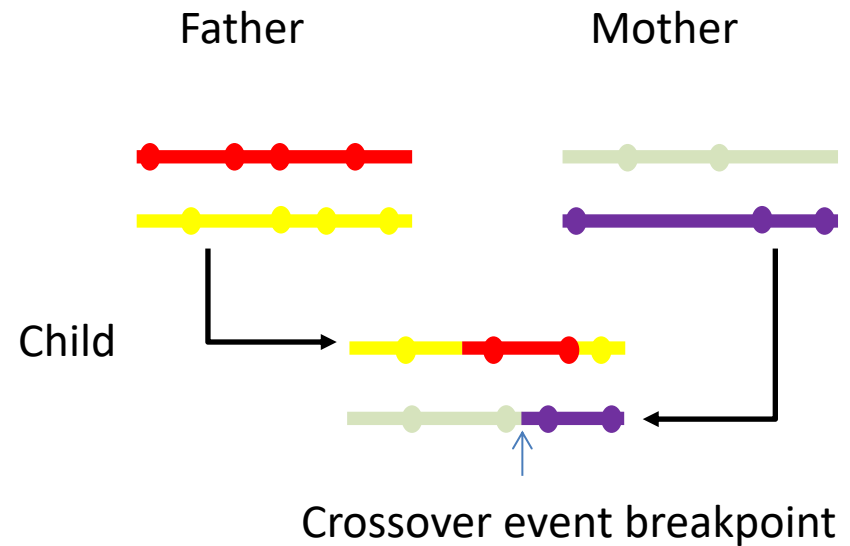
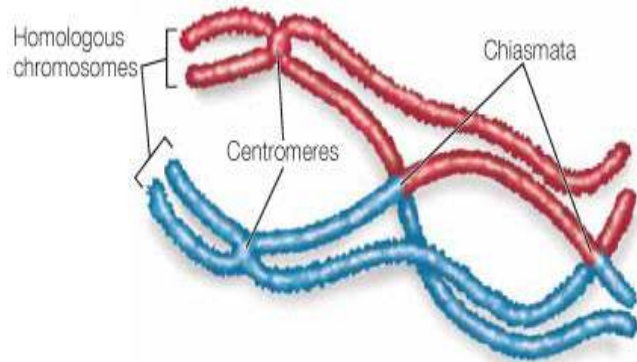
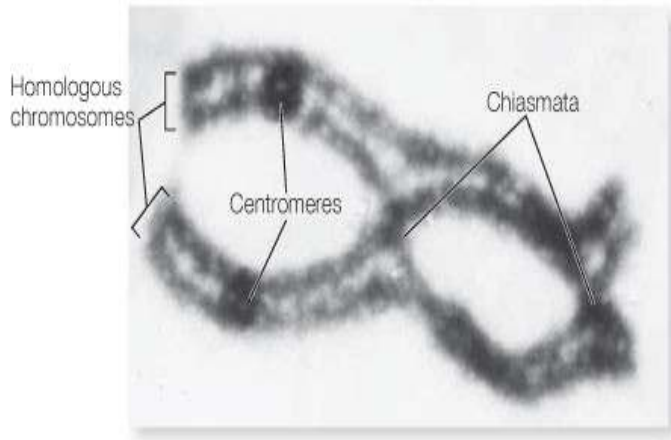


The painting palettes of human ancestry

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

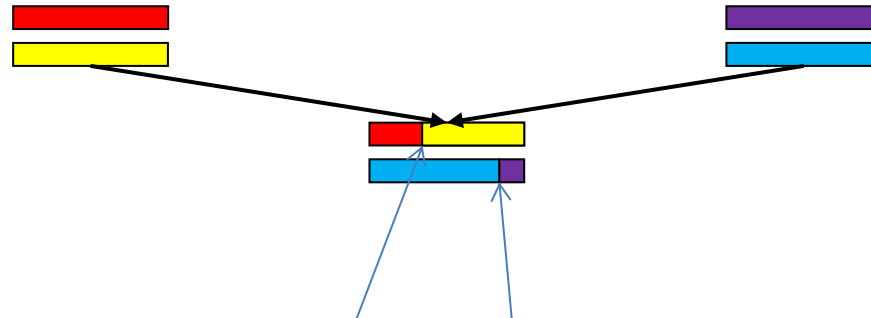
www.paintmychromosomes.com

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



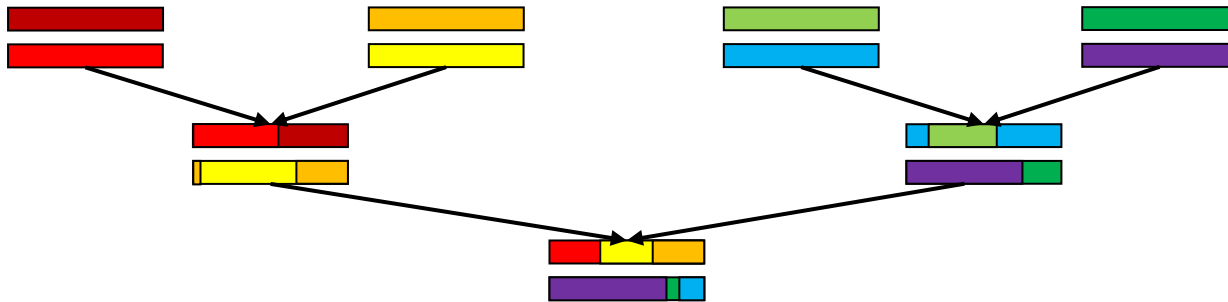
(O'Connor 2008)

Recombination ➡ “mosaic” genomes

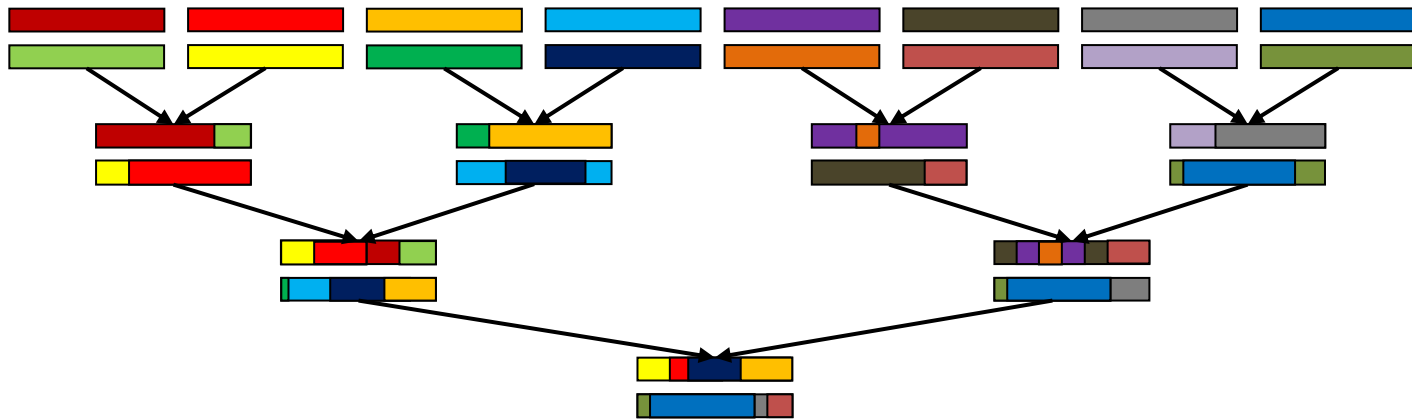


Positions of recombination crossover events

Recombination ➡ “mosaic” genomes



Recombination ➡ “mosaic” genomes



Chromosome “painting”

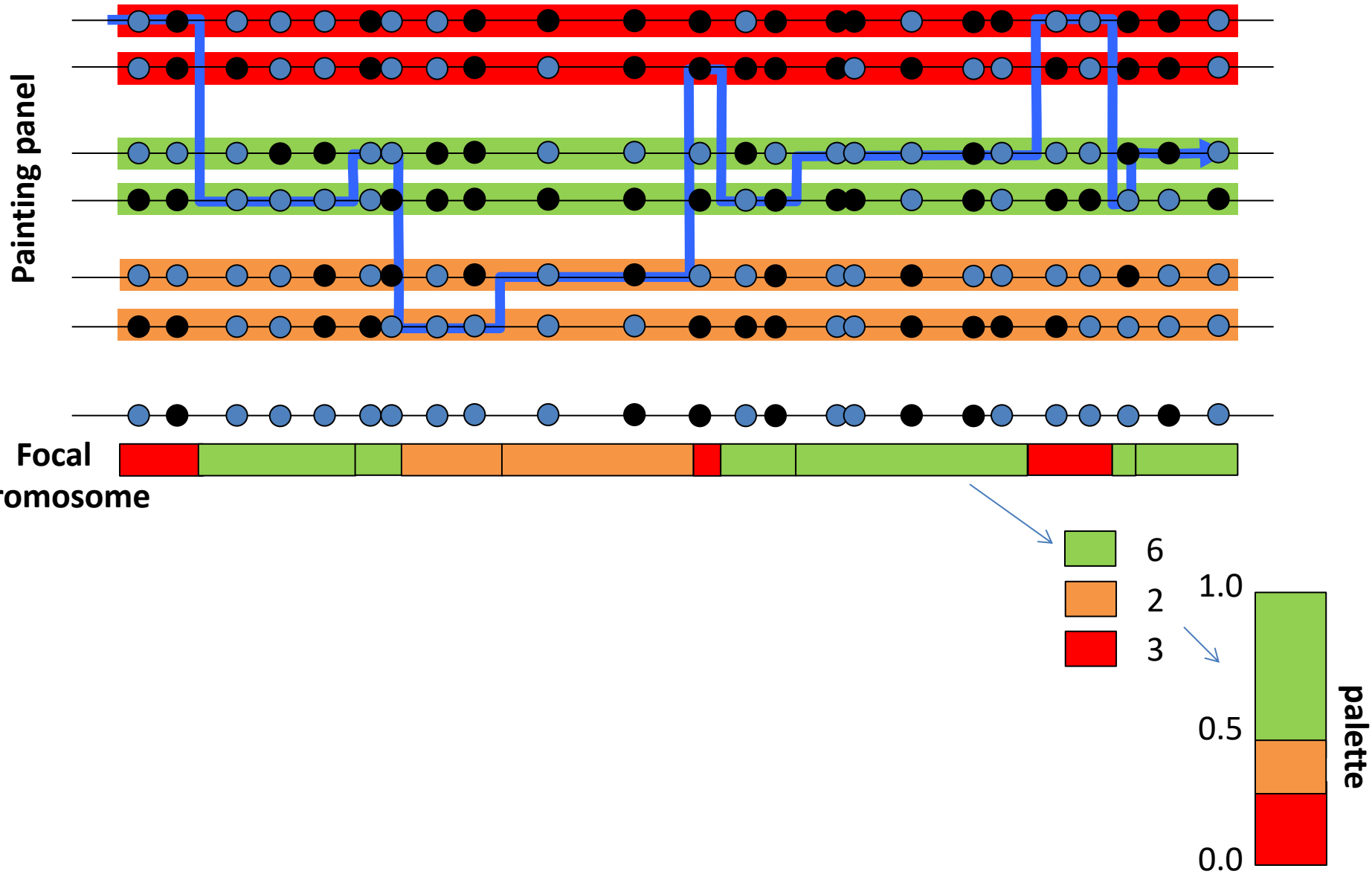


We share segments of DNA today, inherited from shared ancestors living in the distant past.

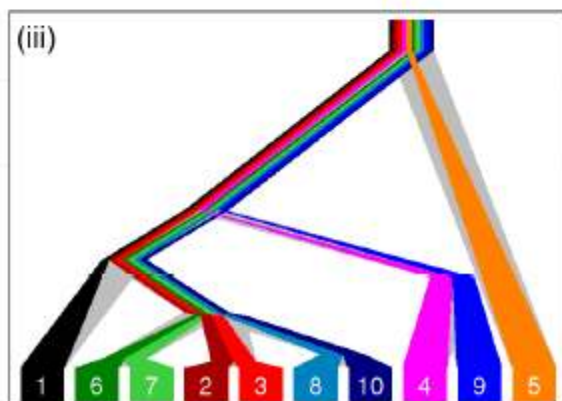
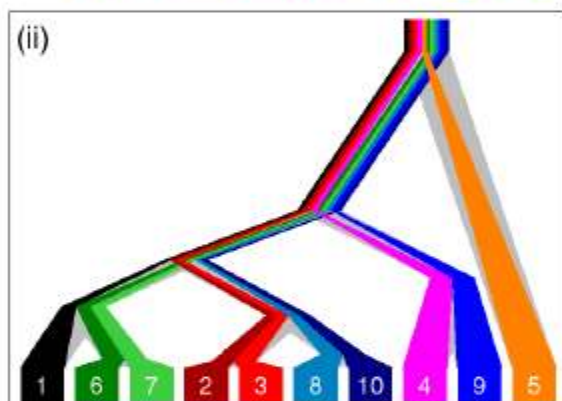
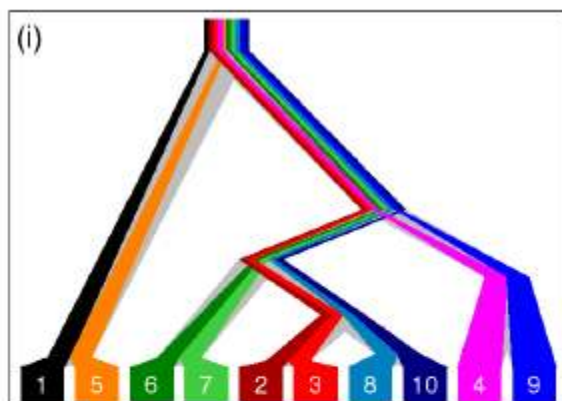
Identifying these shared segments, and insights into the genetics of people from the UK, and worldwide human migrations

Chromosome painting in practice

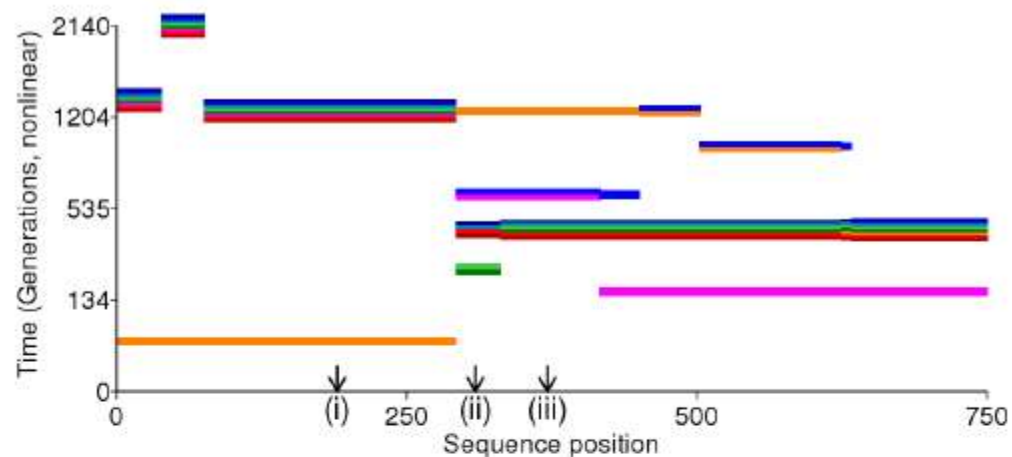
- (Lawson et al. 2012)



A) Local genealogies



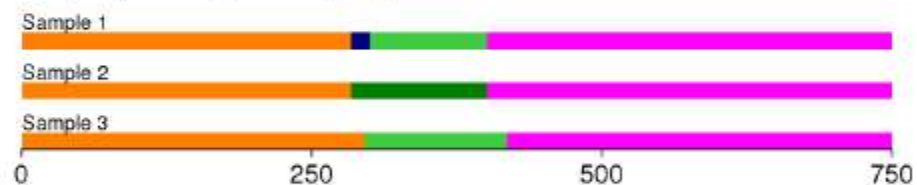
B) Time to MRCA with haplotype 1



C) True 'nearest neighbour' distribution of haplotype 1



D) Sample paintings of haplotype 1



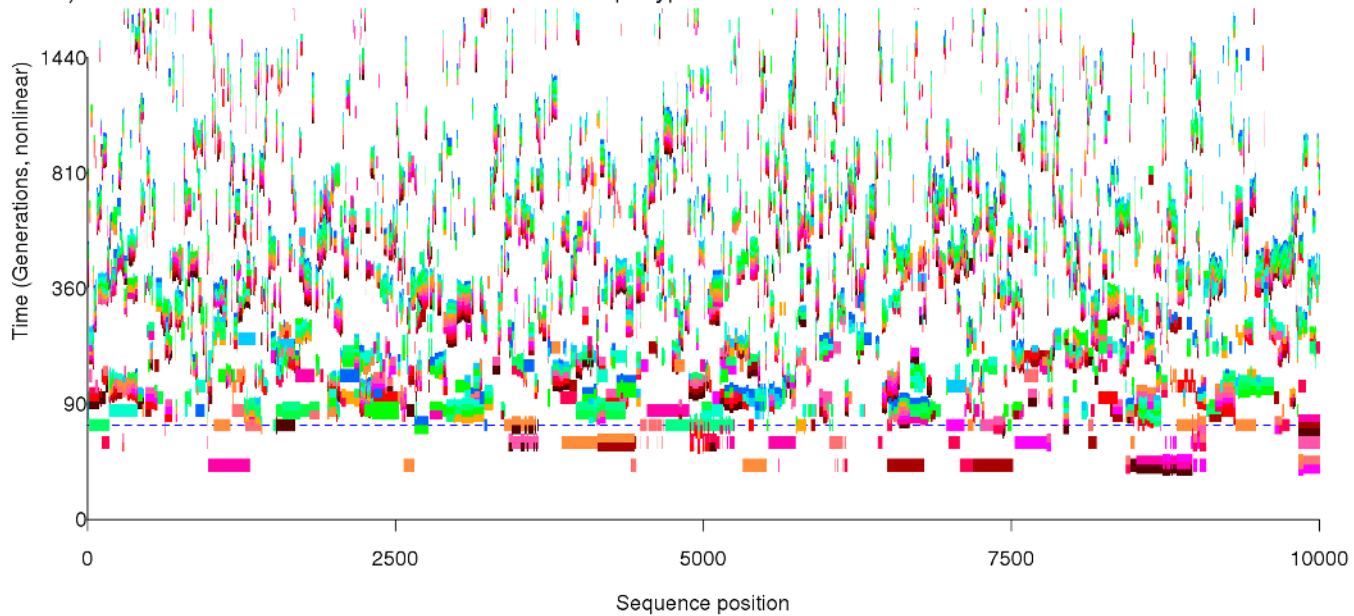
E) Mean painting of haplotype 1



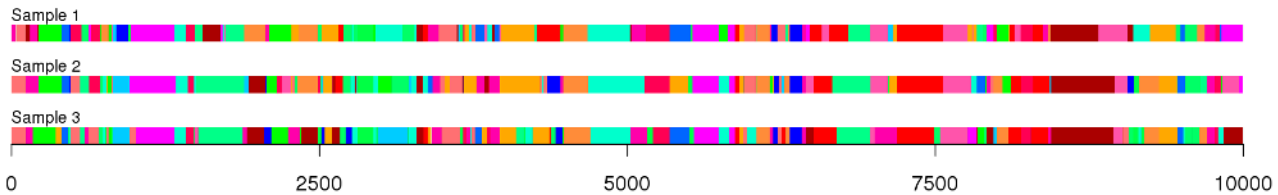
F) Coancestry matrix row for haplotype 1

