Improved ancestry and admixture detection using principal component analysis (PCA) of genetic data

Florian Privé

Aarhus University (DK)

privefl

Genetic Ancestry Deconvolution

All individuals are genetically admixed
from L reference populations

Admixture model and ADMIXTURE method

$$G \approx Q \cdot 2F$$

- Q are the admixture proportions (for each sample i and reference l)
- F are the allele frequencies (for each each reference l and variant j)

Admixture model and ADMIXTURE method

$$G \approx Q \cdot 2F$$

- Q are the admixture proportions (for each sample i and reference l)
- F are the allele frequencies (for each each reference l and variant j)

ADMIXTURE uses Maximum Likelihood Estimation of

$$L(Q, F) = \sum_i \sum_j \left\{ G_{i,j} \log \left[\sum_l Q_{i,l} F_{l,j} \right] + (2 - G_{i,j}) \log \left[1 - \sum_l Q_{i,l} F_{l,j} \right] \right\}$$

with constraints: $0 \leq F_{l,j} \leq 1$ and $Q_{i,l} \geq 0$ and $\sum_l Q_{i,l} = 1$

Admixture model and ADMIXTURE method

$$G \approx Q \cdot 2F$$

- Q are the admixture proportions (for each sample i and reference l)
- F are the allele frequencies (for each each reference l and variant j)

ADMIXTURE uses Maximum Likelihood Estimation of

$$L(Q, F) = \sum_i \sum_j \left\{ G_{i,j} \log \left[\sum_l Q_{i,l} F_{l,j} \right] + (2 - G_{i,j}) \log \left[1 - \sum_l Q_{i,l} F_{l,j} \right] \right\}$$

with constraints: $0 \leq F_{l,j} \leq 1$ and $Q_{i,l} \geq 0$ and $\sum_l Q_{i,l} = 1$

For simplicity, ADMIXTURE iteratively estimates

- each $Q_{i,\cdot}$ independently, with F fixed
- each $F_{\cdot,j}$ independently, with Q fixed

My proposed deconvolution method

$$G \cdot V \approx Q \cdot 2F \cdot V$$

where V are the PC loadings of G

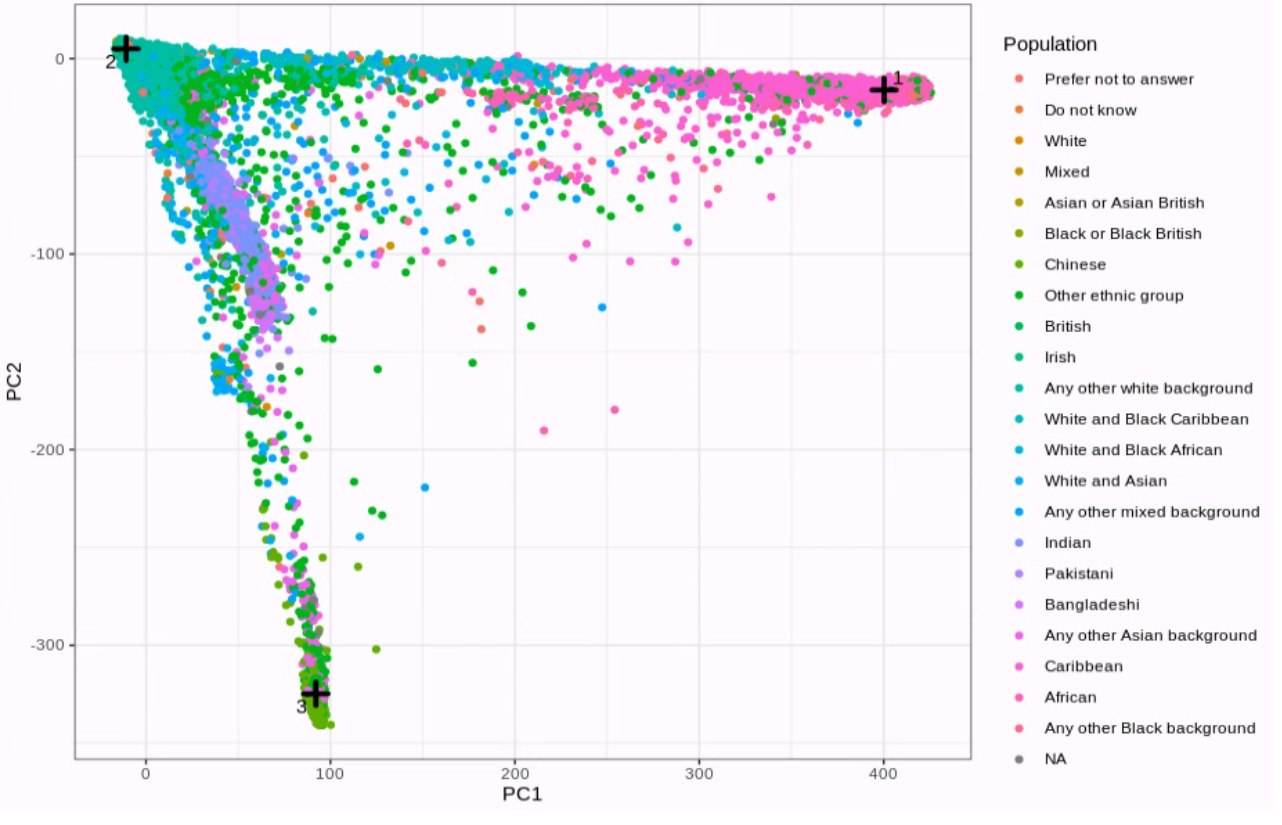
$$\Rightarrow \; PC \approx Q \cdot PC^{\text{ref}}$$

