 3D arrays of emission probabilities

- `r\qtl2` (Broman)
- `DOQTL` (Gatti)

Data

Two *large* 3D arrays of emission probabilities

- `r\qtl2` (Broman)
- `DOQTL` (Gatti)

$500 \times 120,000 \times 36$ (Individual $\times$ Markers $\times$ Diplotype)

Data

Two *large* 3D arrays of emission probabilities

- `r\qtl2` (Broman)
- `DOQTL` (Gatti)

$500 \times 120,000 \times 36$ (Individual $\times$ Markers $\times$ Diplotype)

How are they different?

Data

Two *large* 3D arrays of emission probabilities

- `r\qtl2` (Broman)
- `DOQTL` (Gatti)

$500 \times 120,000 \times 36$ (Individual $\times$ Markers $\times$ Diplotype)

How are they different?
How can we visualize this?

Measure of distance

Measure of distance

For each individual at each marker:

Compute *sum of absolute differences*

$$\sum_{i=1}^{36} |p_{1,i} - p_{2,i}|$$

Measure of distance

For each individual at each marker:

Compute *sum of absolute differences*

$$\sum_{i=1}^{36} |p_{1,i} - p_{2,i}|$$

Each entry represents distance between methods

Measure of distance

For each individual at each marker:

Compute *sum of absolute differences*

$$\sum_{i=1}^{36} |p_{1,i} - p_{2,i}|$$

Each entry represents distance between methods

Reduces problems to two-dimensions:

Measure of distance

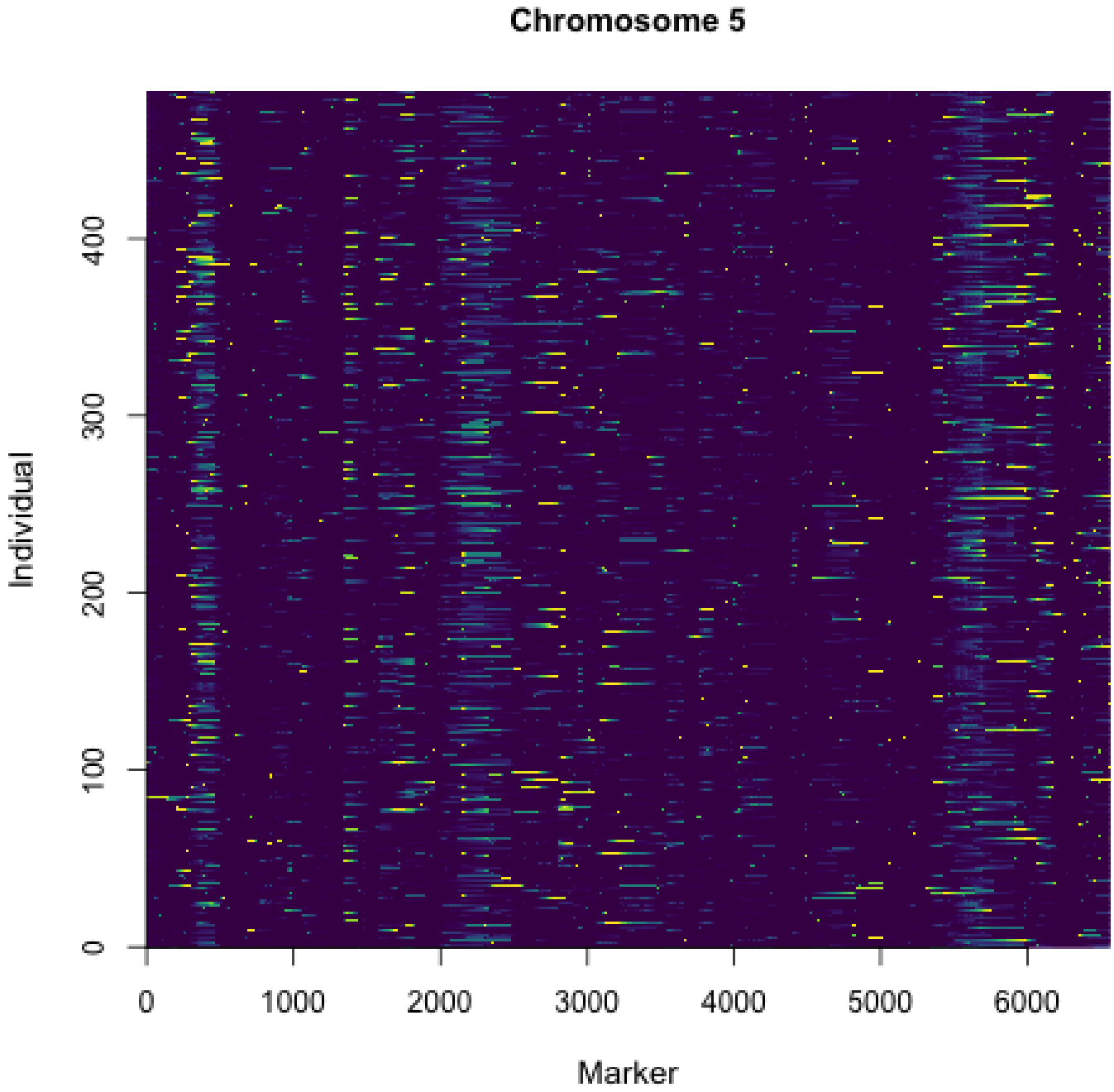
For each individual at each marker:

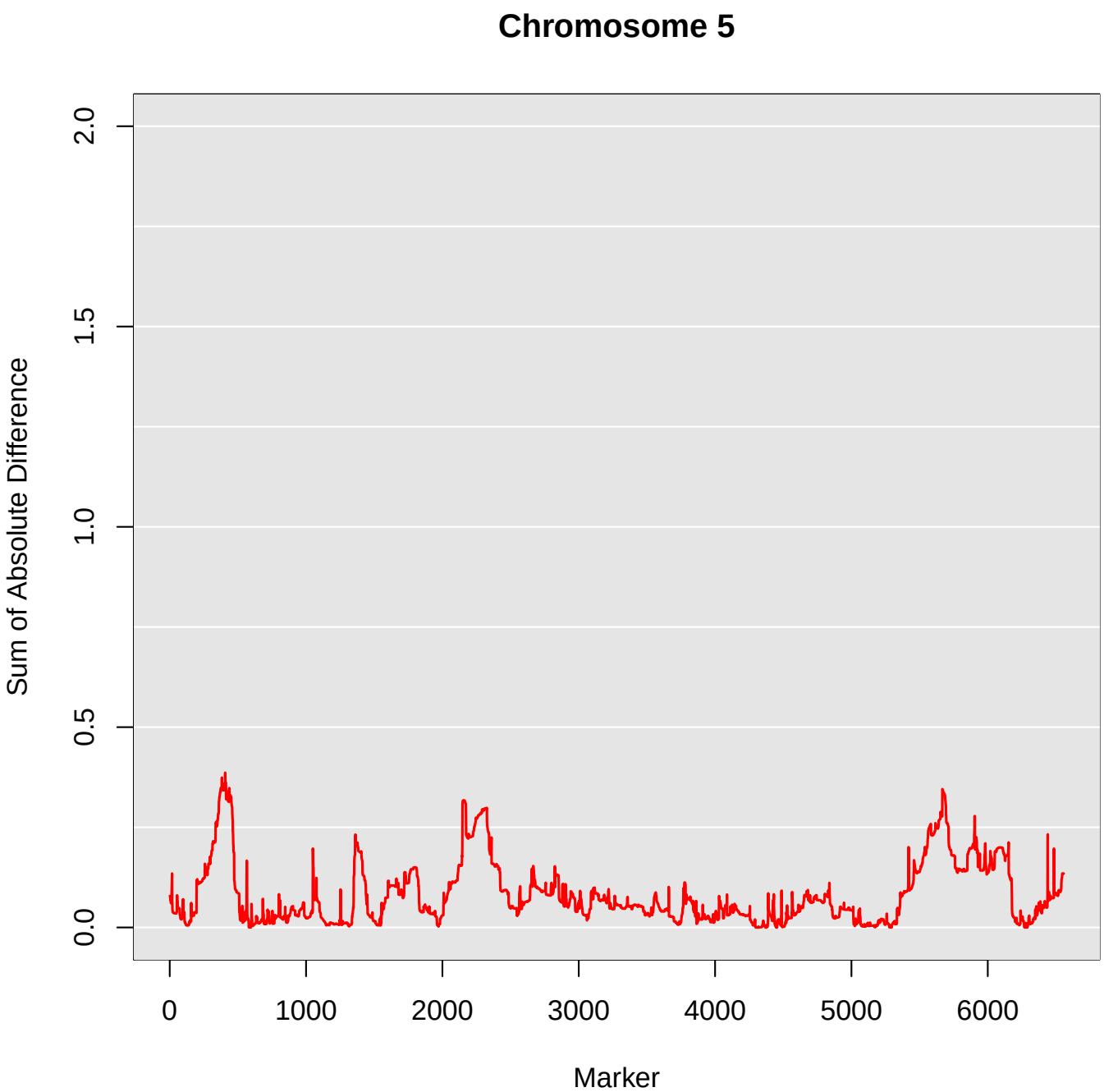
Compute *sum of absolute differences*

