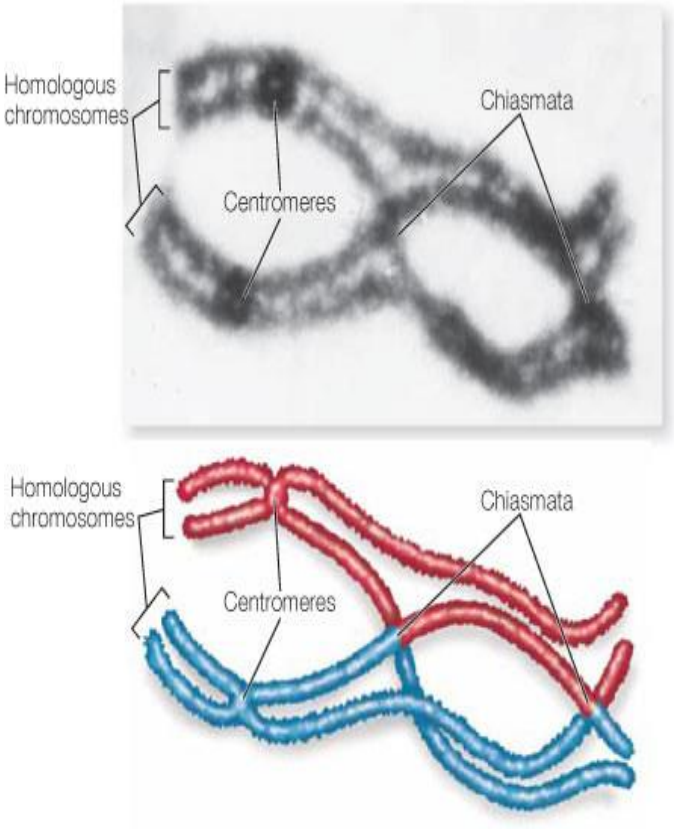
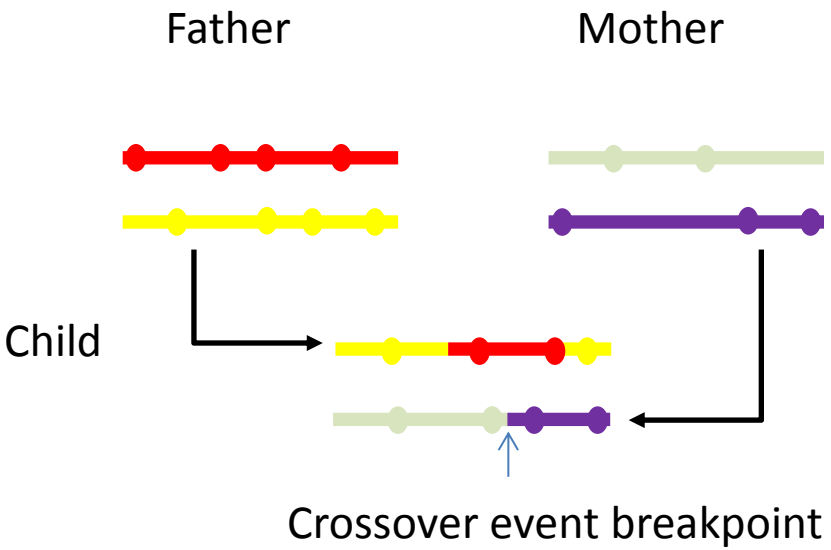
The painting palettes of human ancestry

Daniel Falush, Dan Lawson,
Garrett Hellenthal, Simon Myers.

www.paintmychromosomes.com

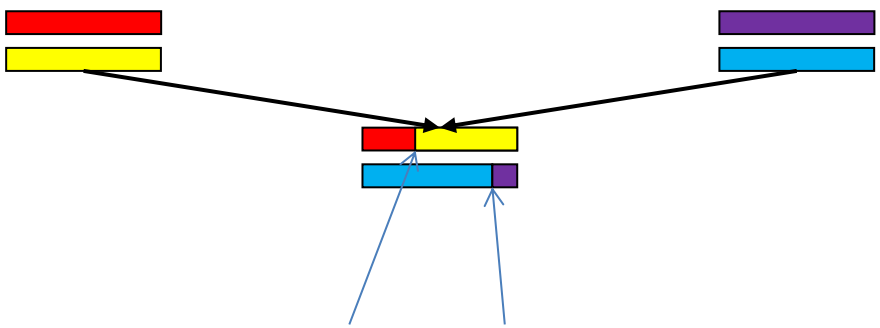
We want to “paint” DNA according to which ancestor it comes from





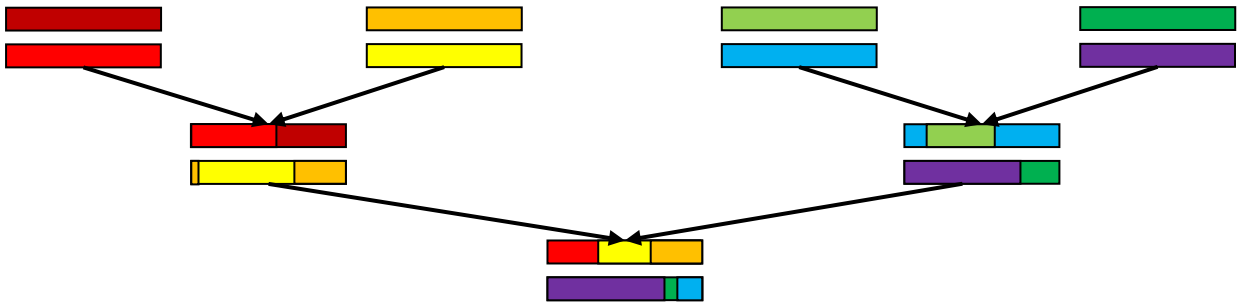
(O’Connor 2008)

Recombination ⇨ “mosaic” genomes

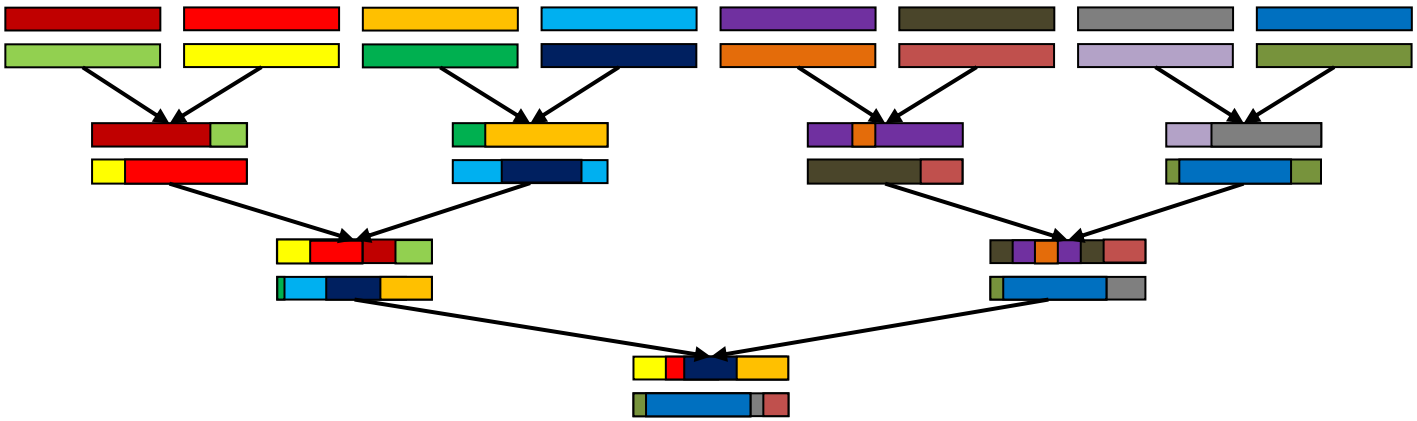


Positions of recombination crossover events

Recombination ⇨ “mosaic” genomes



Recombination ⇨ “mosaic” genomes



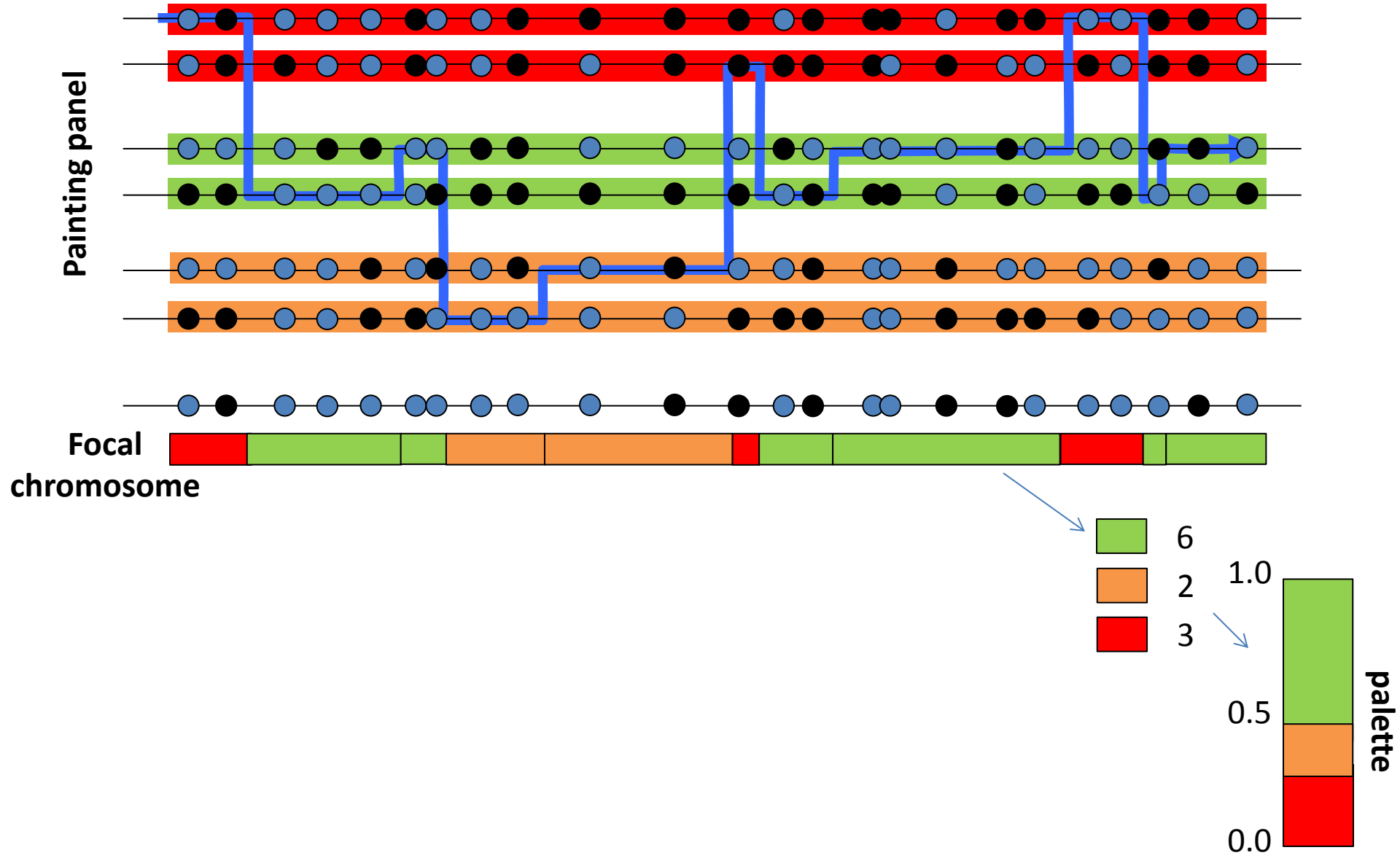
Chromosome “painting”

We share segments of DNA today, inherited from shared ancestors living in the distant past.

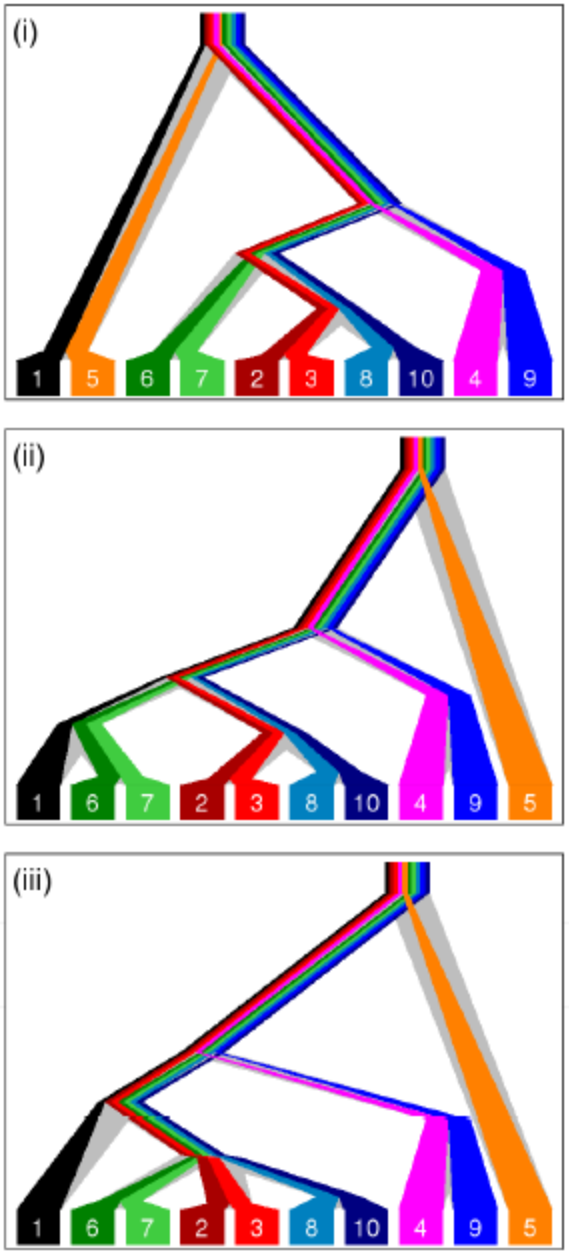
Identifying these shared segments, and insights into the genetics of people from the UK, and worldwide human migrations

Chromosome painting in practice

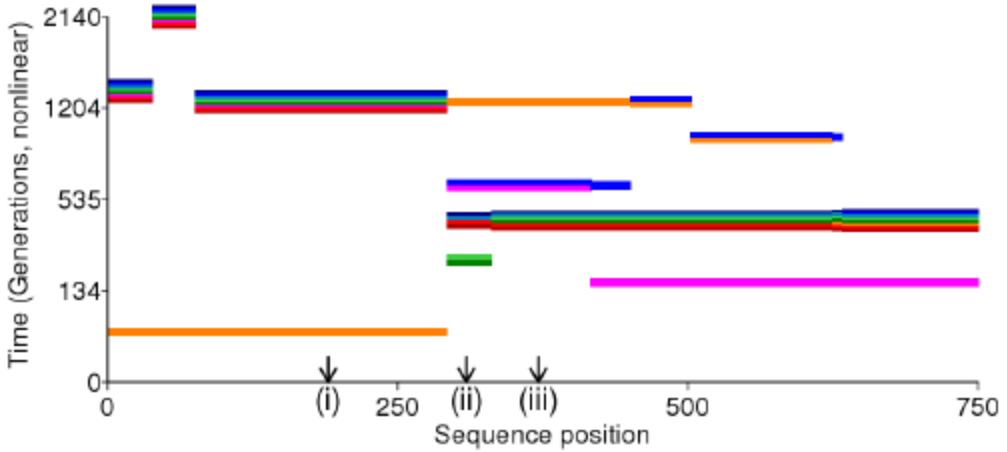
- (Lawson et al. 2012)