My proposed deconvolution method

$$G \cdot V \approx Q \cdot 2F \cdot V$$

where V are the PC loadings of G

$$\Rightarrow \; PC \approx Q \cdot PC^{\text{ref}}$$



Estimating admixture coefficients $Q_{i,.}$ with PC^{ref} fixed

$$\min_{\substack{\forall l,\, Q_{i,l} \geq 0 \\ \sum_l Q_{i,l} = 1}} \quad \sum_{k=1}^{K} \left(PC_{i,k} - \sum_{l=1}^{L} Q_{i,l} PC^{\text{ref}}_{l,k} \right)^2$$

Estimating admixture coefficients $Q_{i,.}$ with PC^{ref} fixed

$$\min_{\substack{\forall l,\, Q_{i,l} \geq 0 \\ \sum_l Q_{i,l} = 1}} \sum_{k=1}^{K} \left(PC_{i,k} - \sum_{l=1}^{L} Q_{i,l} PC^{\text{ref}}_{l,k} \right)^2$$

I've already published this method here:

JOURNAL ARTICLE

Using the UK Biobank as a global reference of worldwide populations: application to measuring ancestry diversity from GWAS summary statistics

Florian Privé

Bioinformatics, Volume 38, Issue 13, July 2022, Pages 3477–3480, https://doi.org/10.1093/bioinformatics/btac348

Published: 23 May 2022

- 18 reference groups curated from the UK Biobank
- provide reference allele frequencies and PC loadings
- work for both individual-level data or GWAS allele frequencies only
- more power when doing the minimization in the PCA space

Estimating admixture coefficients $PC^{\text{ref}}_{l,\cdot}$ with Q fixed

$$PC^{\text{ref}}_{l,\cdot} = \frac{\sum_i {Q_{i,l}}^m \cdot PC_{i,\cdot}}{\sum_i {Q_{i,l}}^m}$$

- this simple formula is used in e.g. fuzzy K-means
- this is also related to archetypal analysis: $PC^{\text{ref}} = W^T \cdot PC$
  (references are weighted combinations of existing samples)

  $\Rightarrow$ Reference allele frequencies: $2F = W^T \cdot G$

Complete deconvolution algorithm

Iterate between

- estimating admixture coefficients $Q_{i,.}$, with PC^{ref} fixed
- estimating reference positions $PC^{\text{ref}}_{l,.}$, with Q fixed

