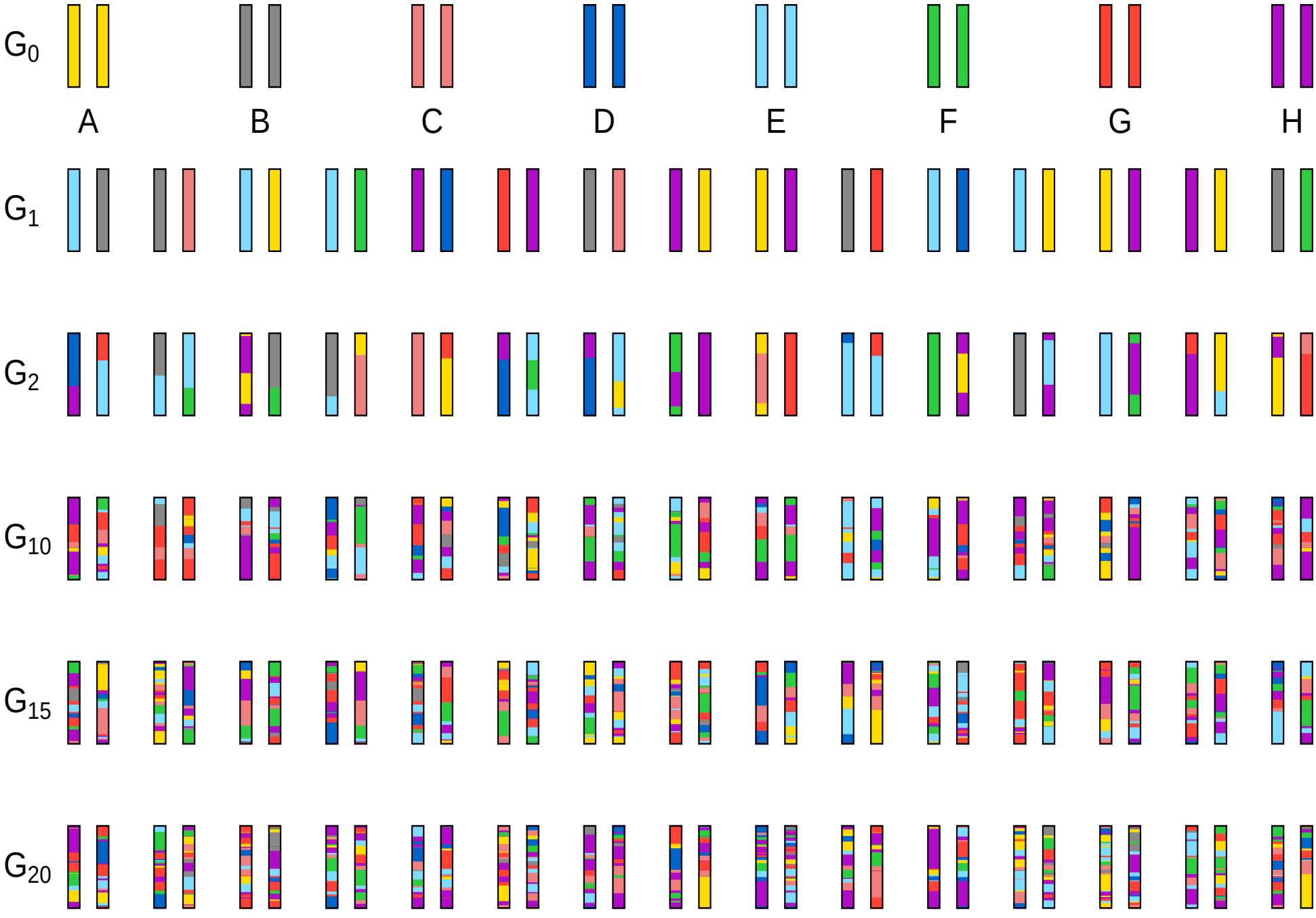
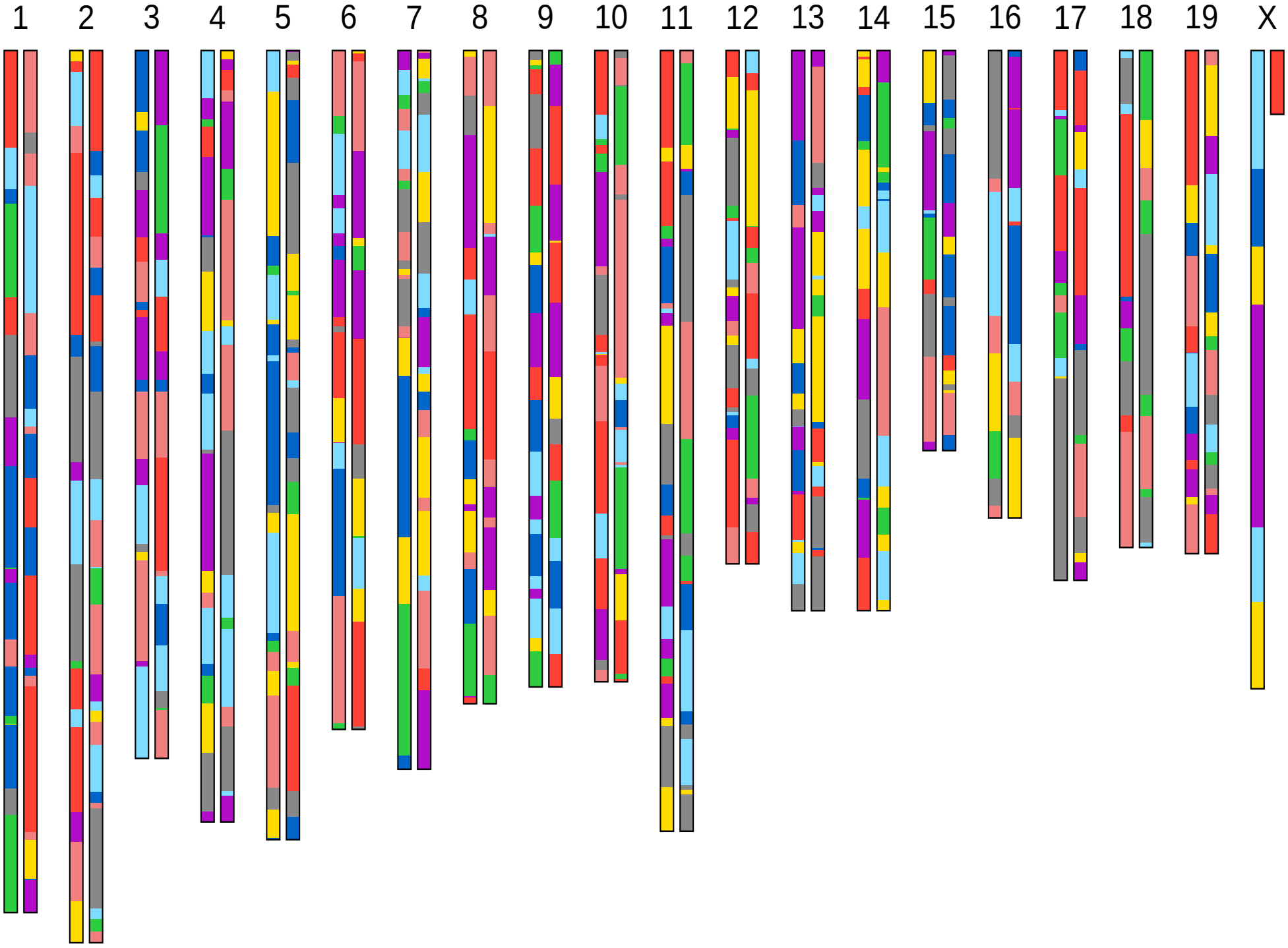
Genotype Reconstruction for Diversity Outbred Mice

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:

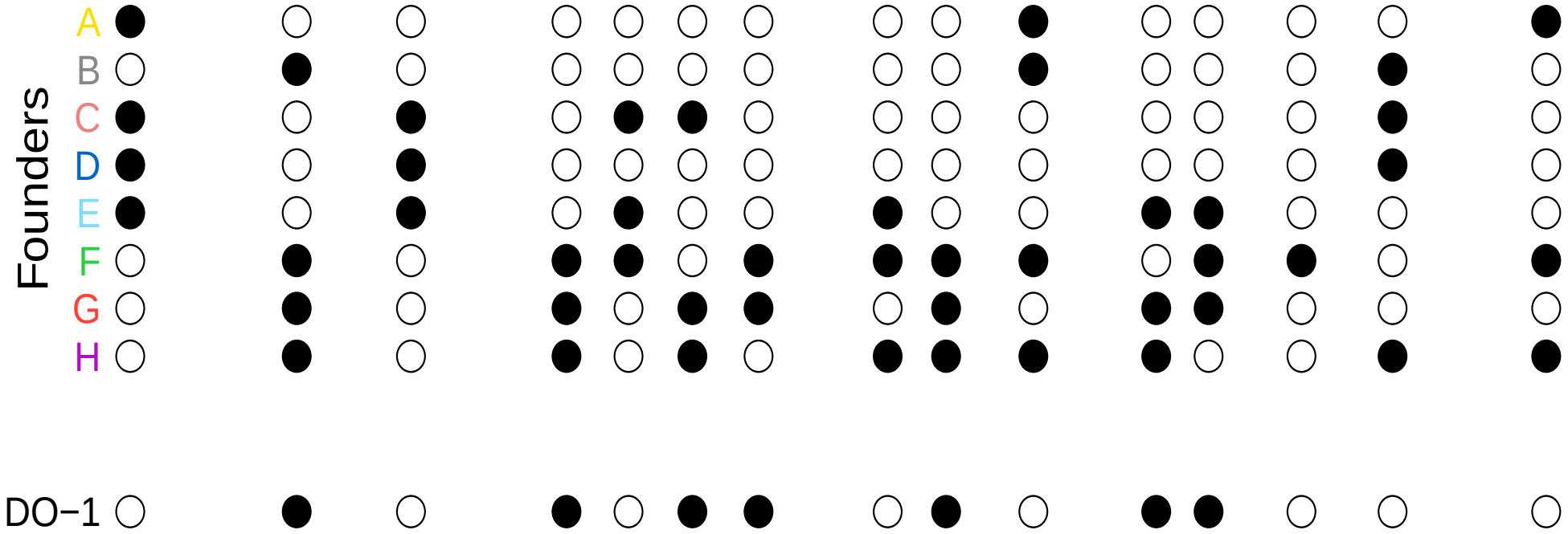
1. Transition Model
2. **Emission Model**