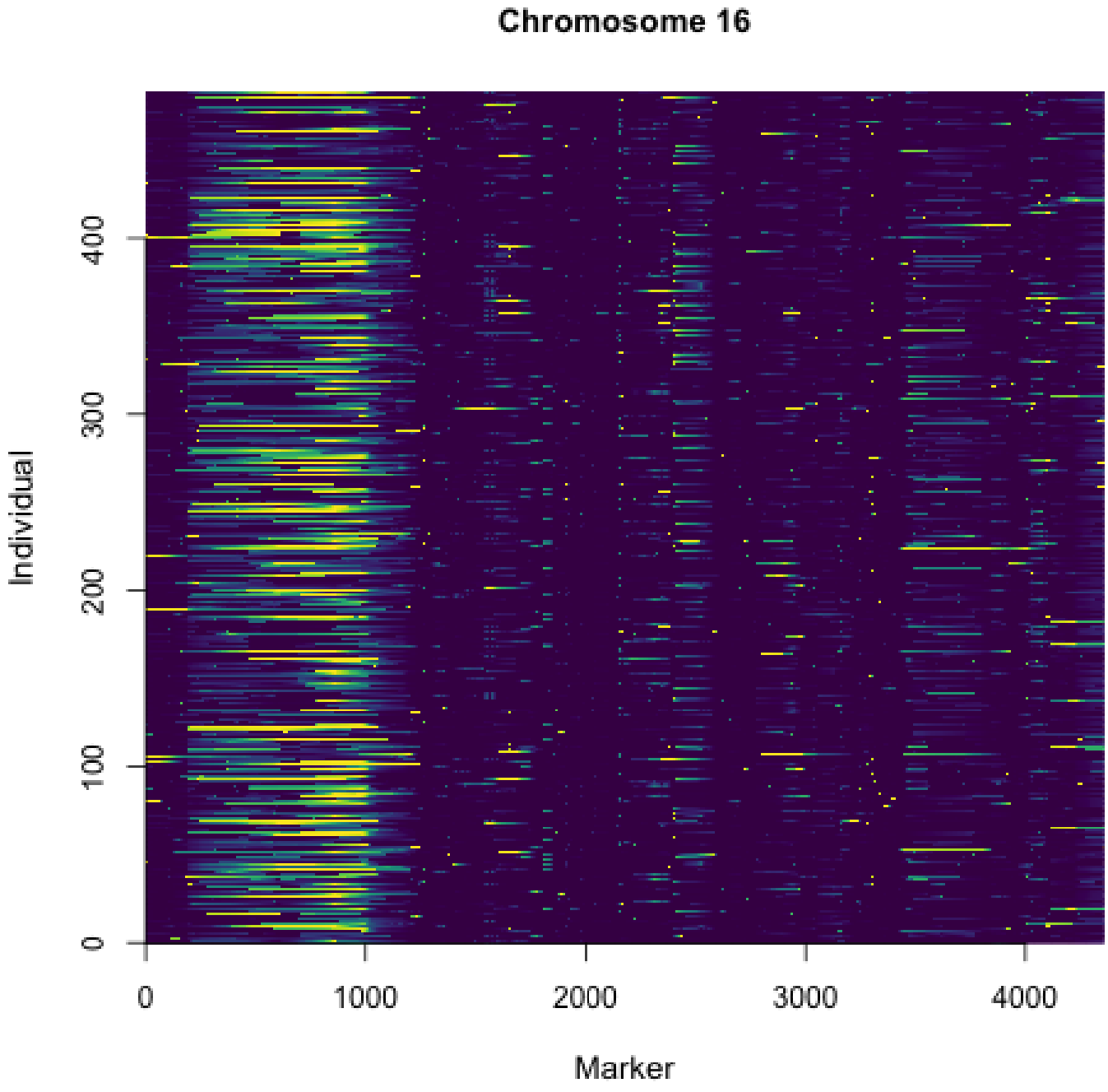
$$\sum_{i=1}^{36} |p_{1,i} - p_{2,i}|$$