Complete deconvolution algorithm

Iterate between

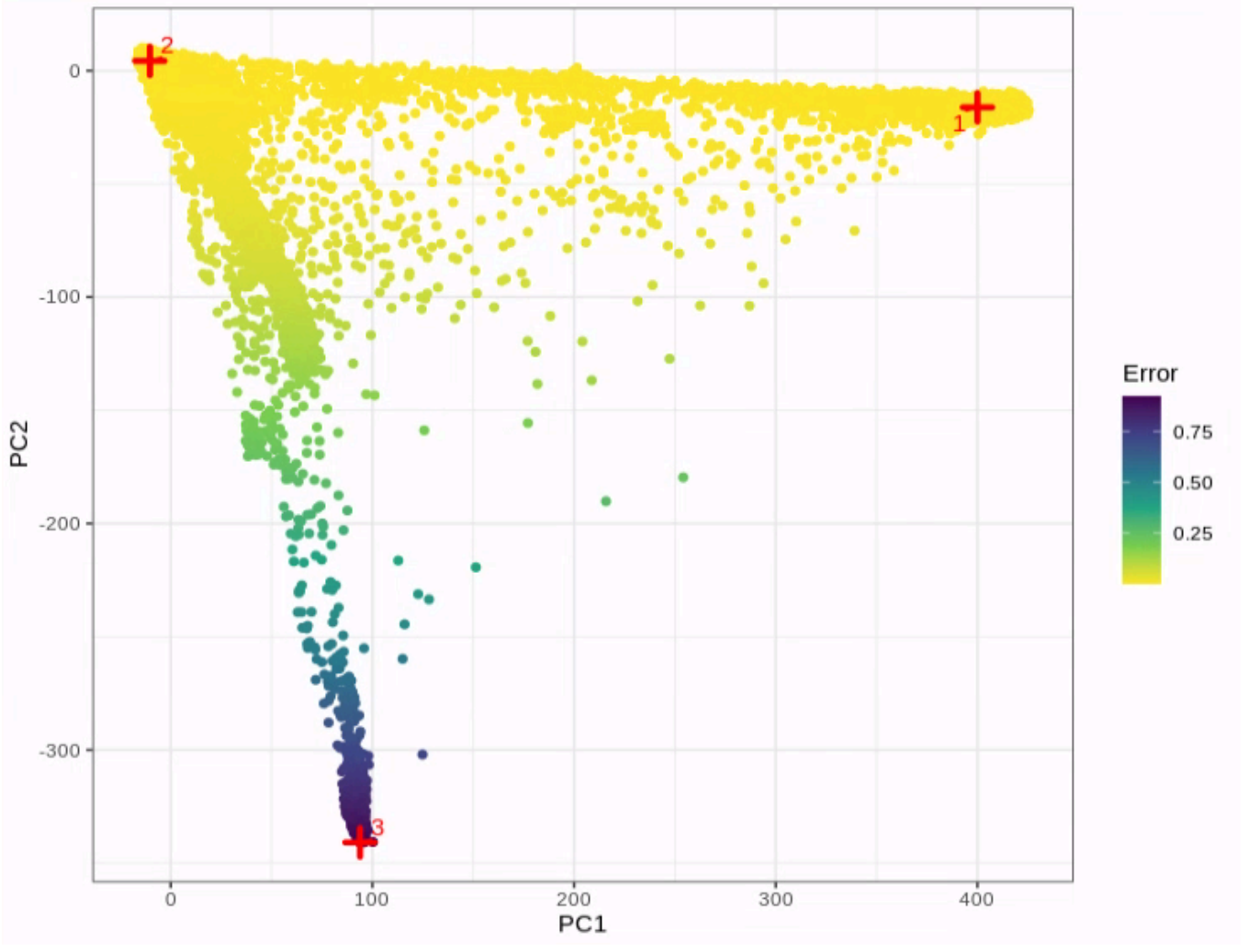
- estimating admixture coefficients $Q_{i,.}$, with PC^{ref} fixed
- estimating reference positions $PC^{\text{ref}}_{l,.}$, with Q fixed

But a starting point is needed..