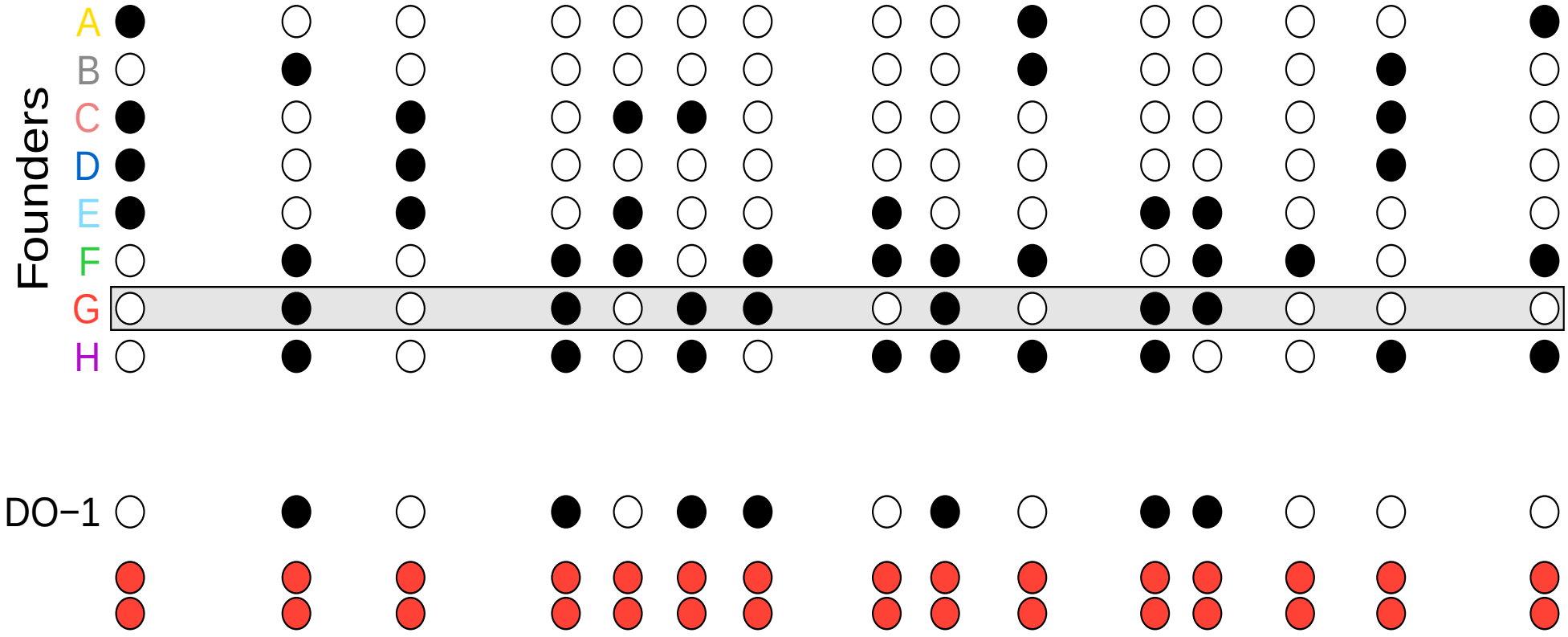
Hidden Markov Model

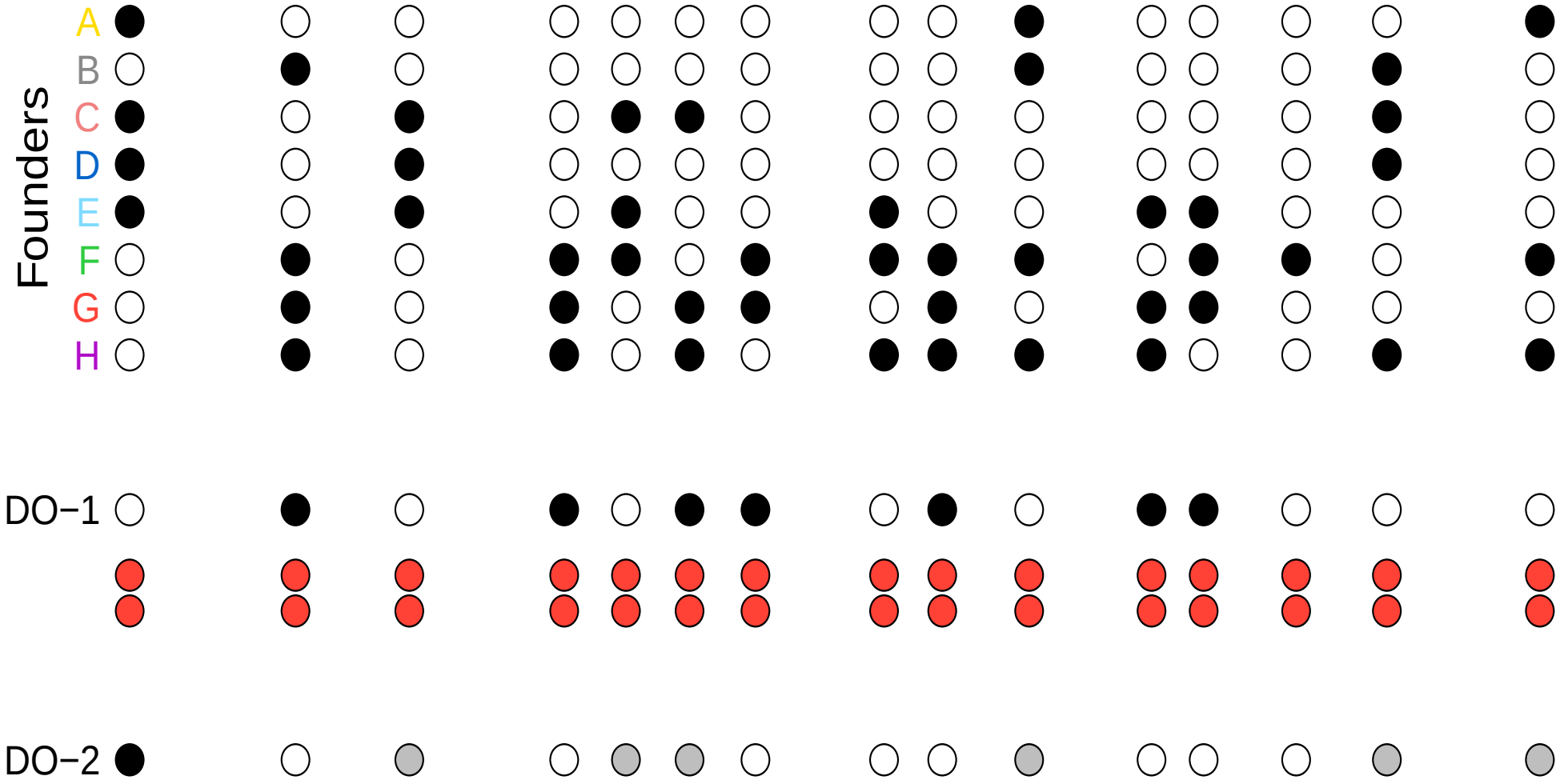
Probabilities are generated from:

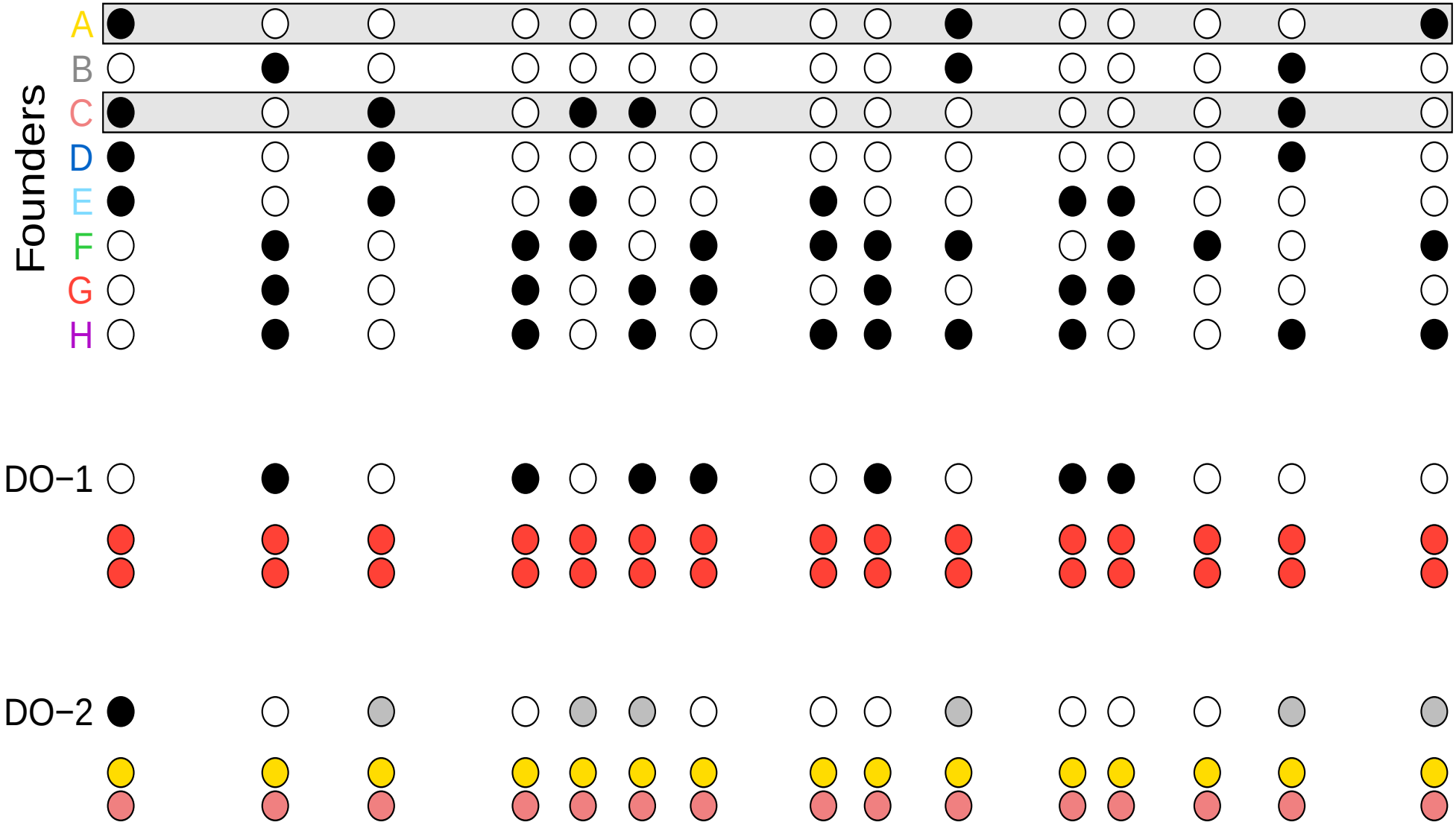
1. Transition Model
2. **Emission Model**

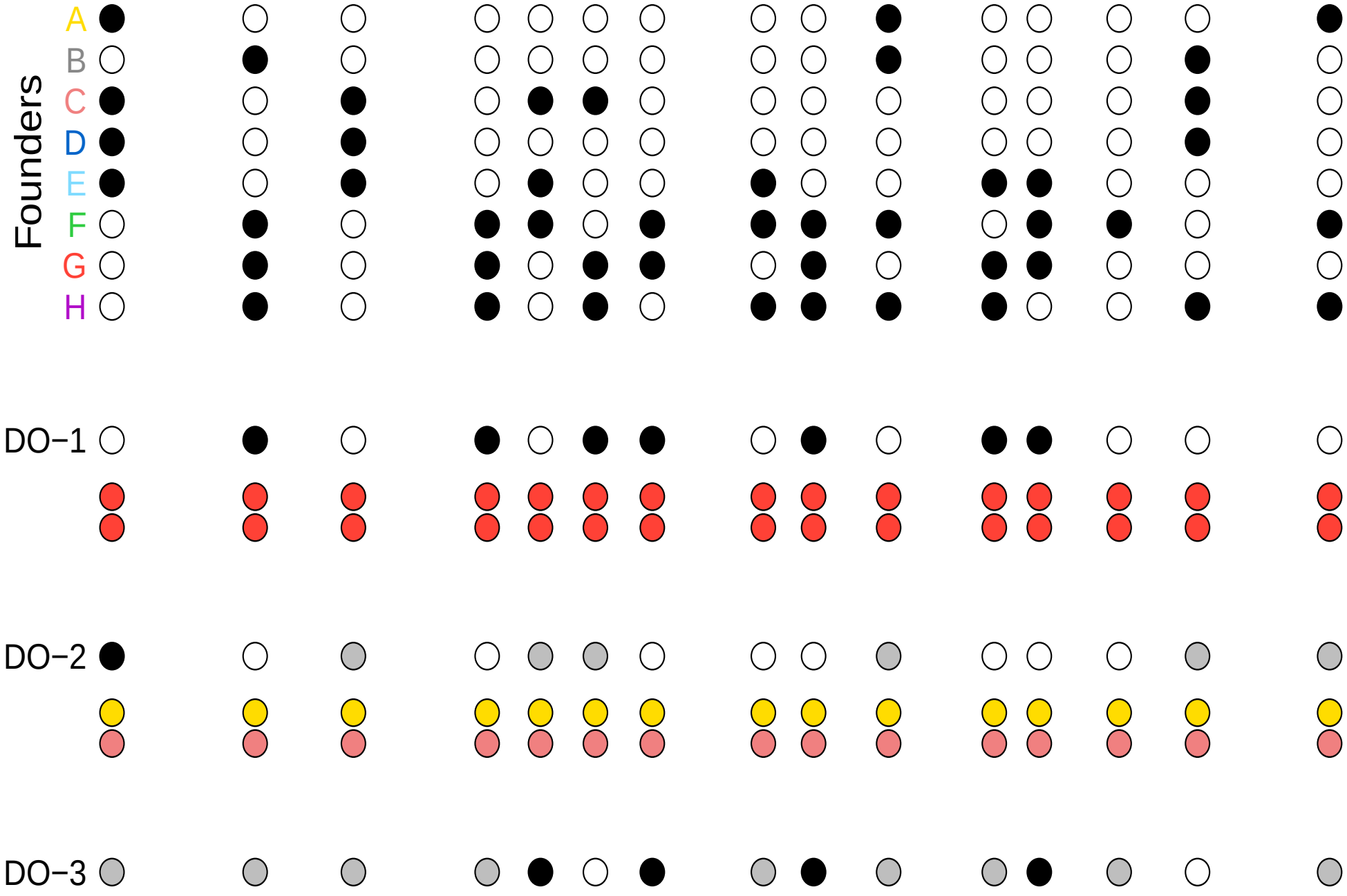
Conditional probability of observed data given underlying diplotype state