A) Local genealogies

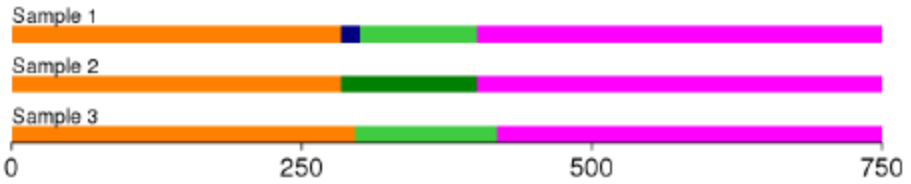


B) Time to MRCA with haplotype 1



C) True 'nearest neighbour' distribution of haplotype 1

D) Sample paintings of haplotype 1

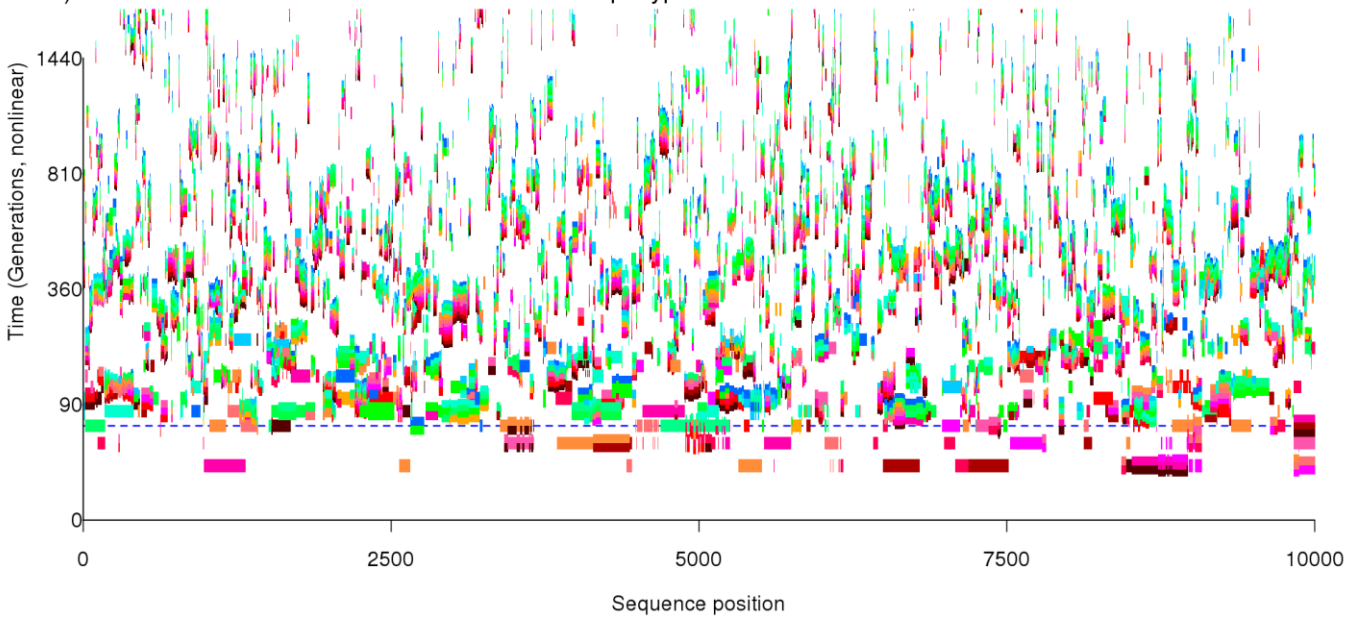


E) Mean painting of haplotype 1

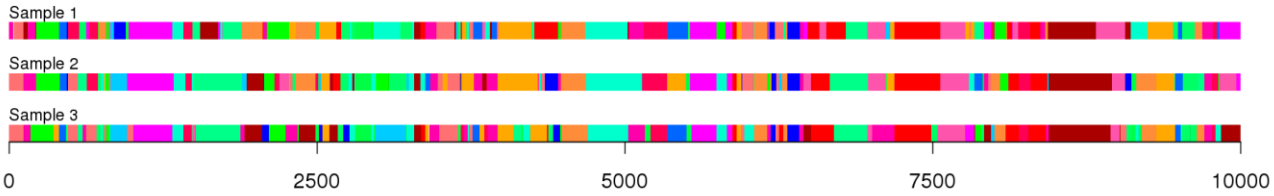
F) Coancestry matrix row for haplotype 1

	Donor haplotype									
	1	2	3	4	5	6	7	8	9	10
Haplotype 1	0	0.08	0.09	1.1	1.24	0.52	0.52	0.06	0.01	0.06

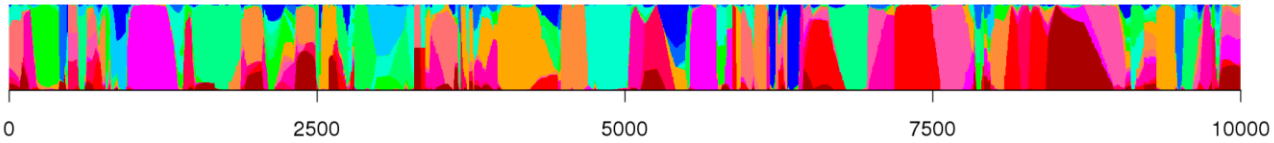
A) Time to Most Recent Common Ancestor with haplotype 1



B) Sample paintings of haplotype 1



C) Expectation painting of haplotype 1



D) Coancestry matrix row for haplotype 1

	Donor haplotype																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Haplotype 1	0	8.06	9.59	8.27	11.64	6.51	7.17	7.68	9.25	10.94	4.11	7	5.91	5.62	5.68	6.53	6.72	5.03	5.86	7.05
	Population 1										Population 2									

Genealogical interpretation of painting

- The painting of each segment indicates who in the painting panel has the has the most recent shared common ancestor with the focal chromosome for that stretch of DNA.
- Boundaries between segments correspond to the positions of ancestral recombination events.
- The time at which the ancestor lived varies from segment to segment.
- Longer segments are associated with more recent shared ancestors

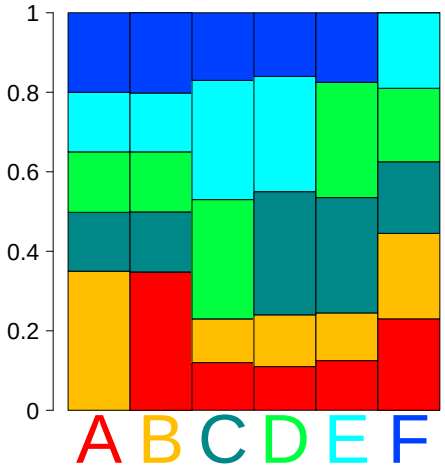
FineSTRUCTURE

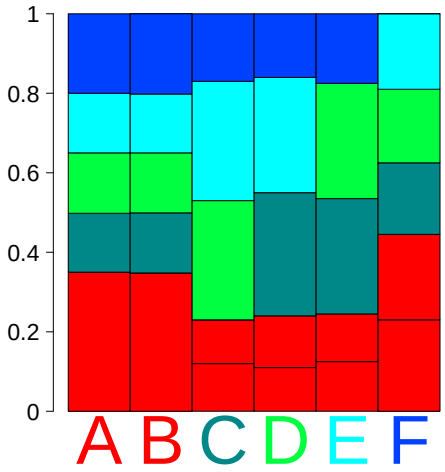
1. Paint each individual using every **other** individual in the sample as the painting panel and record the painting palettes.

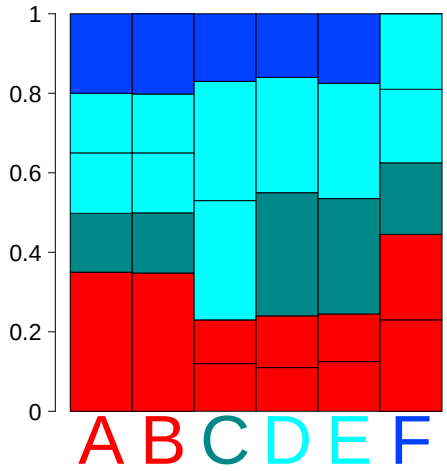
2. MCMC algorithm is used to find clustering such that:
   (a) members of the same cluster paint with the same colour
   (b) individuals within clusters have similar palettes
   (c) palettes are enriched for their own colour (mostly).

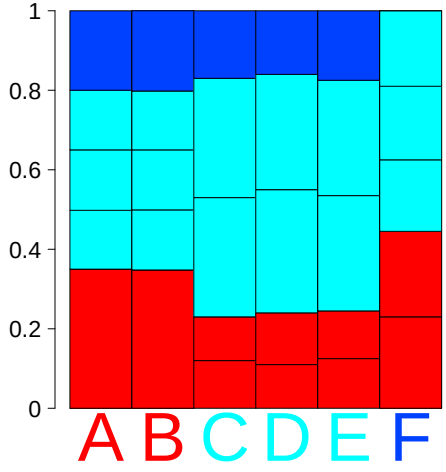
(Algorithm takes into account the fact that individuals are not used to paint themselves).

toy FineSTRUCTURE example





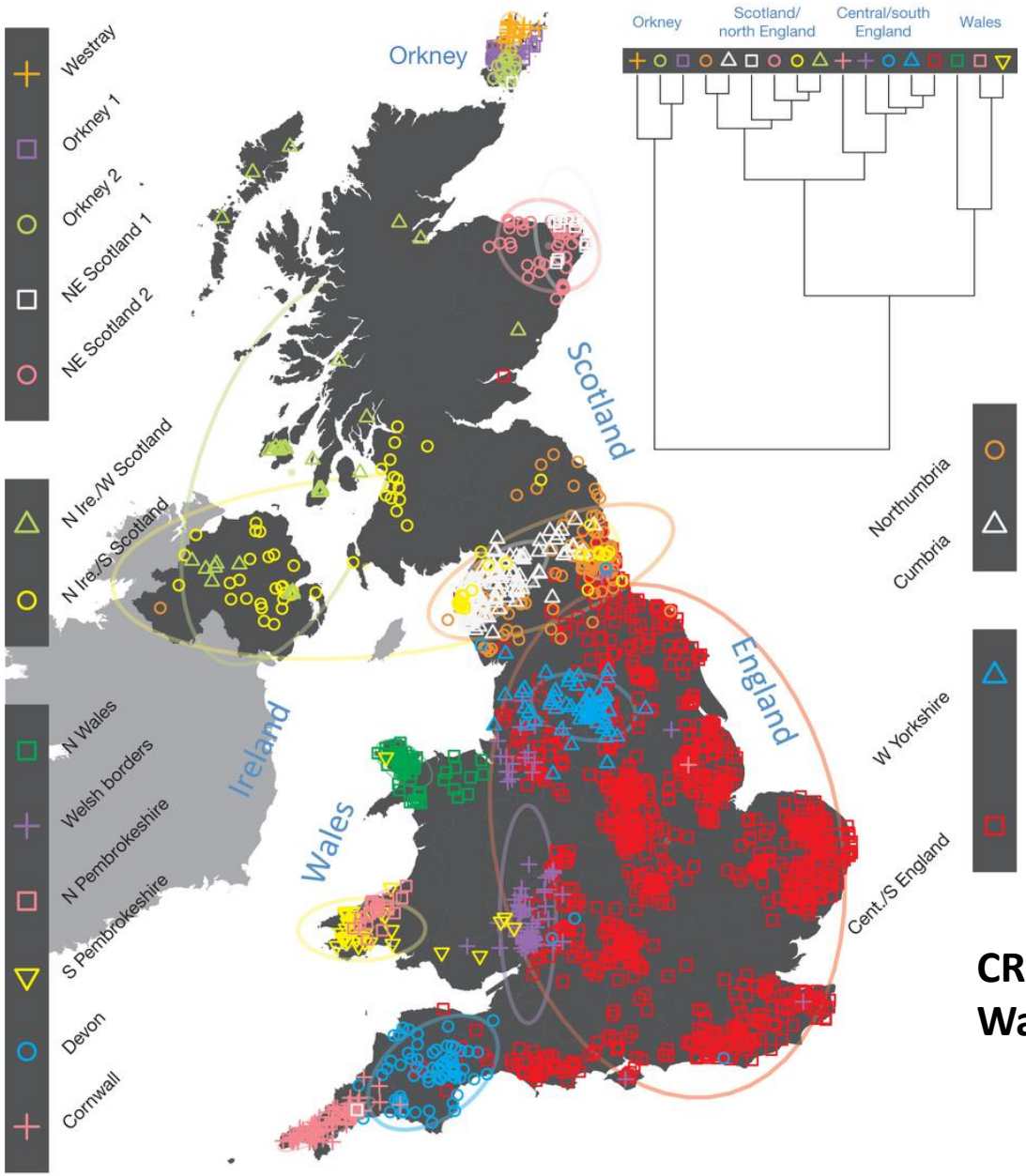




Notes about FineSTRUCTURE

- MCMC algorithm includes both merges and splits.
- Good convergence properties in practice.
- Number and membership of clusters is inferred based only on DNA.
- We call the collection of palettes obtained in the all-versus-all painting the “coancestry matrix”.
- If the markers are treated as unlinked, the coancestry matrix is equivalent to the covariance matrix used by SMARTPCA.
- Likelihood approximately equivalent to that of STRUCTURE for weak genetic drift

Application to Peopling of British Isles project



Sampled individuals had all four grandparents living within 50 miles of each other.

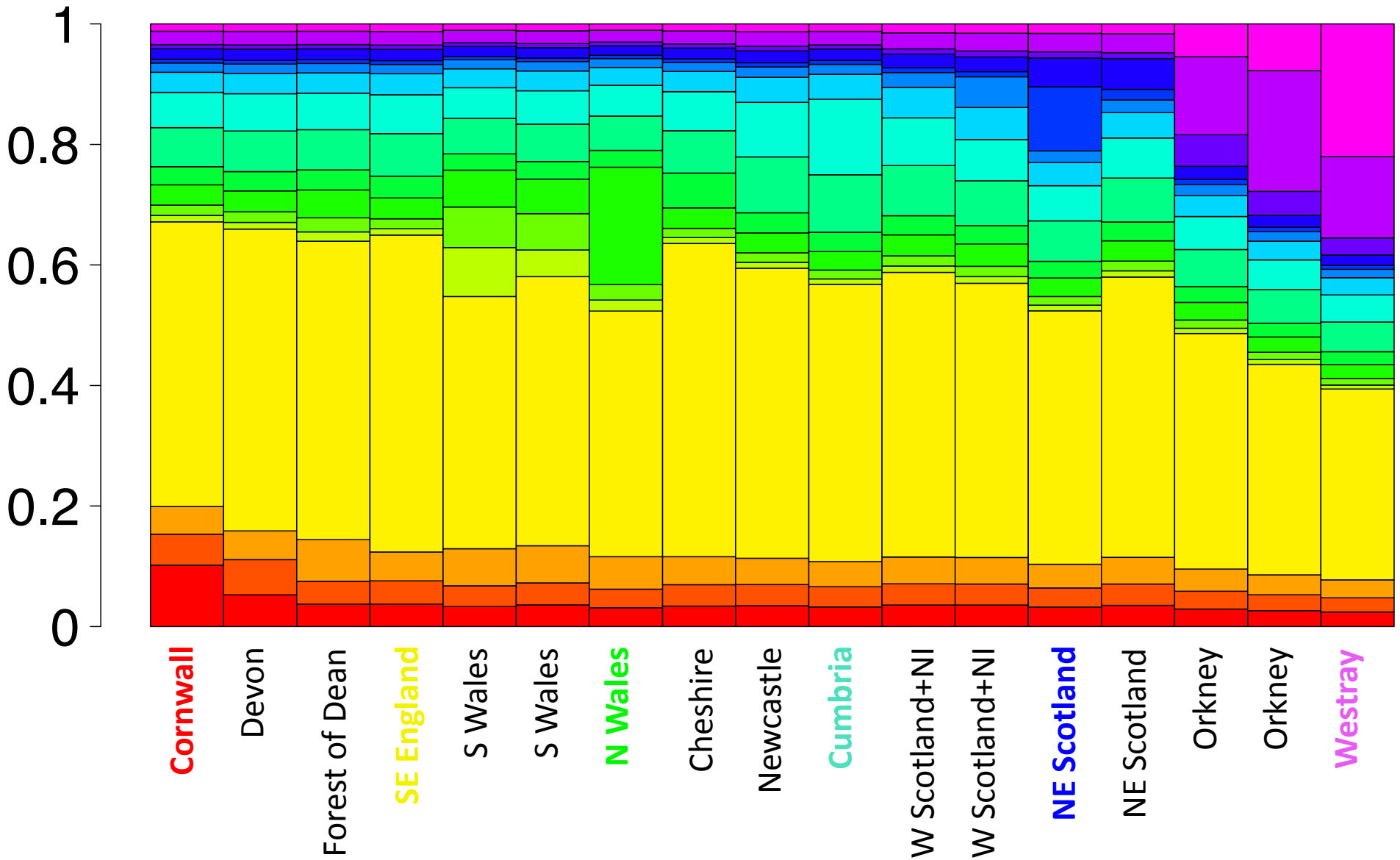
Genotyped using SNP chip. (500,000 markers)

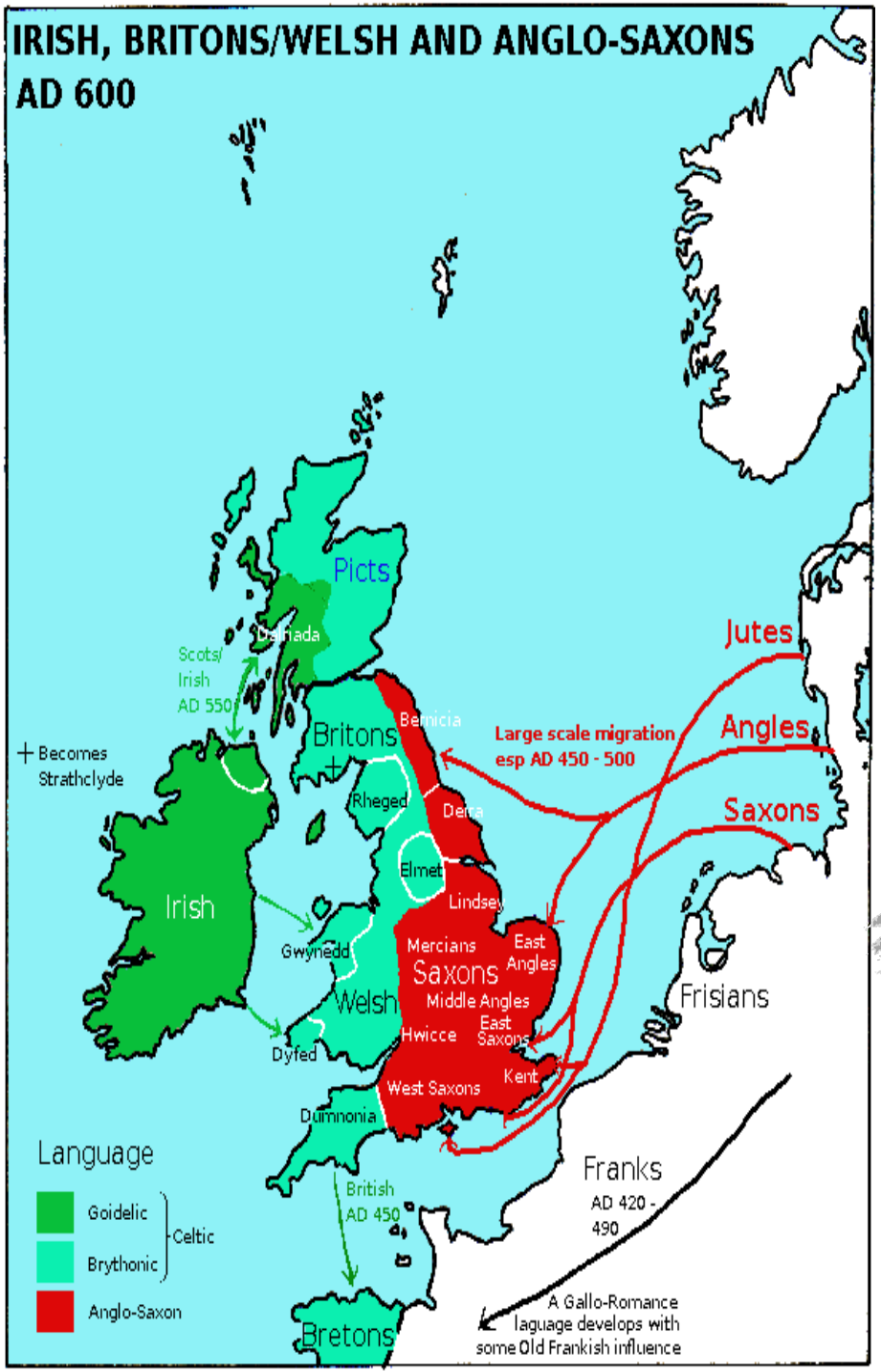
54 distinct palettes inferred by fineSTRUCTURE

Build a tree: by successively joining groups with the most similar palettes.