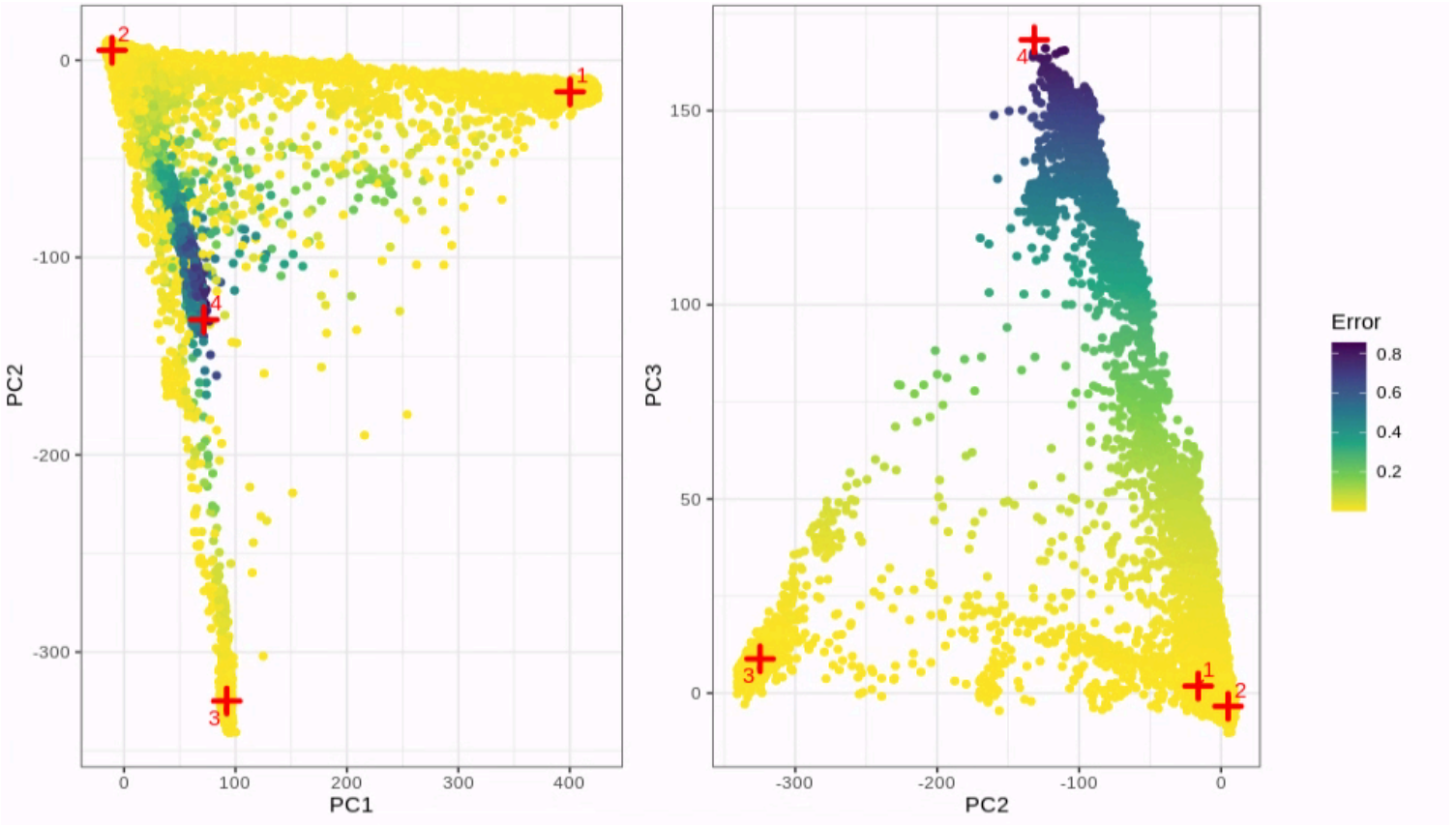
A naive approach would pick L initial $PC^{\text{ref}}_{l,.}$ at random.

Instead, I use an **iterative procedure with warm starts**
to make the algorithm **deterministic** and much **faster to converge**.