	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

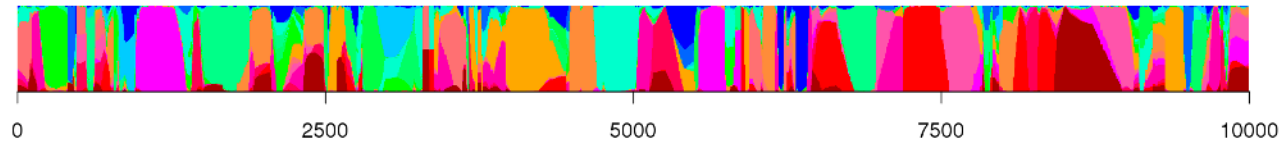
A) Time to Most Recent Common Ancestor with haplotype 1



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									

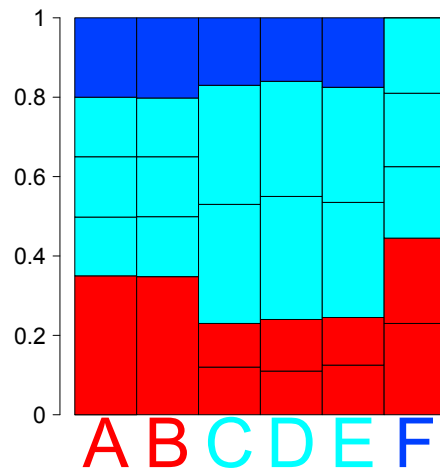
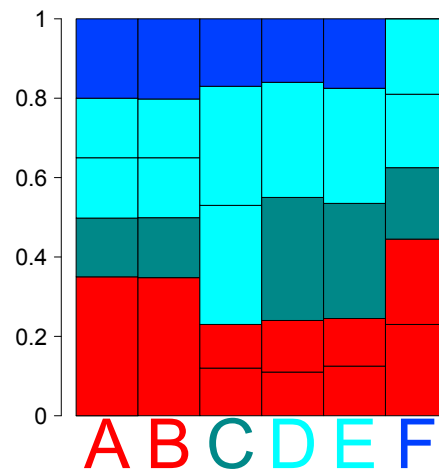
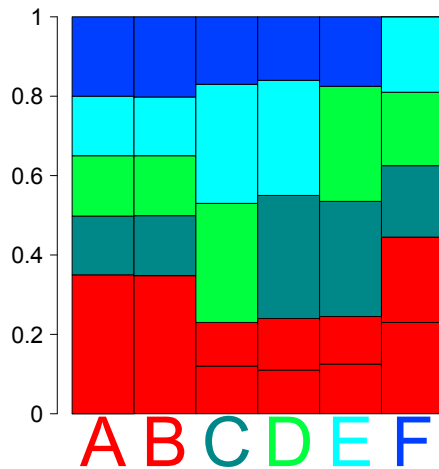
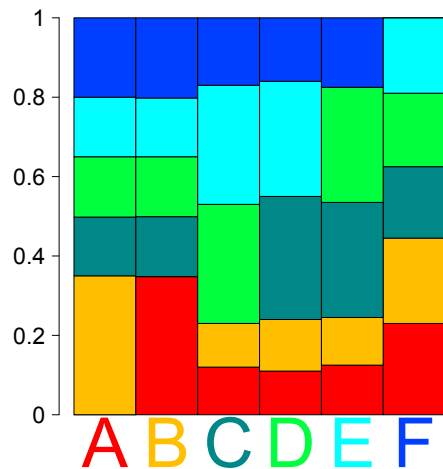
Genealogical interpretation of painting

- The painting of each segment indicates who in the painting panel has the 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).

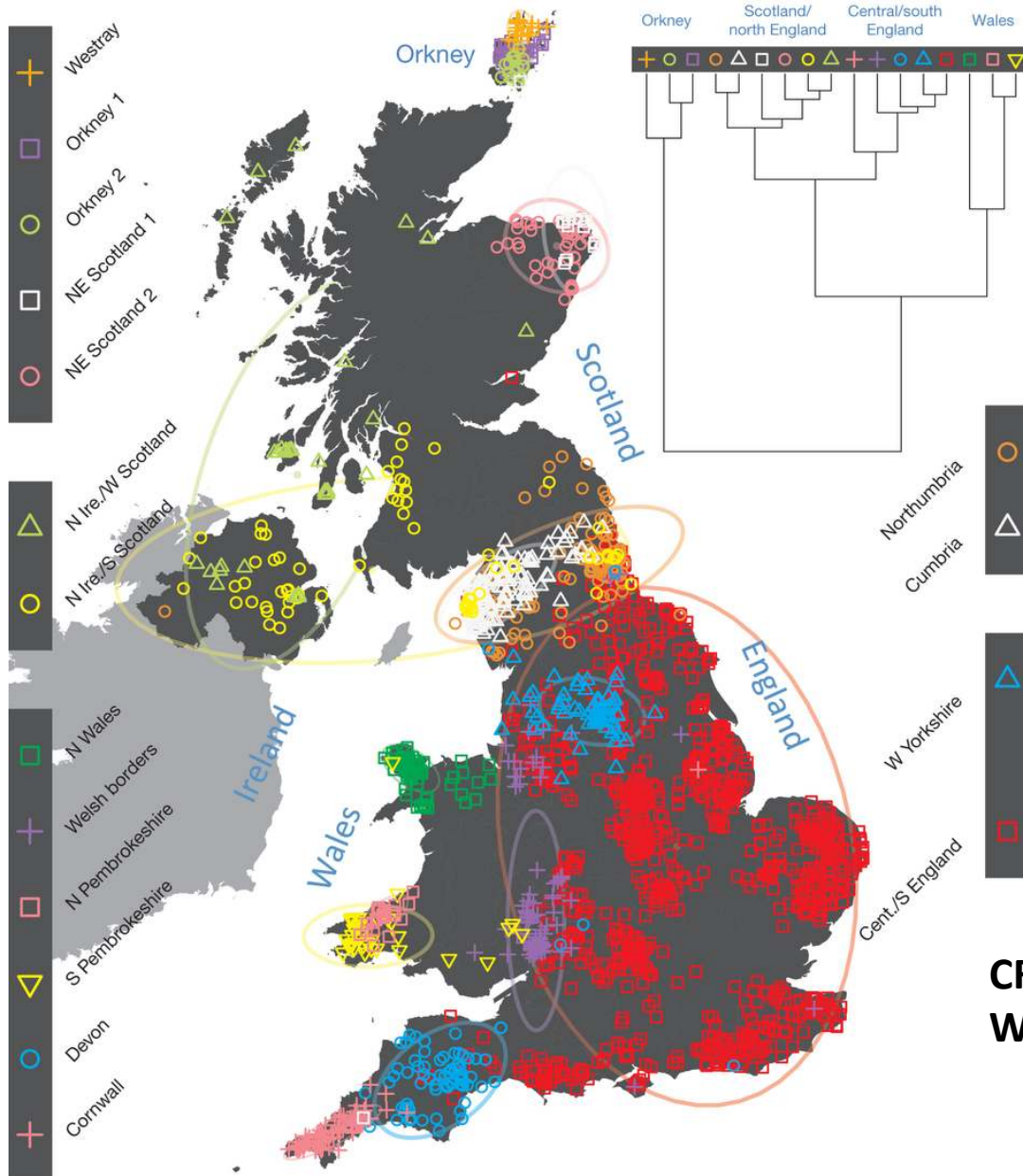
toy FineSTRUCTURE example



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.

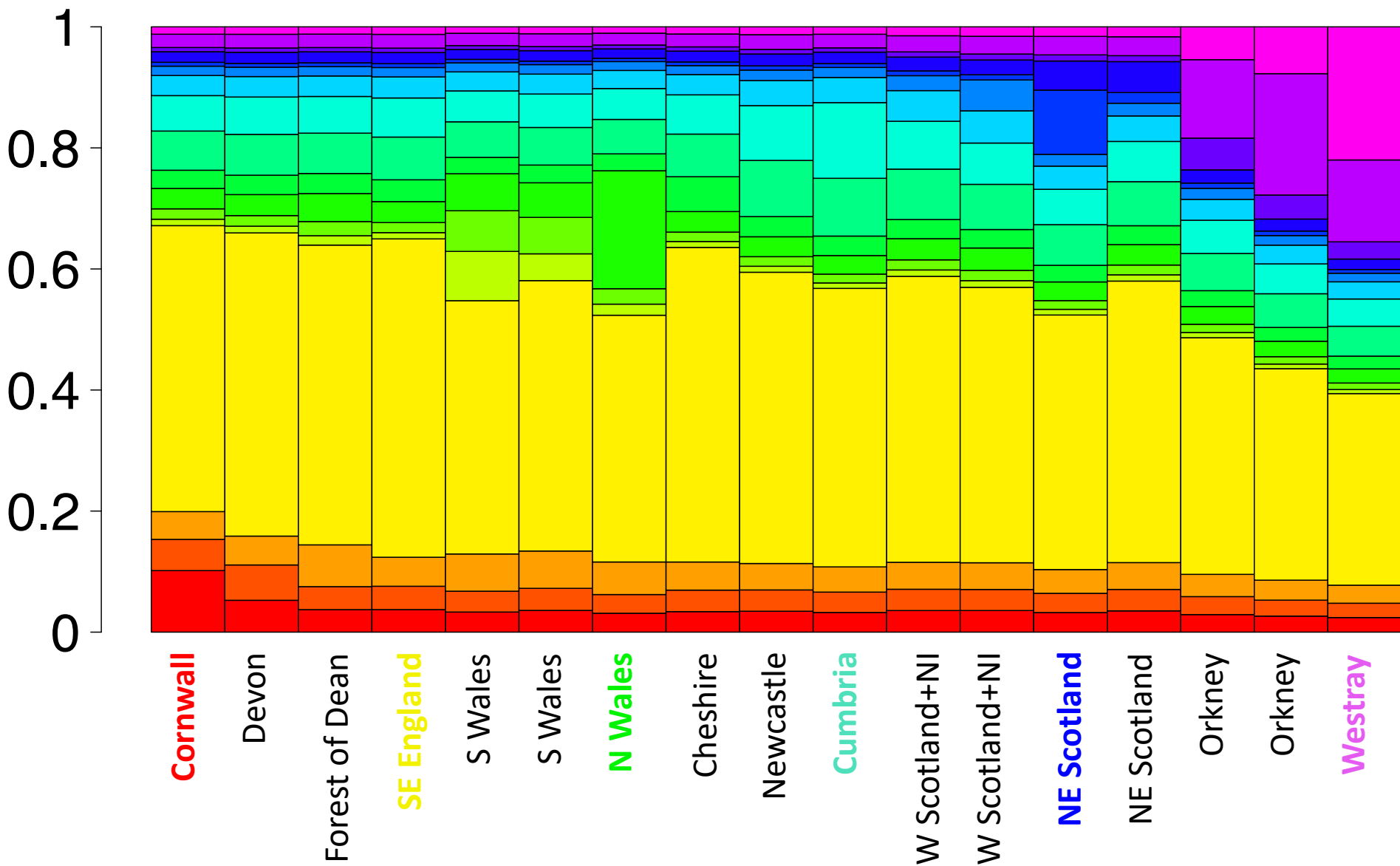
Genotyped using SNP chip. (500,000 markers)

54 distinct palettes inferred by fineSTRUCTURE

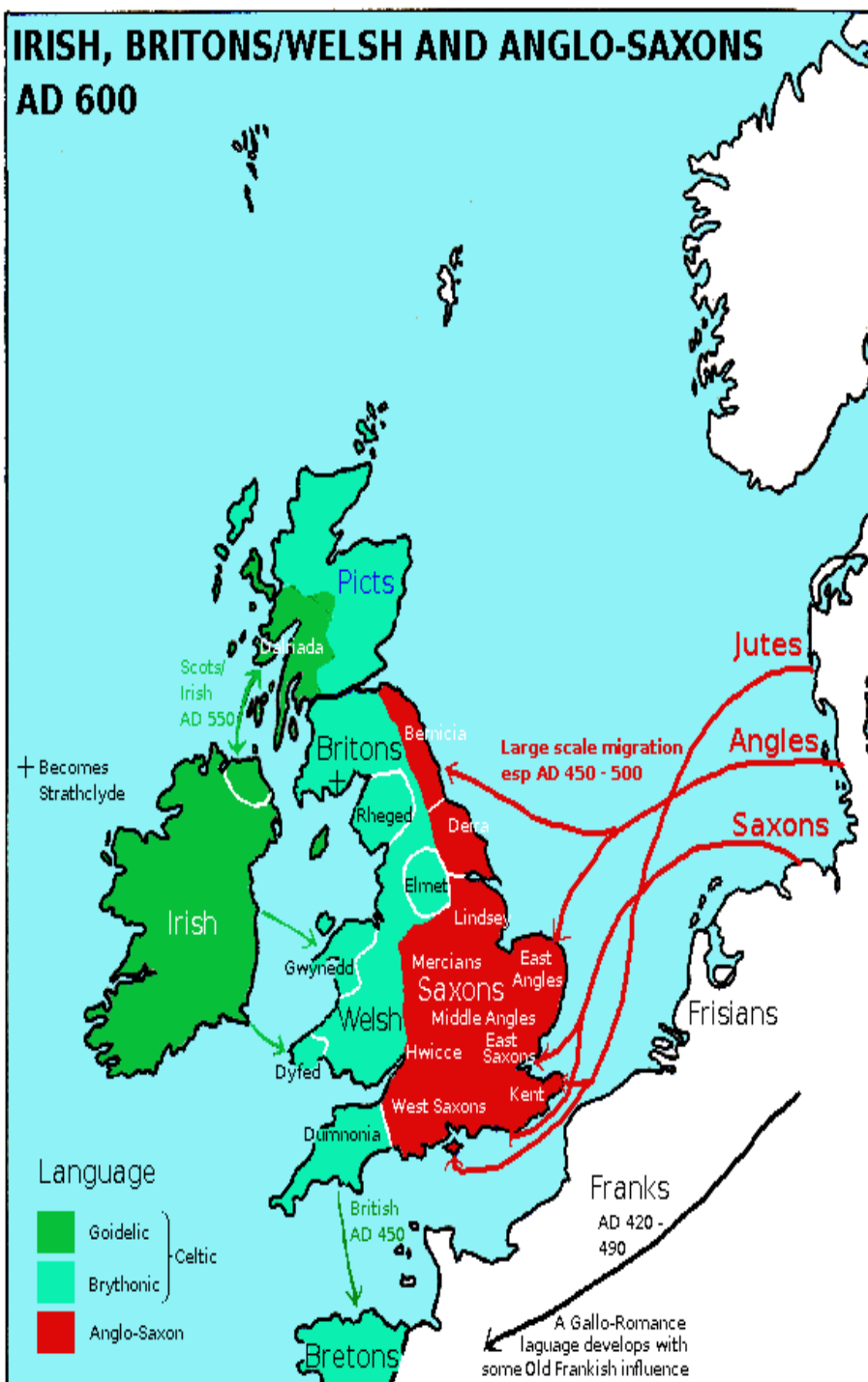
Build a tree: by successively joining groups with the most similar palettes.