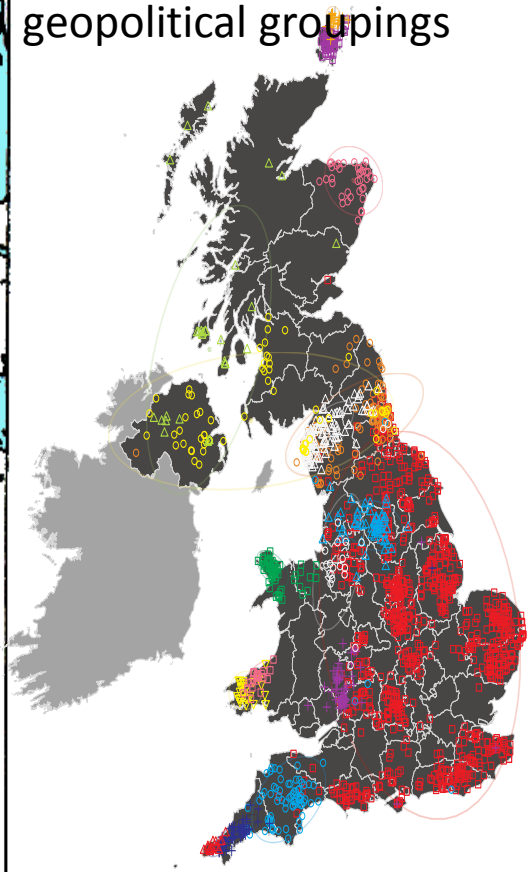
CREDIT: Bruce Winney, Stephen Leslie, Walter Bodmer, Peter Donnelly

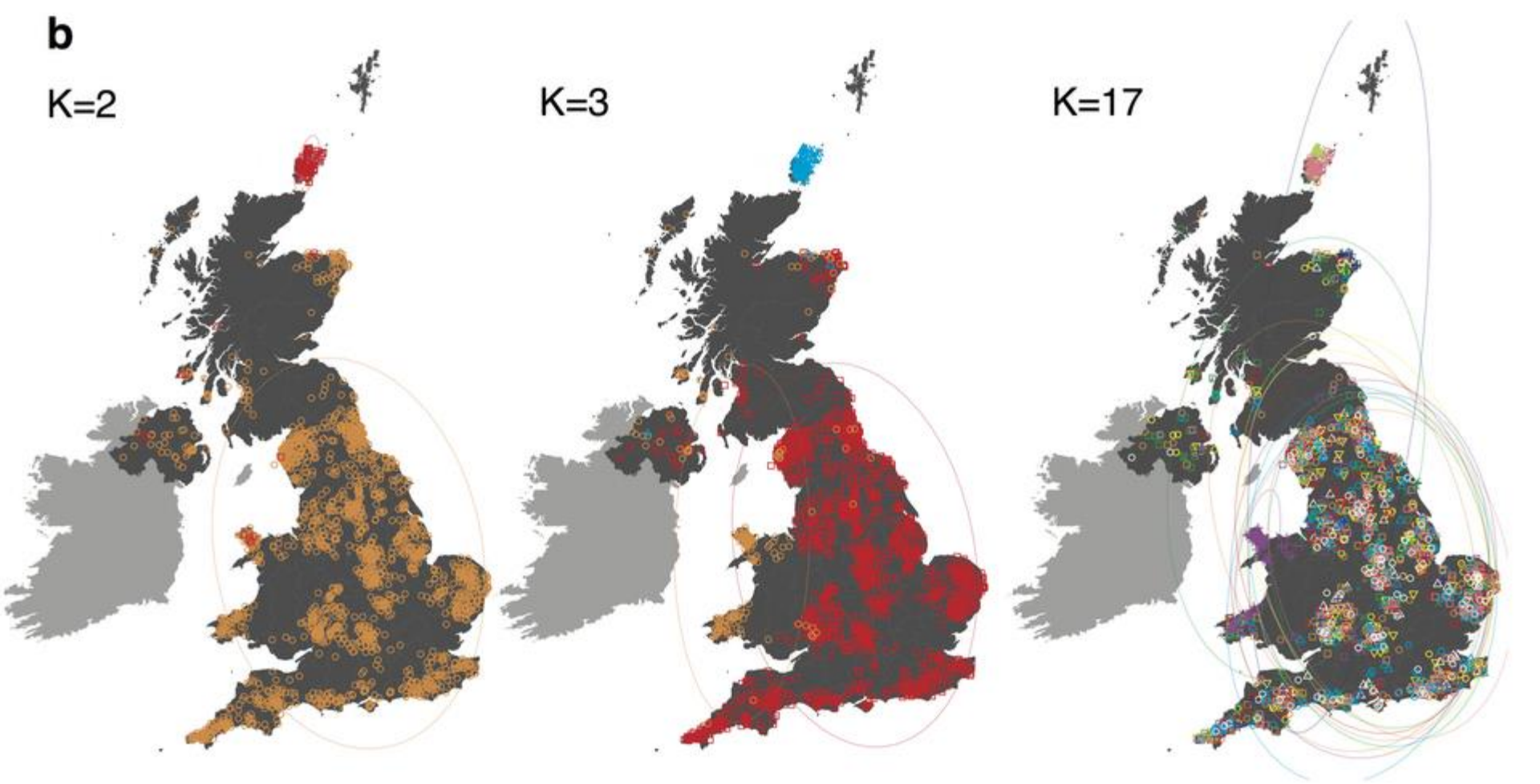
British palettes



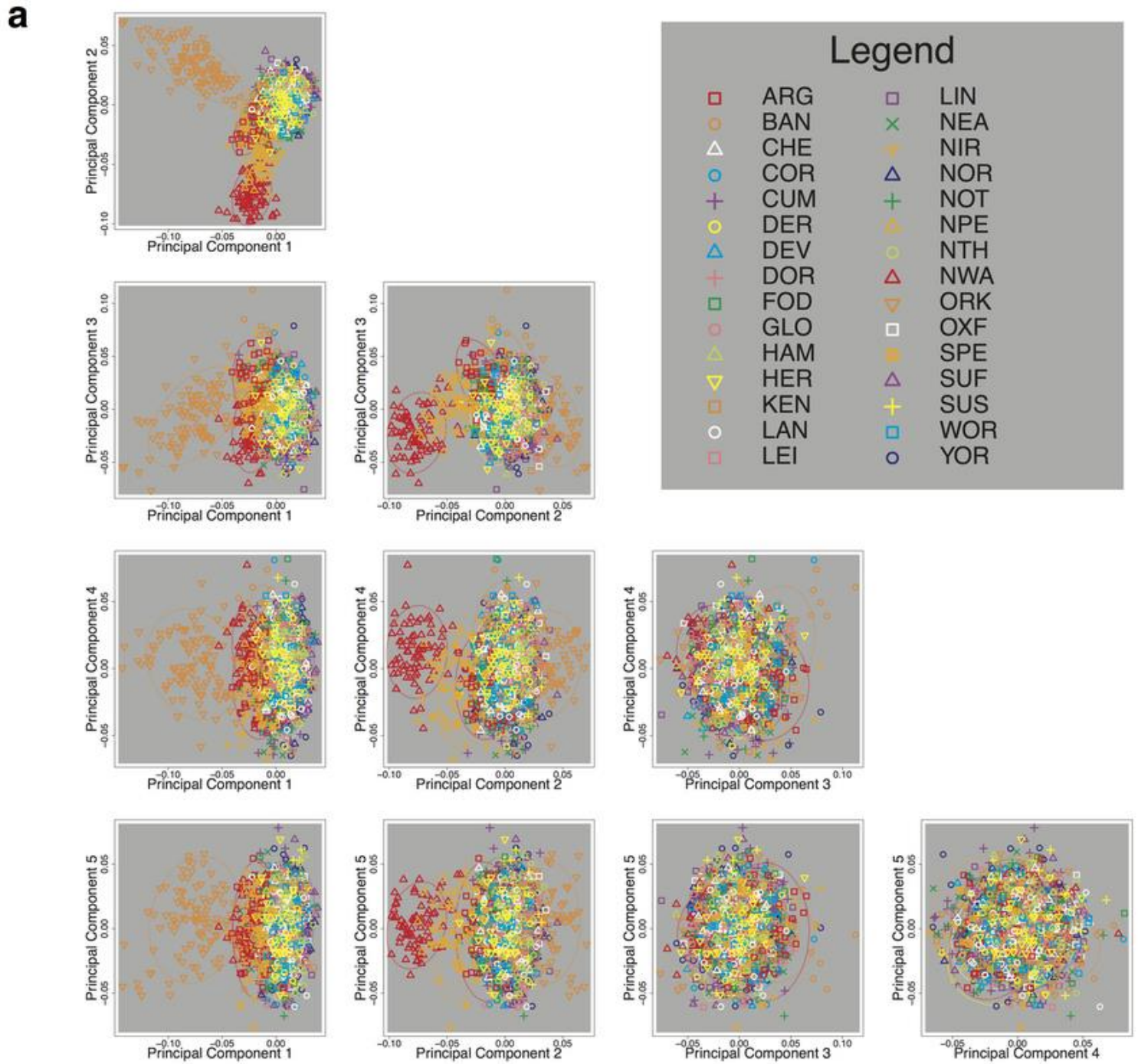


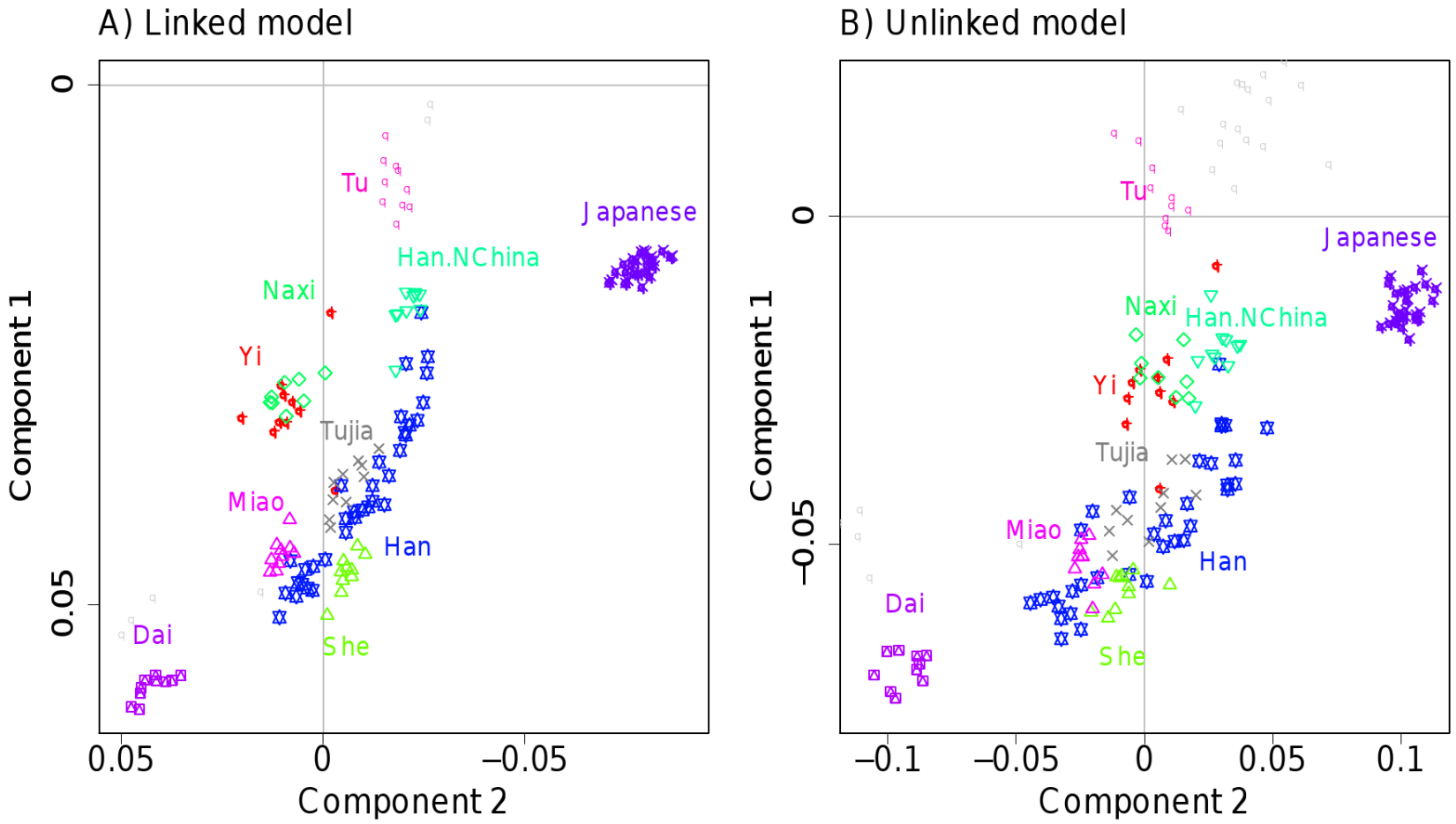
Clusters very strongly resemble ancient geopolitical groupings