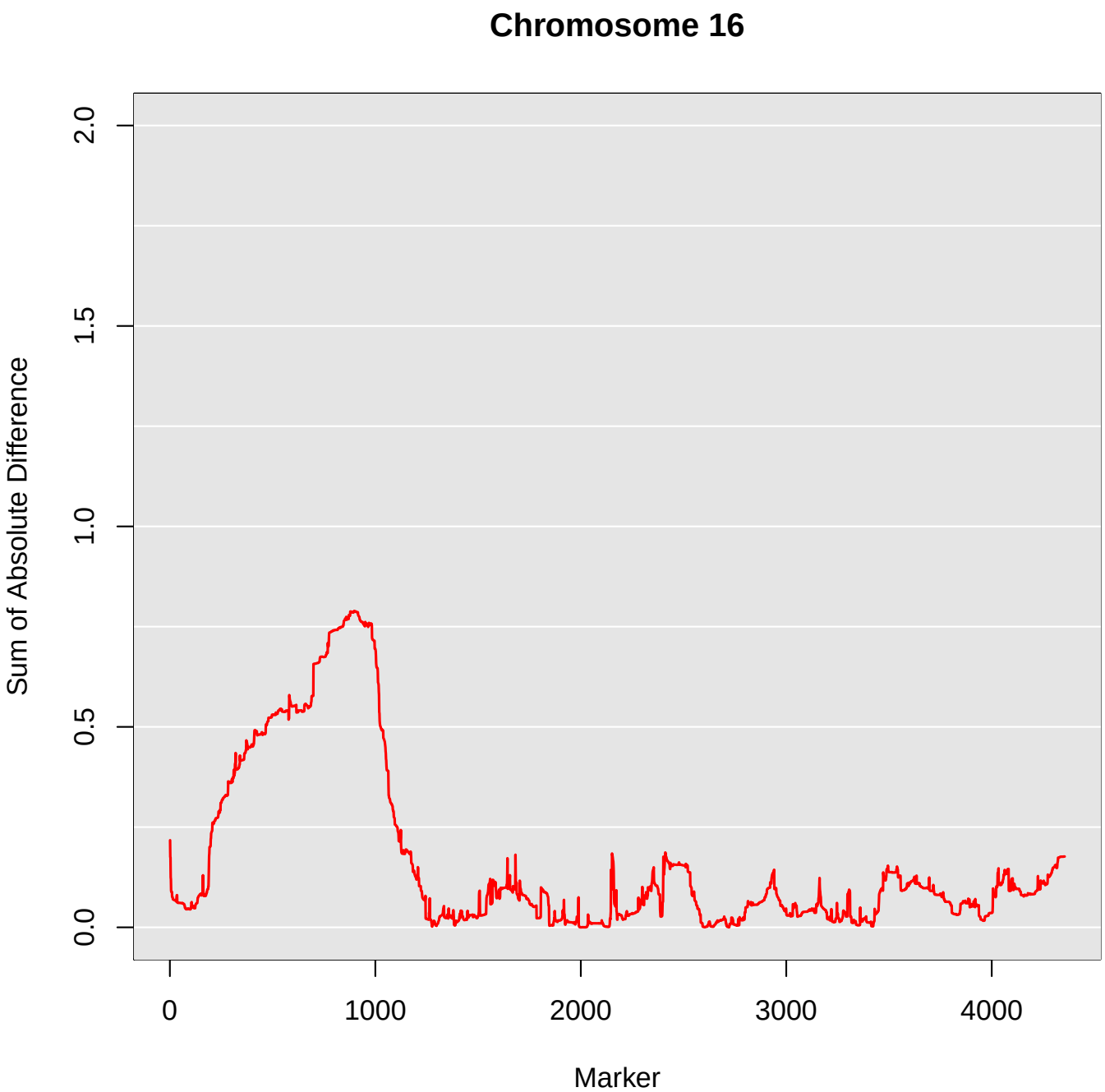
Each entry represents distance between methods