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

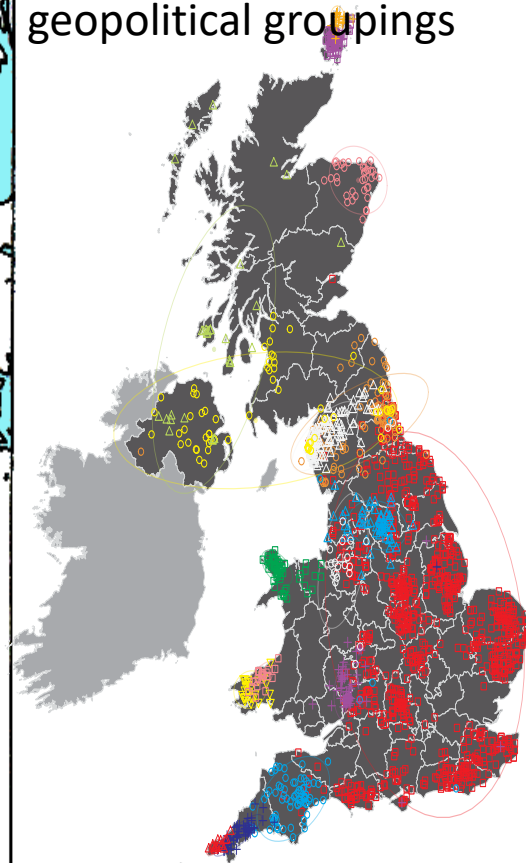
British palettes



IRISH, BRITONS/WELSH AND ANGLO-SAXONS AD 600



Clusters very strongly resemble ancient geopolitical groupings

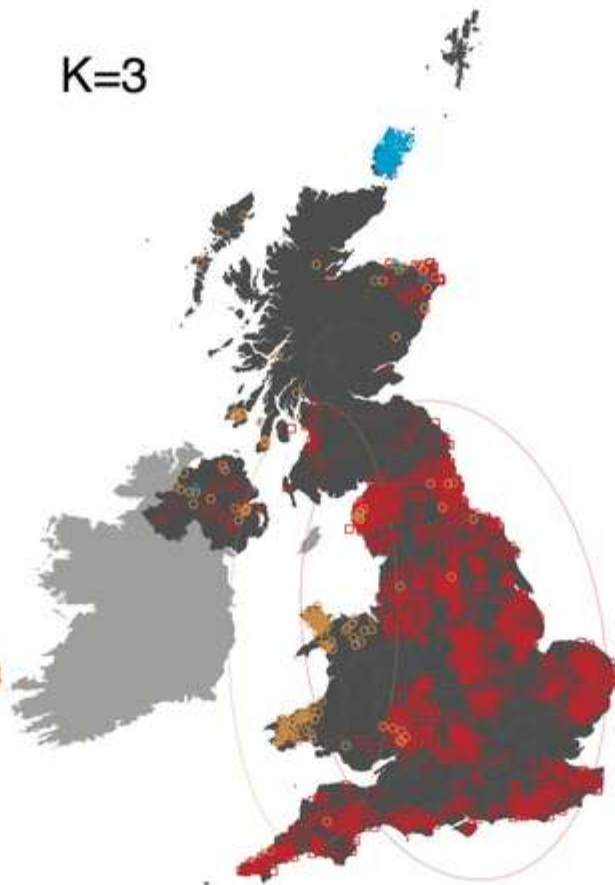


b

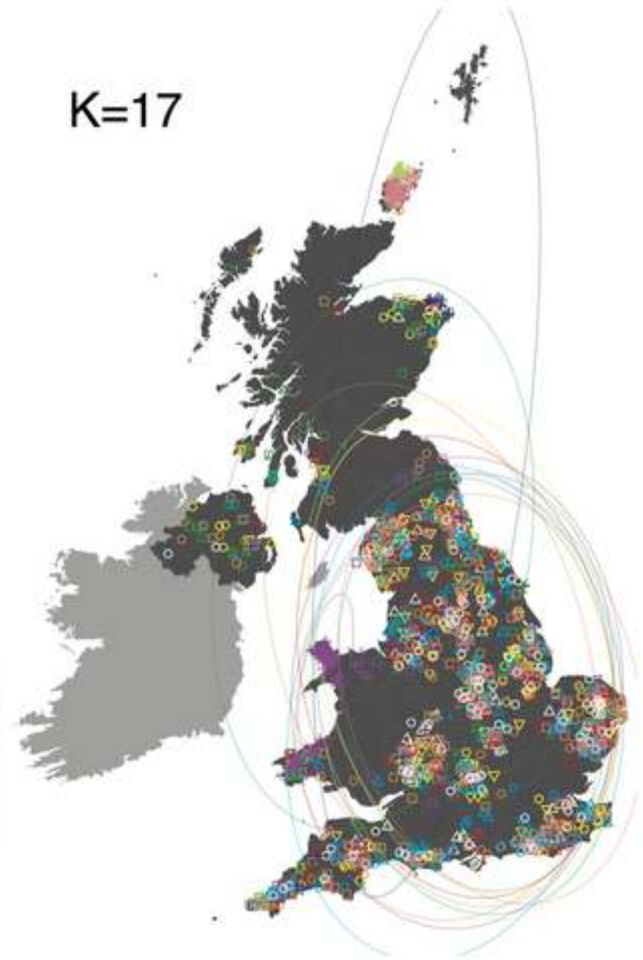
K=2



K=3

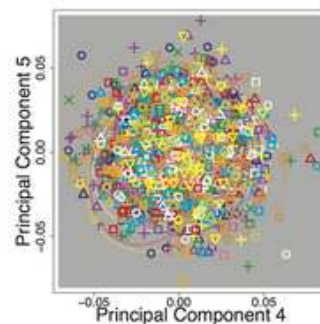
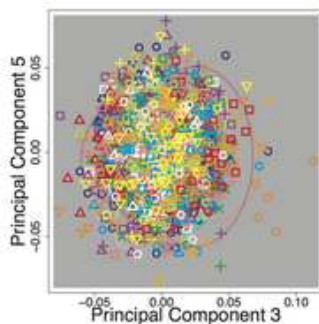
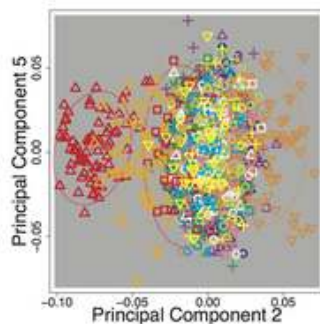
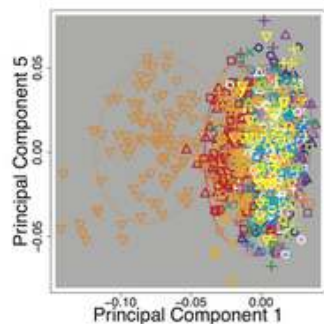
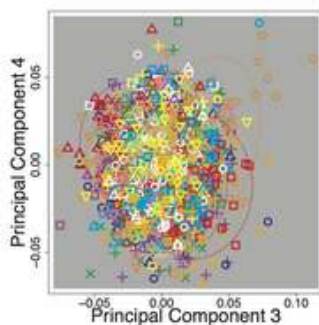
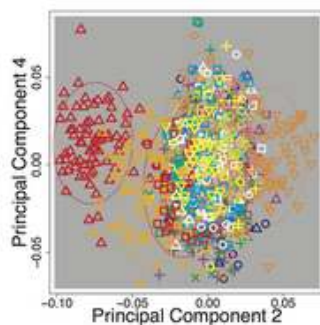
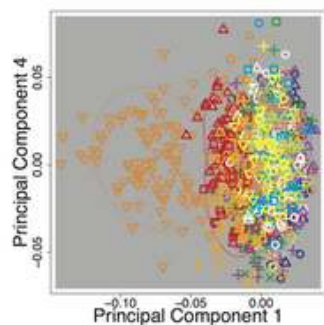
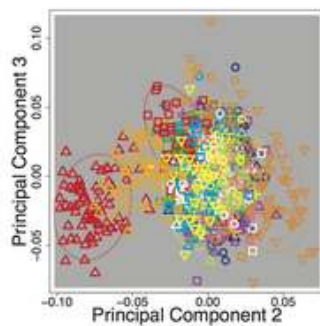
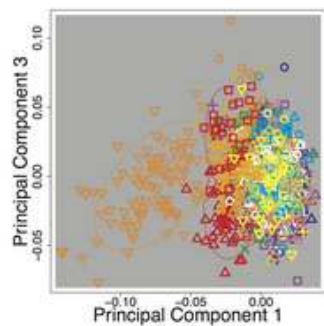
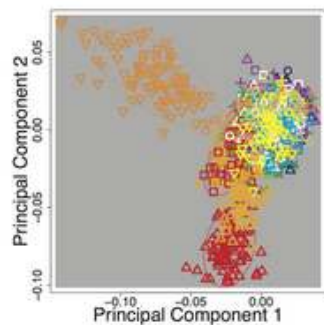


K=17



Traditional software (Admixture)

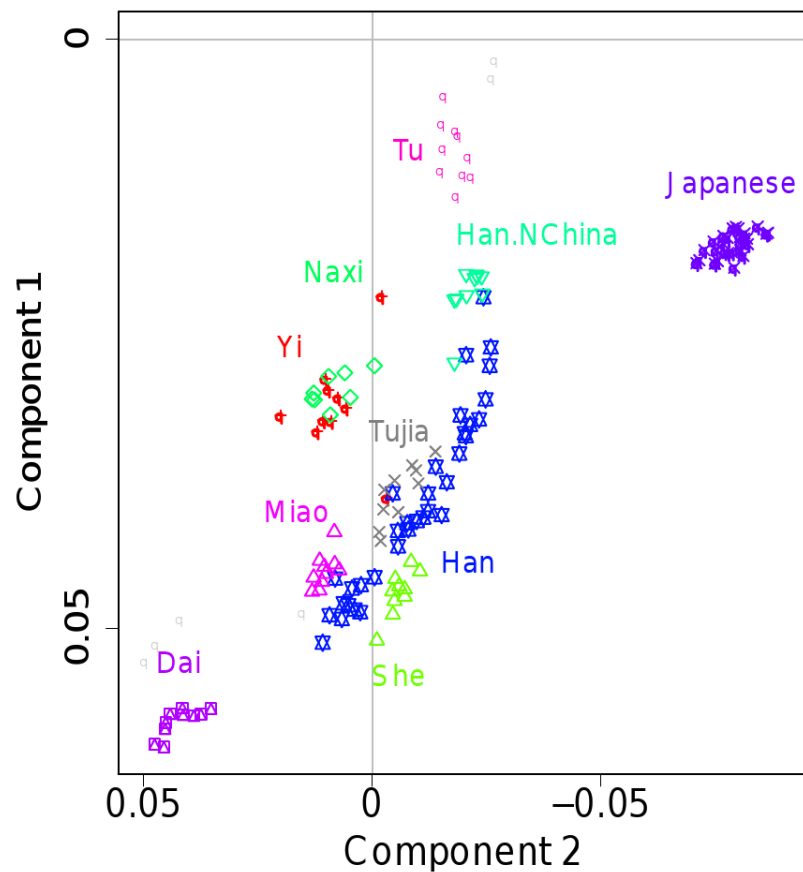
a



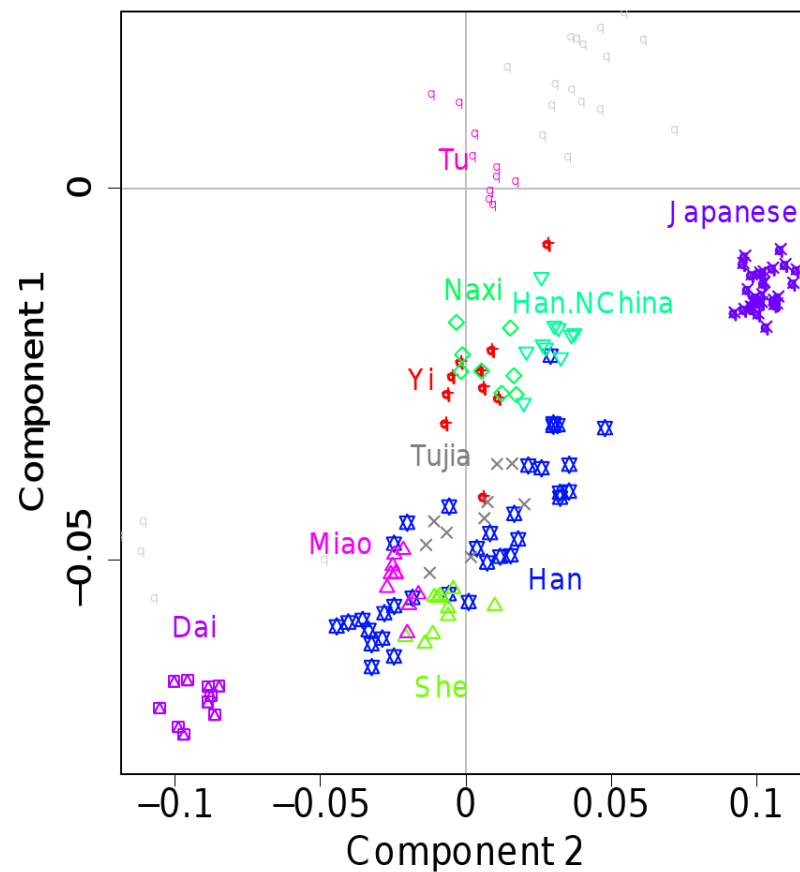
Legend

□	ARG	□	LIN
○	BAN	×	NEA
△	CHE	●	NIR
○	COR	△	NOR
+	CUM	+	NOT
○	DER	●	NPE
△	DEV	○	NTH
+	DOR	△	NWA
□	FOD	▽	ORK
○	GLO	□	OXF
△	HAM	●	SPE
▽	HER	△	SUF
□	KEN	+	SUS
○	LAN	□	WOR
□	LEI	○	YOR

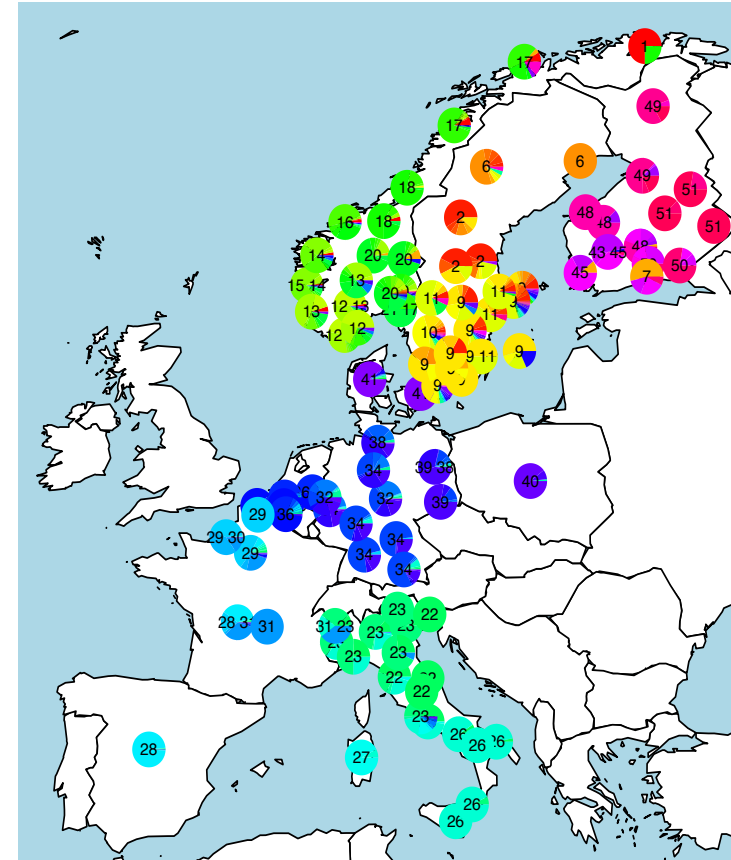
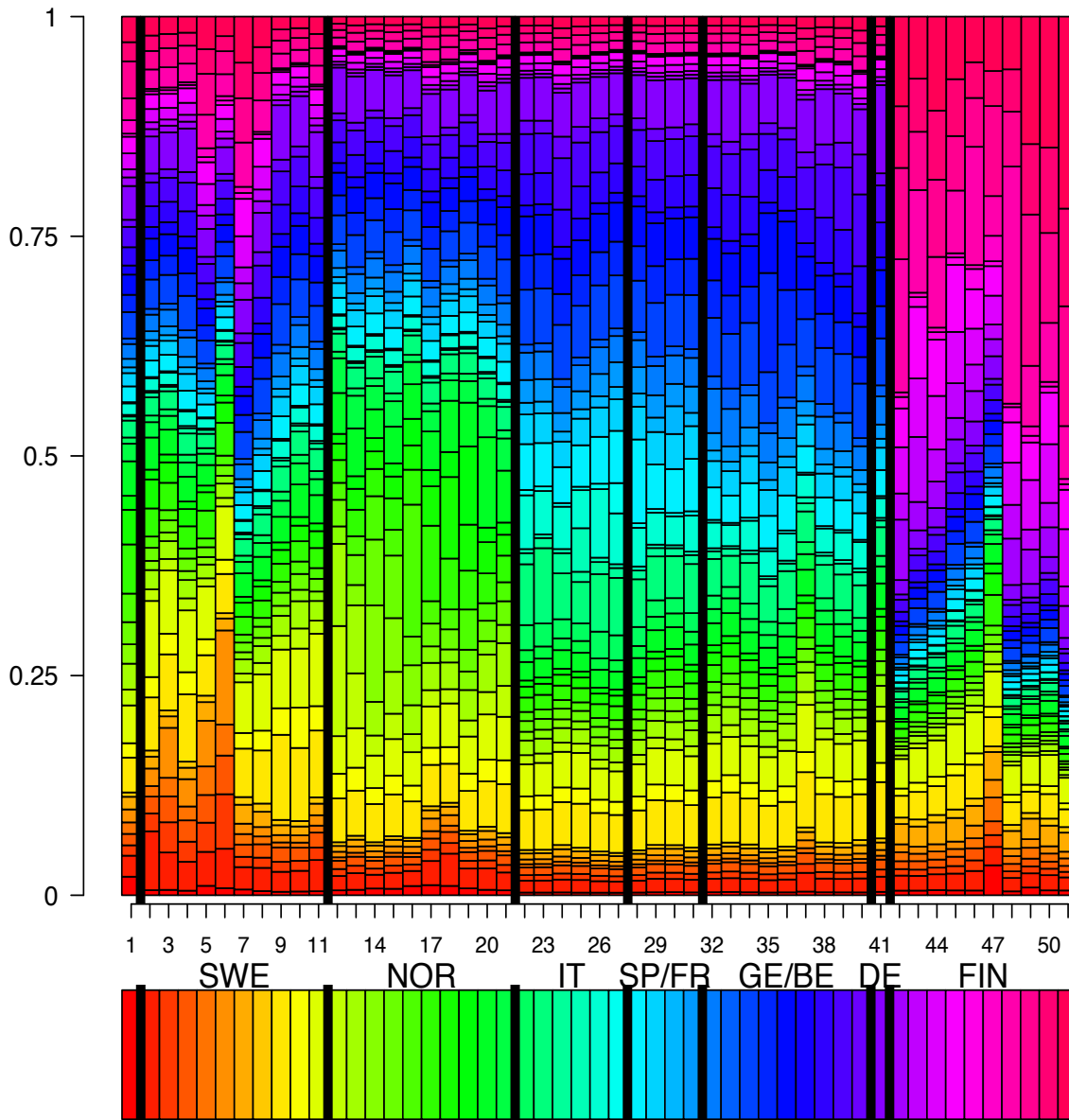
A) Linked model