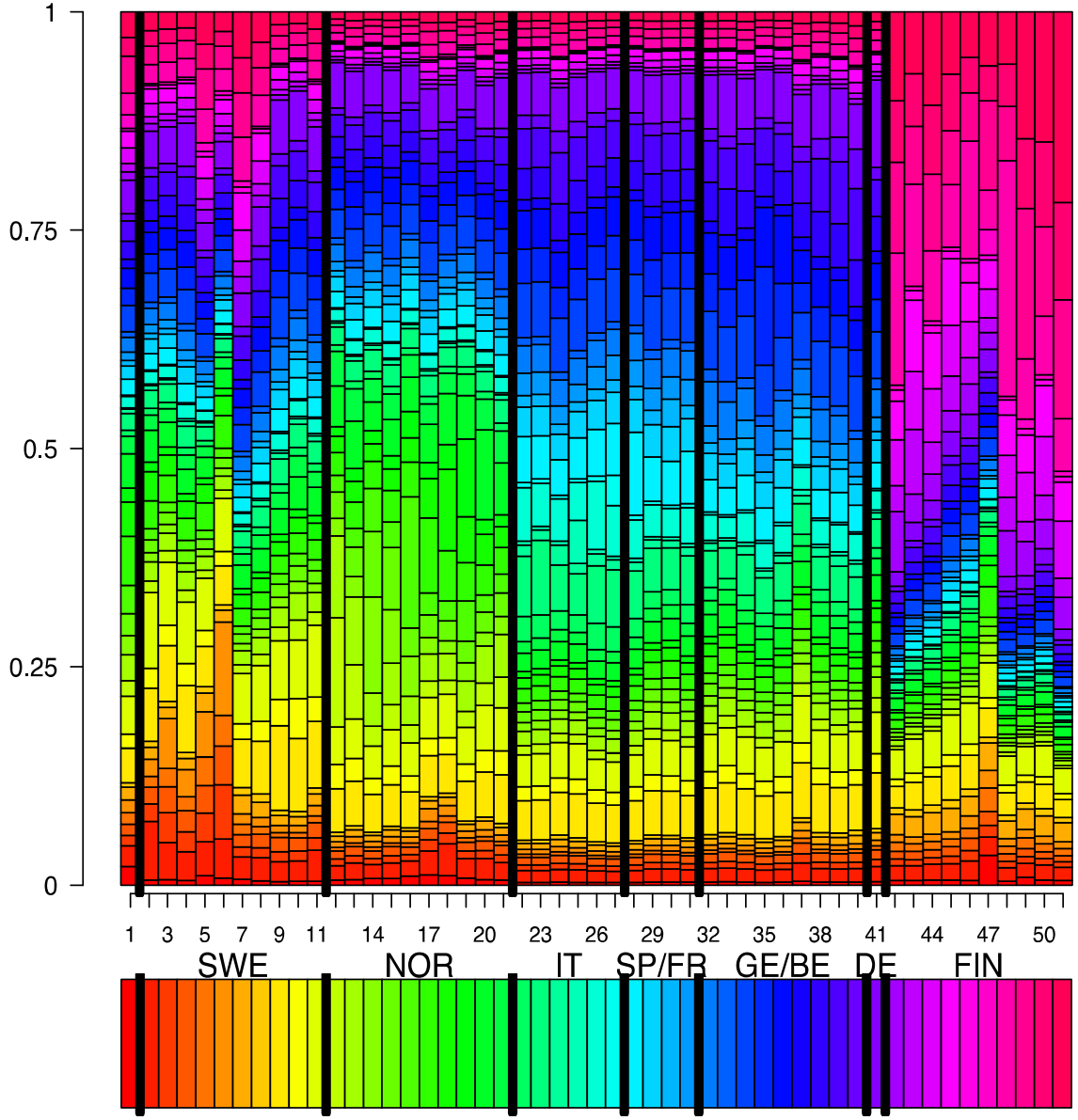


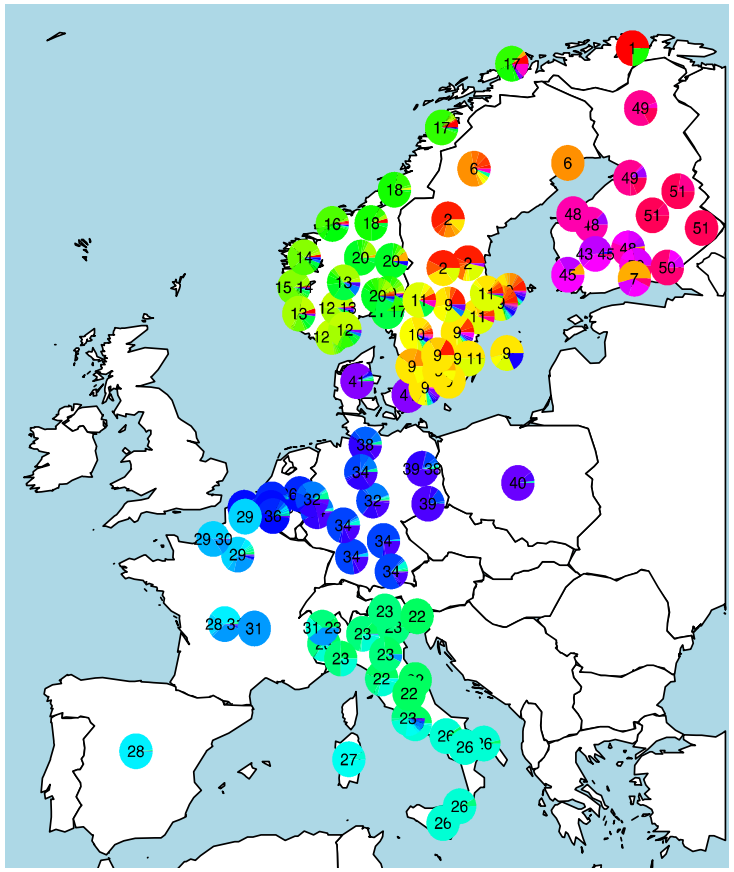
Traditional software (Admixture)


a
Legend
ARG
BAN
CHE
COR
CUM
DER
DEV
DOR
FOD
GLO
HAM
HER
KEN
LAN
LEI
LIN
NEA
NIR
NOR
NOT
NPE
NTH
NWA
ORK
OXF
SPE
SUF
SUS
WOR
YOR
Principal Component 1
Principal Component 2
Principal Component 3
Principal Component 4
Principal Component 5


A) Linked model
B) Unlinked model
Component 1
Component 2
0
0.05
−0.05
0.05
0
−0.05
−0.1
0.1
Tu
Japanese
Han.NChina
Naxi
Yi
Tujia
Miao
Han
Dai
She

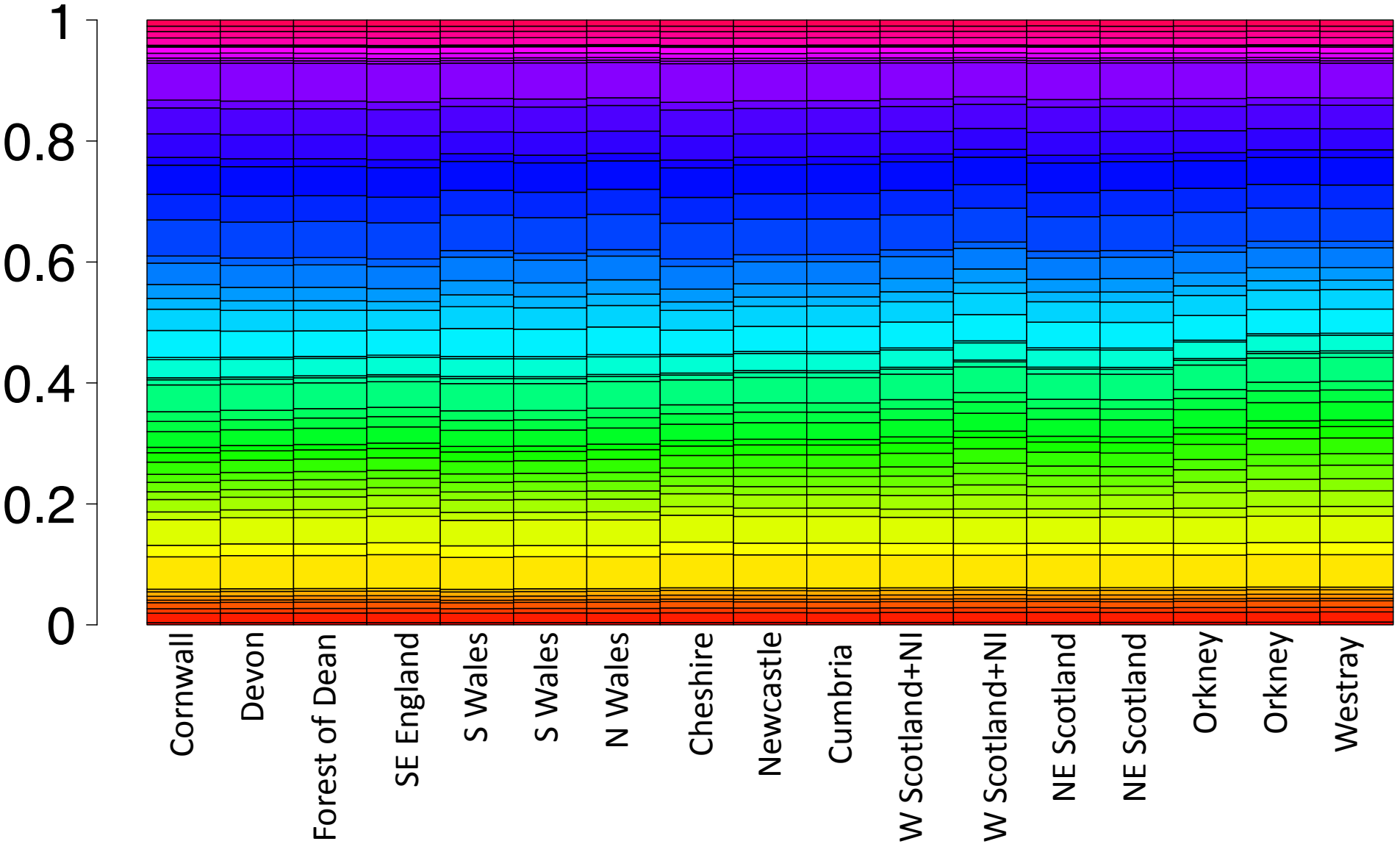
Continental European palettes



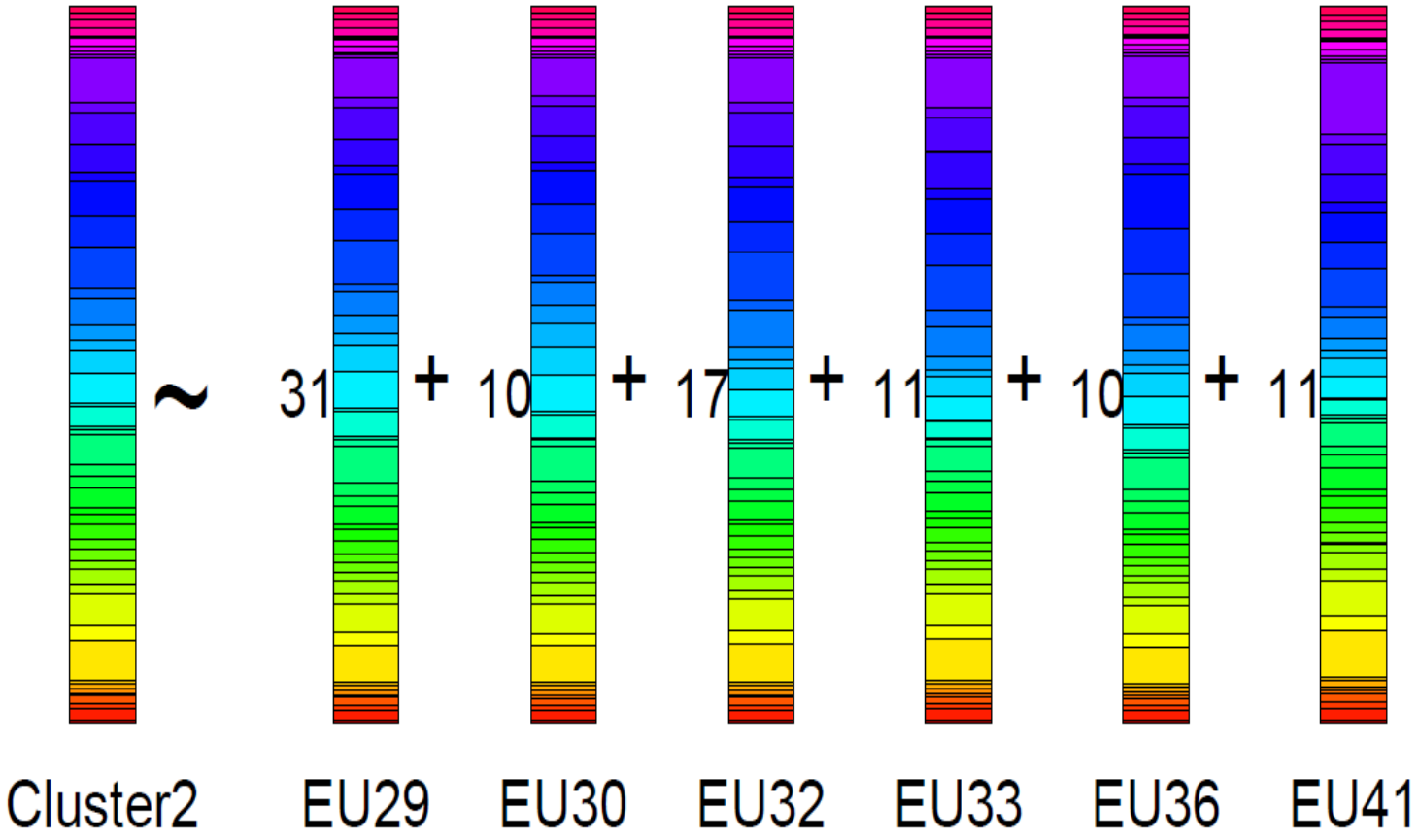


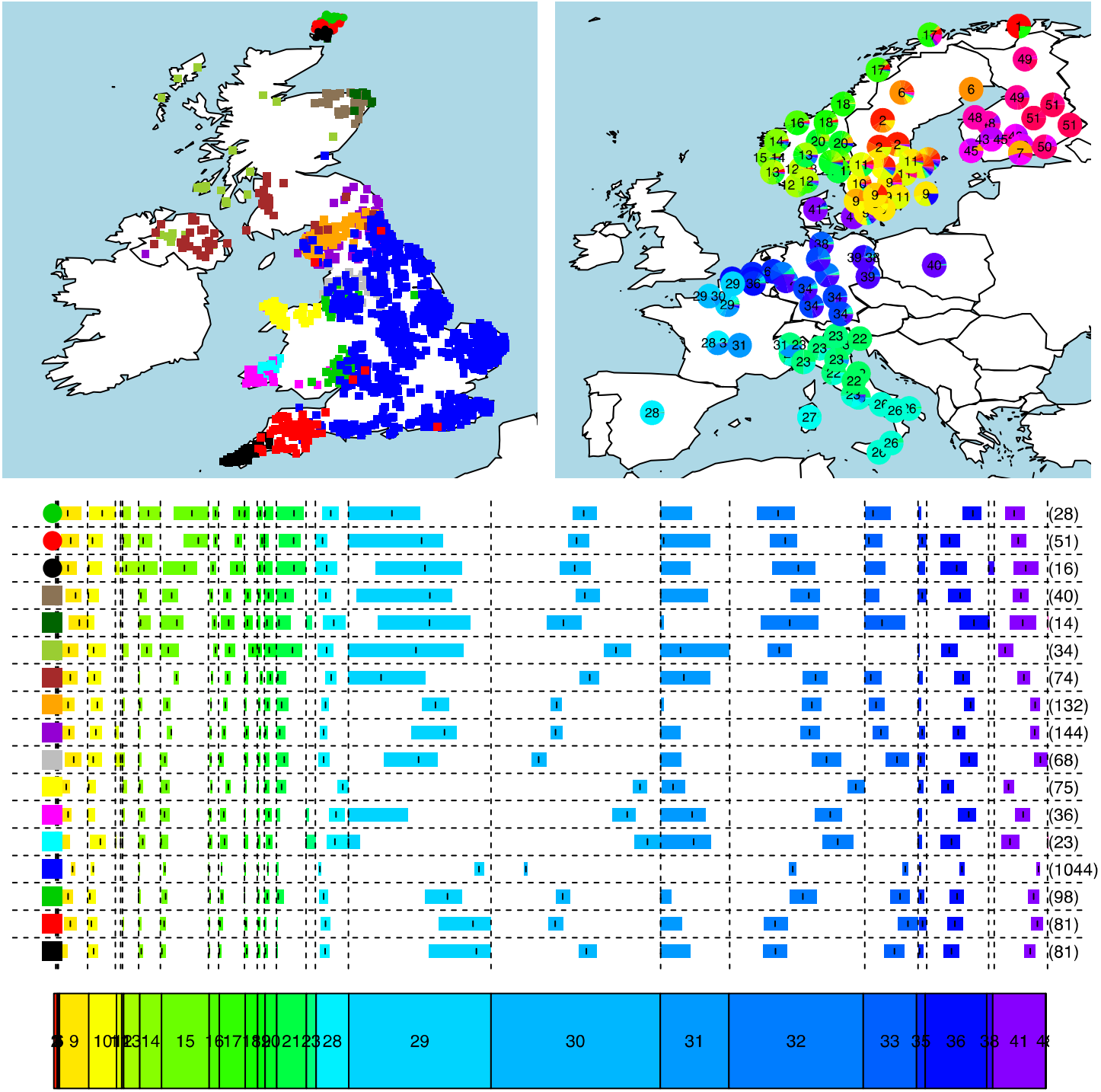
Hospital based sampling as controls for an association study with FineSTRUCTURE used to identify populations