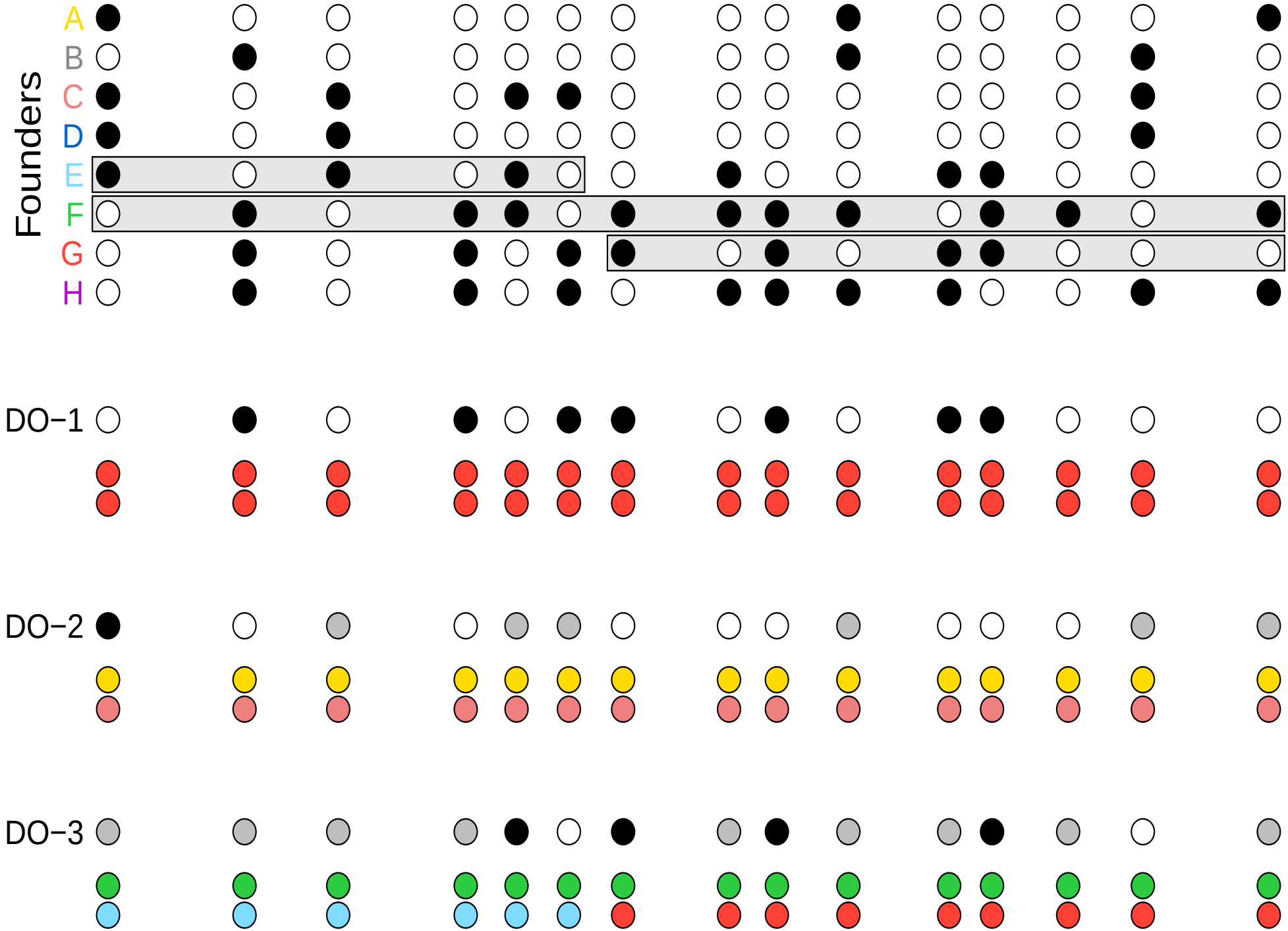

Founders
A
B
C
D
E
F
G
H
DO−1


Founders
A
B
C
D
E
F
G
H
DO−1

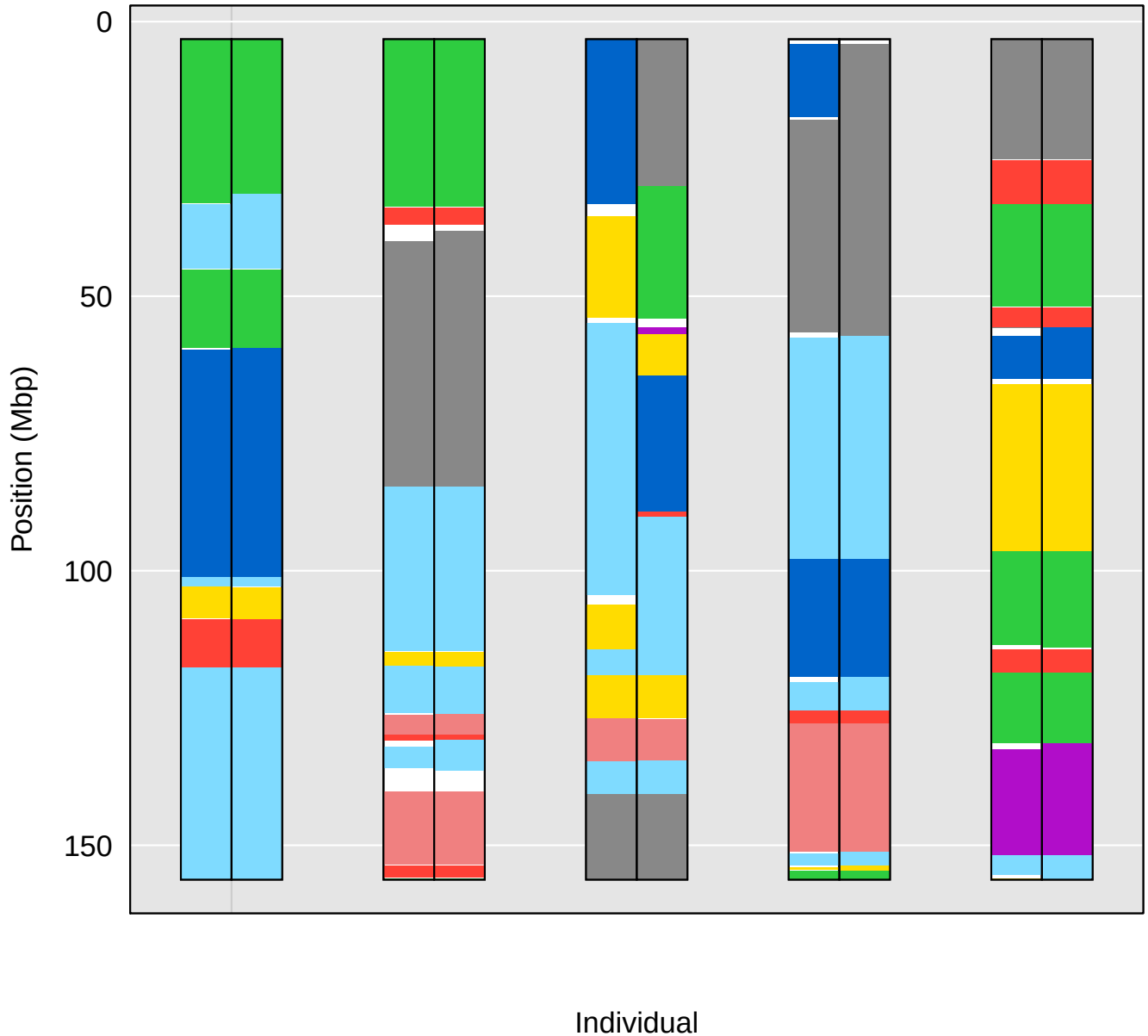

Founders
A
B
C
D
E
F
G
H
DO−1
DO−2


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

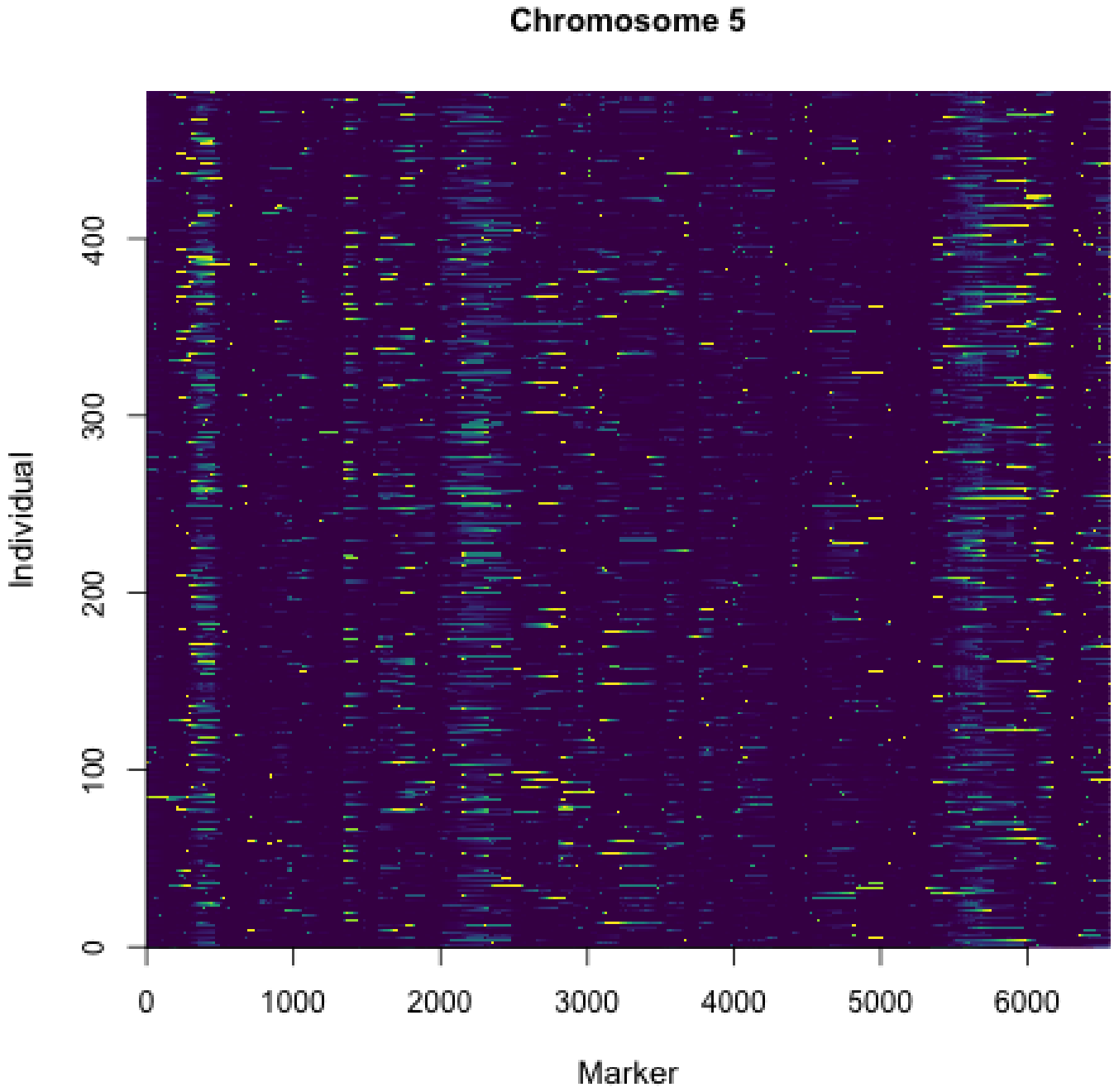
For each individual at each marker:

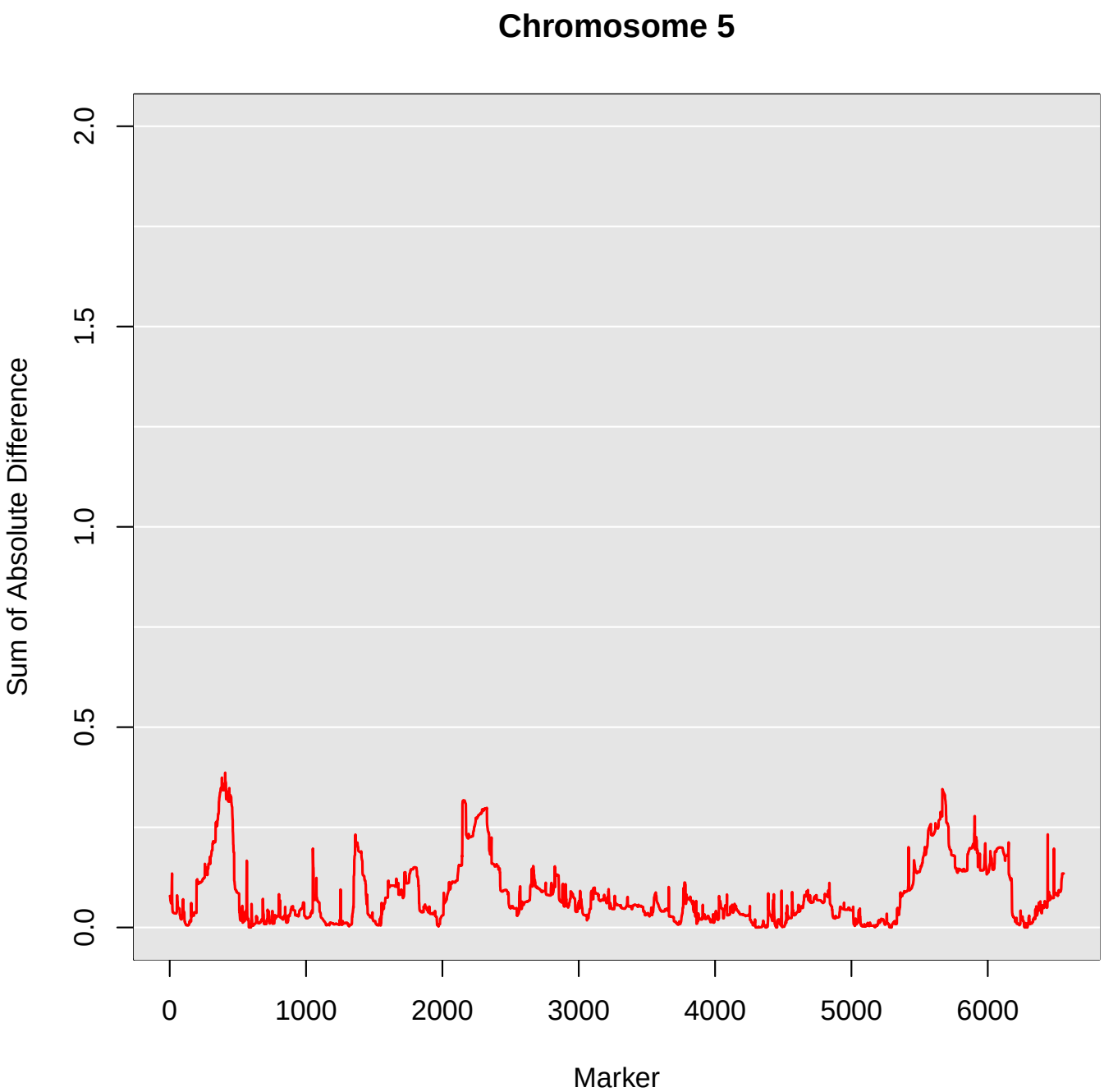
Compute *sum of absolute differences*

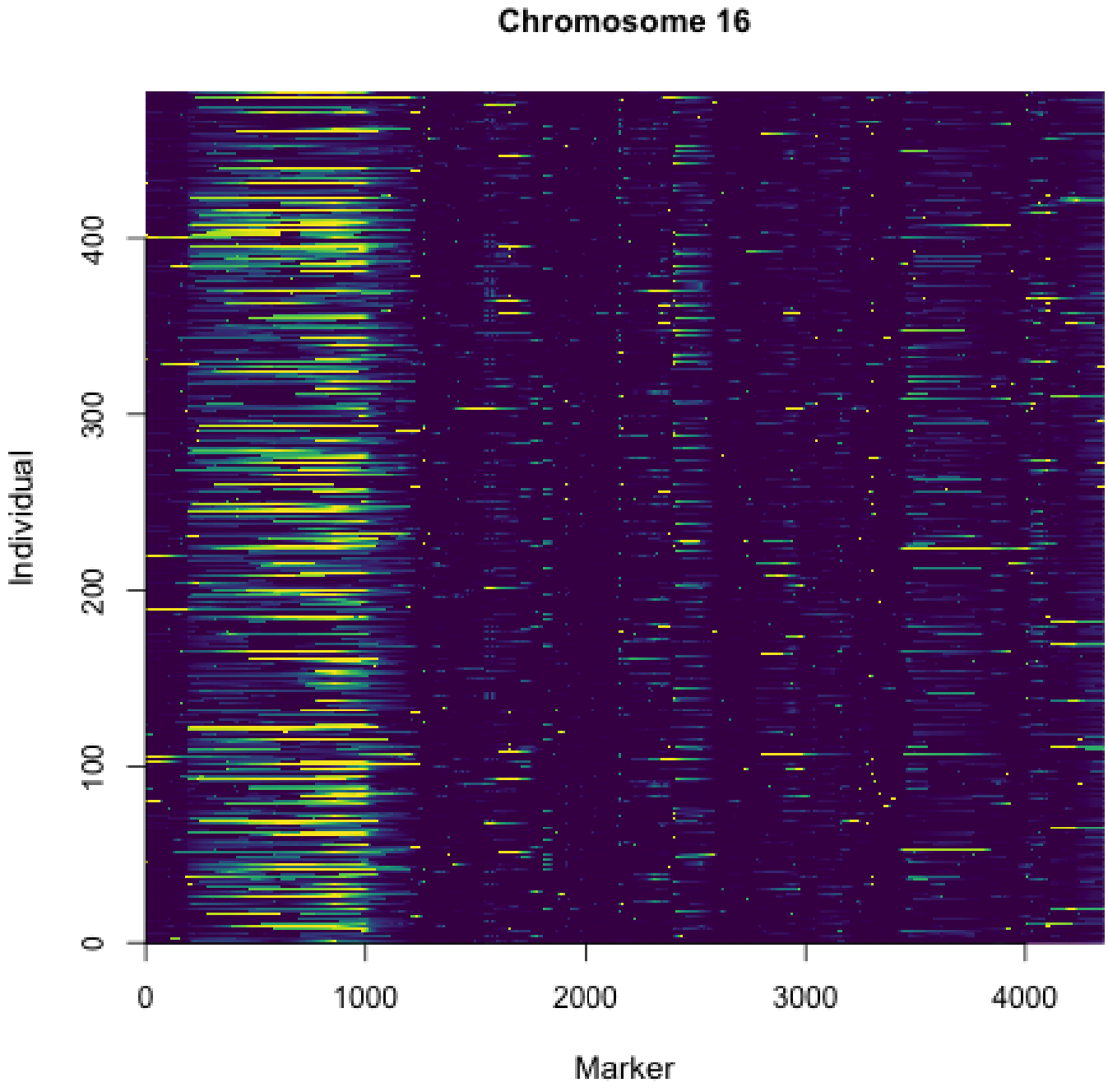
$$\sum_{i=1}^{36} |p_{1,i} - p_{2,i}|$$

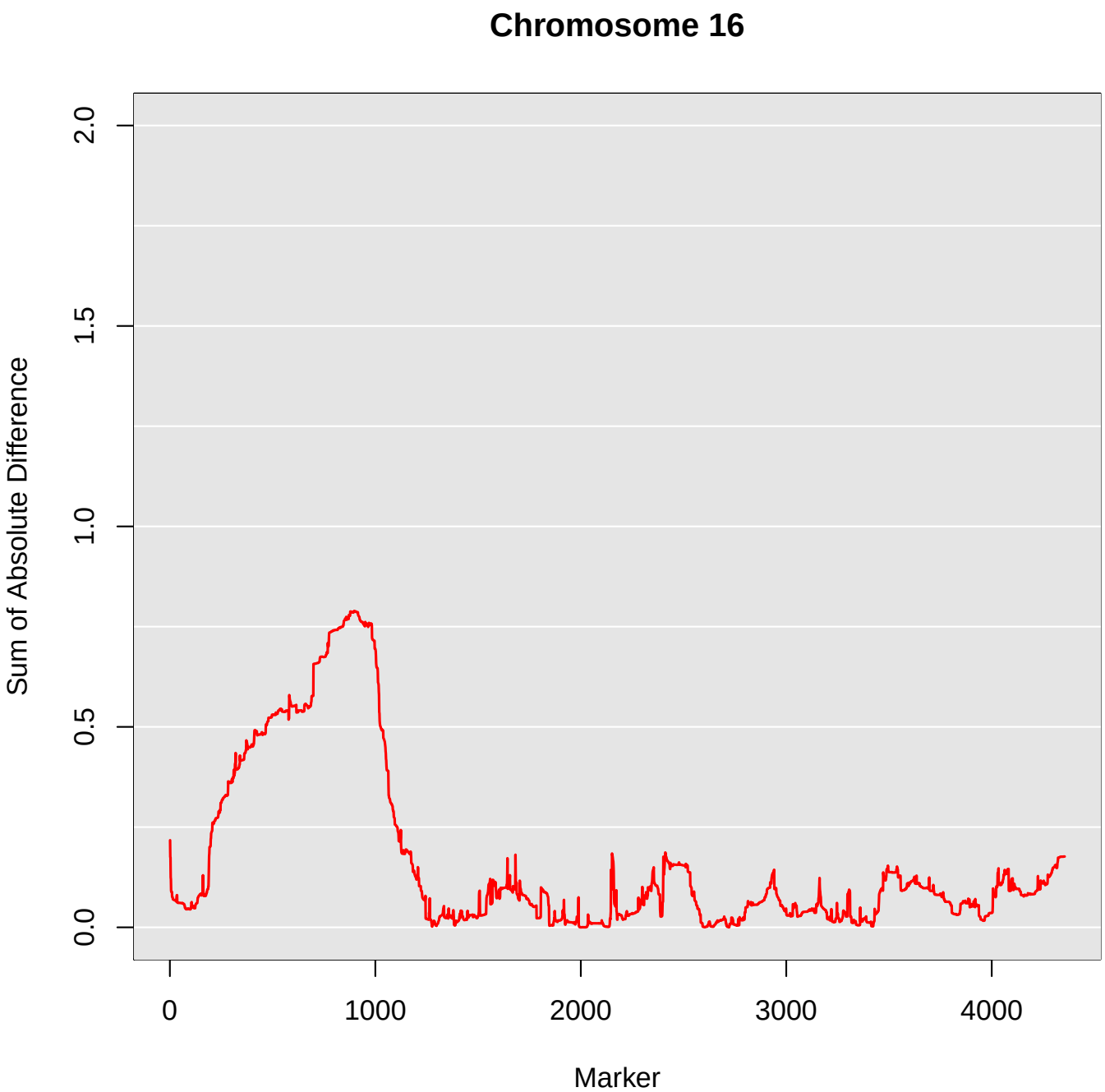
Each entry represents distance between methods