B) Unlinked model



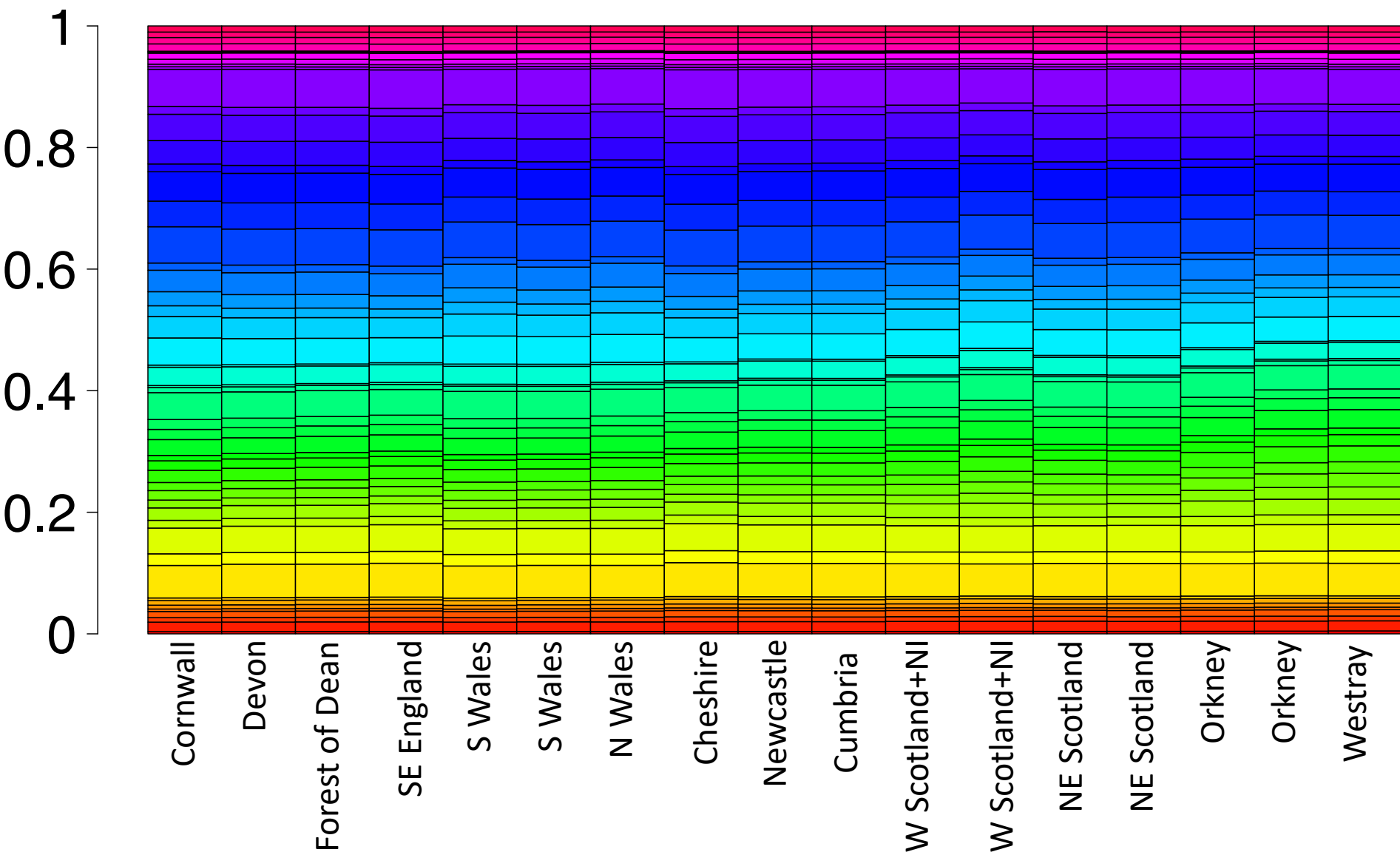
Continental European palettes



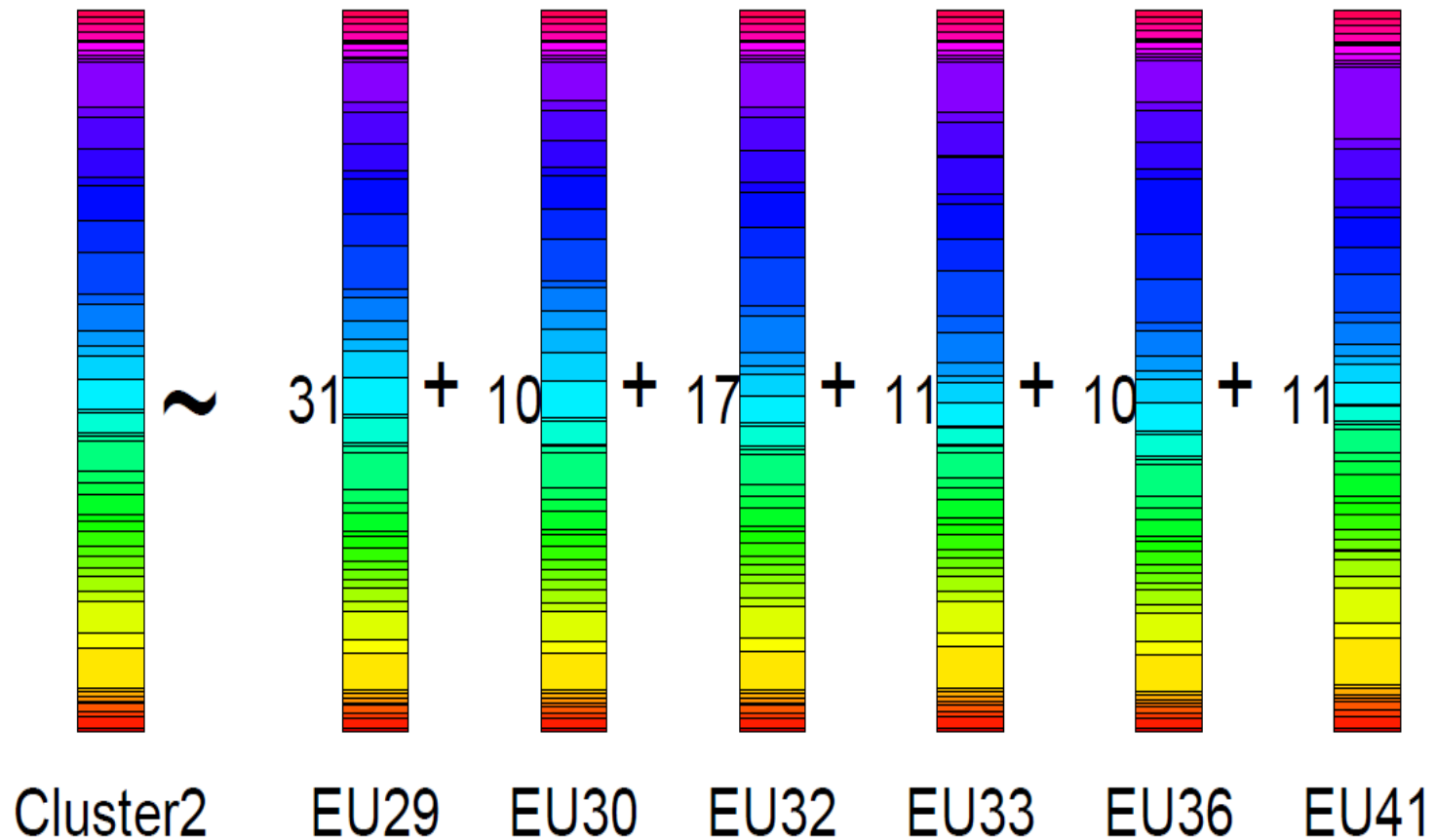
Hospital based sampling as controls for an association study with FineSTRUCTURE used to identify populations

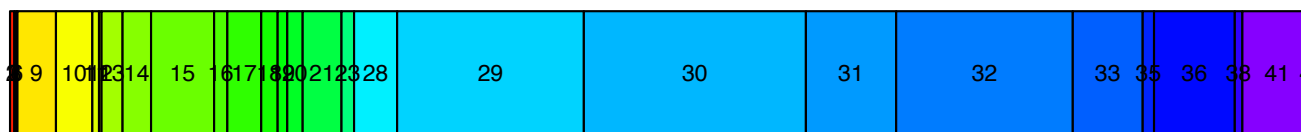
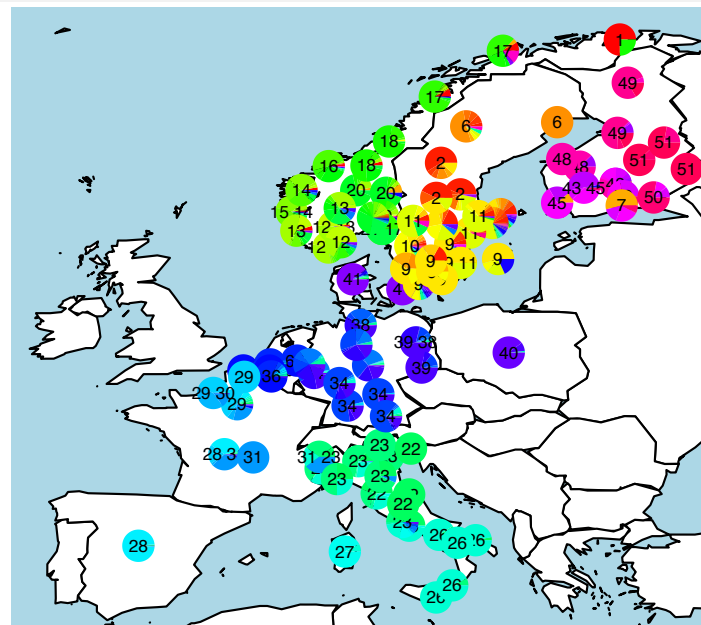
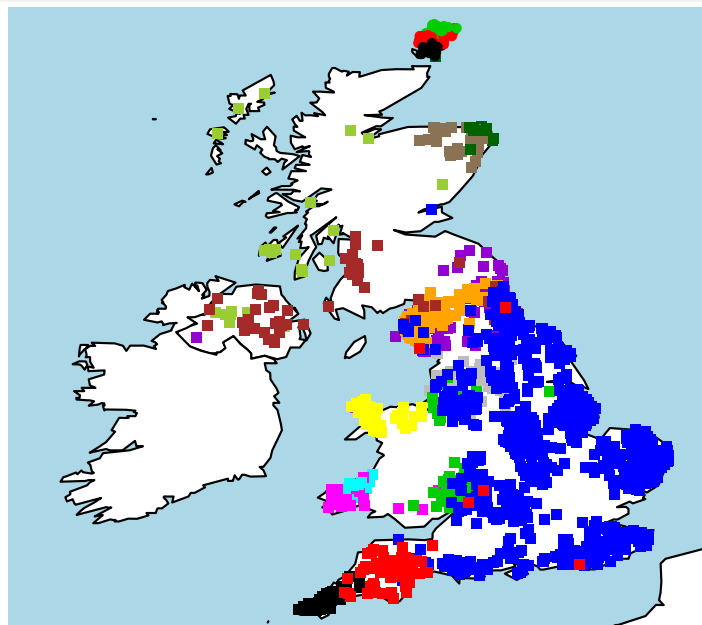
British palettes

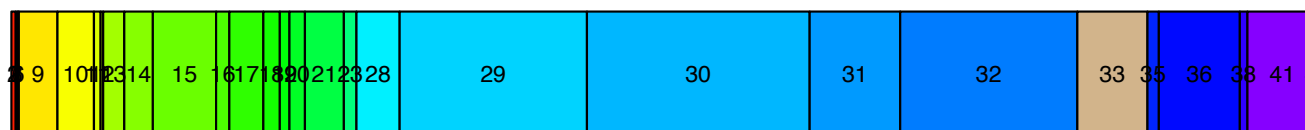
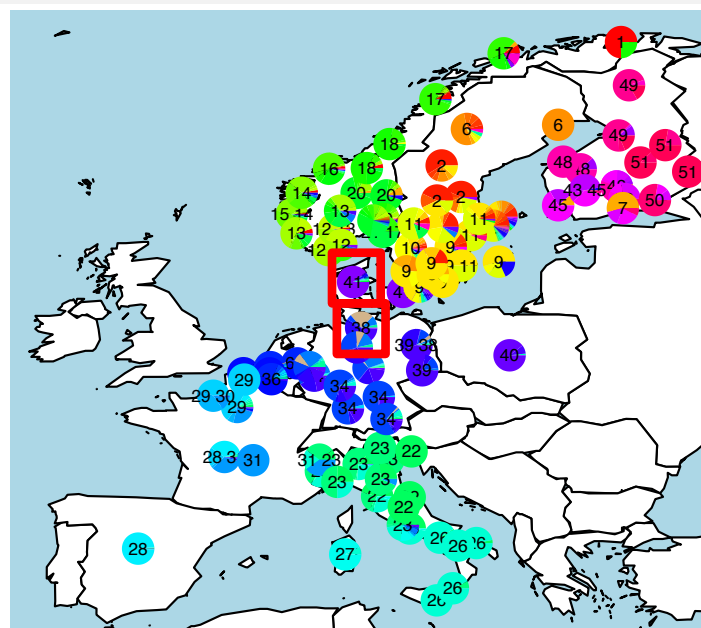
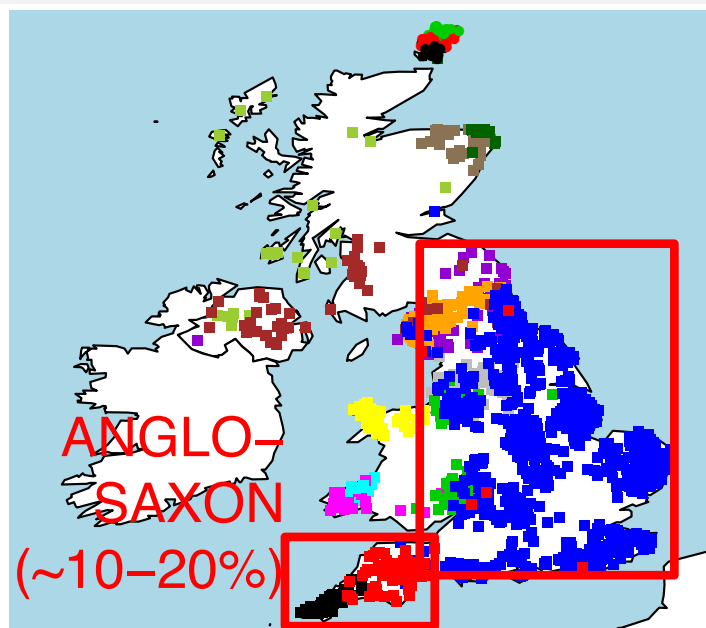
painted with a continental European panel

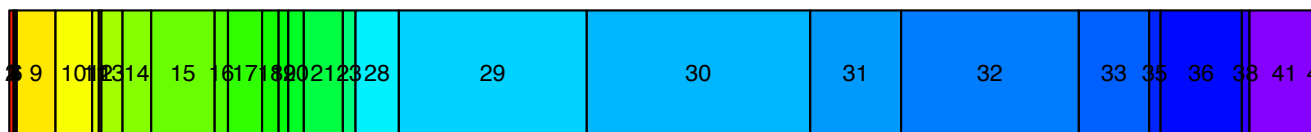
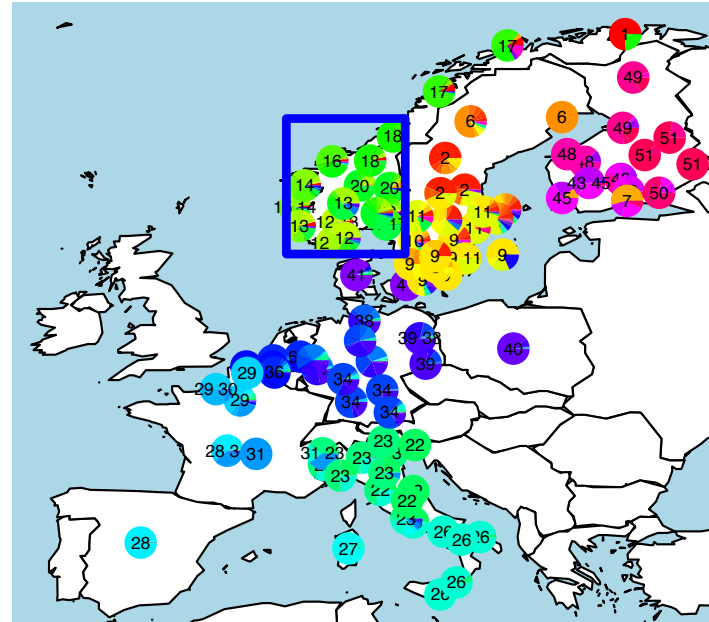
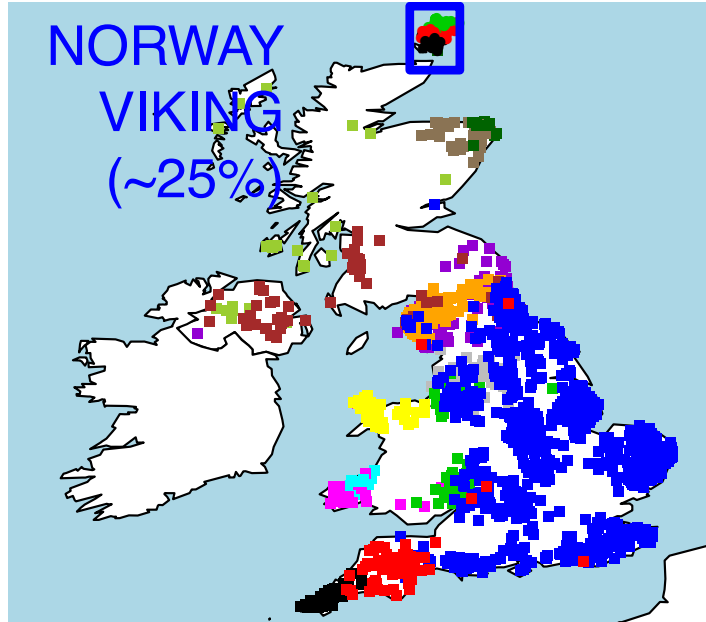


Mixture modelling of an English palette based on European palettes using **Non-Negative Least Squares**