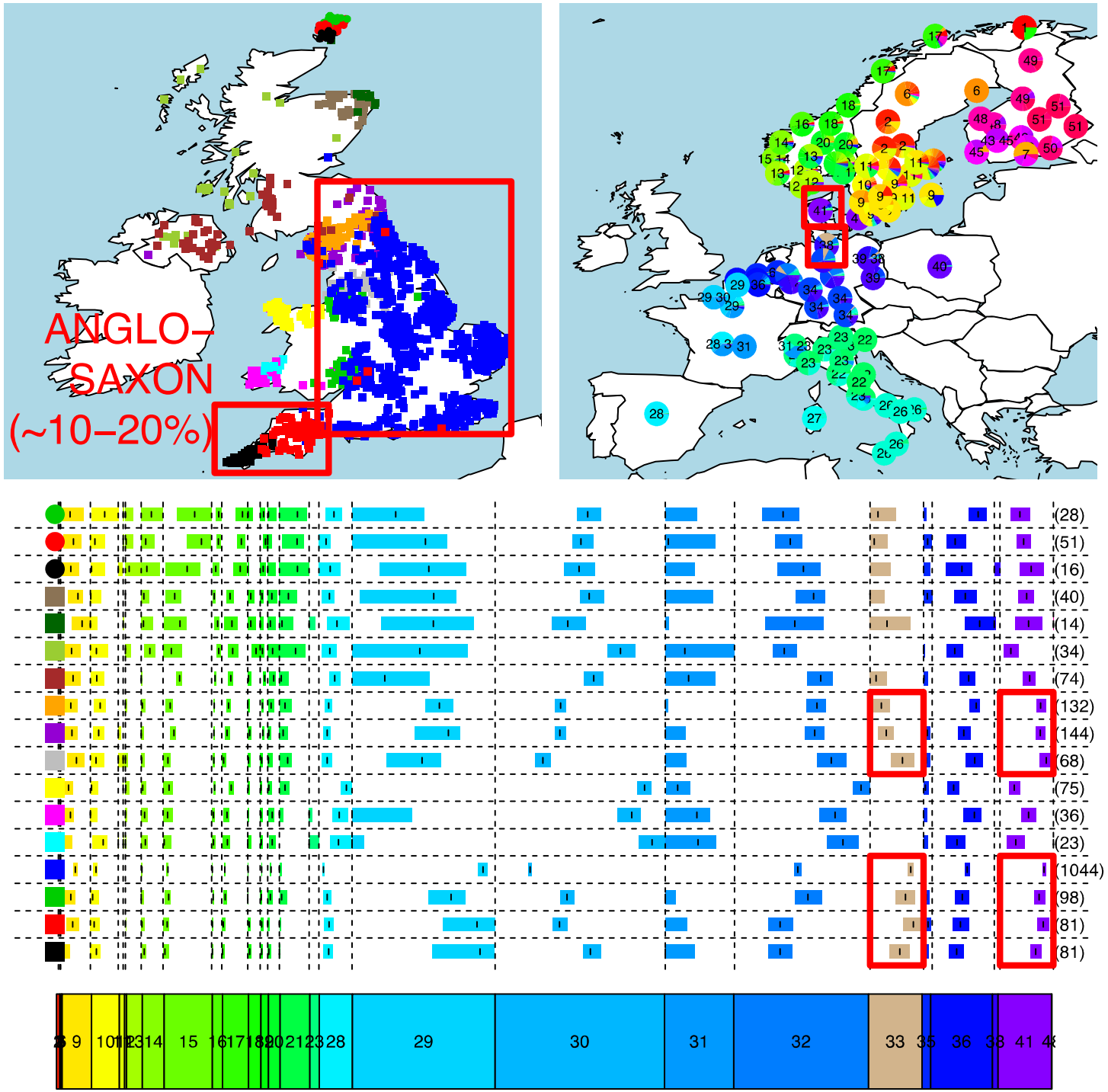
British palettes
painted with a continental European panel

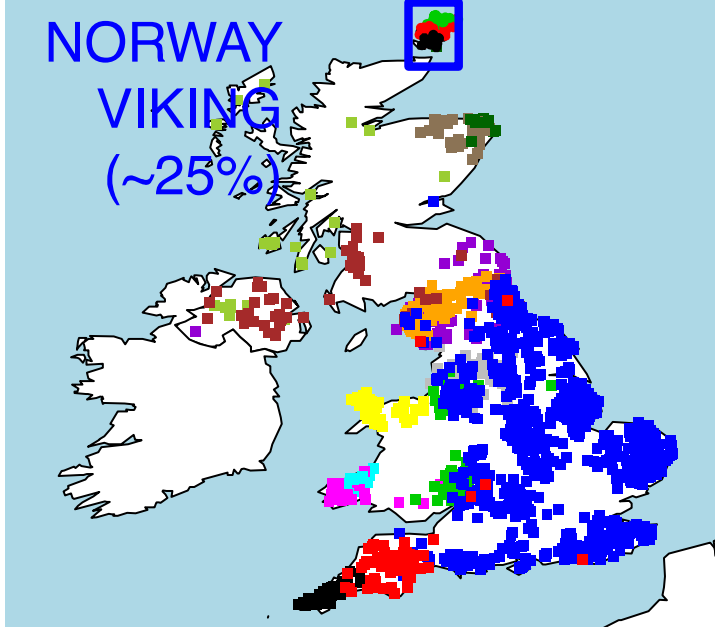


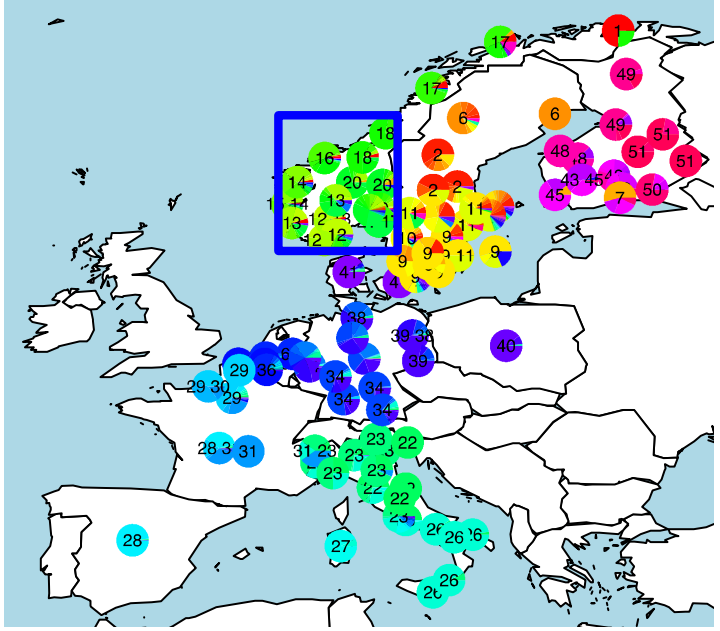
Mixture modelling of an English palette based on European palettes using **N**on-**N**egative **L**east **S**quares

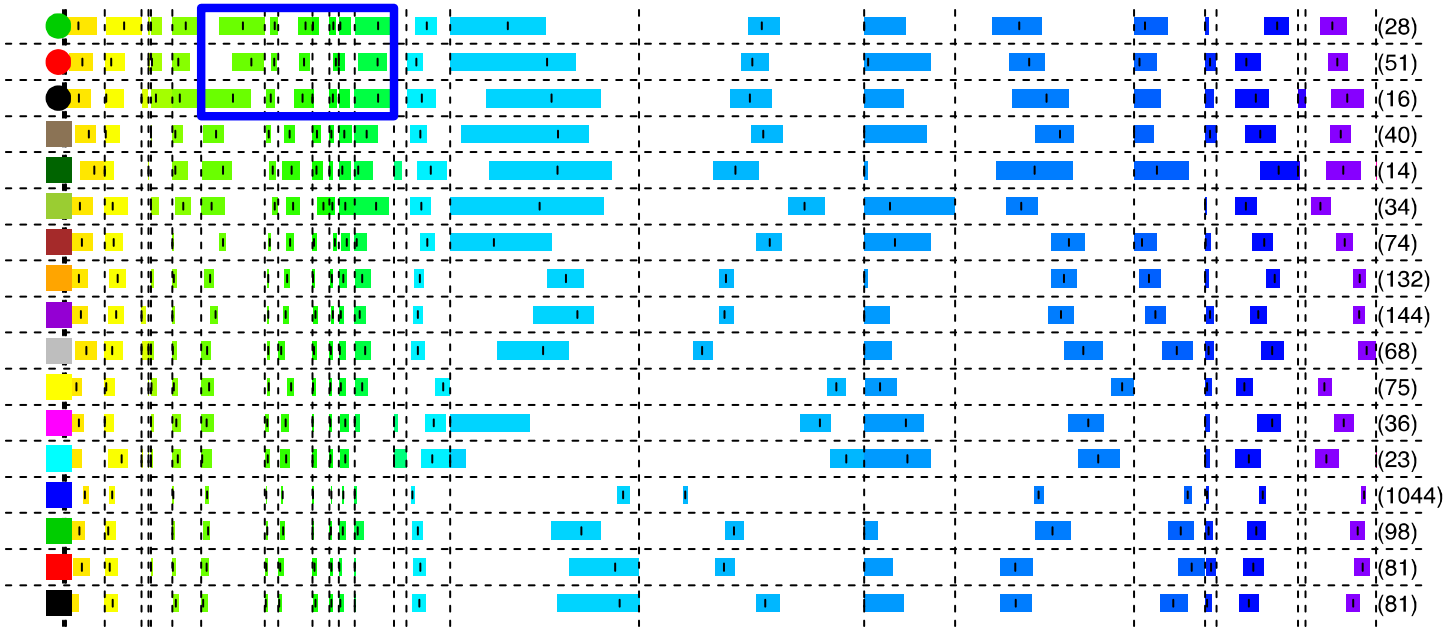


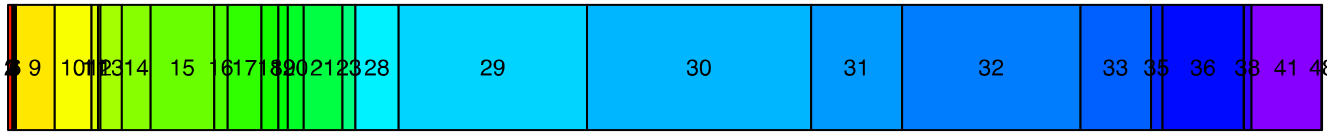

(28)
(51)
(16)
(40)
(14)
(34)
(74)
(132)
(144)
(68)
(75)
(36)
(23)
(1044)
(98)
(81)
(81)
29
30
31
32


ANGLO–
SAXON
(~10–20%)
(28)
(51)
(16)
(40)
(14)
(34)
(74)
(132)
(144)
(68)
(75)
(36)
(23)
(1044)
(98)
(81)
(81)
29
30
31
32
33
36

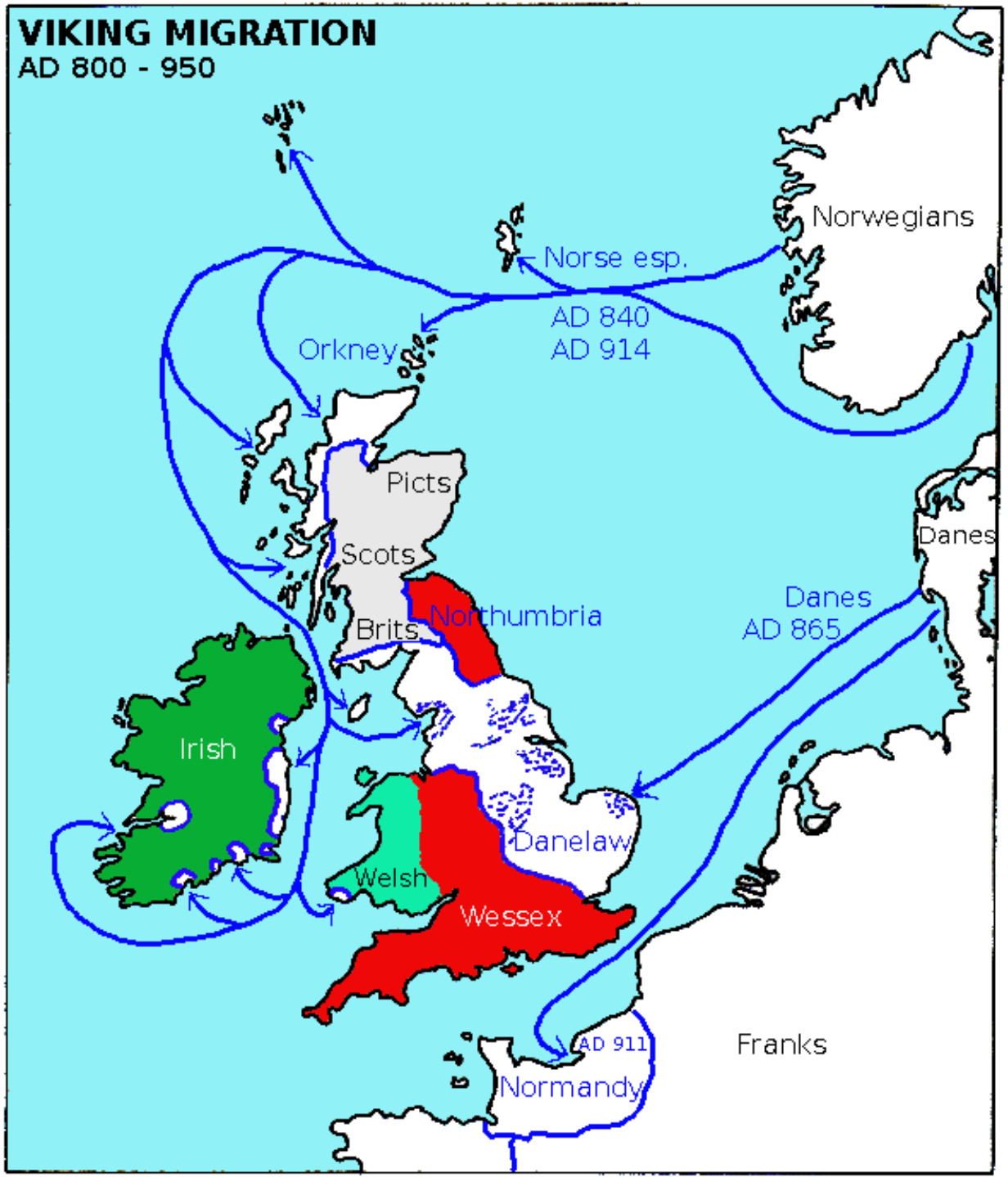

NORWAY
VIKING
(~25%)




(28)
(51)
(16)
(40)
(14)
(34)
(74)
(132)
(144)
(68)
(75)
(36)
(23)
(1044)
(98)
(81)
(81)


8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 28 29 30 31 32 33 35 36 38 41

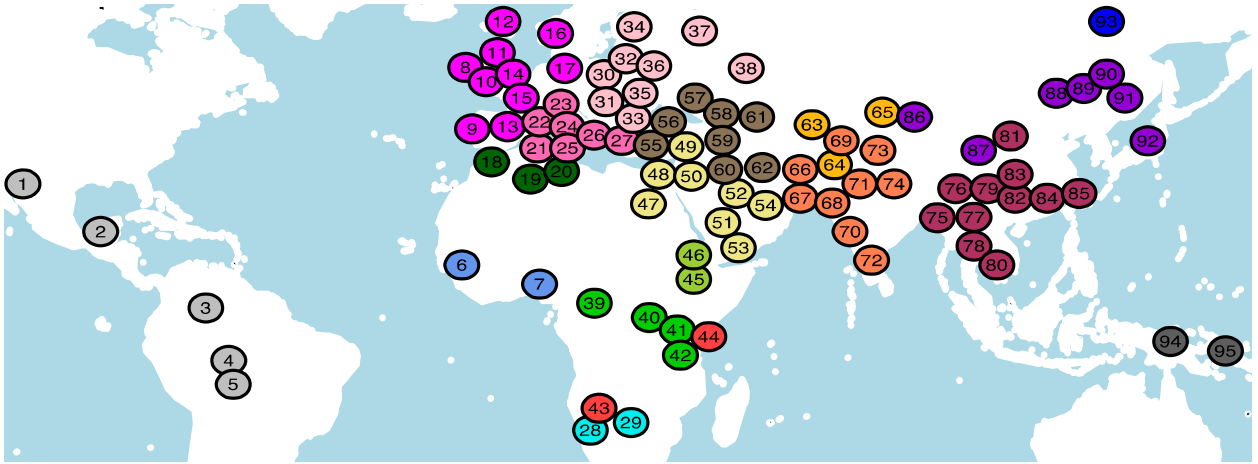
e



Danish Vikings colonisations from v. similar region to previous Anglo-Saxons

World-wide dataset application (HGDP+more)

≈475K SNPs on 1490 individuals from 95 pops (5-45 inds/pop)



"AMERICAS"
1. Pima
2. Maya
3. Colombian
4. Karitiana
5. Surui

"W.AFRICA"
6. Mandenka
7. Yoruba

"N.W.EUROPE"
8. Ireland
9. Spanish
10. Welsh
11. Scottish
12. Orcadian
13. Basque
14. English
15. French
16. Norwegian
17. GermanyAustria

"N.AFRICA"
18. Moroccan
19. Mozabite
20. Tunisian

"S.EUROPE"
21. WestSicilian
22. Sardinian
23. NorthItalian
24. Tuscan
25. EastSicilian
26. SouthItalian
27. Greek

"SAN"
28. SanNamibia
29. SanKhomani

"E.EUROPE"
30. Polish
31. Hungarian
32. Lithuanian
33. Bulgarian
34. Finnish
35. Romanian
36. Belorussian
37. Russian
38. Chuvash

"C.AFRICA"
39. BiakaPygmy
40. MbutiPygmy
41. Hadza
42. Sandawe

"BANTU"
43. BantuSouthAfrica
44. BantuKenya

"ETHIOPIAN"
45. Ethiopian
46. EthiopianJewII

"S.MIDDLEEAST"
47. Egyptian
48. Palestinian
49. Syrian
50. Jordanian
51. Saudi
52. Bedouin
53. Yemeni
54. UAE

"W.ASIA"
55. Cypriot
56. Turkish
57. Adygei
58. Georgian
59. Armenian
60. Druze
61. Lezgin
62. Iranian

"C.SOUTHASIA2"
63. Uzbekistani
64. Hazara
65. Uygur

"C.SOUTHASIA"
66. Brahui
67. Makrani
68. Sindhi
69. Kalash
70. IndianJew
71. Balochi
72. Indian
73. Burusho
74. Pathan

"S.EASTASIA"
75. Myanmar
76. Naxi
77. Dai
78. Lahu
79. Yi
80. Cambodian
81. HanNchina
82. Miao
83. Tujia
84. Han
85. She