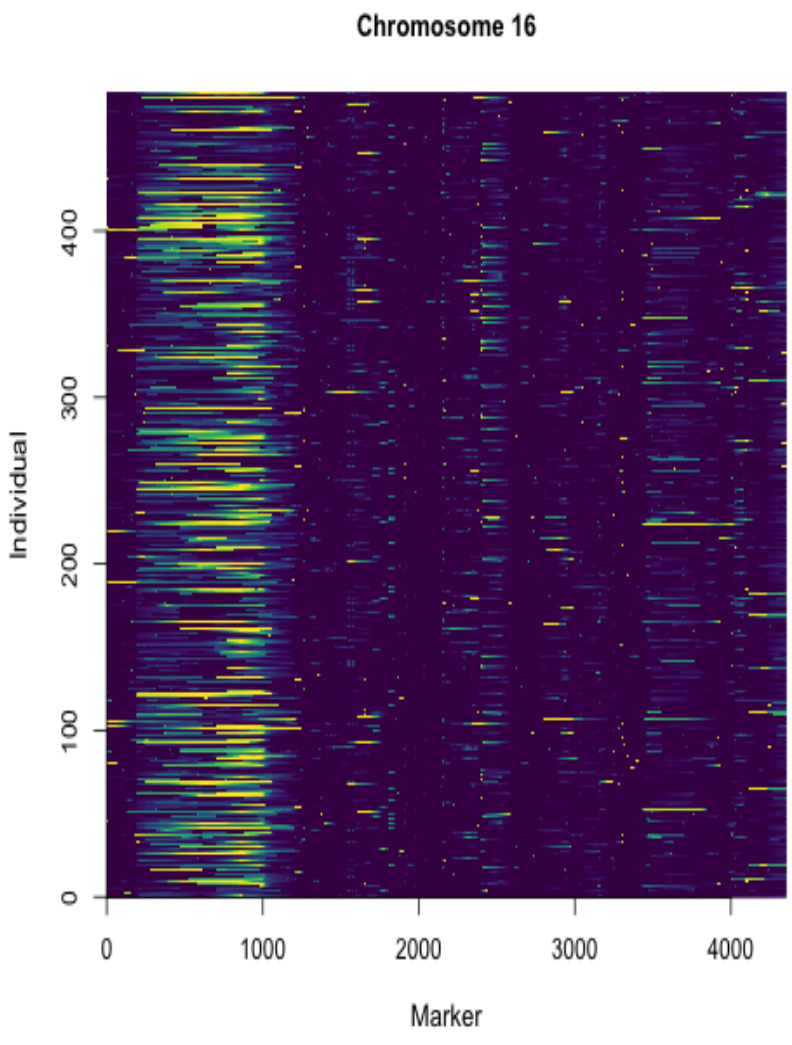
Reduces problems to two-dimensions: $500 \times 120,000$

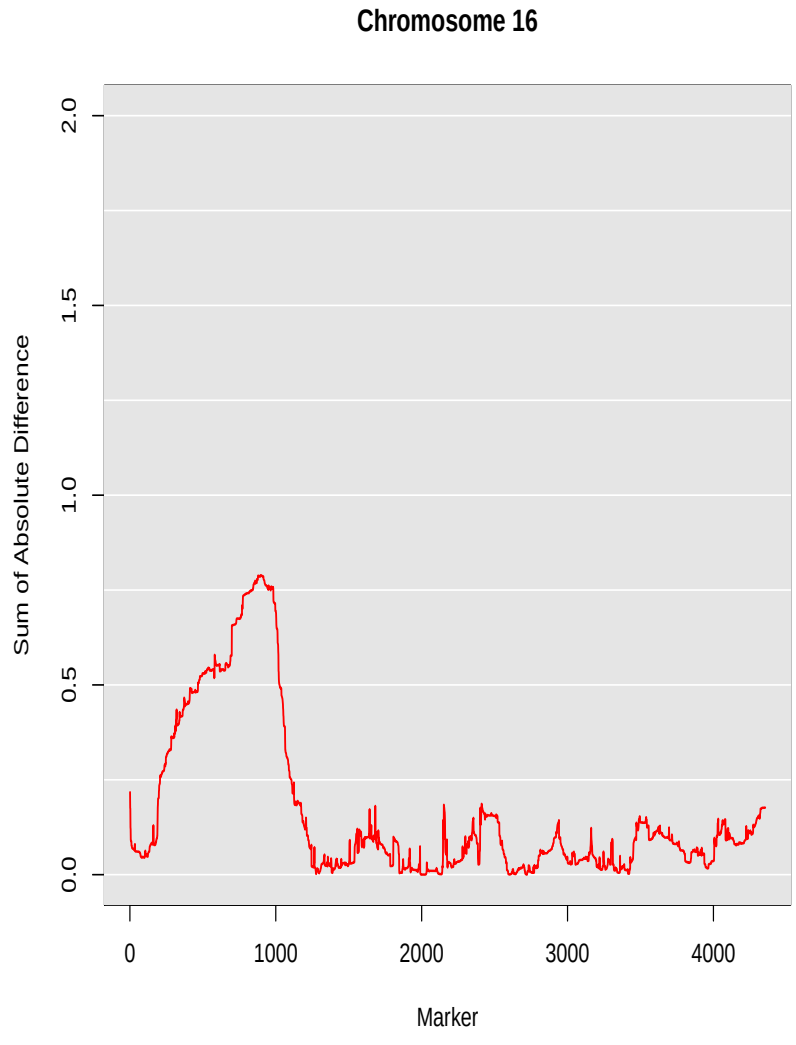

Chromosome 5
Individual
0
100
200
300
400
0
1000
2000
3000
4000
5000
6000
Marker

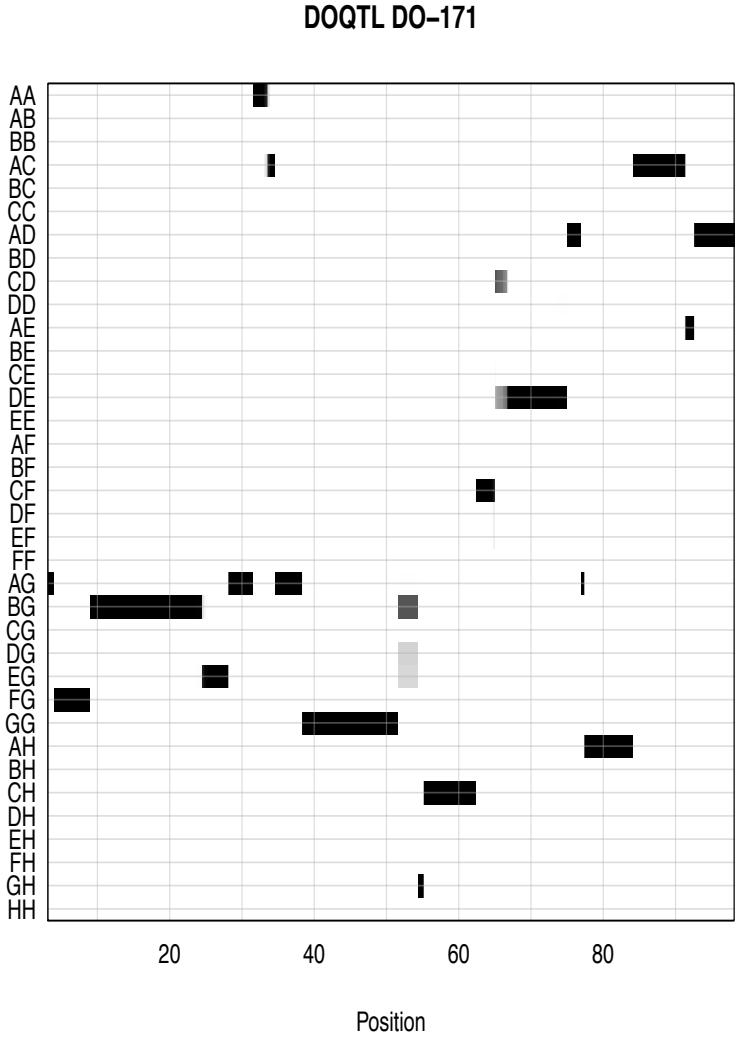

Chromosome 5
Sum of Absolute Difference
0.0
0.5
1.0
1.5
2.0
0
1000
2000
3000
4000
5000
6000
Marker