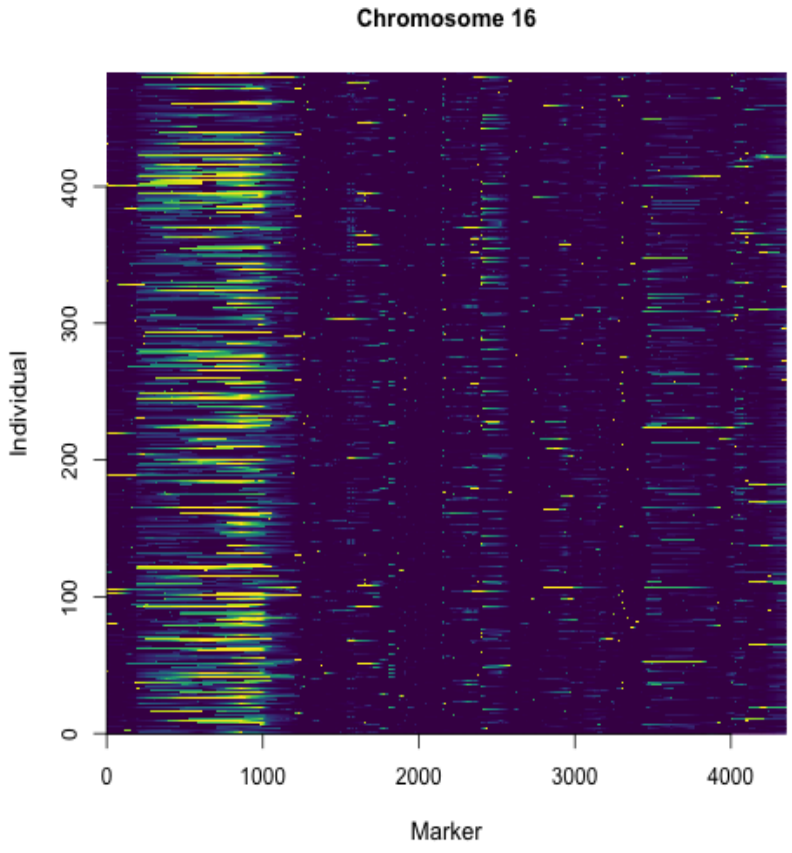
Reduces problems to two-dimensions: $500 \times 120,000$