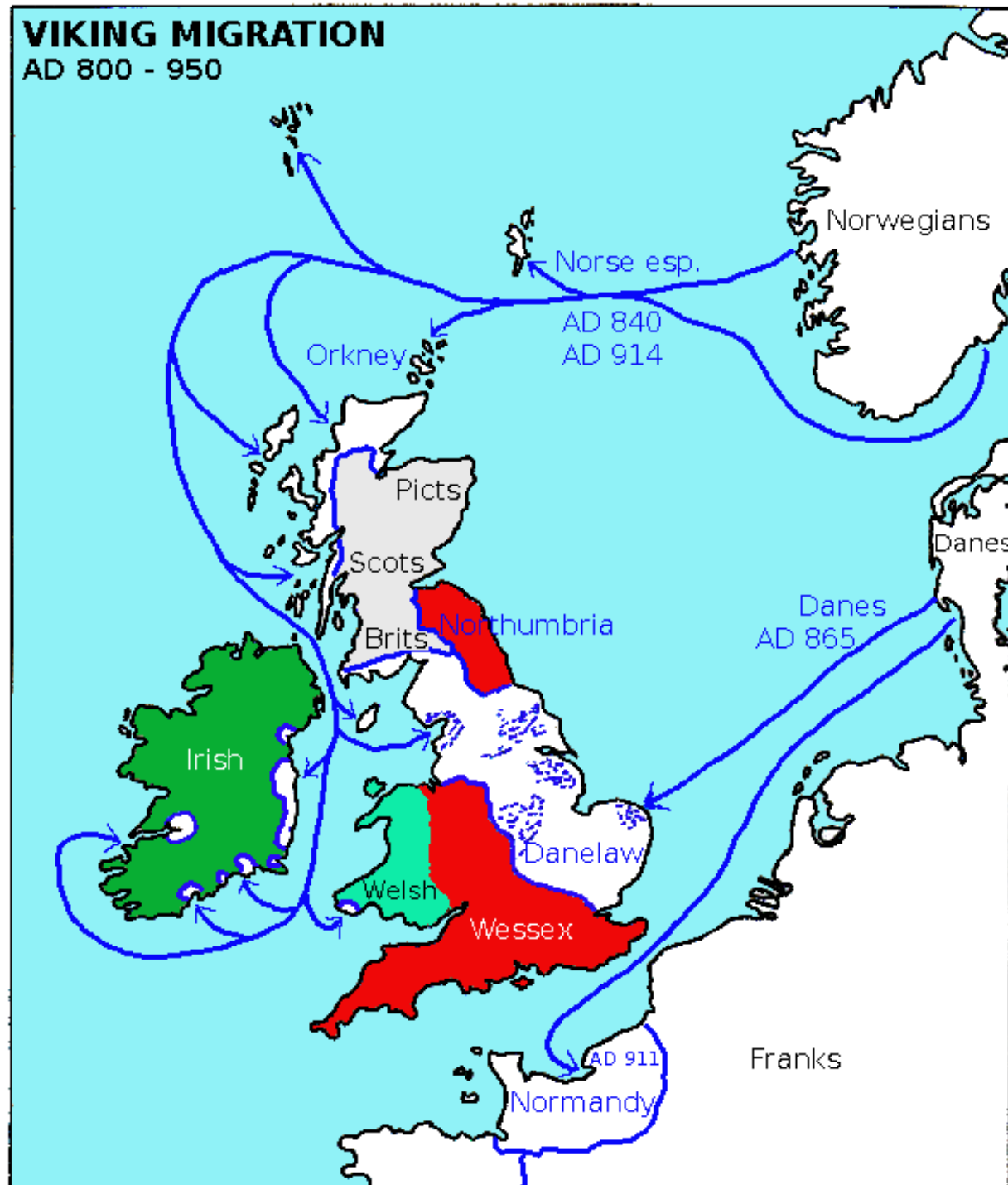




e

VIKING MIGRATION

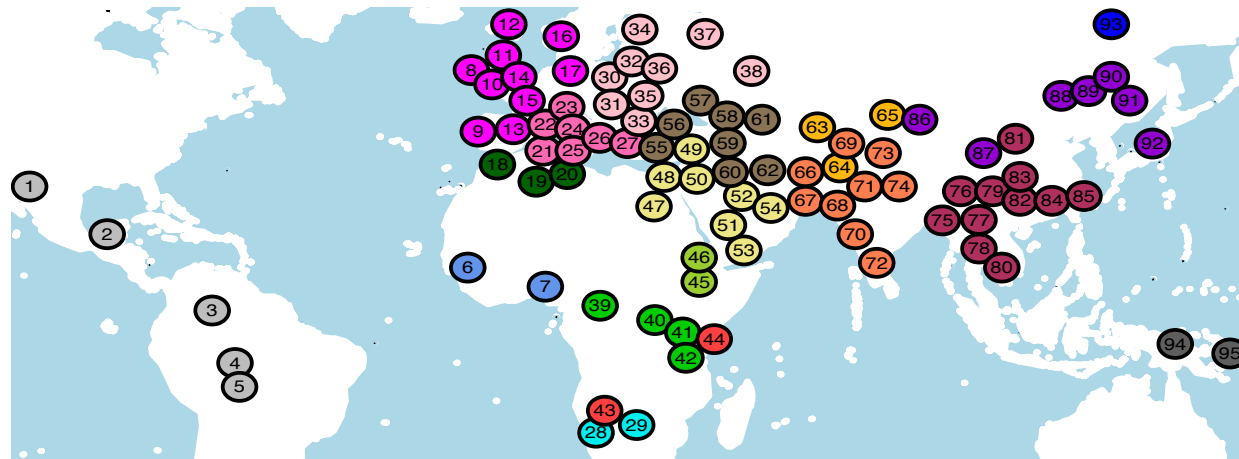
AD 800 - 950



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. Uyгур

"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

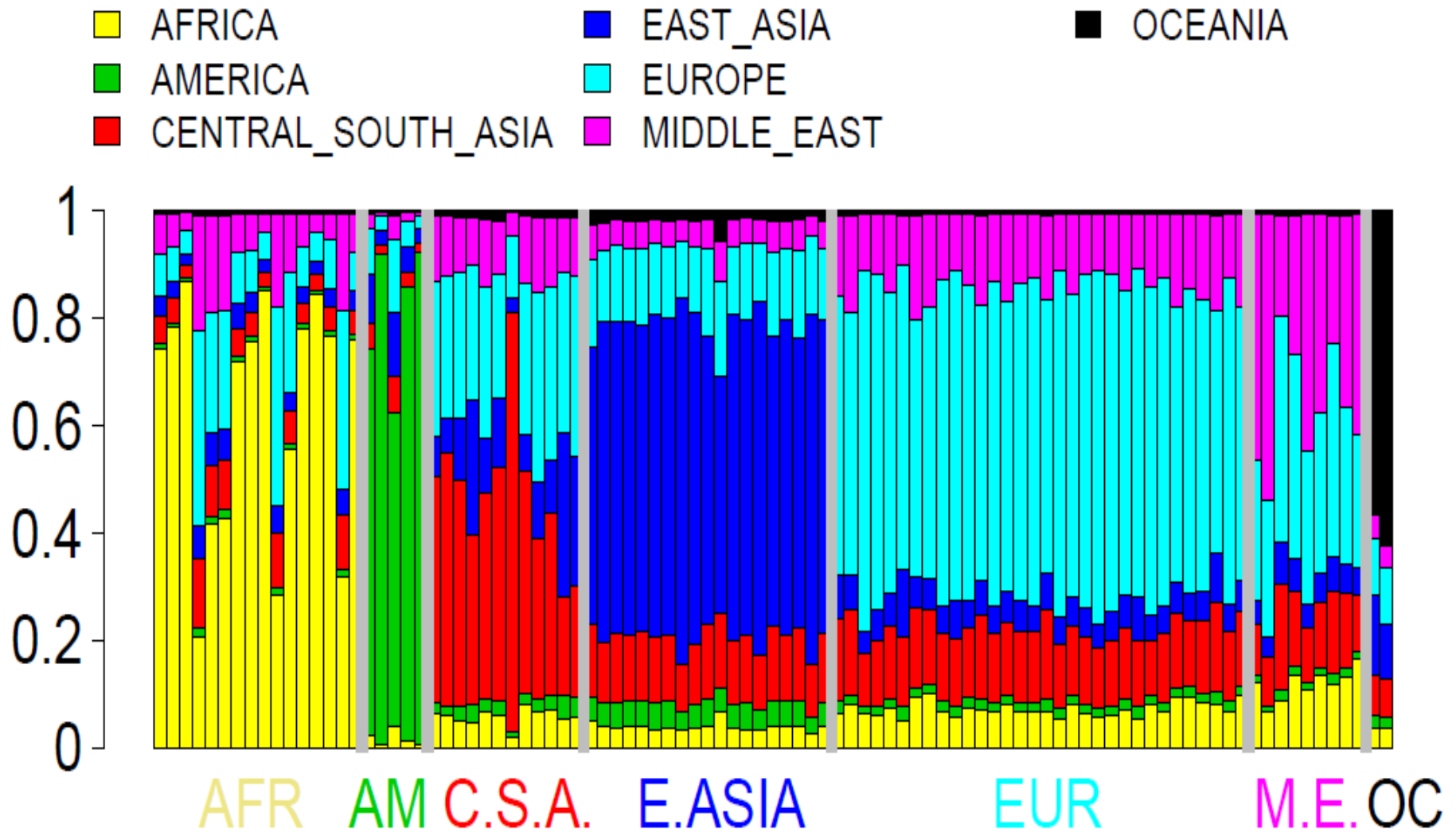
"YAKUT"

93. Yakut

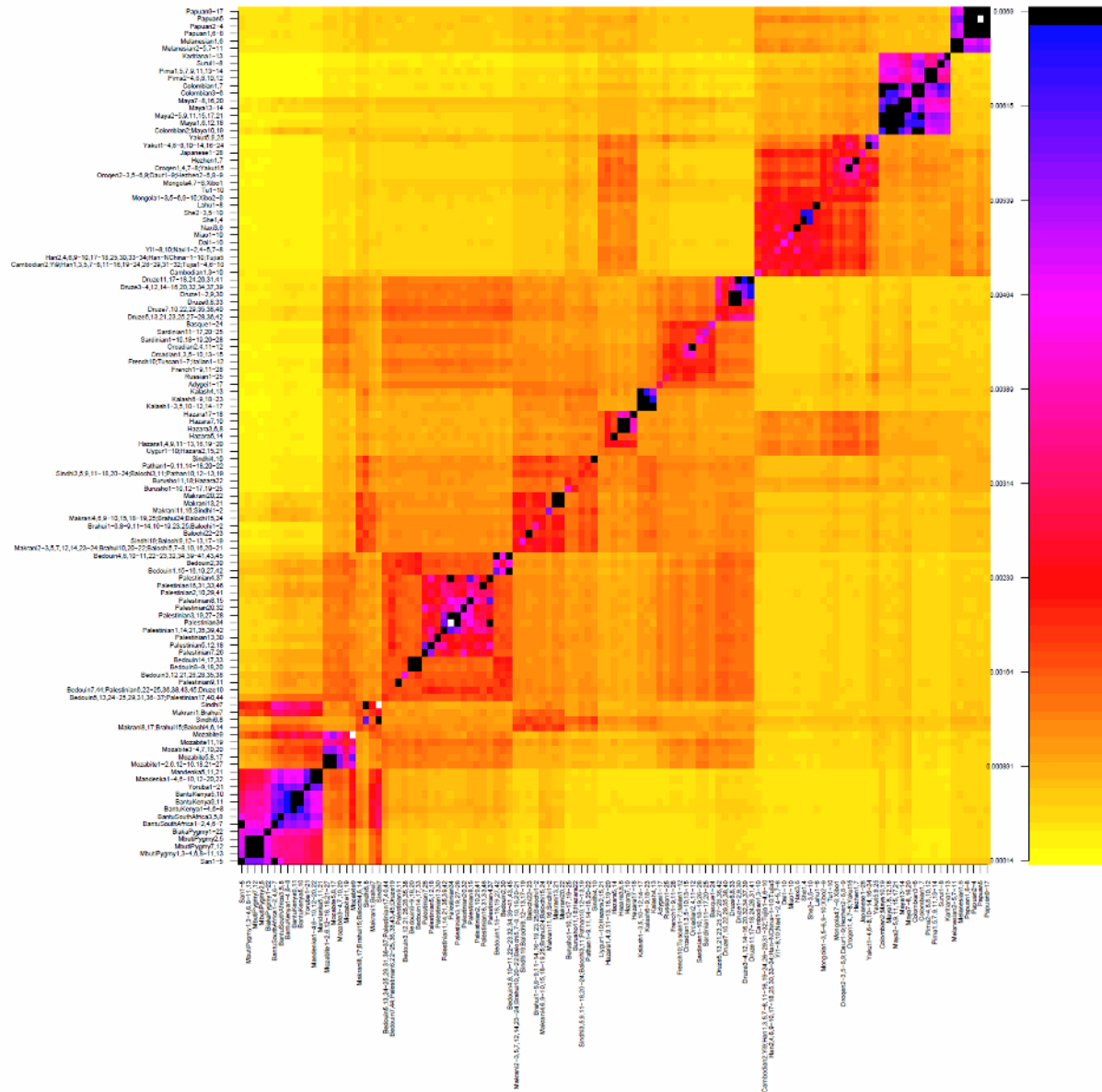
"OCEANIA"

94. Papuan
95. Melanesian

Worldwide palettes

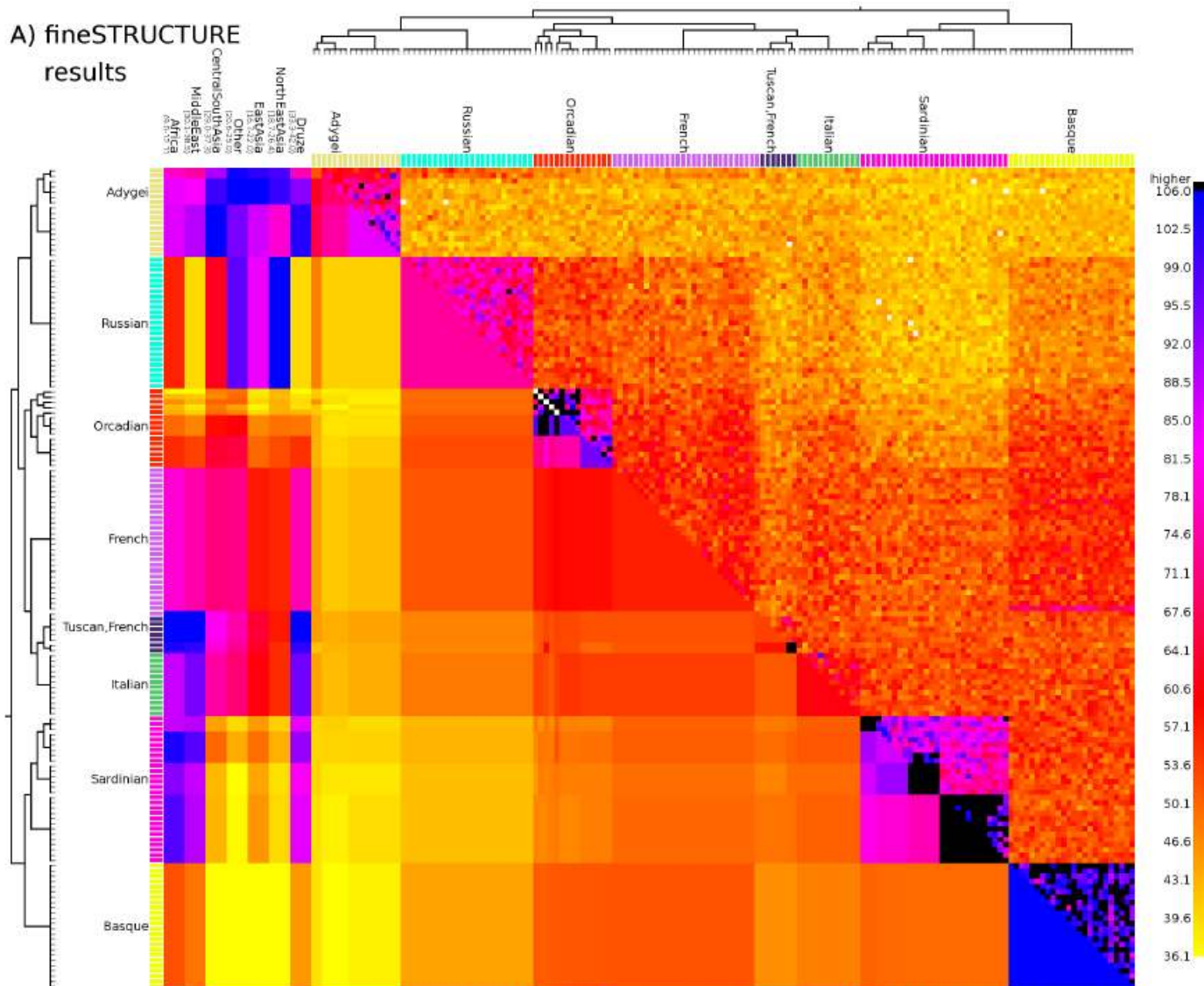


169 populations inferred in HGDP data

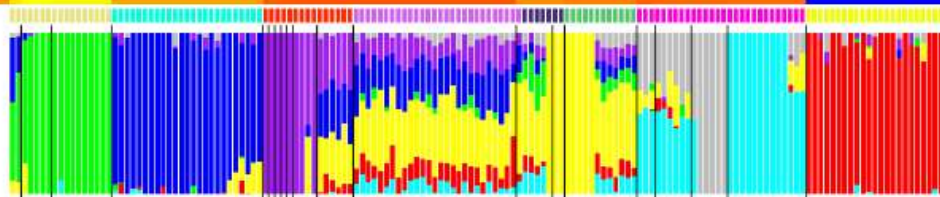


A) fineSTRUCTURE results

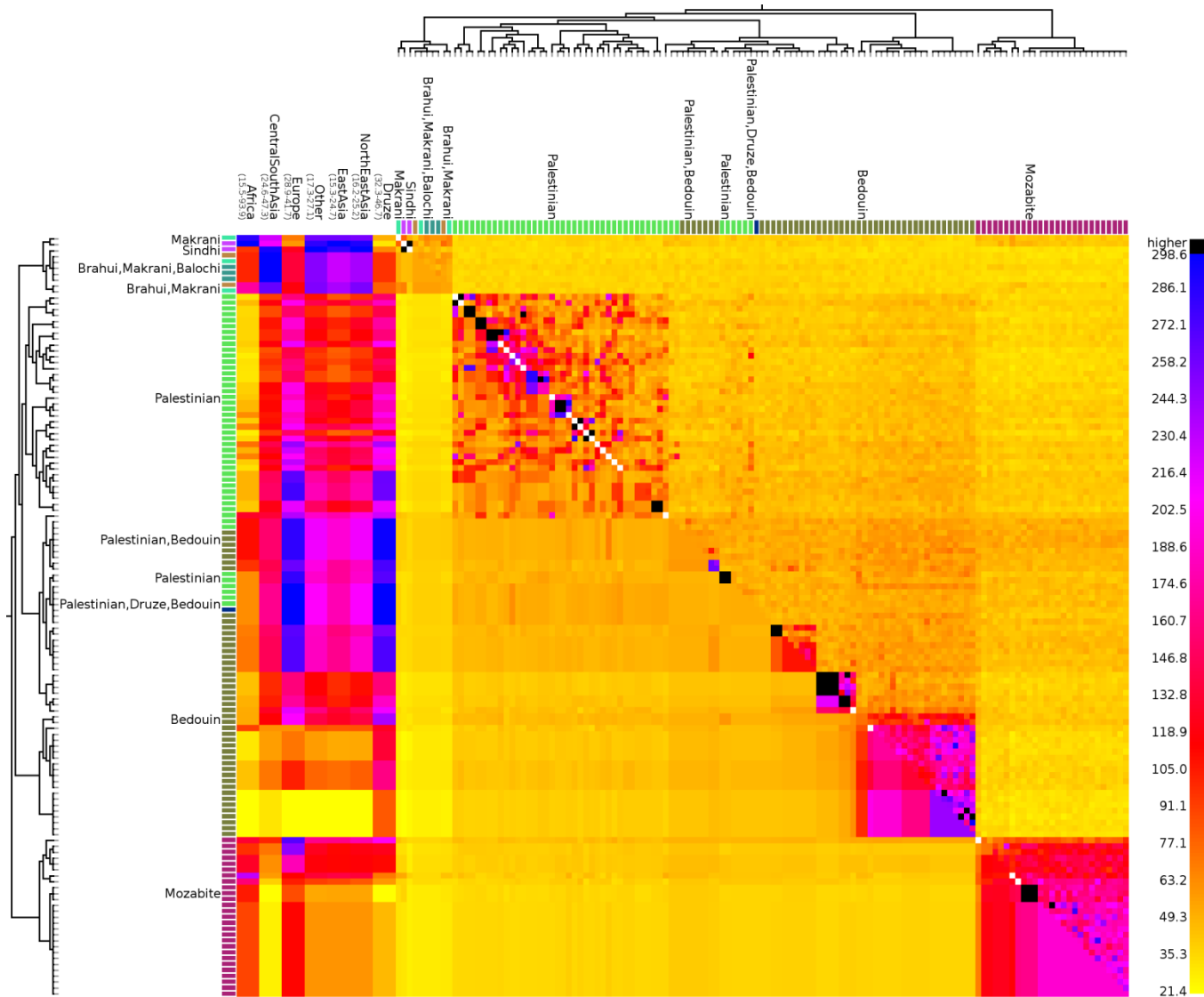
Adygei
Russian
Orcadian
French
Tuscan
Bergamo
Sardinian
Basque