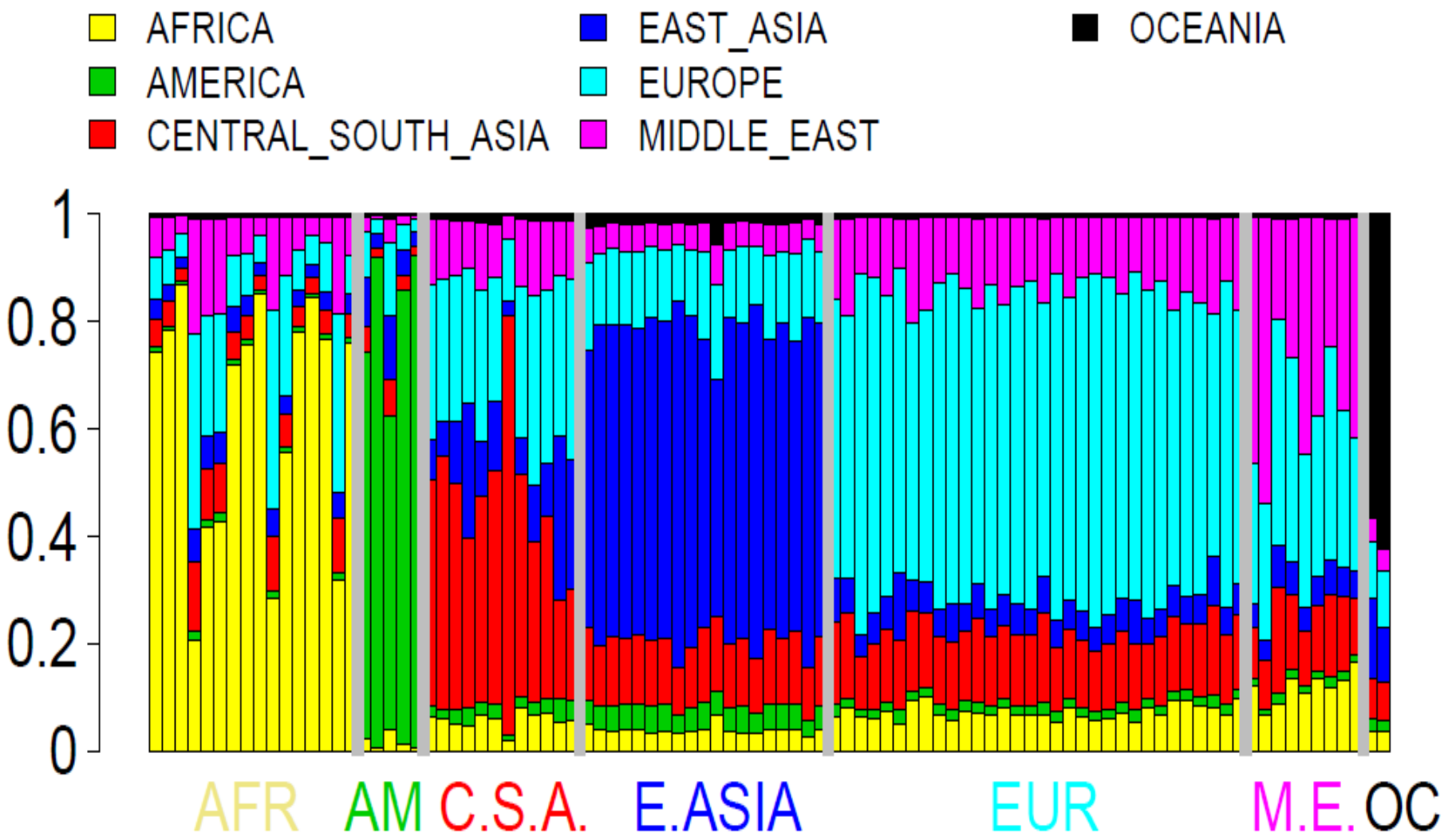
"N.EASTASIA"
86. Xibo
87. Tu
88. Mongola
89. Daur
90. Oroqen
91. Hezhen
92. Japanese

"YAKUT"
93. Yakut

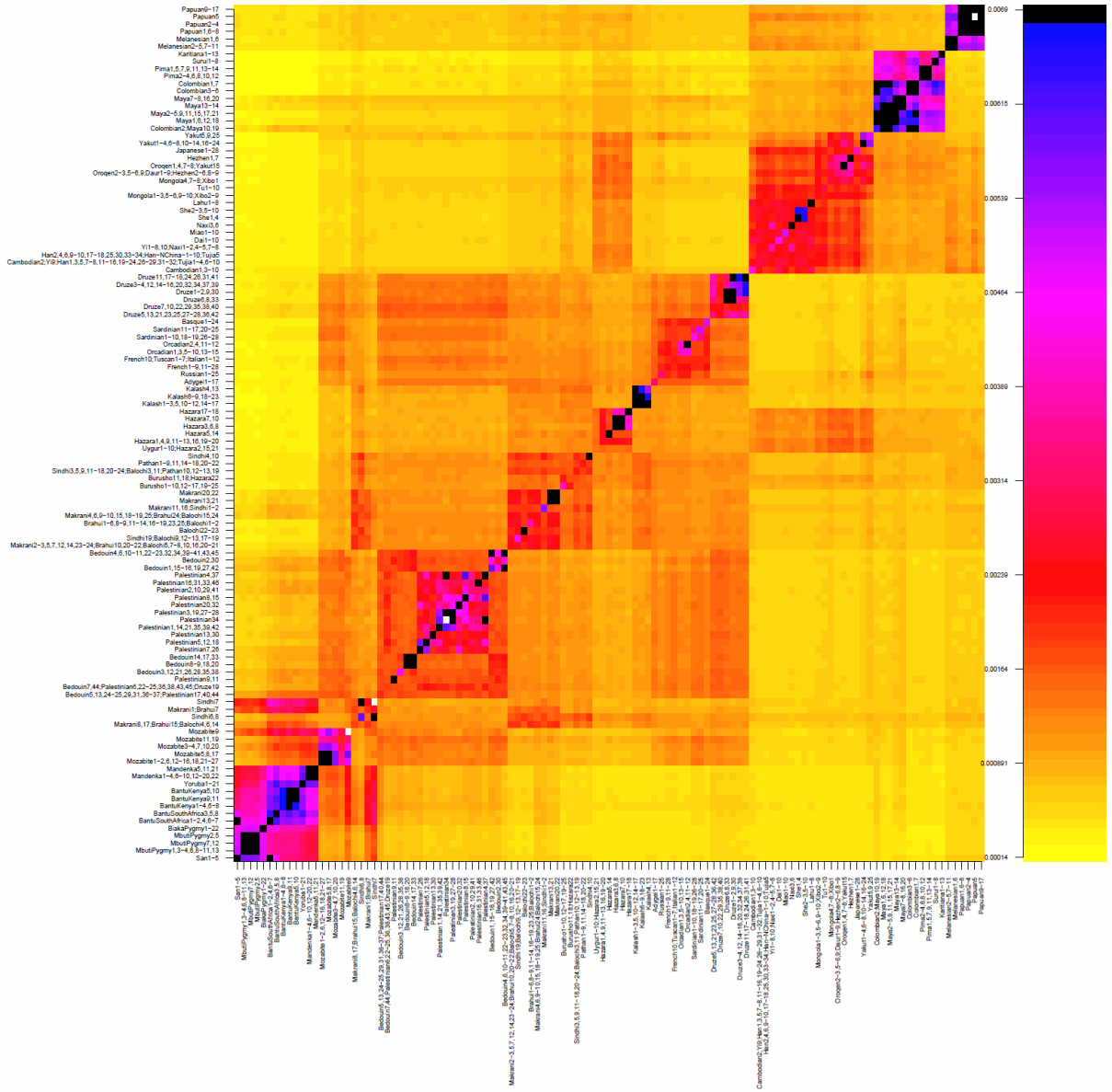
"OCEANIA"
94. Papuan
95. Melanesian

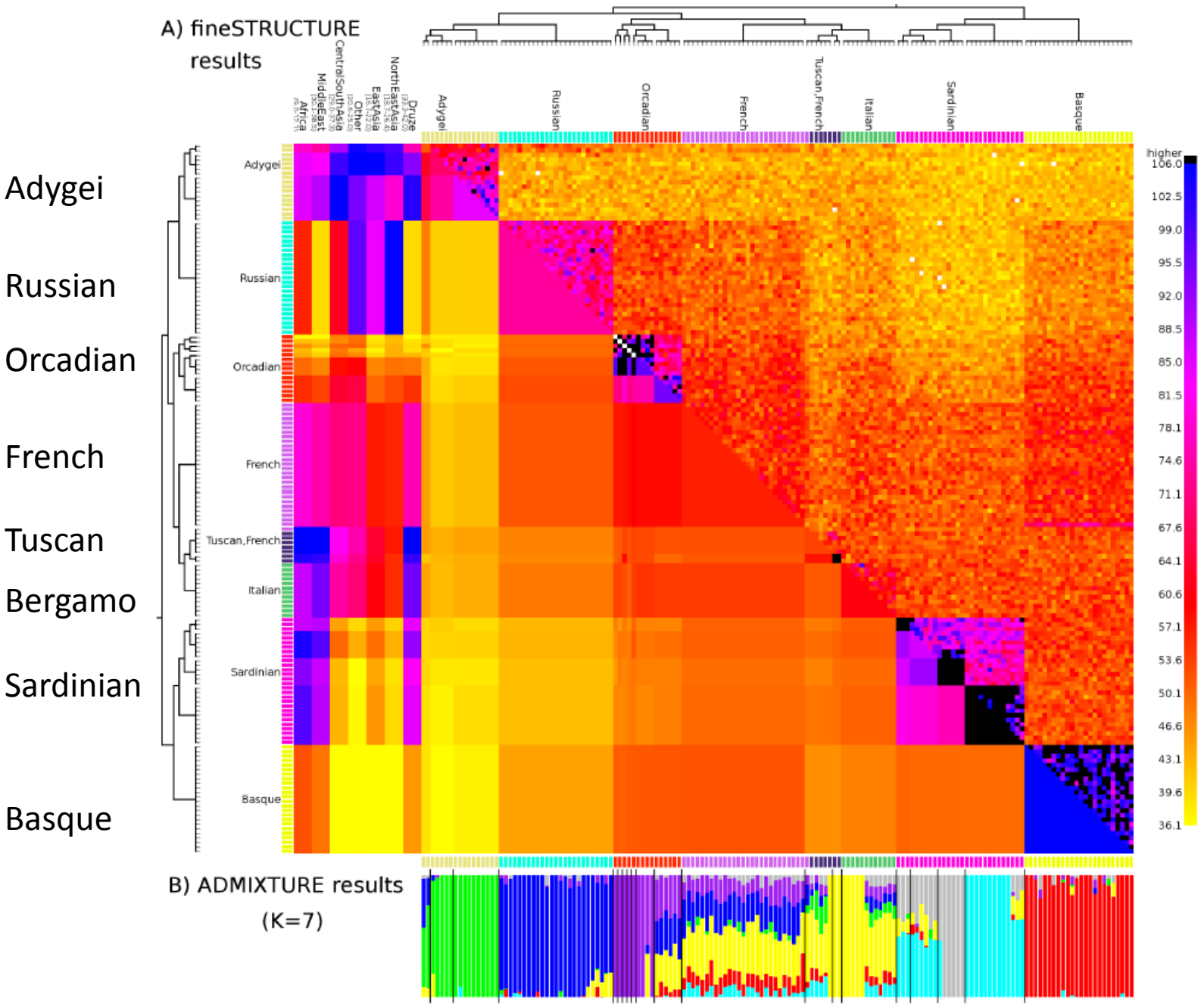
CREDIT: Christian Capelli, George Busby

Worldwide palettes

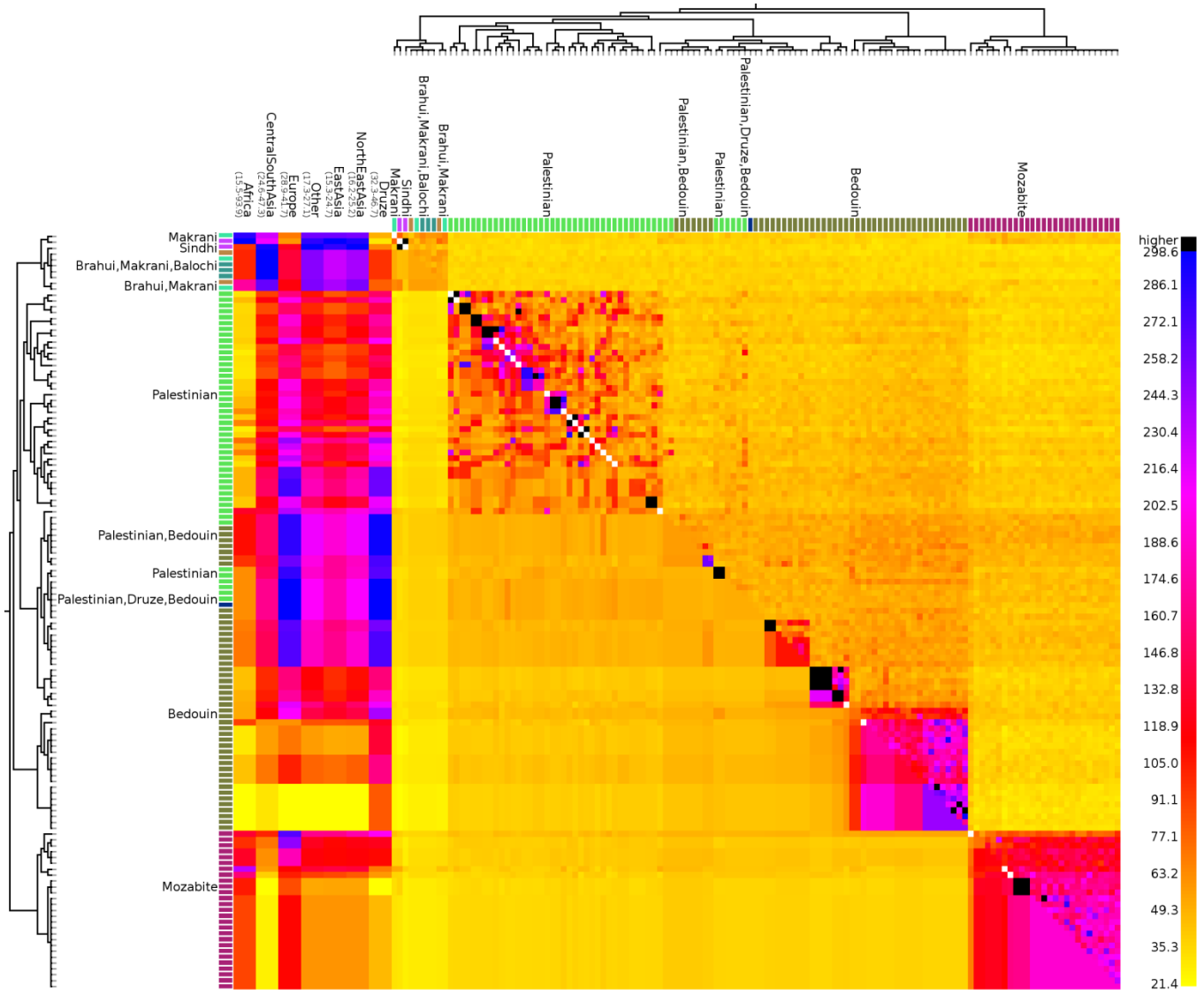


169 populations inferred in HGDP data

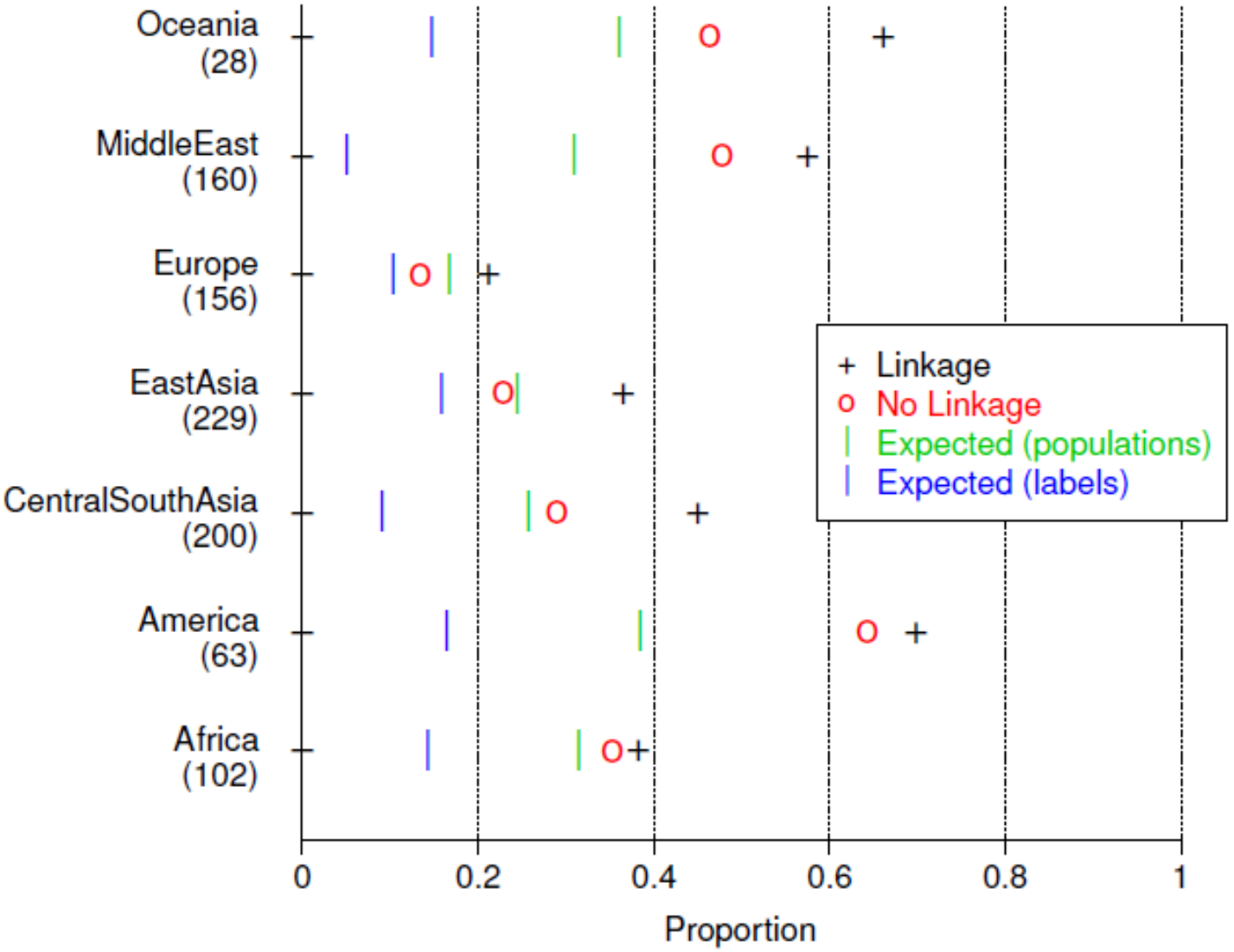



A) fineSTRUCTURE
results
Adygei
Russian
Orcadian
French
Tuscan
Bergamo
Sardinian
Basque
Adygei
Russian
Orcadian
French
Tuscan,French
Italian
Sardinian
Basque
Druze
NorthEastAsia
EastAsia
Other
CentralSouthAsia
MiddleEast
Africa
higher
106.0
102.5
99.0
95.5
92.0
88.5
85.0
81.5
78.1
74.6
71.1
67.6
64.1
60.6
57.1
53.6
50.1
46.6
43.1
39.6
36.1
B) ADMIXTURE results
(K=7)