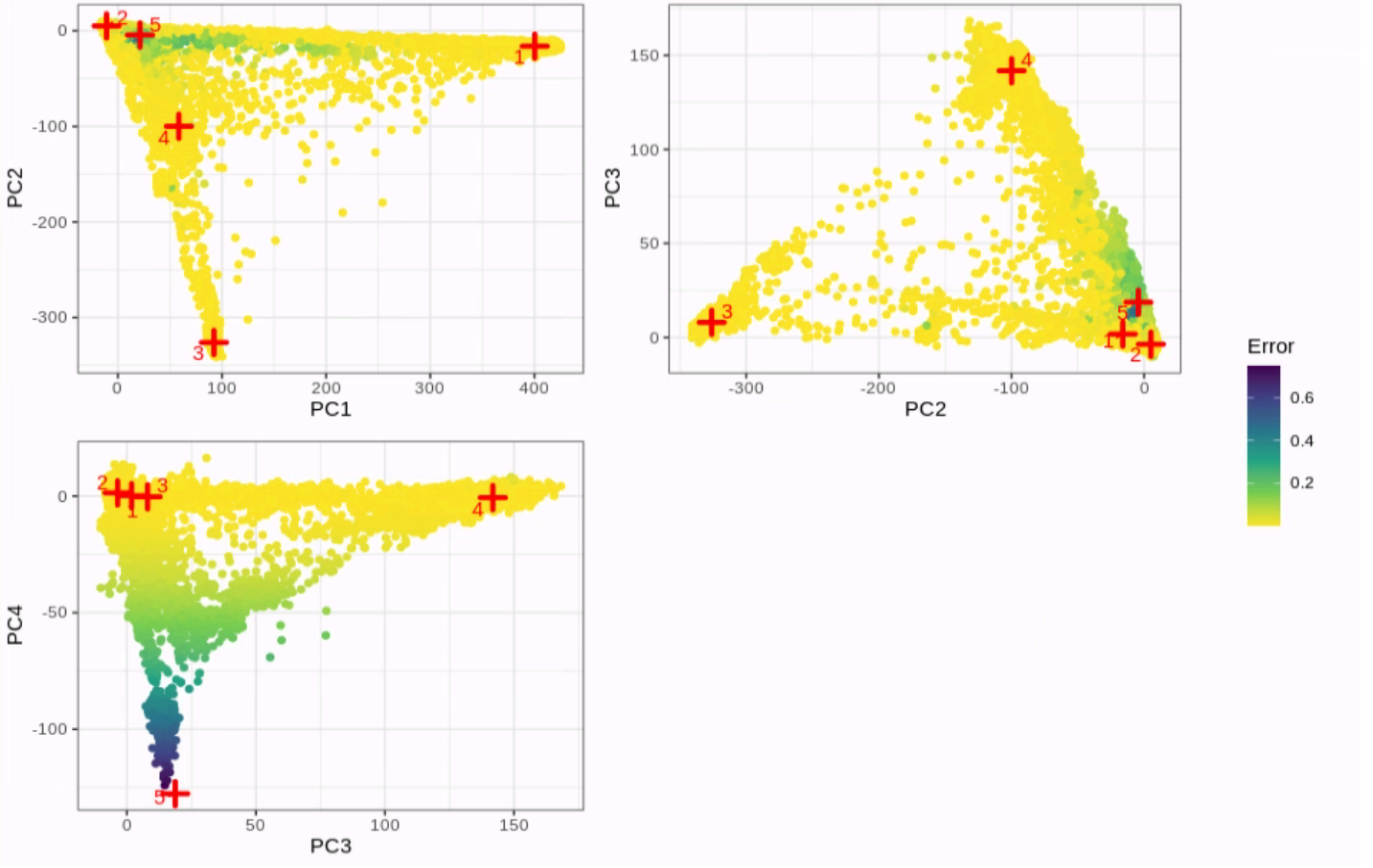
Start with PC1 and 2 refs, then add 3rd ref when considering 2 PCs