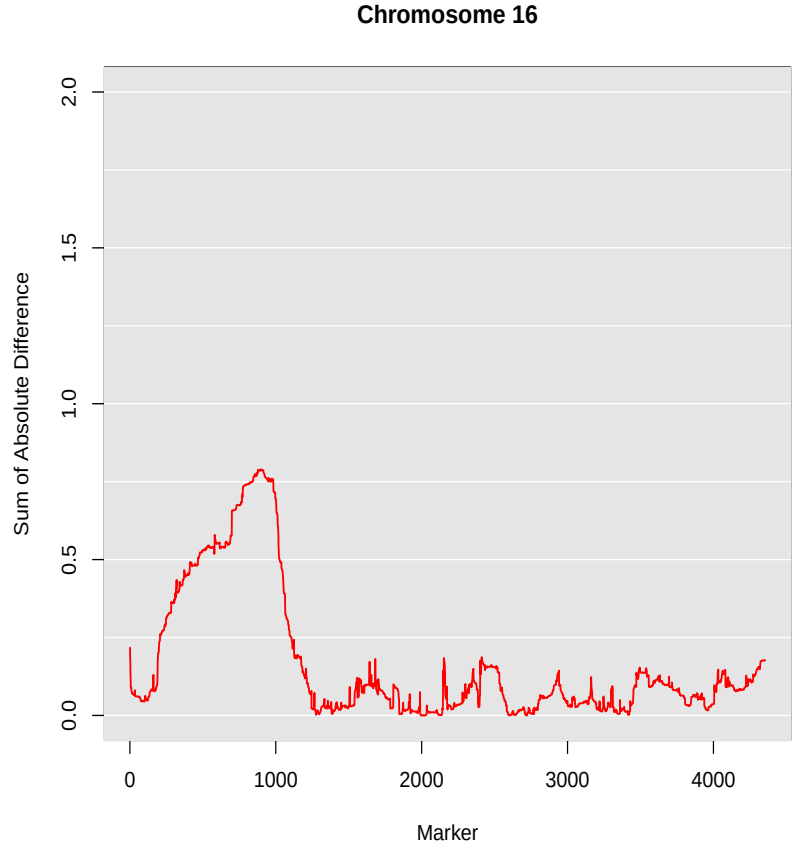

Chromosome 5
Individual
0
100
200
300
400
0
1000
2000
3000
4000
5000
6000
Marker

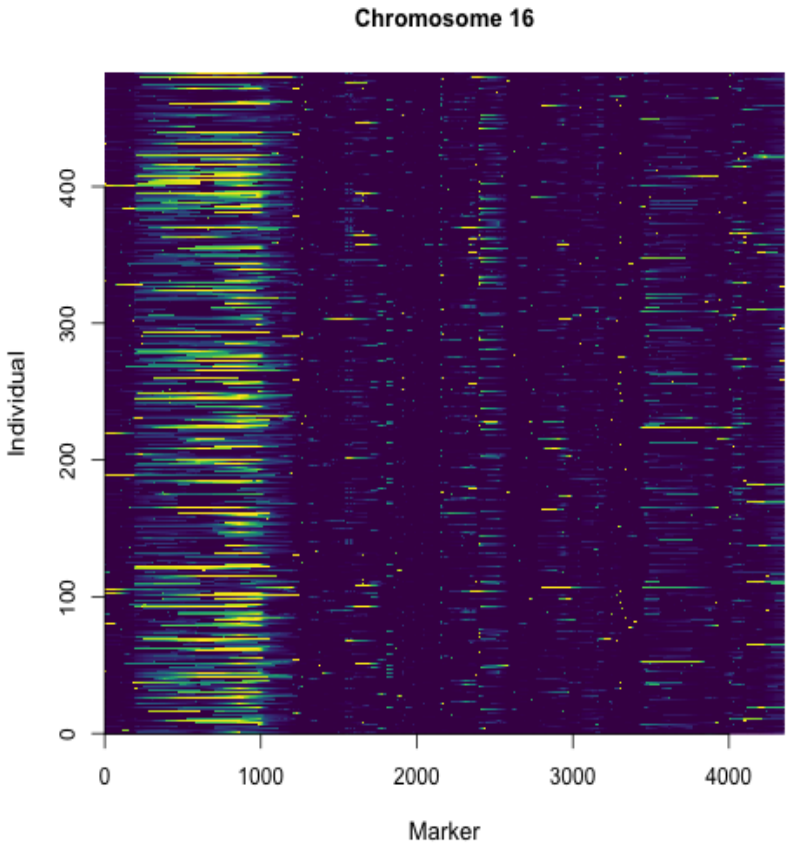

Chromosome 5
Sum of Absolute Difference
0.0
0.5
1.0
1.5
2.0
0
1000
2000
3000
4000
5000
6000
Marker

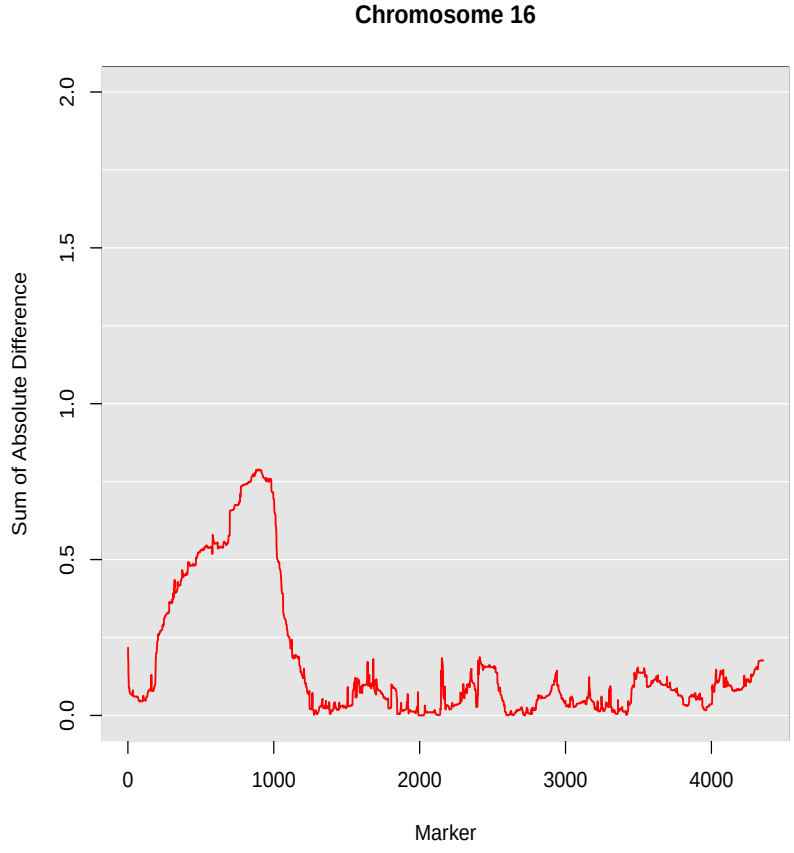

Chromosome 16
Individual
0
100
200
300
400
0
1000
2000
3000
4000
Marker


Chromosome 16
Sum of Absolute Difference
0.0
0.5
1.0
1.5
2.0
0
1000
2000
3000
4000
Marker


Chromosome 16
Individual
0
100
200
300
400
0
1000
2000
3000
4000
Marker


Chromosome 16
Sum of Absolute Difference
0.0
0.5
1.0
1.5
2.0
0
1000
2000
3000
4000
Marker





Marker region specific issues