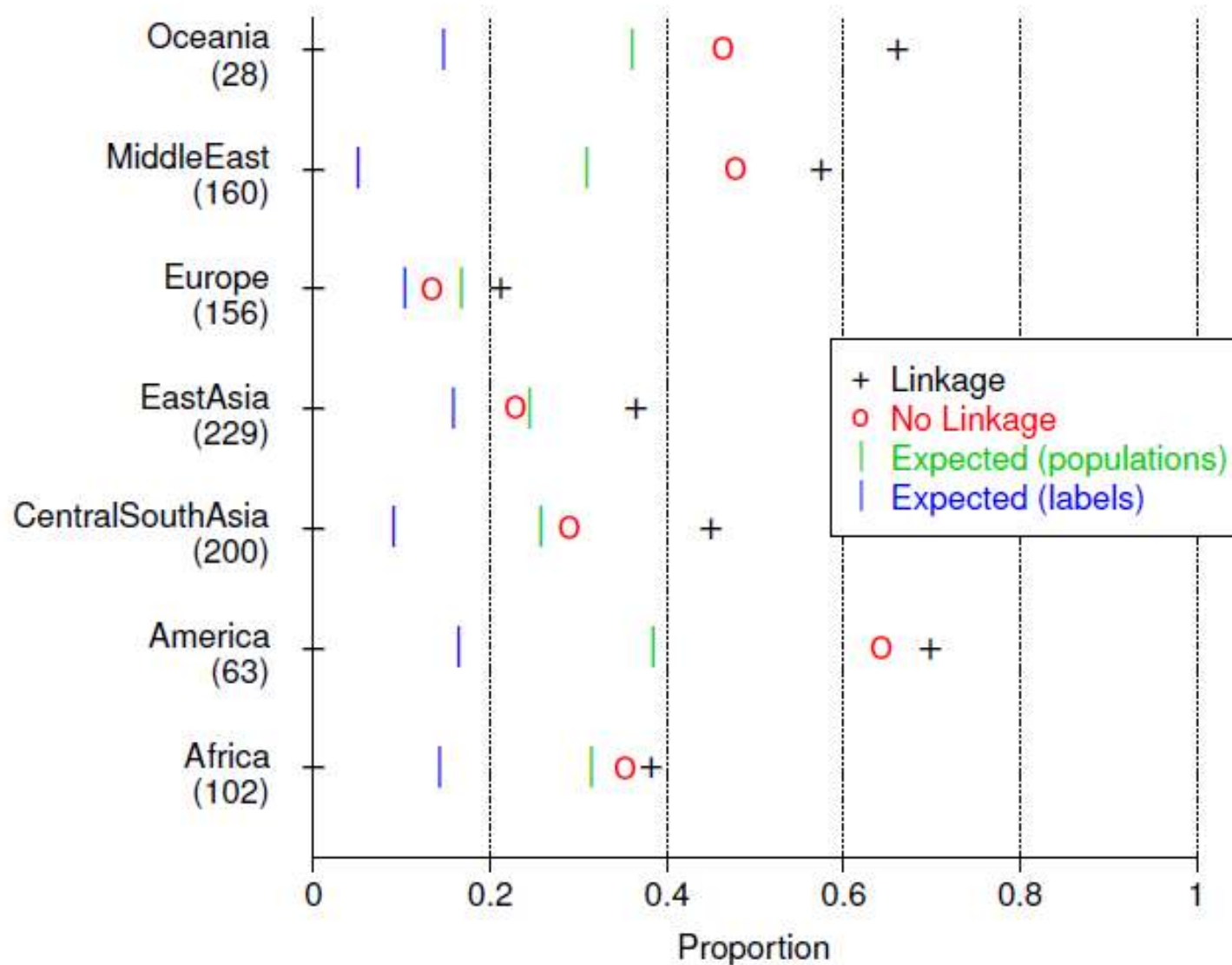
B) ADMIXTURE results (K=7)



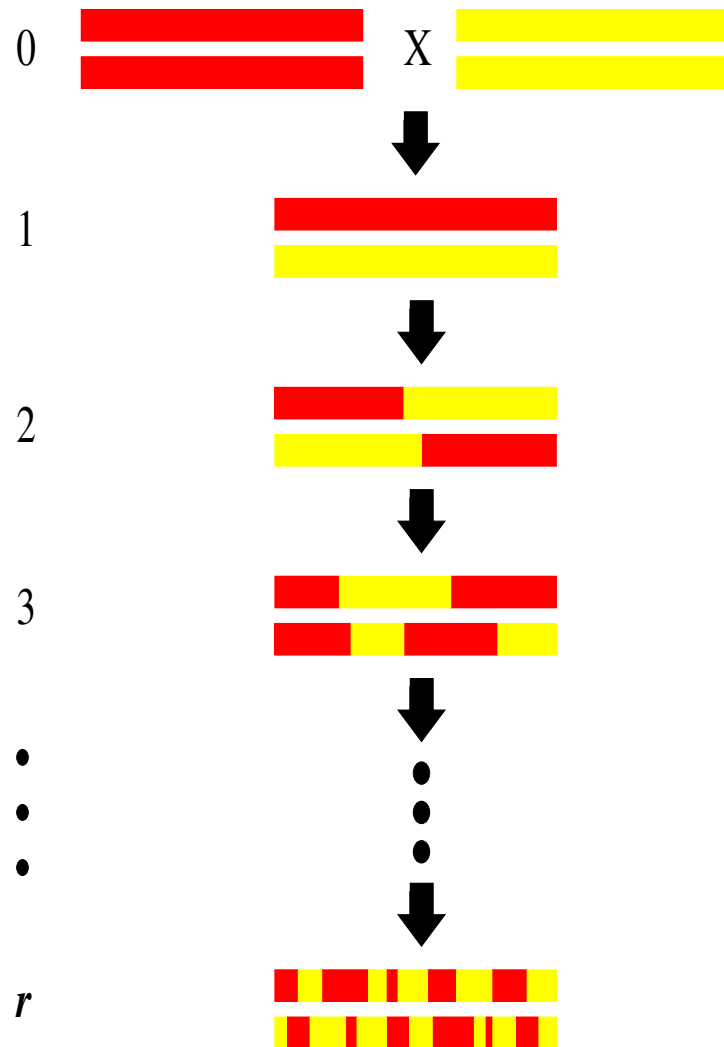
HGDP Middle east



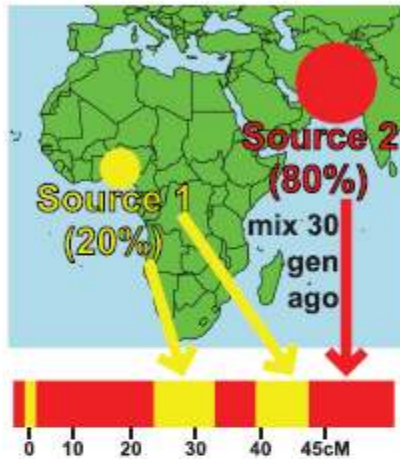
Half-matching based on coancestry matrix



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



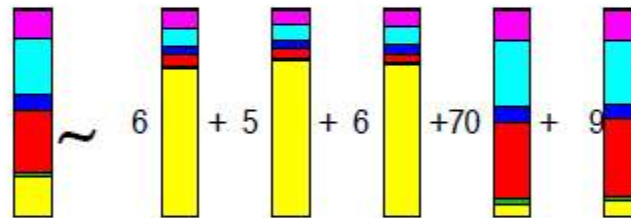
Globetrotter



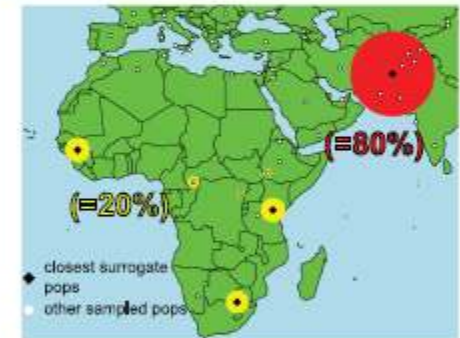
Simulated admixture
30 generations ago



Raw painting

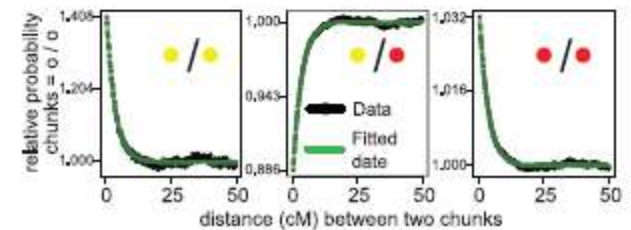
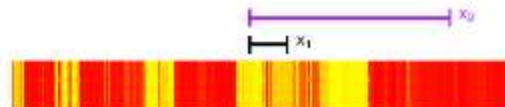


Mixture modelling of genomic palettes



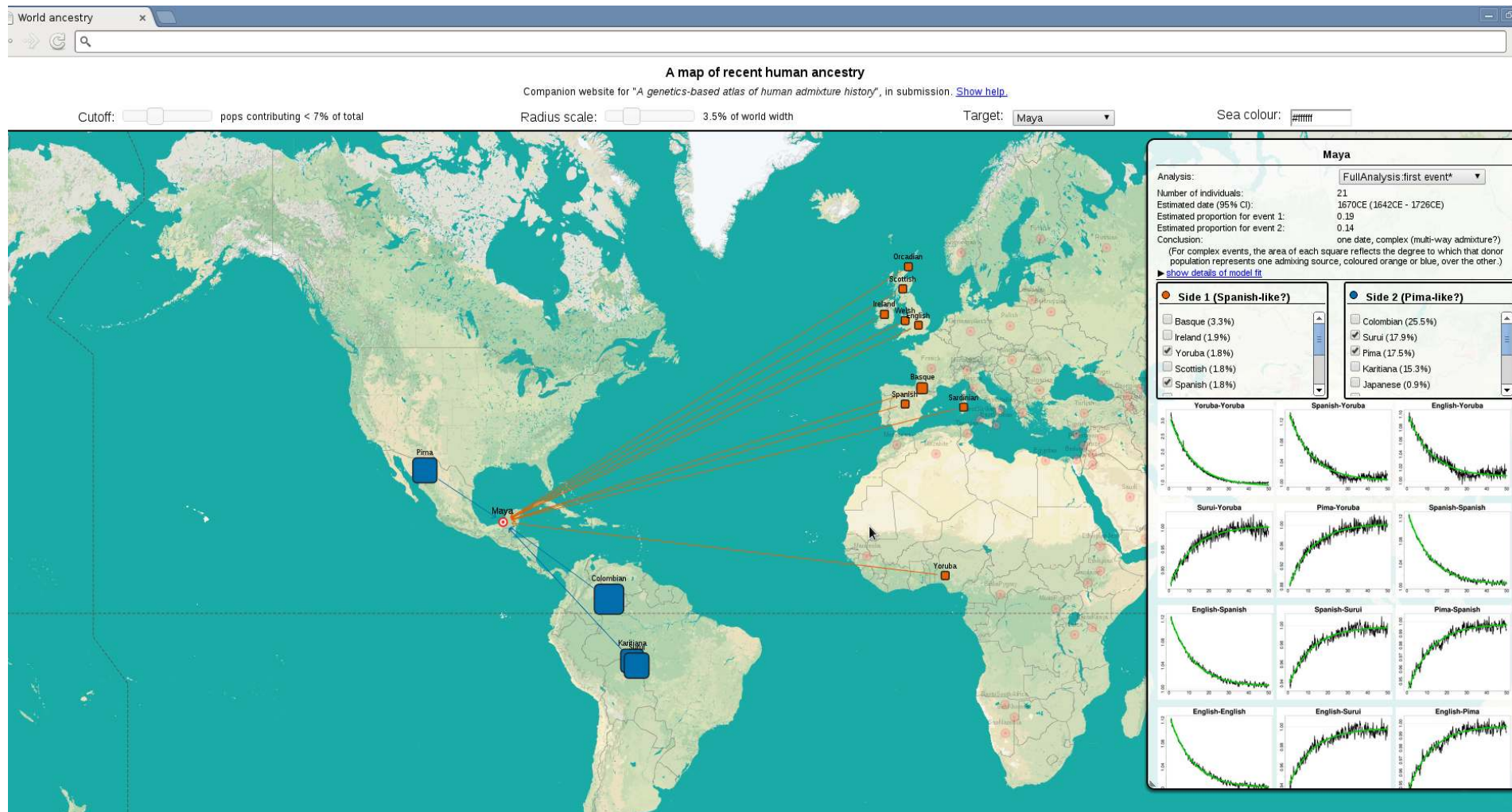
Cleaned painting

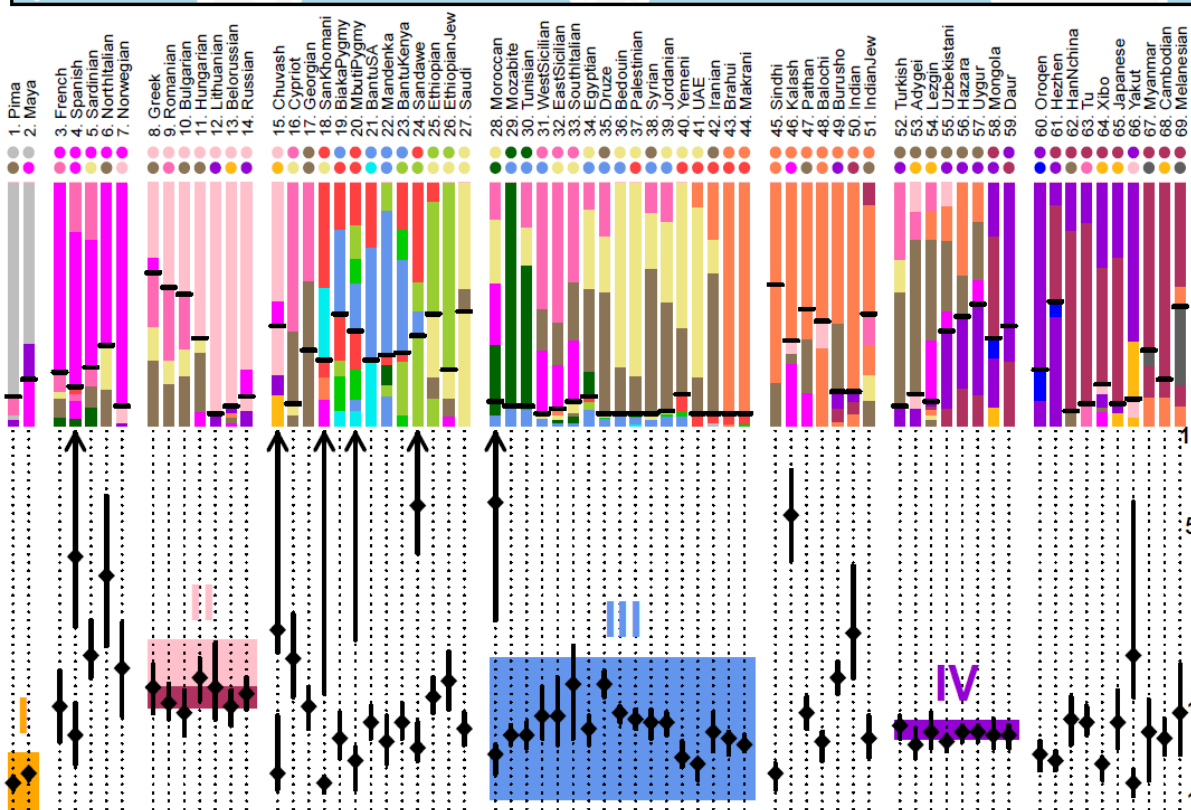
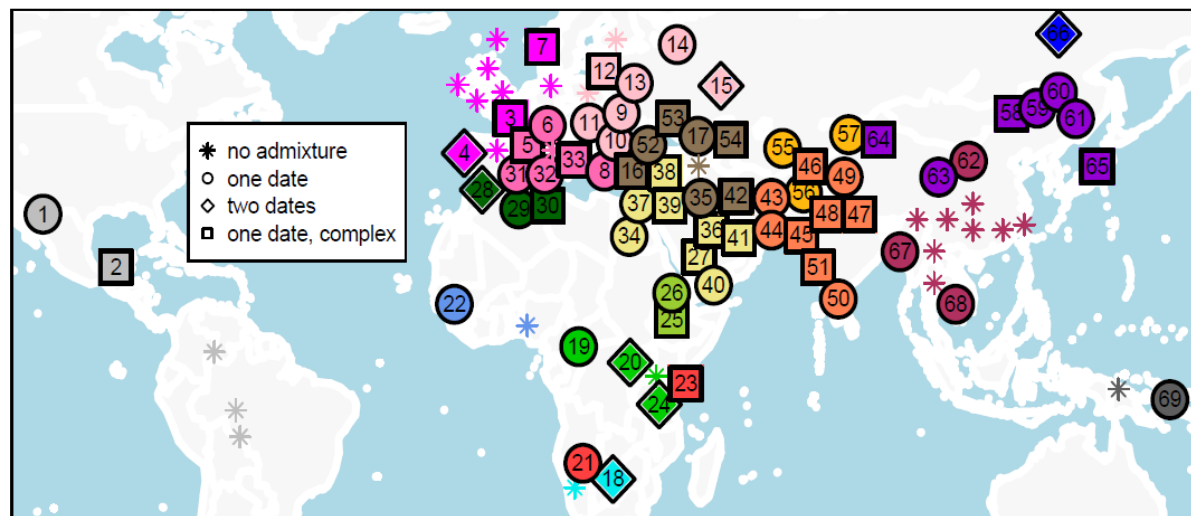
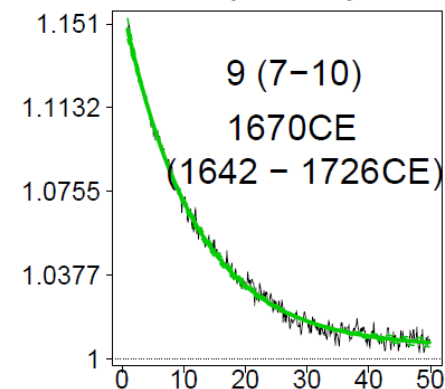
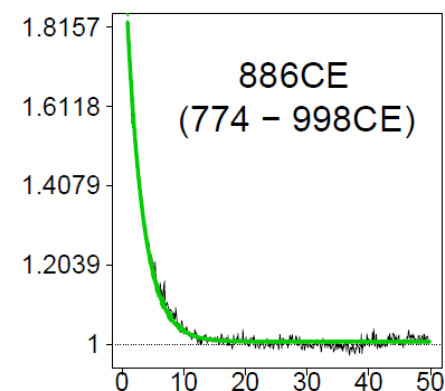
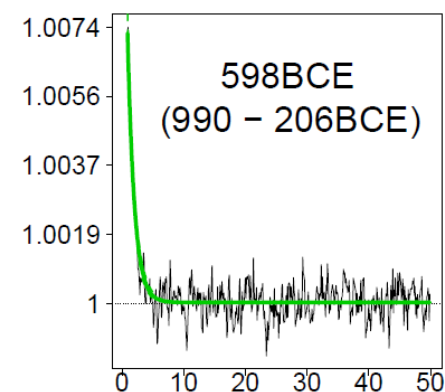
Fitting of spatial structure of variation in palettes along the chromosome