Add 4th reference when considering 3 PCs

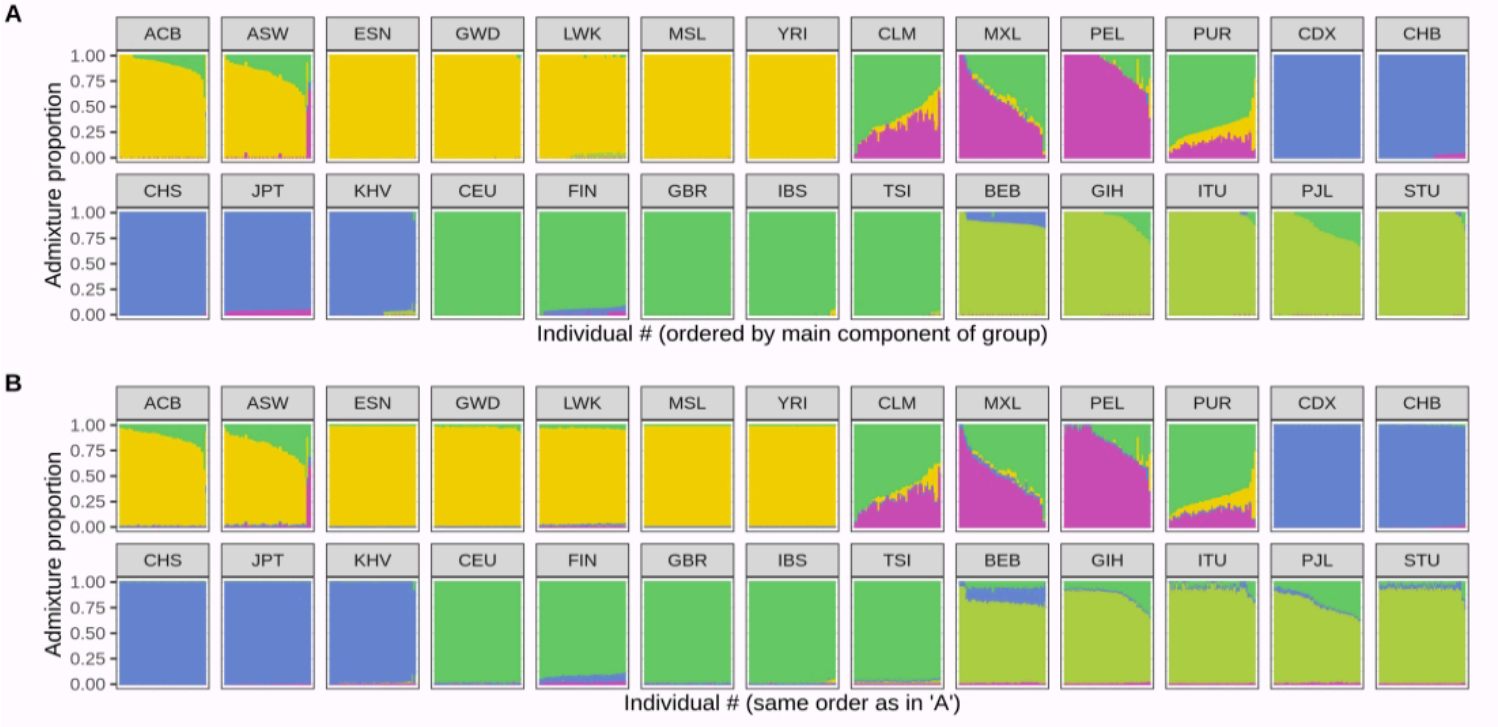


Add 5th reference when considering 4 PCs



Comparing to ADMIXTURE in the 1000 Genomes data

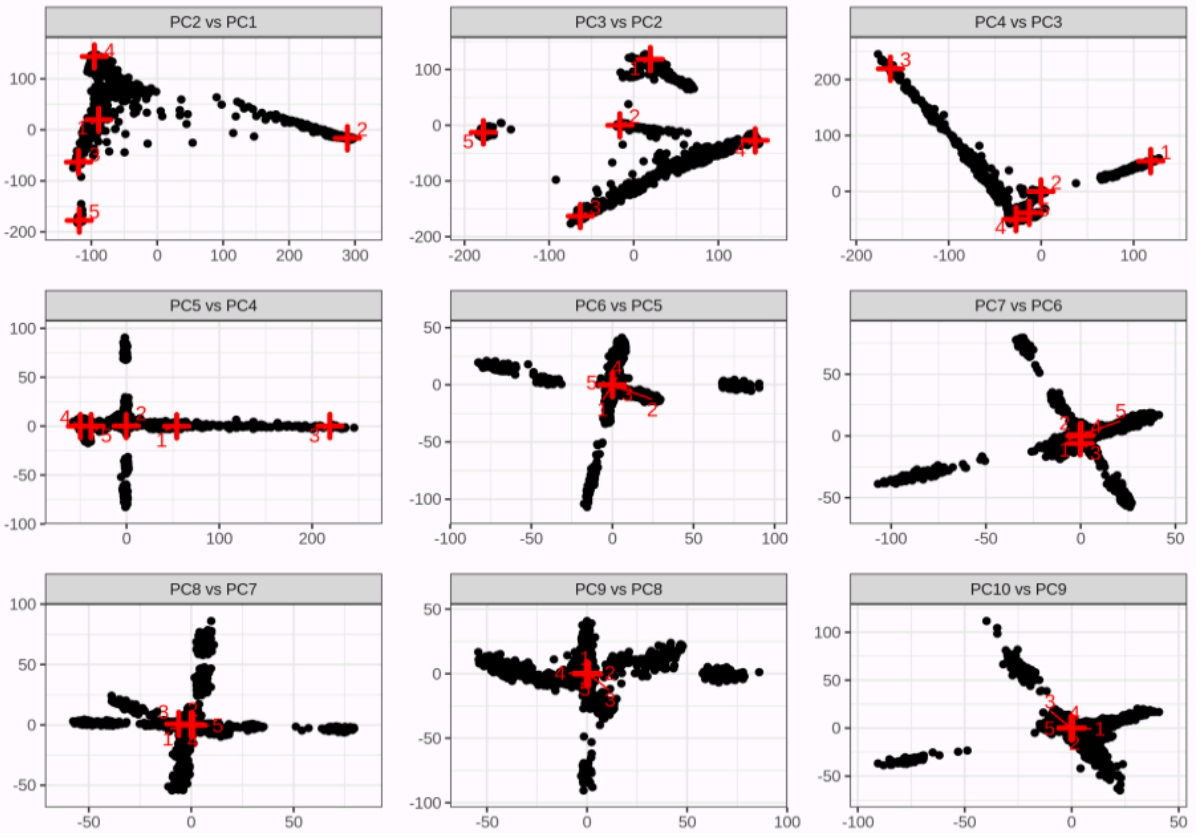
Admixture coefficients Q using L=5 reference populations



A: with my method; B: with ADMIXTURE

ACB: African Caribbean in Barbados; **ASW**: African Ancestry in Southwest US; **ESN**: Esan in Nigeria; **GWD**: Gambian in Western Division, The Gambia; **LWK**: Luhya in Webuye, Kenya; **MSL**: Mende in Sierra Leone; **YRI**: Yoruba in Ibadan, Nigeria; **CLM**: Colombian in Medellin, Colombia; **MXL**: Mexican Ancestry in Los Angeles, California; **PEL**: Peruvian in Lima, Peru; **PUR**: Puerto Rican in Puerto Rico; **CDX**: Chinese Dai in Xishuangbanna, China; **CHB**: Han Chinese in Bejing, China; **CHS**: Southern Han Chinese, China; **JPT**: Japanese in Tokyo, Japan; **KHV**: Kinh in Ho Chi Minh City, Vietnam; **CEU**: Utah residents with Northern and Western European ancestry; **FIN**: Finnish in Finland; **GBR**: British in England and Scotland; **IBS**: Iberian populations in Spain; **TSI**: Toscani in Italy; **BEB**: Bengali in Bangladesh; **GIH**: Gujarati Indian in Houston,TX; **ITU**: Indian Telugu in the UK; **PJL**: Punjabi in Lahore,Pakistan; **STU**: Sri Lankan Tamil in the UK