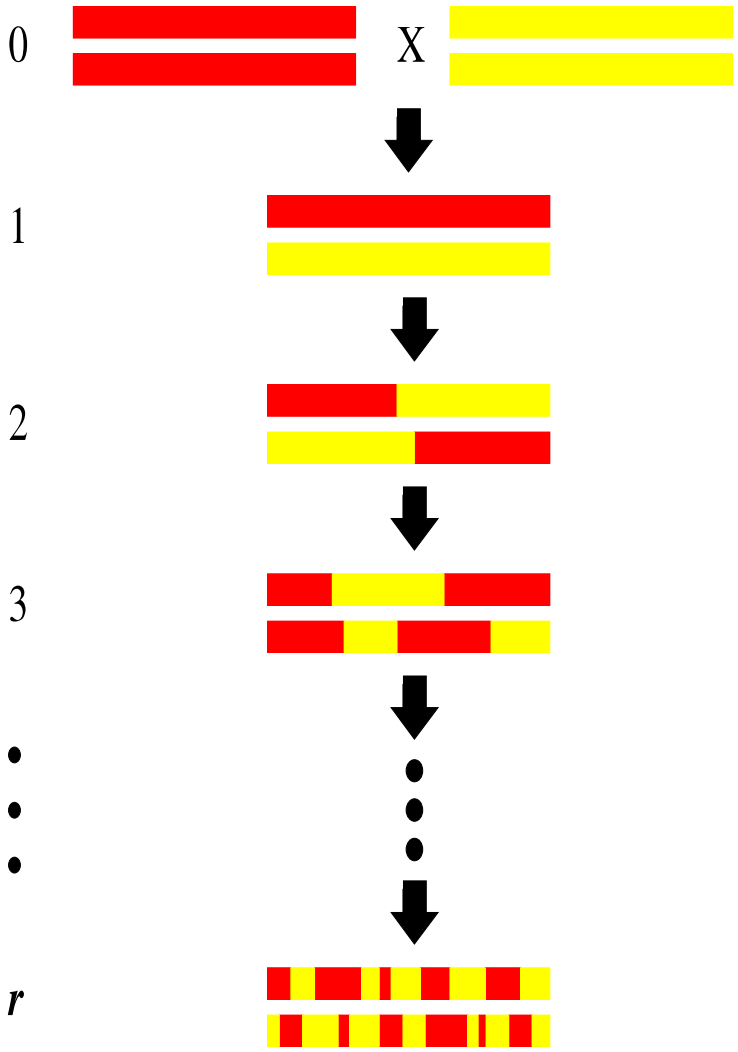
HGDP Middle east



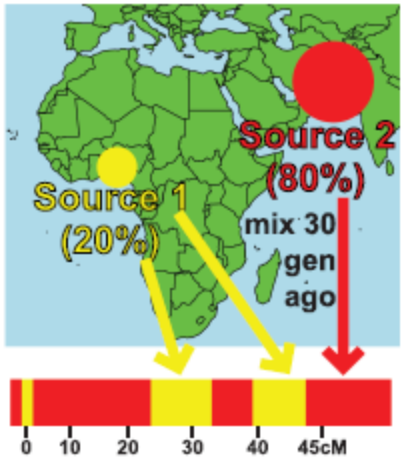
Half-matching based on coancestry matrix



Information on time since mixture given by spatial structure of ancestry along the chromosome

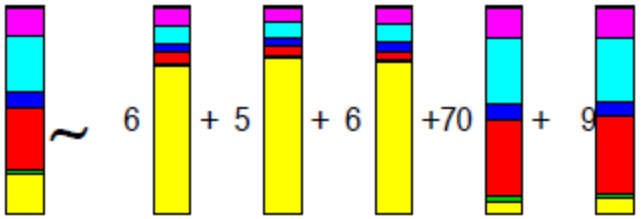


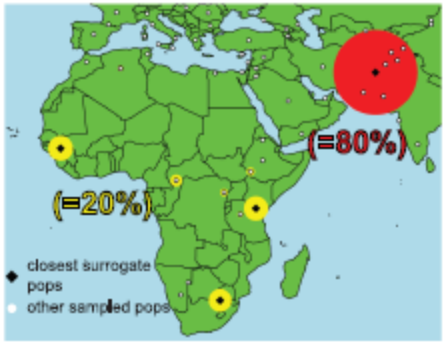
Globetrotter



Simulated admixture
30 generations ago

Mixture modelling of genomic palettes

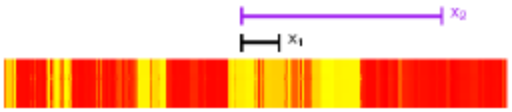


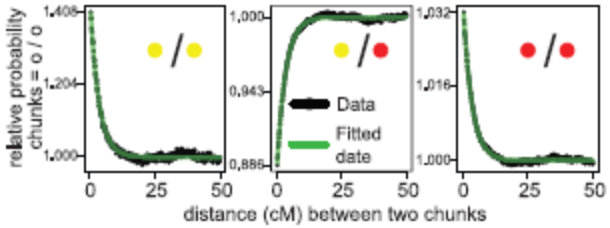


Raw painting

Cleaned painting

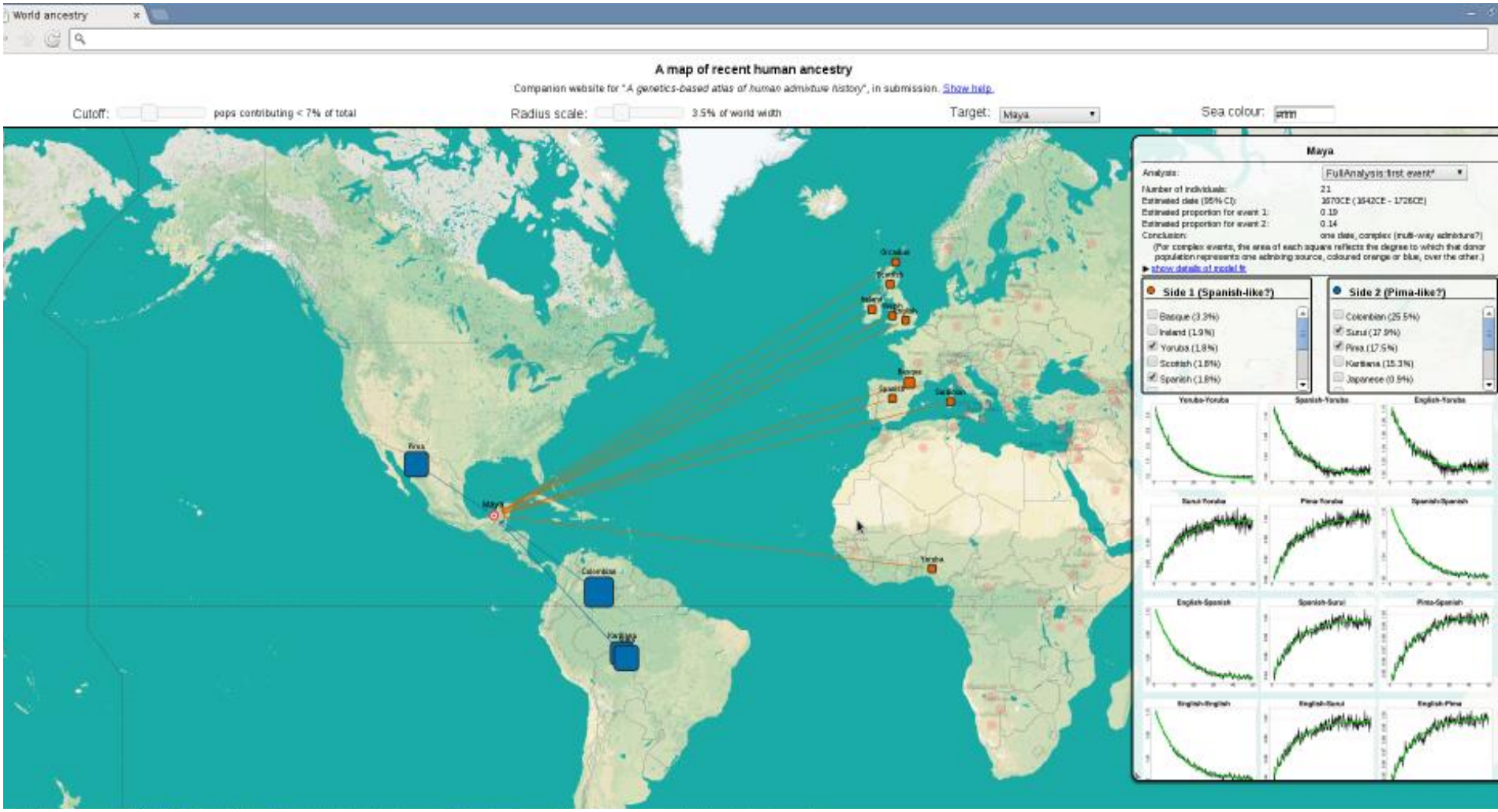
Fitting of spatial structure of variation in palettes along the chromosome

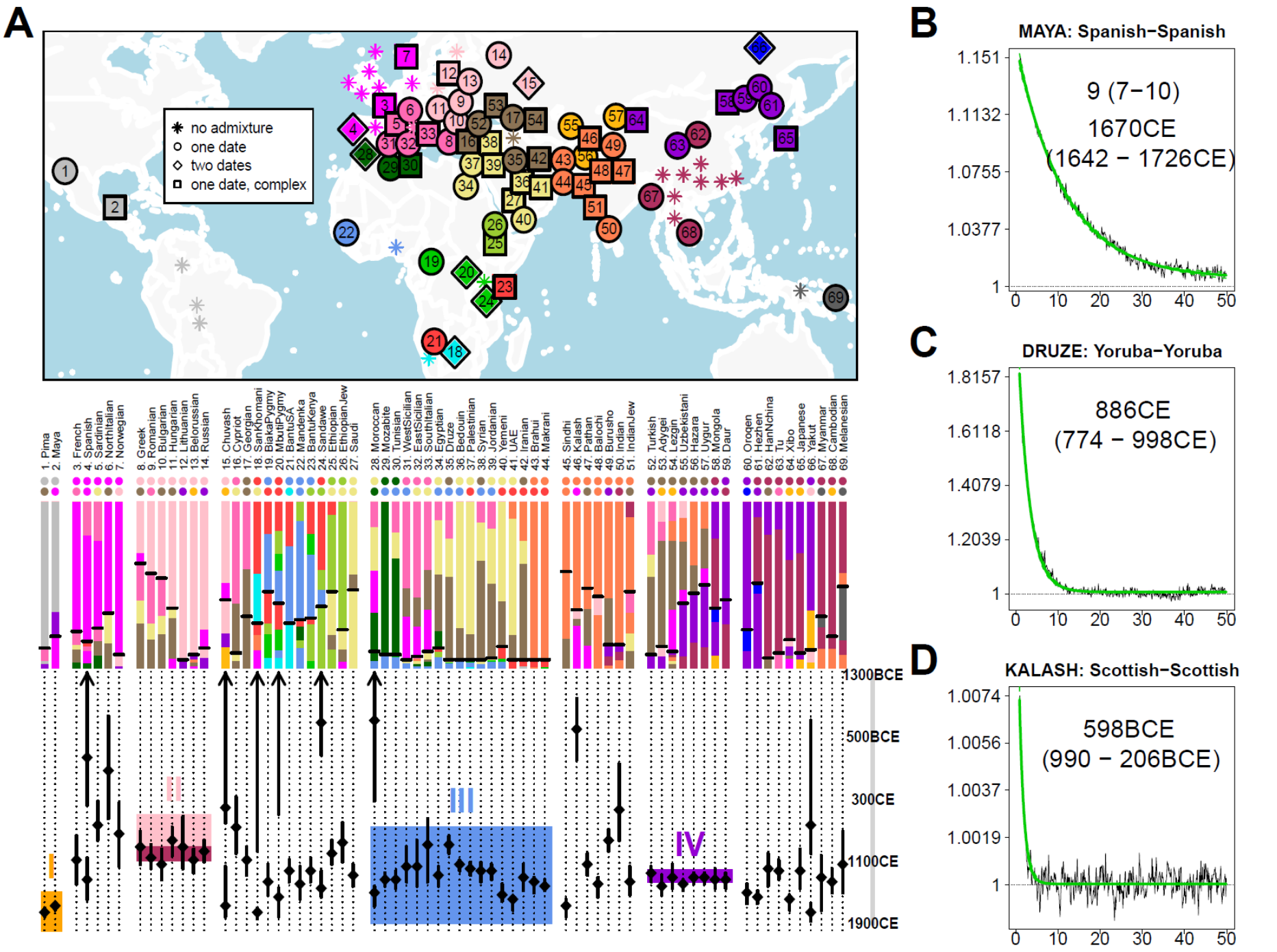




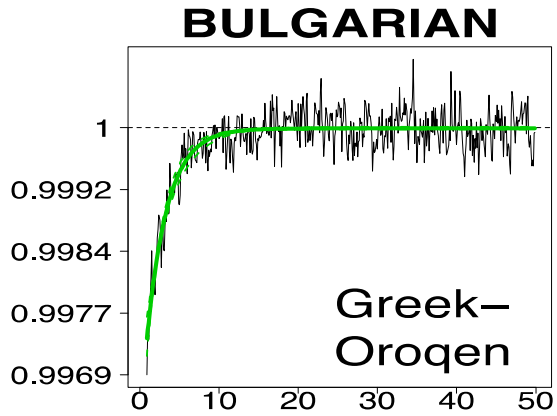
http://admixturemap.paintmychromosomes.com

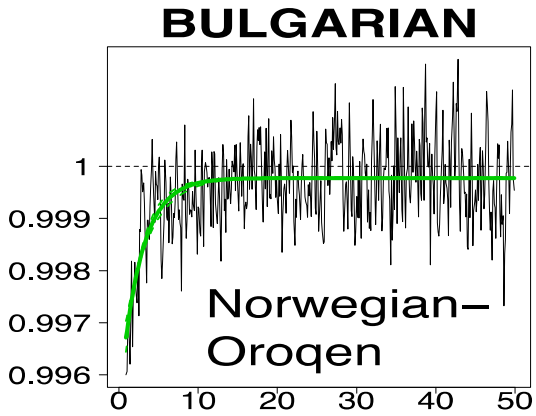
(Hellenthal et al 2014)