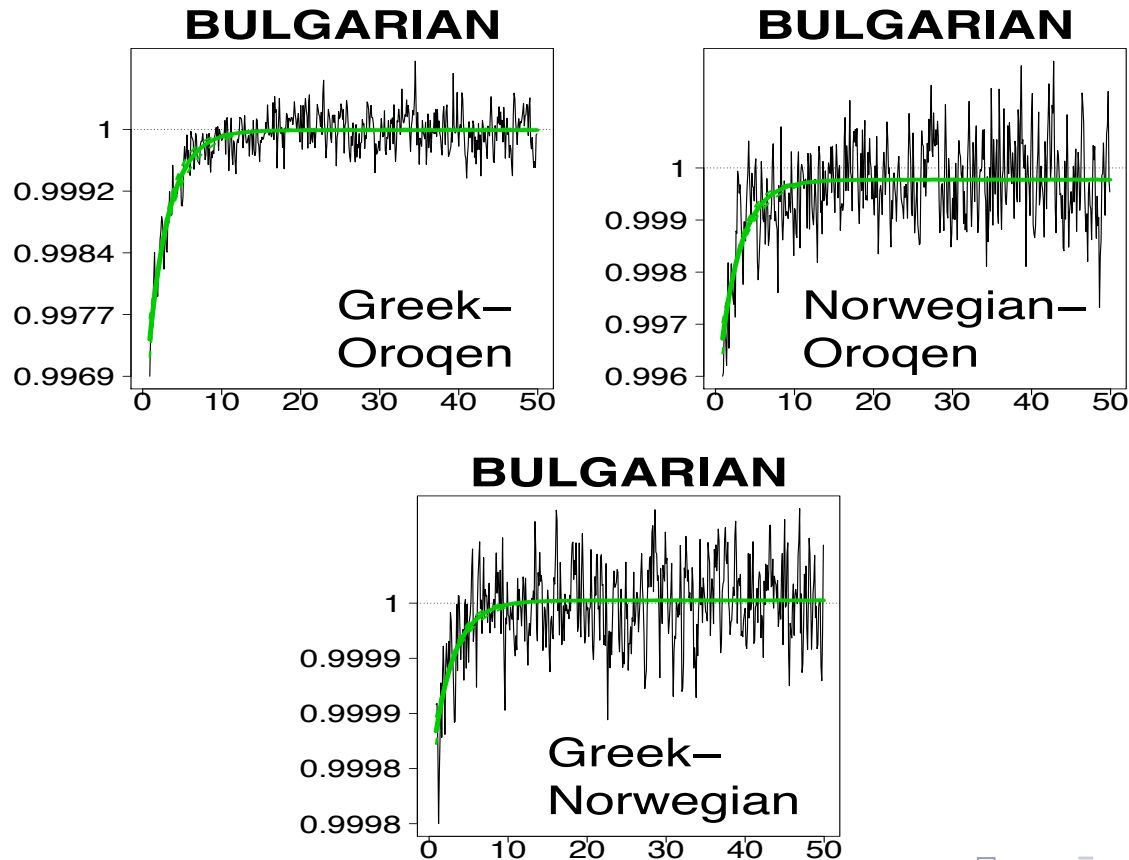
http://admixturemap.paintmychromosomes.com

(Hellenthal et al 2014)

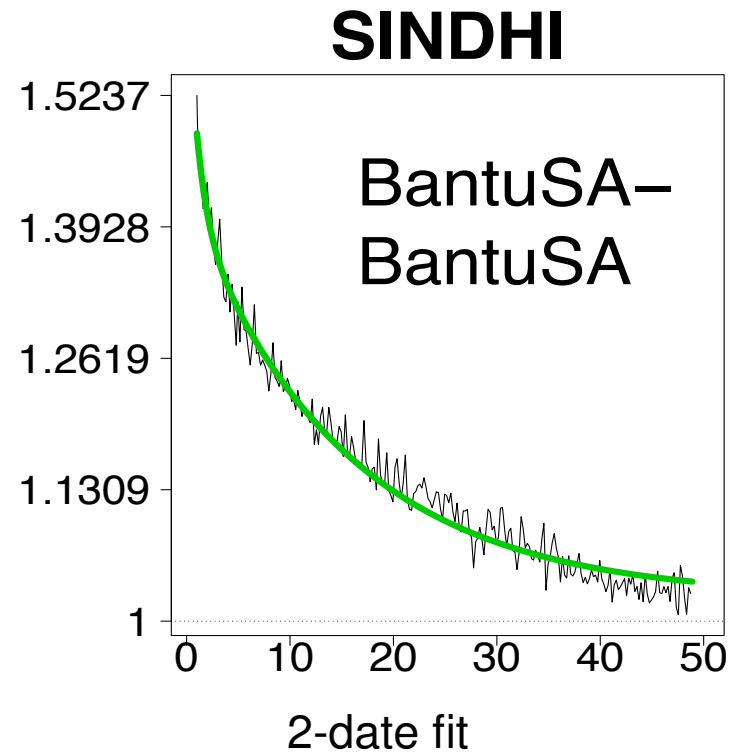
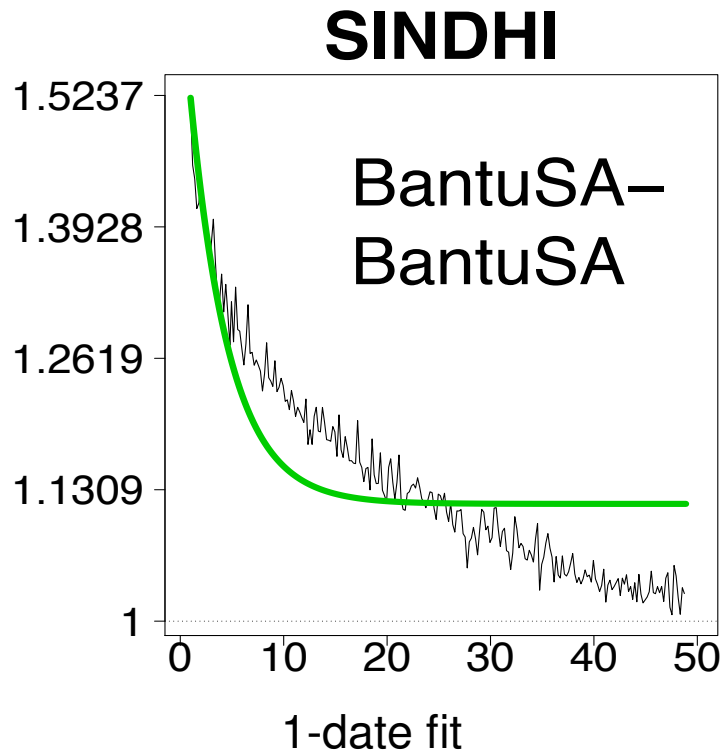


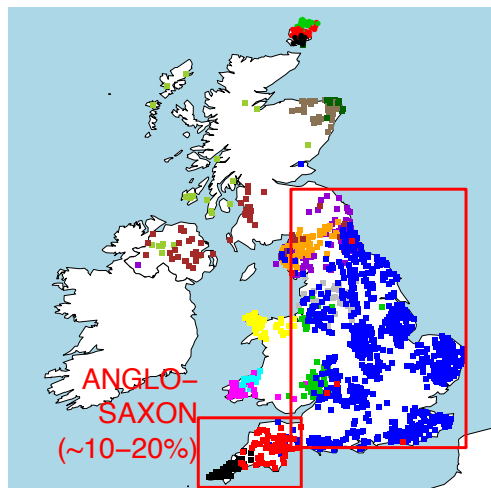
A**B****MAYA: Spanish-Spanish****C****DRUZE: Yoruba-Yoruba****D****KALASH: Scottish-Scottish**

Can also identify complex events such as multi-way admixture

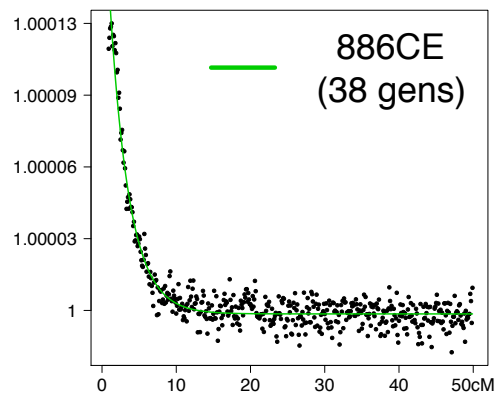


- can fit curves with sum of two exponential distributions

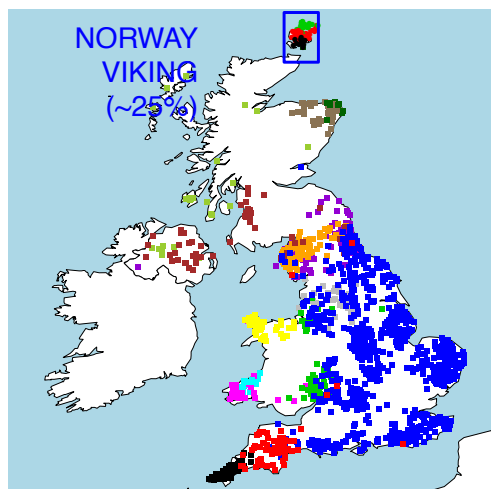
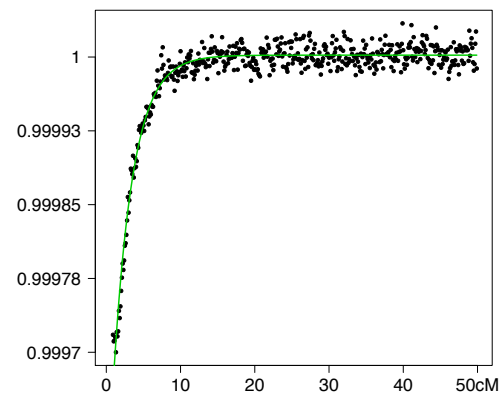




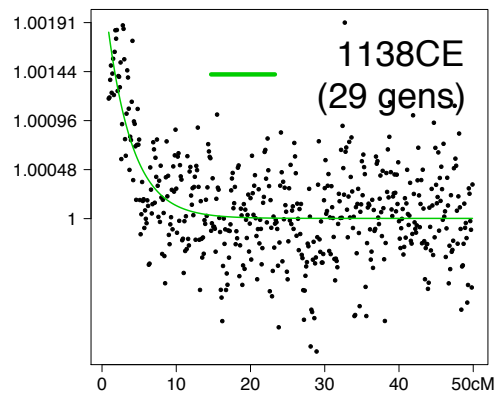
EU33 – EU33



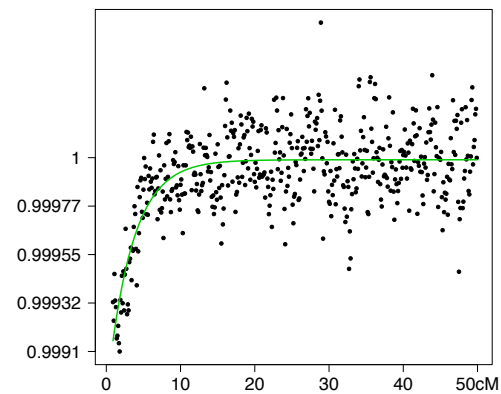
EU33 – EU28



EU17 – EU17



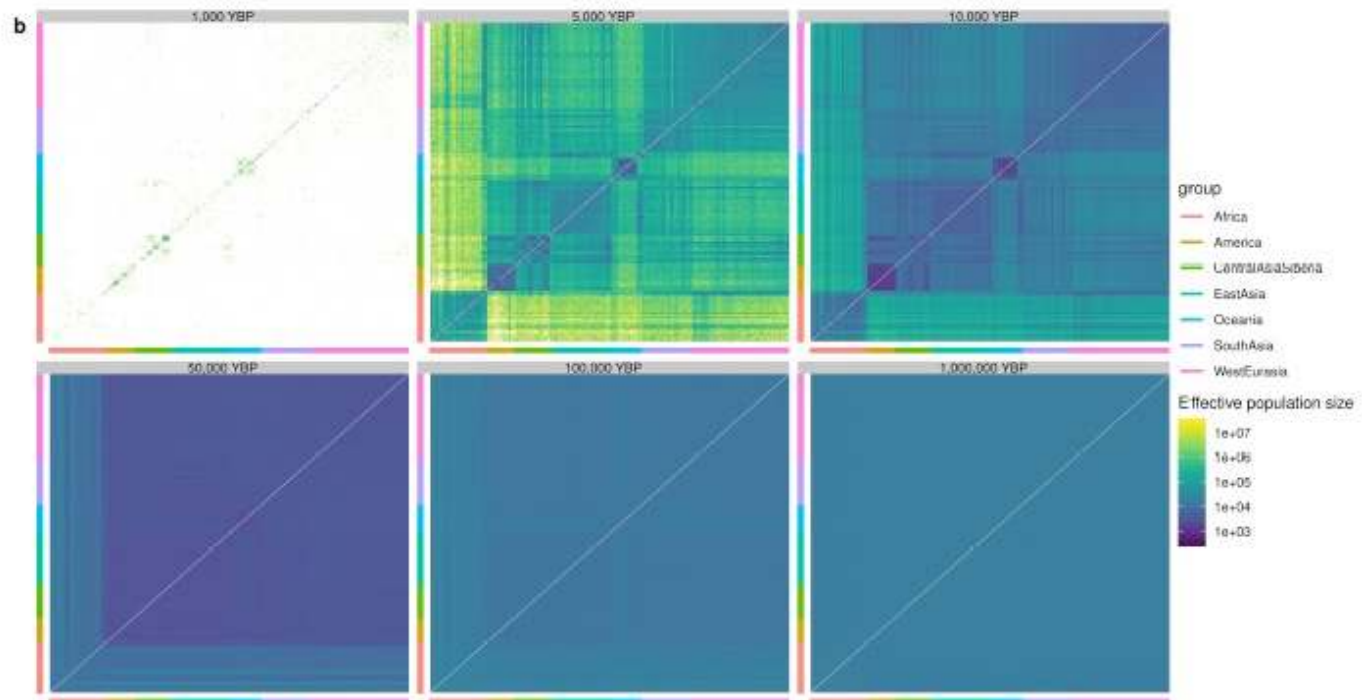
EU17 – EU31



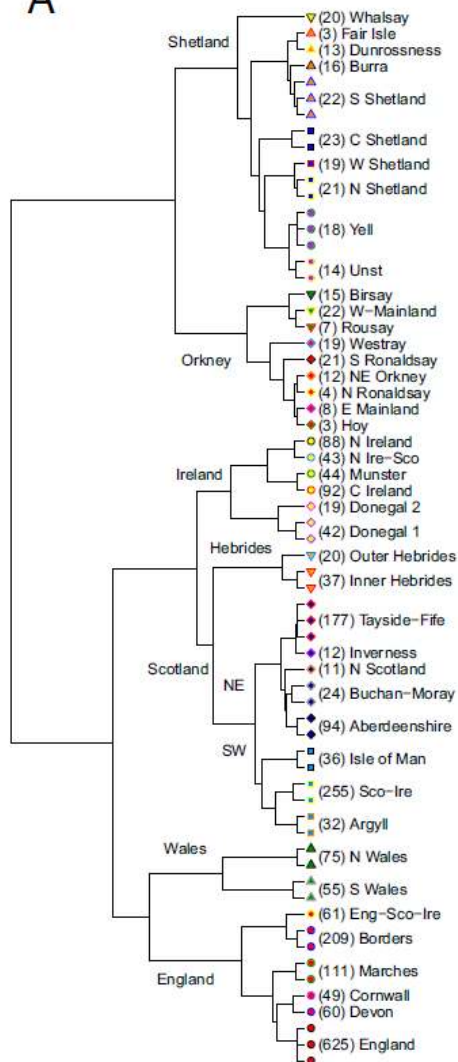
Conclusions

Can use chromosome painting to

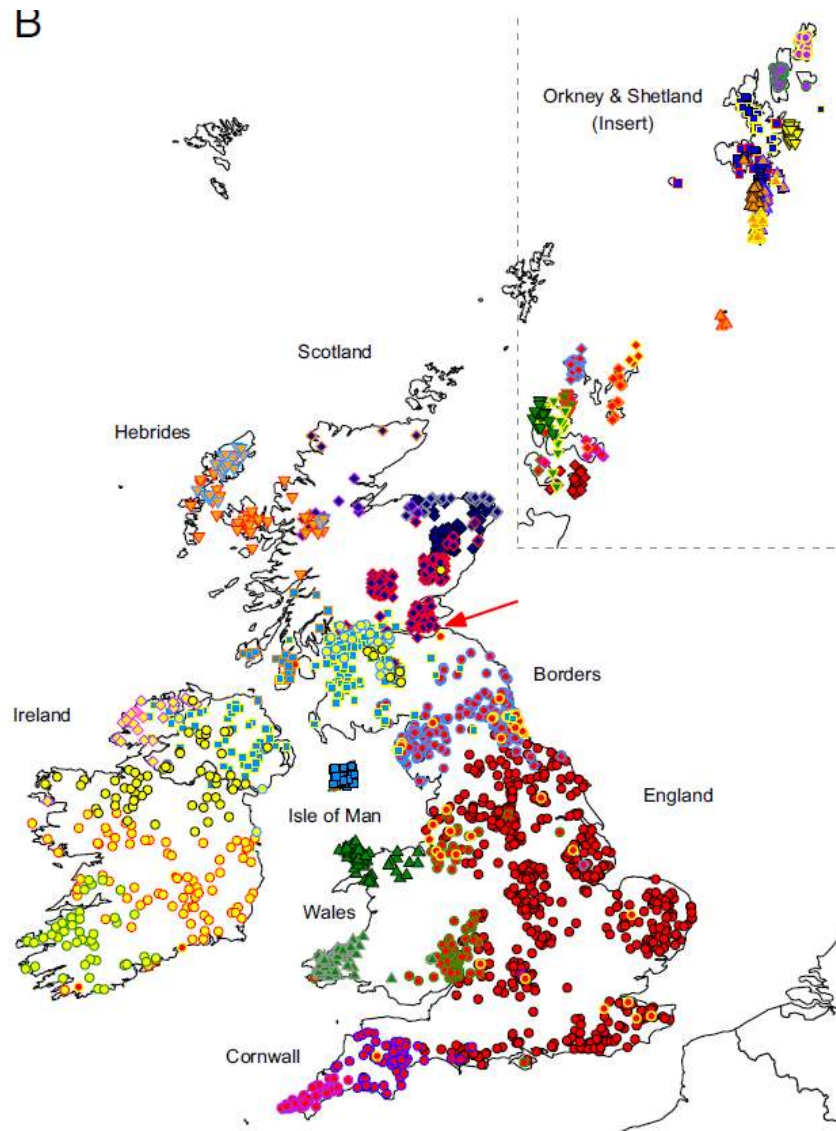
- Distill ancestry information
- Cluster based on genetic similarity
- Visualise genetic drift
- Identify mixtures
- Date mixture events
- Reconstruct history