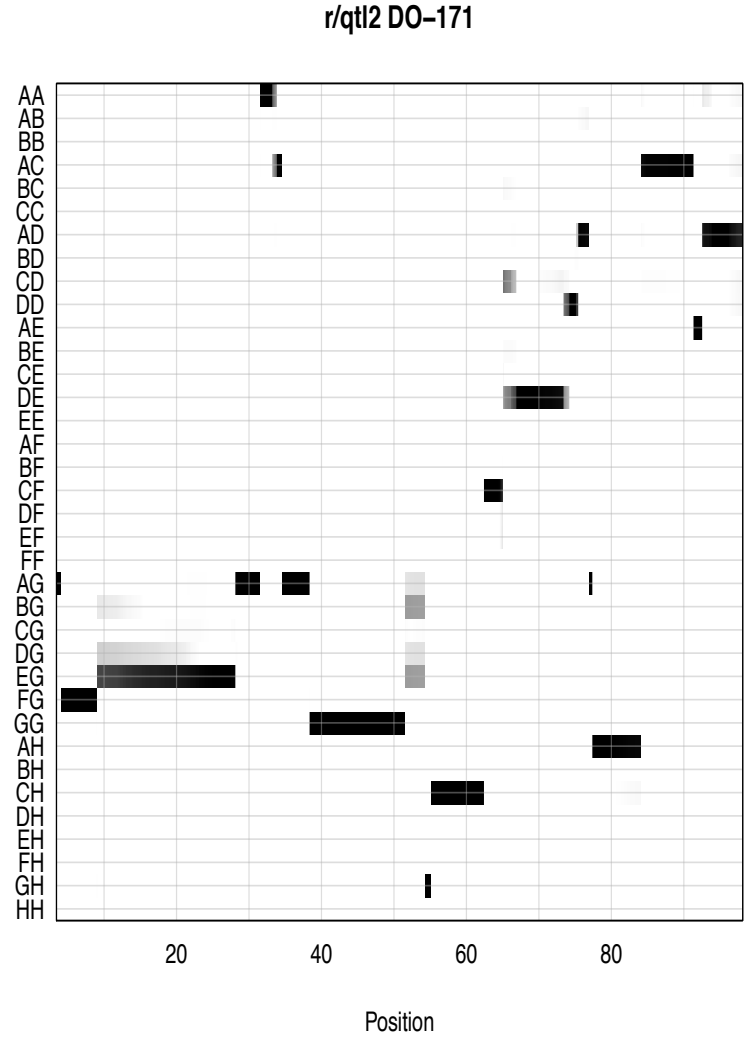

Chromosome 16
Individual
0
100
200
300
400
0
1000
2000
3000
4000
Marker

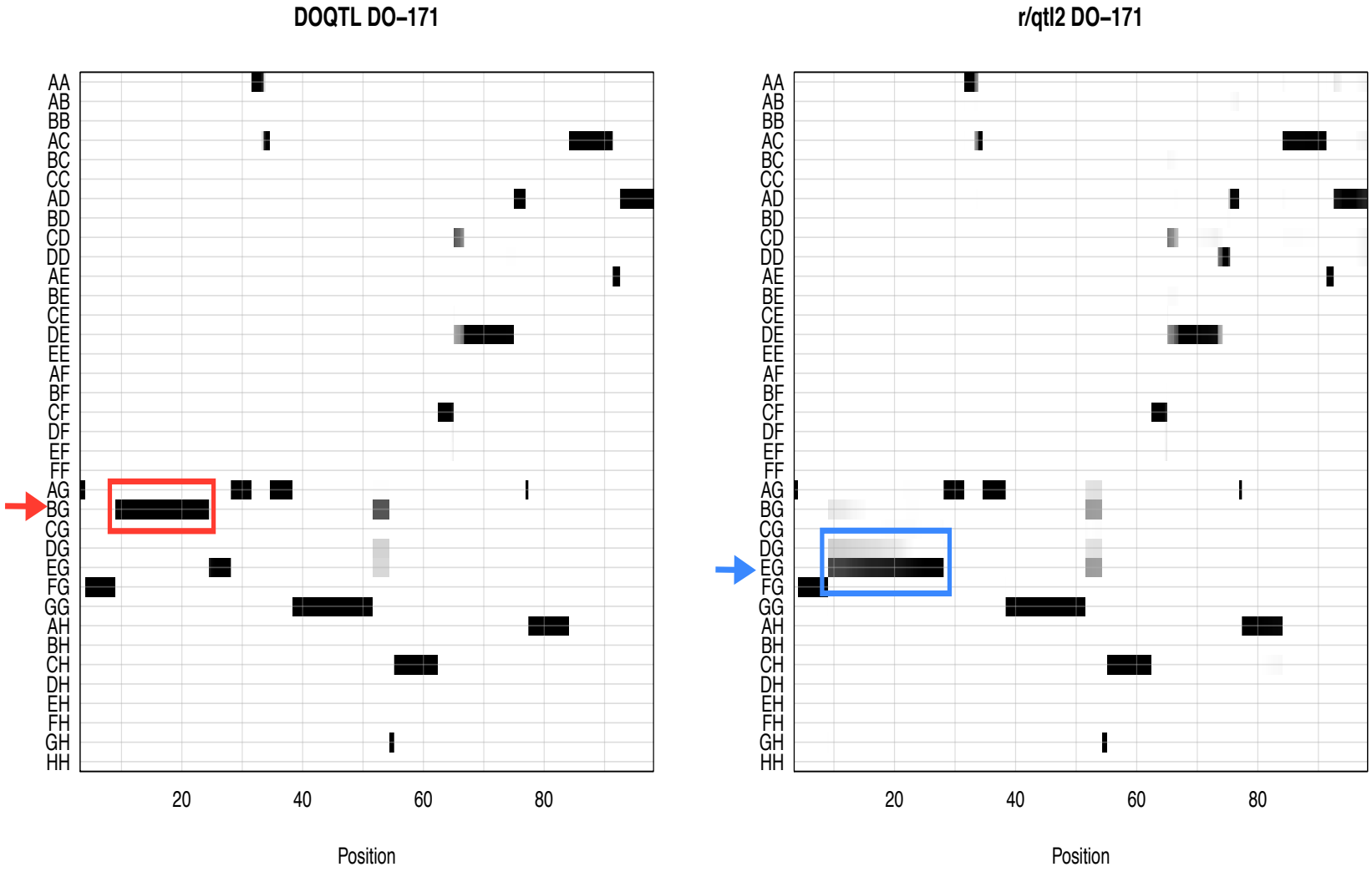

Chromosome 16
Sum of Absolute Difference
0.0
0.5
1.0
1.5
2.0
0
1000
2000
3000
4000
Marker