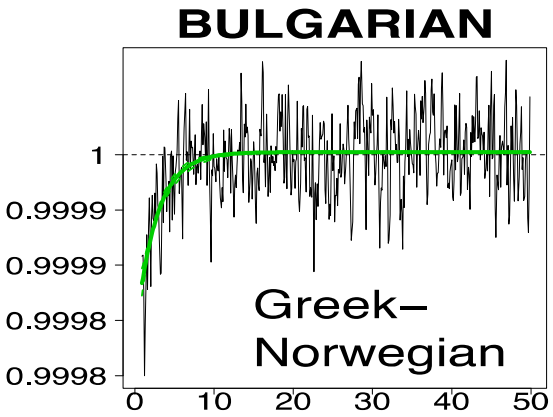



A
no admixture
one date
two dates
one date, complex
1. Pima
2. Maya
3. French
4. Spanish
5. Sardinian
6. NorthItalian
7. Norwegian
8. Greek
9. Romanian
10. Bulgarian
11. Hungarian
12. Lithuanian
13. Belorussian
14. Russian
15. Chuvash
16. Cypriot
17. Georgian
18. SanKhomani
19. BiakaPygmy
20. MbutiPygmy
21. BantuSA
22. Mandenka
23. BantuKenya
24. Sandawe
25. Ethiopian
26. Ethiopian-Jew
27. Saudi
28. Moroccan
29. Mozabite
30. Tunisian
31. WestSicilian
32. EastSicilian
33. SouthItalian
34. Egyptian
35. Druze
36. Bedouin
37. Palestinian
38. Syrian
39. Jordanian
40. Yemeni
41. UAE
42. Iranian
43. Brahui
44. Makrani
45. Sindhi
46. Kalash
47. Pathan
48. Balochi
49. Burusho
50. Indian
51. IndianJew
52. Turkish
53. Adygei
54. Lezgin
55. Uzbekistani
56. Hazara
57. Uygur
58. Mongola
59. Daur
60. Orogen
61. Hezhen
62. HanNchina
63. Tu
64. Xibo
65. Japanese
66. Yakut
67. Myanmar
68. Cambodian
69. Melanesian
1300BCE
500BCE
300CE
1100CE
1900CE
I
II
III
IV
B
MAYA: Spanish−Spanish
9 (7−10)
1670CE
(1642 − 1726CE)
1.151
1.1132
1.0755
1.0377
1
0
10
20
30
40
50
C
DRUZE: Yoruba−Yoruba
886CE
(774 − 998CE)
1.8157
1.6118
1.4079
1.2039
1
0
10
20
30
40
50
D
KALASH: Scottish−Scottish
598BCE
(990 − 206BCE)
1.0074
1.0056
1.0037
1.0019
1
0
10
20
30
40
50

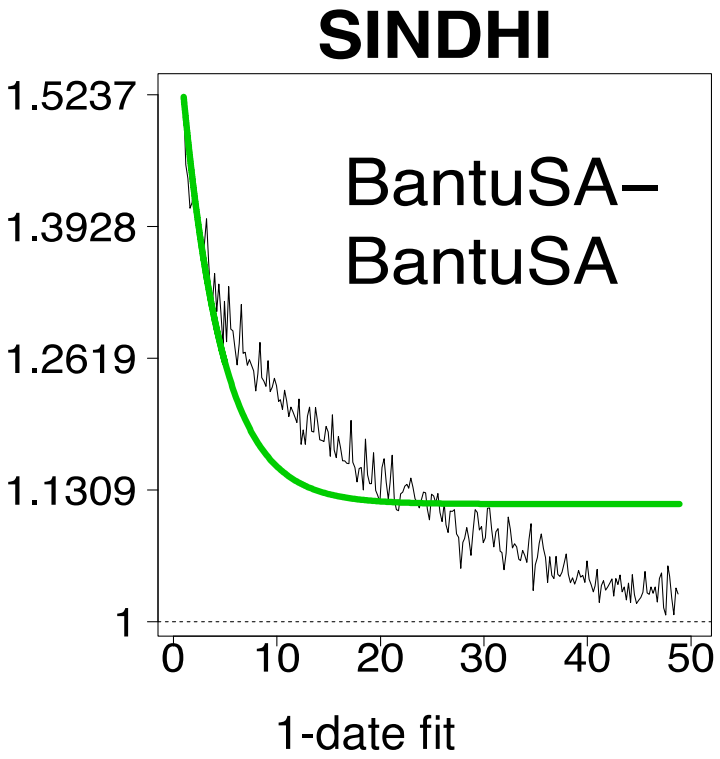
Can also identify complex events such as multi-way admixture

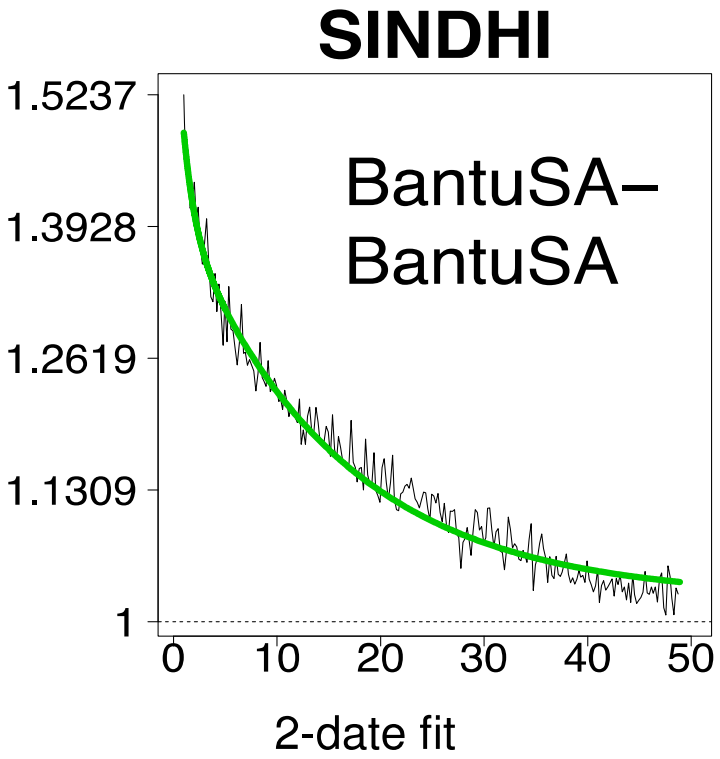


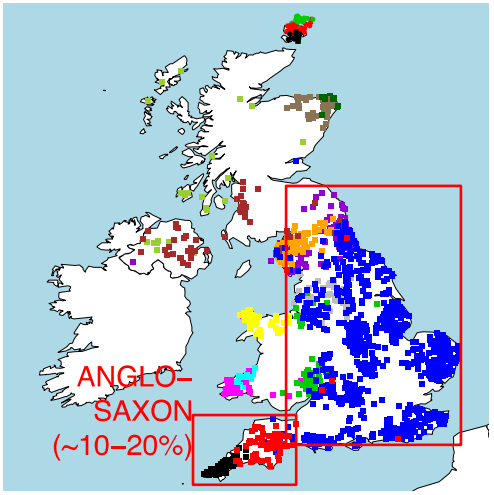


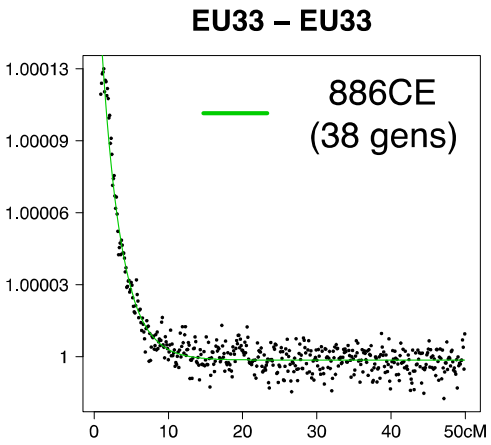


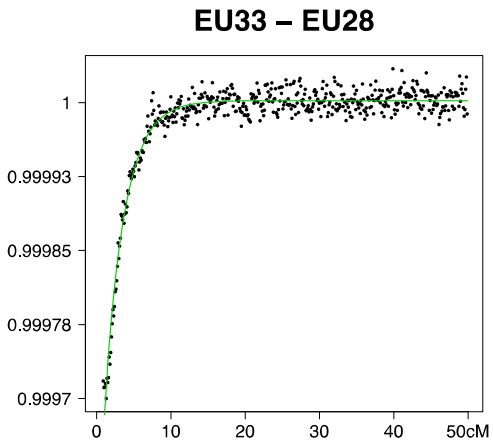
- can fit curves with sum of two exponential distributions

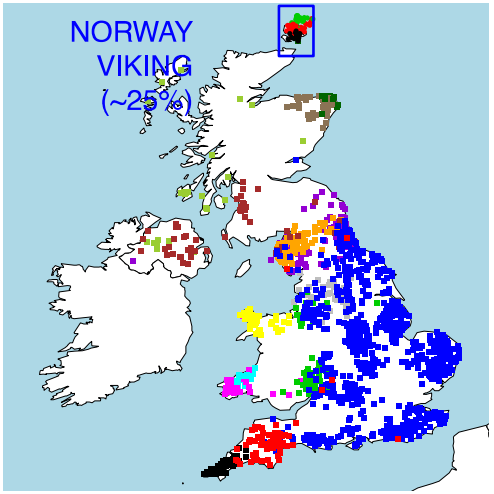


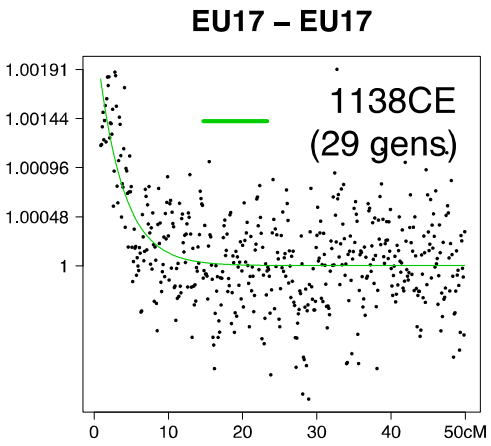


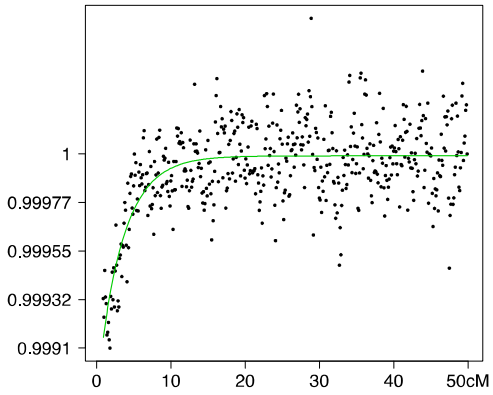

ANGLO–
SAXON
(~10–20%)


EU33 – EU33
886CE
(38 gens)
1.00013
1.00009
1.00006
1.00003
1
0
10
20
30
40
50cM


EU33 – EU28
1
0.99993
0.99985
0.99978
0.9997
0
10
20
30
40
50cM


NORWAY
VIKING
(~25%)


EU17 – EU17
1138CE
(29 gens)
1.00191
1.00144
1.00096
1.00048
1
0
10
20
30
40
50cM


EU17 – EU31
1
0.99977
0.99955
0.99932
0.9991
0
10
20
30
40
50cM

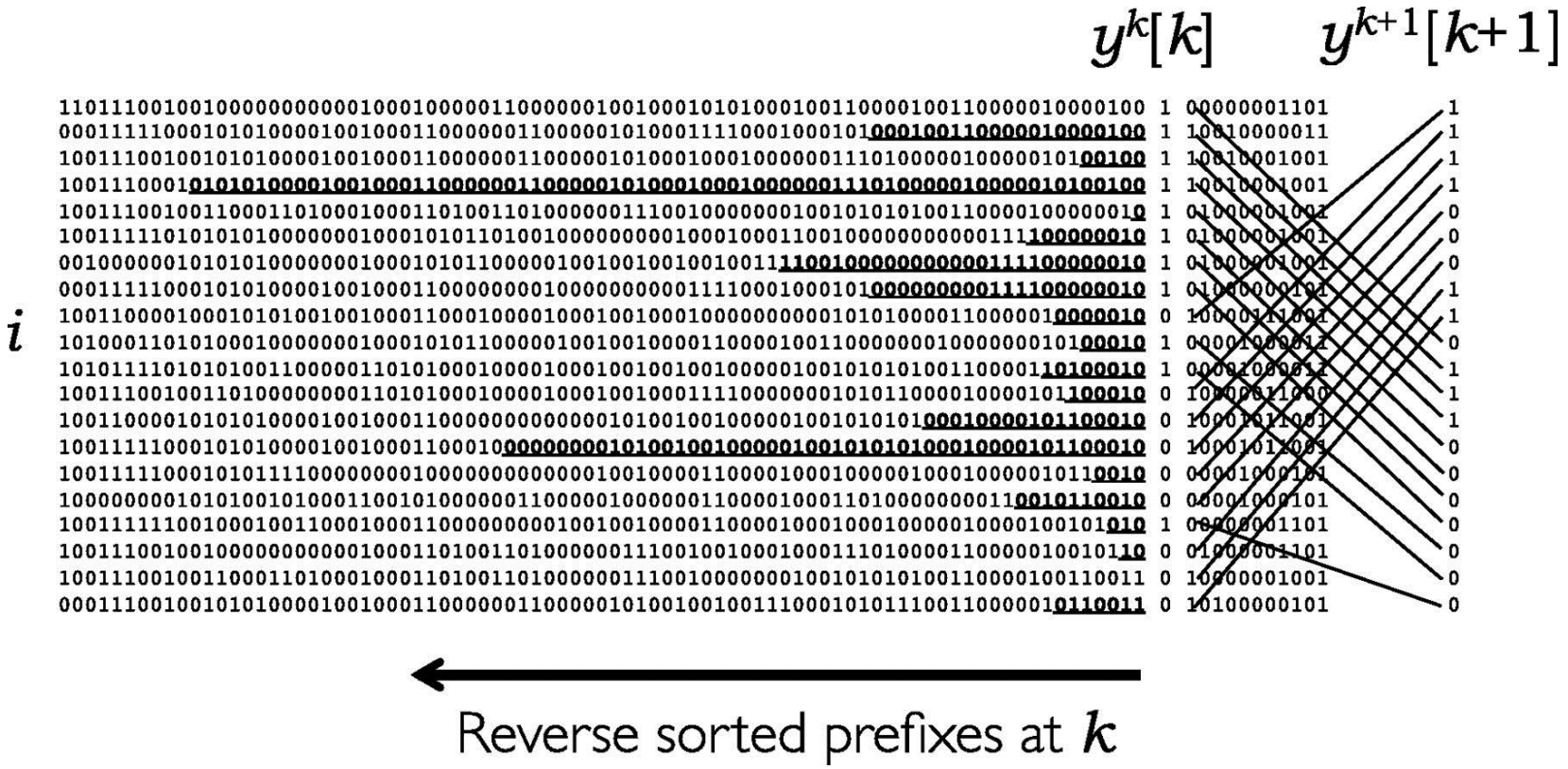
Conclusions

Can use chromosome painting to

- Distill ancestry information
- Cluster based on genetic similarity
- Visualise genetic drift
- Identify mixtures
- Date mixture events
- Reconstruct history

Positional Burrows Wheeler Transform

A set of haplotype sequences sorted in order of reversed prefixes at position k, showing the set of values at k isolated from those before and after, and on the right hand side how the order at position (k + 1) is derived from that at k as in Algorithm 1.



Richard Durbin Bioinformatics 2014;30:1266-1272

Bioinformatics

CPU Time

seconds

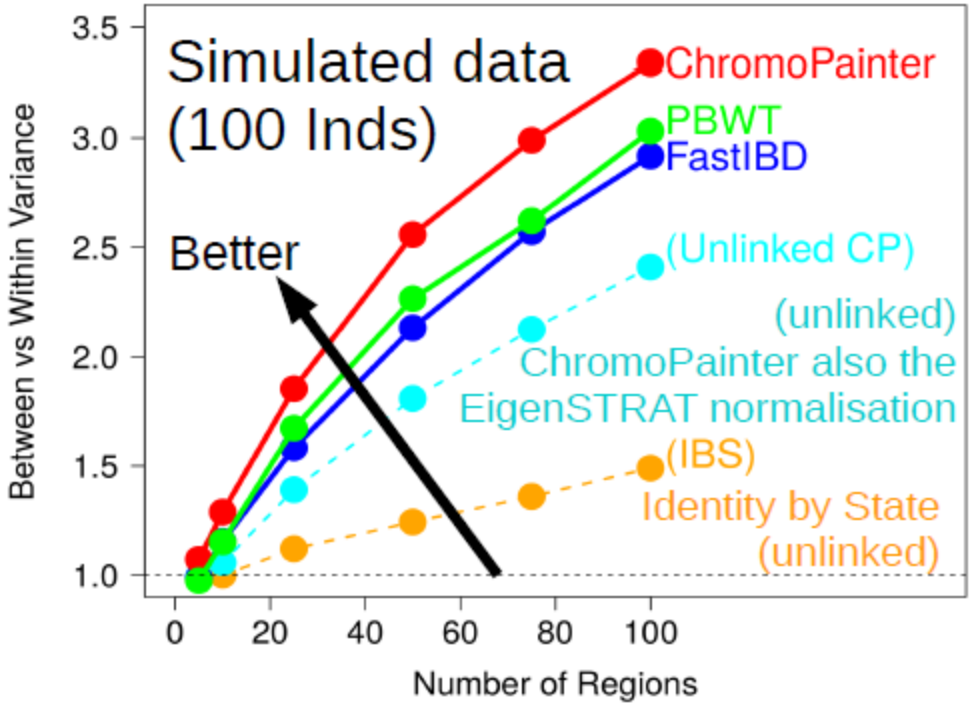
88K SNPs

100 Inds

ChromoPainter	3130
FastIBD	2686
PBWT	**1**

500 Inds

ChromoPainter	264000
FastIBD	169093
PBWT	**19**



HGDP Data (938 Inds, 800K SNPs)

ChromoPainter	FastIBD	PBWT	CP Unlinked
0.35		0.24	0.21

Daniel J. Lawson (Heilbronn Institute, University of Bristol), **Daniel Falush** (Max Planck, Leipzig)

Simon Myers (Oxford), **Garrett Hellenthal** (UCL), **Richard Durbin** (Wellcome Trust Sanger Institute)