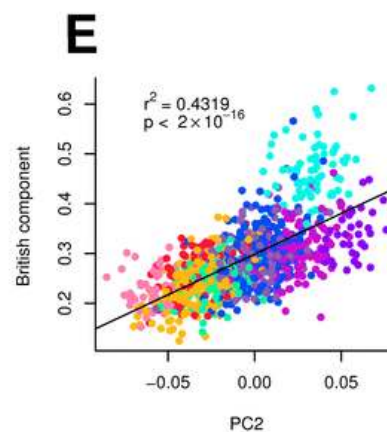
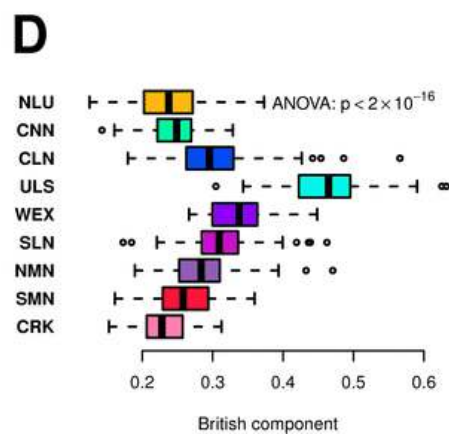
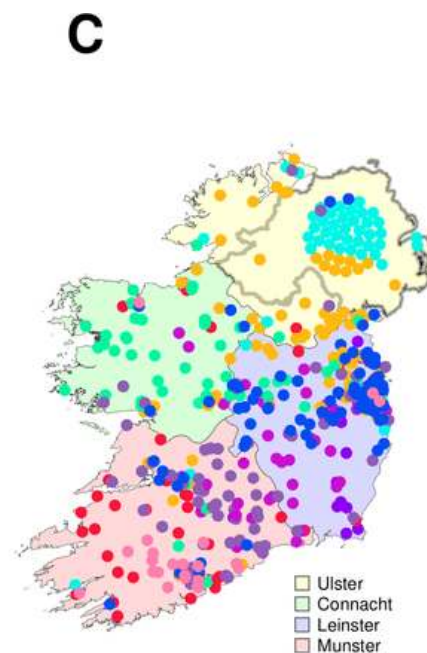
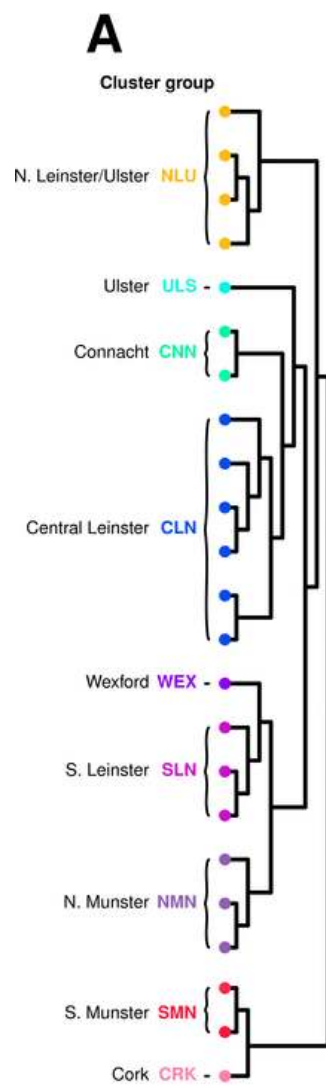


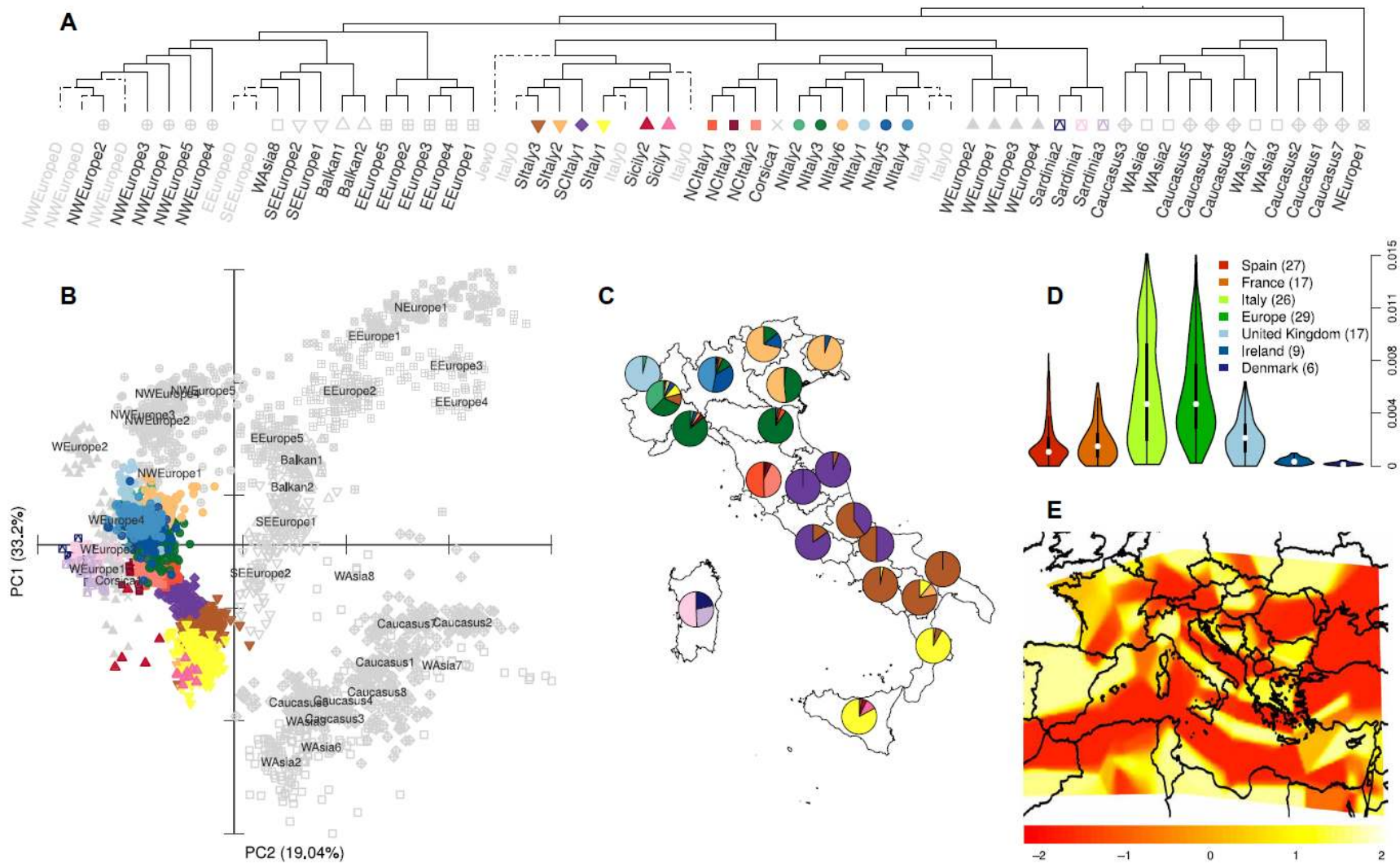
A

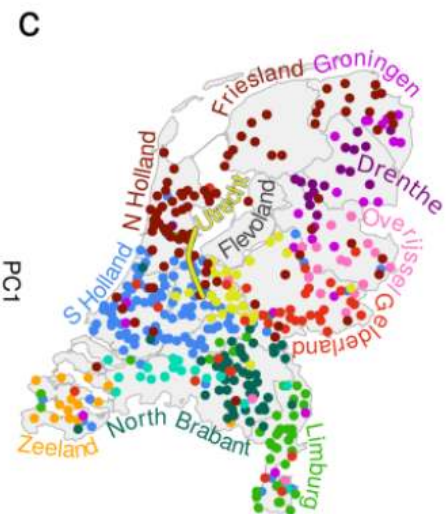
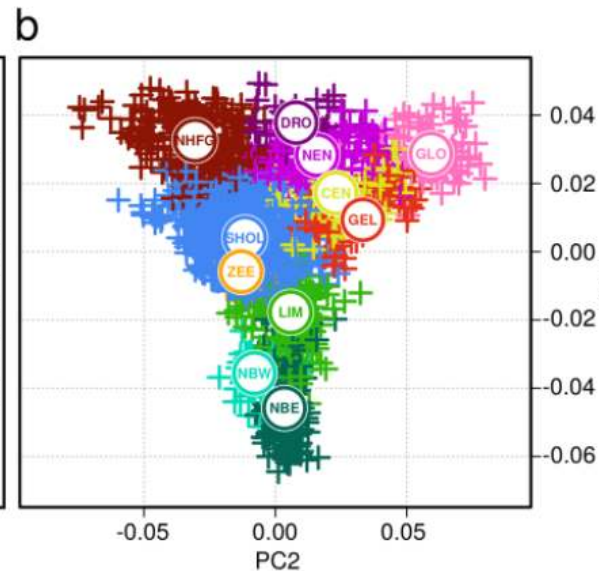
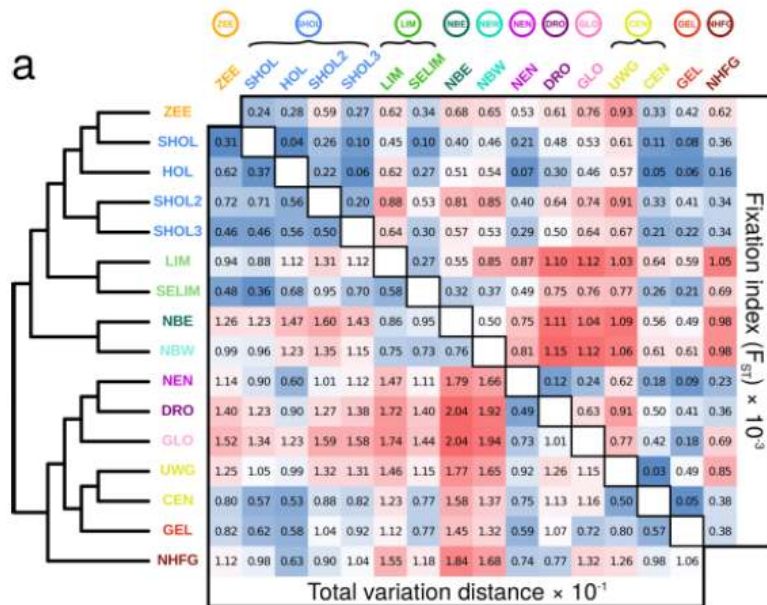


B









Cluster labels

ZEE Zeeland

SHOL South Holland

HOL Holland

SHOL2 South Holland 2

SHOL3 South Holland 3

LIM Limburg

SELIM Southeast Limburg

NBE North Brabant (east)

NBW North Brabant (west)

NEN Northeast Netherlands

DRO Drenthe and Overijssel

GLO Gelderland and Overijssel

UWG Utrecht and west Gelderland

CEN Central Netherlands

GEL Gelderland

NHFG North Holland, Friesland and Groningen

Cluster group labels

ZEE Zeeland

SHOL South Holland

LIM Limburg

NBE North Brabant (east)

NBW North Brabant (west)

NEN Northeast Netherlands

DRO Drenthe and Overijssel

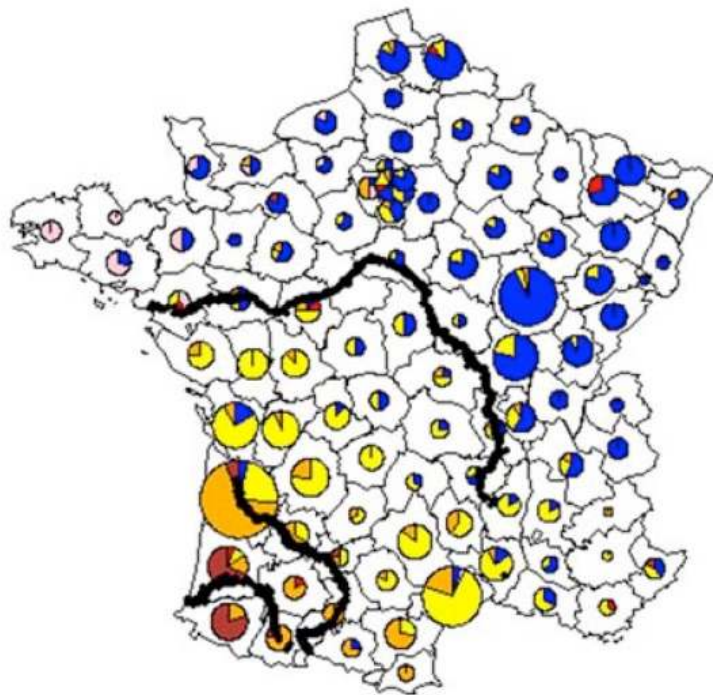
GLO Gelderland and Overijssel

CEN Central Netherlands

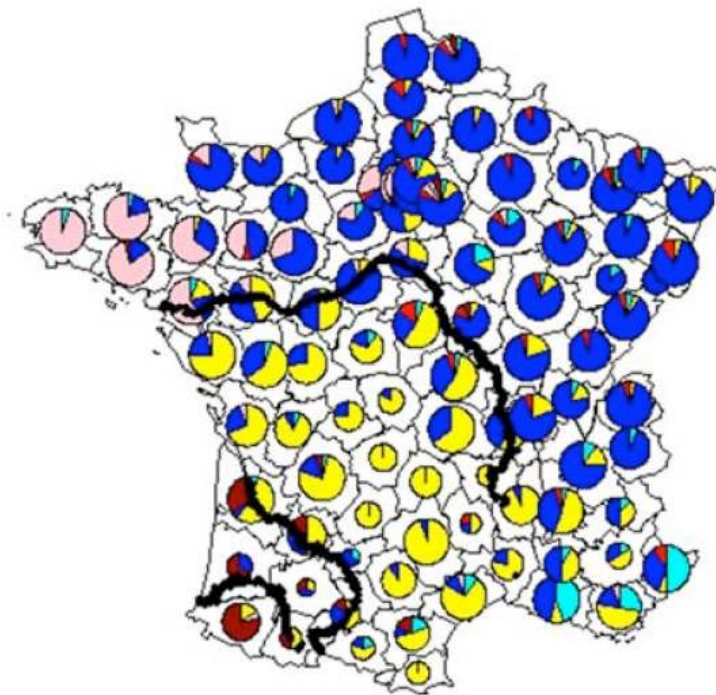
GEL Gelderland

NHFG North Holland, Friesland and Groningen

3C



SU.VI.MAX



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.

	$y^k[k]$	$y^{k+1}[k+1]$
1	1101110010010000000000100010000011000000100100010101000100110000100110000010000100	1 00000001101
2	00011111000101010000100100011000000110000010100011110001000101000100110000100	1 10010000011
3	10011100100101010000100100011000000110000010100010001000000111010000010000010100100	1 10010001001
4	10011100010101010000100100011000000110000010100010001000000111010000010000010100100	1 10010001001
5	10011100100110001101000100011010011010000001110010000000100101010100110000100000010	1 01000001001
6	10011111010101010000000100010101101001000000000100010001100100000000000111100000010	1 01000001001
7	00100000010101010000000100010101100000100100100100100111100100000000000111100000010	1 01000001001
8	000111110001010100001001000110000000010000000001111000100010100000000111100000010	1 01000001001
9	1001100001000101010010010001100010000100010010001000000000101010000110000010000010	0 1000011001
10	10100011010100010000000100010101100000100100100001100001001100000001000000010100010	1 00001000011
11	1010111101010011000001101010001000010001001001001000000100101010100110000110100010	1 00001000011
12	10011100100110100000000110101000100001000100100011100000000101011000010000101100010	0 1000011001
13	1001100001010101000010010001100000000000010100100100000100101010100010000101100010	0 1000011001
14	1001111100010101000010010001100010000000010001000100001000100001000100000101100010	0 10001011001
15	1001111100010101111000000001000000000000100100001100001000100000100010000010110010	0 00001000101
16	10000000010101001010001100101000000110000010000001100001000110100000000110010110010	0 00001000101
17	1001111100100010011000100011000000000100100100001100001000100001000001000100101010	1 00000001101
18	10011100100100000000000100011010011010000001110010001000111010000110000010010110	0 01000001101
19	1001110010011000110100010001101001101000000111001000000100101010100110000100110011	0 10000001001
20	00011100100101010000100100011000000110000010100100100111000101011100110000010110011	0 1010000101

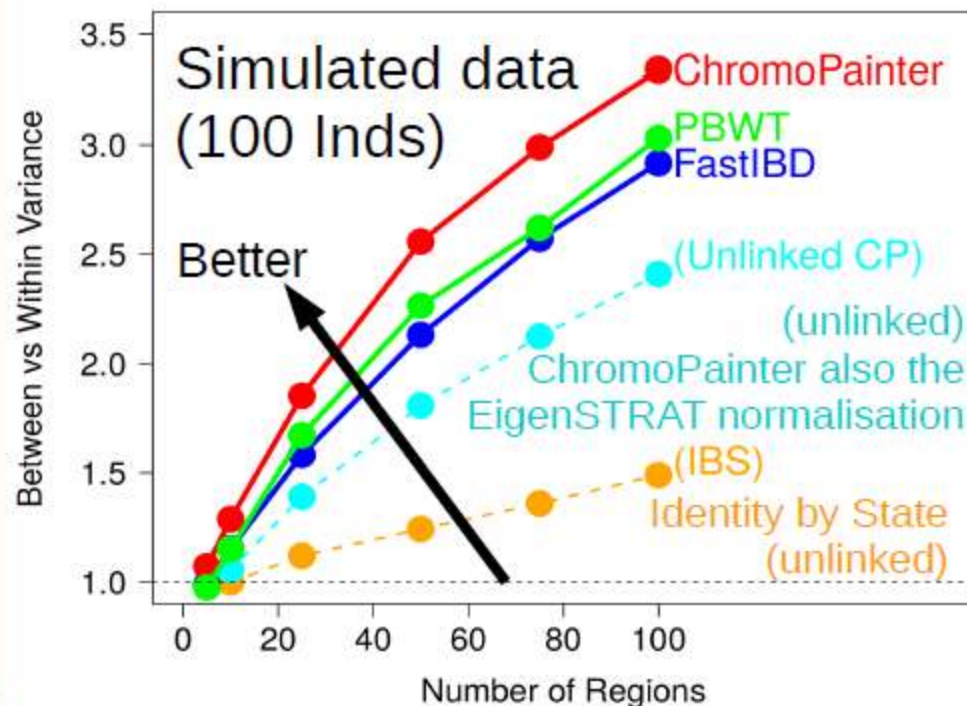
←
Reverse sorted prefixes at k

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)