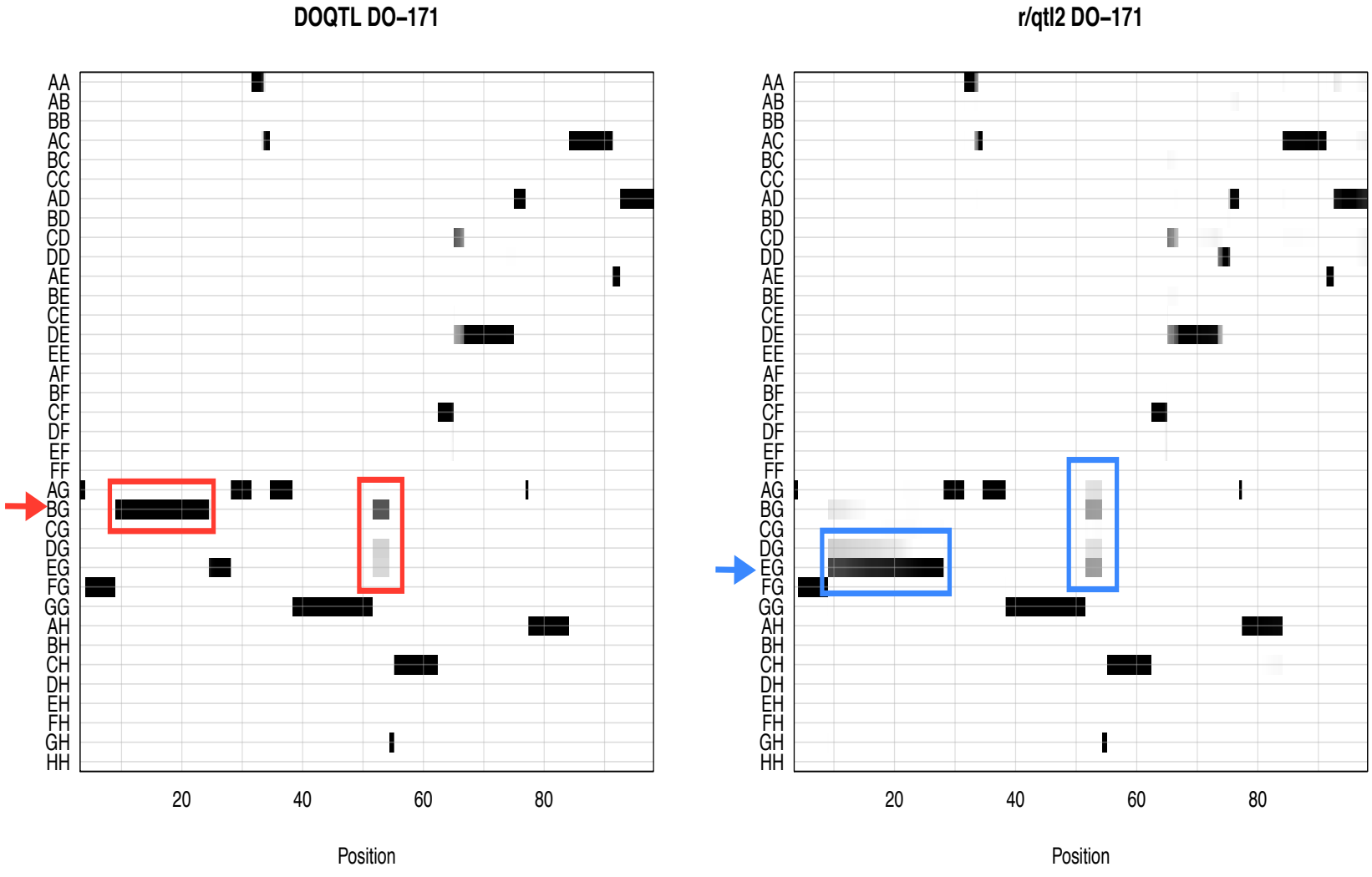

Chromosome 16
Individual
0
100
200
300
400
0
1000
2000
3000
4000
Marker

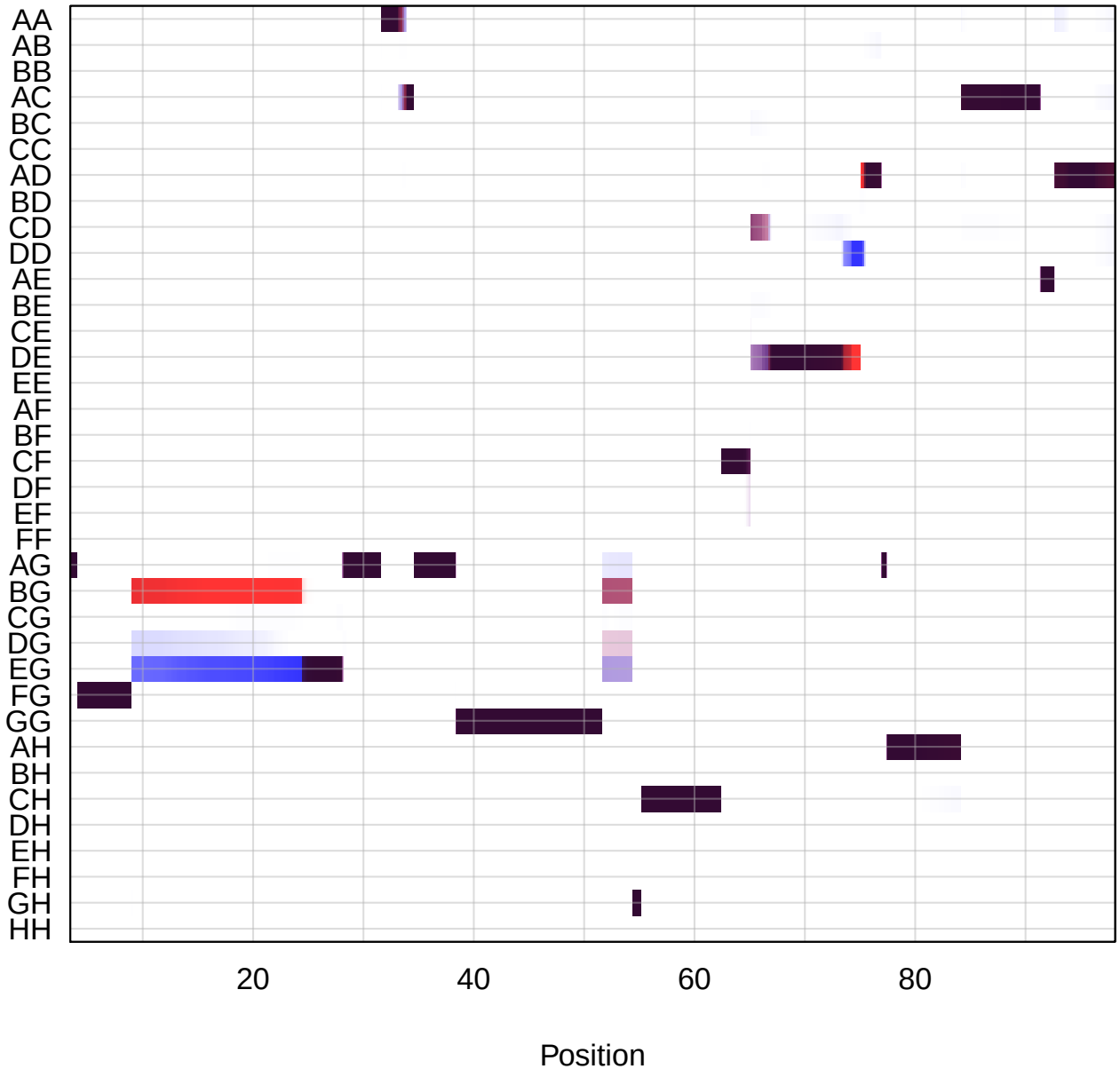

Chromosome 16
Sum of Absolute Difference
0.0
0.5
1.0
1.5
2.0
0
1000
2000
3000
4000
Marker


DOQTL DO−171
AA
AB
BB
AC
BC
CC
AD
BD
CD
DD
AE
BE
CE
DE
EE
AF
BF
CF
DF
EF
FF
AG
BG
CG
DG
EG
FG
GG
AH
BH
CH
DH
EH
FH
GH
HH
20
40
60
80
Position


r/qtl2 DO−171
AA
AB
BB
AC
BC
CC
AD
BD
CD
DD
AE
BE
CE
DE
EE
AF
BF
CF
DF
EF
FF
AG
BG
CG
DG
EG
FG
GG
AH
BH
CH
DH
EH
FH
GH
HH
20
40
60
80
Position


DOQTL DO−171
r/qtl2 DO−171
AA
AB
BB
AC
BC
CC
AD
BD
CD
DD
AE
BE
CE
DE
EE
AF
BF
CF
DF
EF
FF
AG
BG
CG
DG
EG
FG
GG
AH
BH
CH
DH
EH
FH
GH
HH
20
40
60
80
Position


DOQTL DO−171
r/qtl2 DO−171
AA
AB
BB
AC
BC
CC
AD
BD
CD
DD
AE
BE
CE
DE
EE
AF
BF
CF
DF
EF
FF
AG
BG
CG
DG
EG
FG
GG
AH
BH
CH
DH
EH
FH
GH
HH
20
40
60
80
Position


AA
AB
BB
AC
BC
CC
AD
BD
CD
DD
AE
BE
CE
DE
EE
AF
BF
CF
DF
EF
FF
AG
BG
CG
DG
EG
FG
GG
AH
BH
CH
DH
EH
FH
GH
HH
20
40
60
80
Position

Which answer is ‘correct’?

Which answer is ‘correct’?

- Pick an interval with differing, confident answers

Which answer is ‘correct’?

- Pick an interval with differing, confident answers
- Compare SNP genotypes with expected genotypes

Which answer is ‘correct’?

- Pick an interval with differing, confident answers
- Compare SNP genotypes with expected genotypes
- Compare SNP genotypes between discordant founders

Which answer is ‘correct’?

Which answer is ‘correct’?

Individual: DO-171 **Chromosome:** 16

Which answer is 'correct'?

Individual: DO-171 **Chromosome:** 16

	Prediction	**Probability**	**Match Proportion**
r\qtl2	EG	0.892	0.9958848
doqtl	BG	0.999	0.9917695

Which answer is 'correct'?

Individual: DO-171 **Chromosome:** 16

	Prediction	**Probability**	**Match Proportion**
r\qtl2	EG	0.892	0.9958848
doqtl	BG	0.999	0.9917695

Best match: EG

Which answer is 'correct'?

Individual: DO-171 **Chromosome:** 16

	Prediction	**Probability**	**Match Proportion**
r\qtl2	EG	0.892	0.9958848
doqtl	BG	0.999	0.9917695

Best match: EG

Disagreement between **E** and **B**

Which answer is 'correct'?

Individual: DO-171 **Chromosome:** 16

	Prediction	**Probability**	**Match Proportion**
r\qtl2	EG	0.892	0.9958848
doqtl	BG	0.999	0.9917695

Best match: EG

Disagreement between **E** and **B** $\rightarrow$ Proportion match $= 0.997$

Which answer is 'correct'?

Individual: DO-171 **Chromosome:** 16

	Prediction	**Probability**	**Match Proportion**
r\qtl2	EG	0.892	0.9958848
doqtl	BG	0.999	0.9917695

Best match: EG

Disagreement between **E** and **B** $\rightarrow$ Proportion match $= 0.997$

We see this consistently across the marker region

Overall Results

Overall Results

Goal: Compare large 3D arrays

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes
  - Heat map of individuals $\times$ markers

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes
  - Heat map of individuals $\times$ markers
  - Plot average distance by marker position

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes
  - Heat map of individuals $\times$ markers
  - Plot average distance by marker position
- Visually compare probabilities *at individual level*

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes
  - Heat map of individuals $\times$ markers
  - Plot average distance by marker position
- Visually compare probabilities *at individual level*
  - Univariate probability heatmap

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes
  - Heat map of individuals $\times$ markers
  - Plot average distance by marker position
- Visually compare probabilities *at individual level*
  - Univariate probability heatmap
  - Bivariate probability heatmap

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes
  - Heat map of individuals $\times$ markers
  - Plot average distance by marker position
- Visually compare probabilities *at individual level*
  - Univariate probability heatmap
  - Bivariate probability heatmap
- Compare observed SNP genotypes to expected

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes
  - Heat map of individuals $\times$ markers
  - Plot average distance by marker position
- Visually compare probabilities *at individual level*
  - Univariate probability heatmap
  - Bivariate probability heatmap
- Compare observed SNP genotypes to expected
  - r\qtl2 consistently gives better answers

Overall Results

Goal: Compare large 3D arrays

- Measure of distance to identify problem regions
  - Sum of absolute differences across haplotypes
  - Heat map of individuals $\times$ markers
  - Plot average distance by marker position
- Visually compare probabilities *at individual level*
  - Univariate probability heatmap
  - Bivariate probability heatmap
- Compare observed SNP genotypes to expected
  - r\qtl2 consistently gives better answers
  - Disagreement often related to similarity of founders