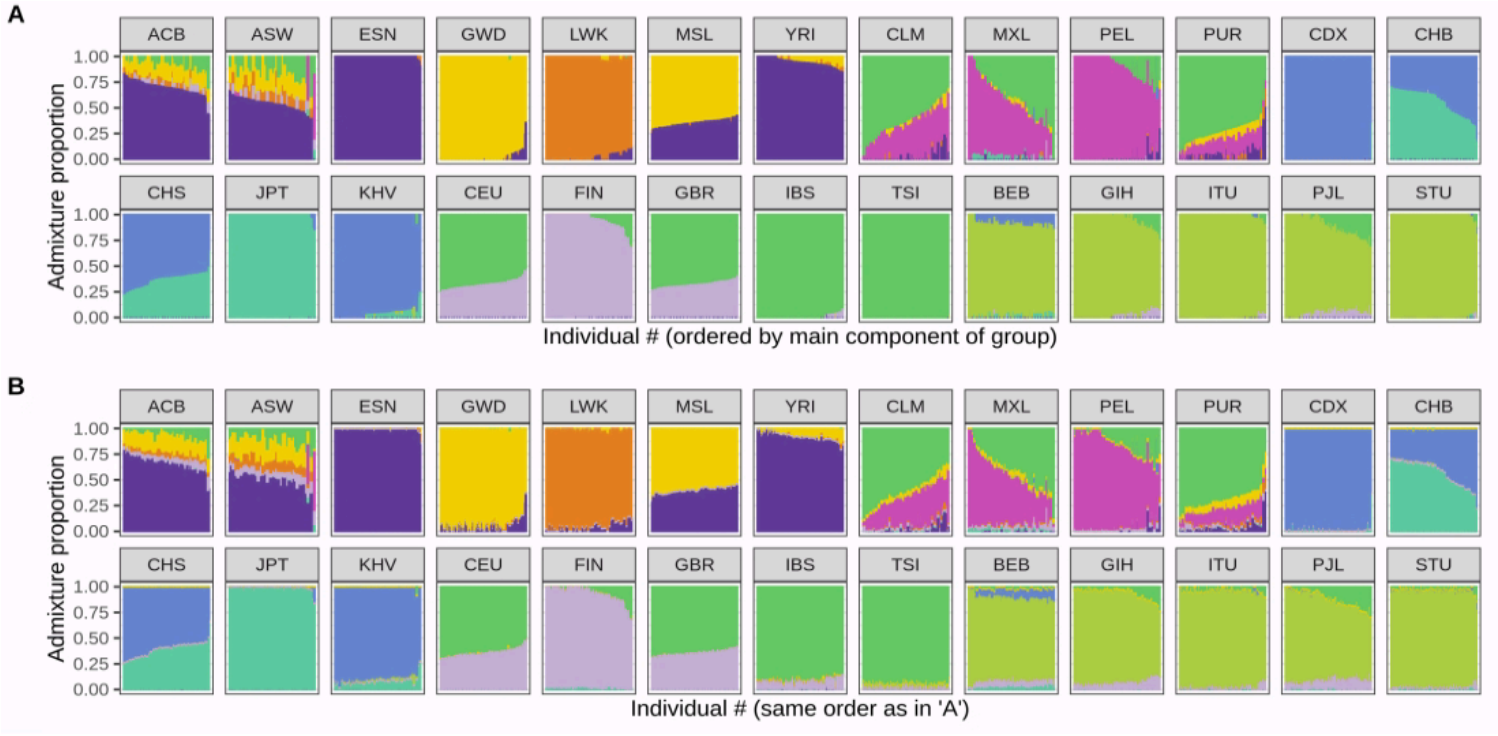
Projected reference allele frequencies ($2F \cdot V$) from ADMIXTURE



$\Rightarrow$ **highly similar to what I get with my method**

$\Rightarrow$ **supports using K PCs only for $L = K + 1$ reference populations**

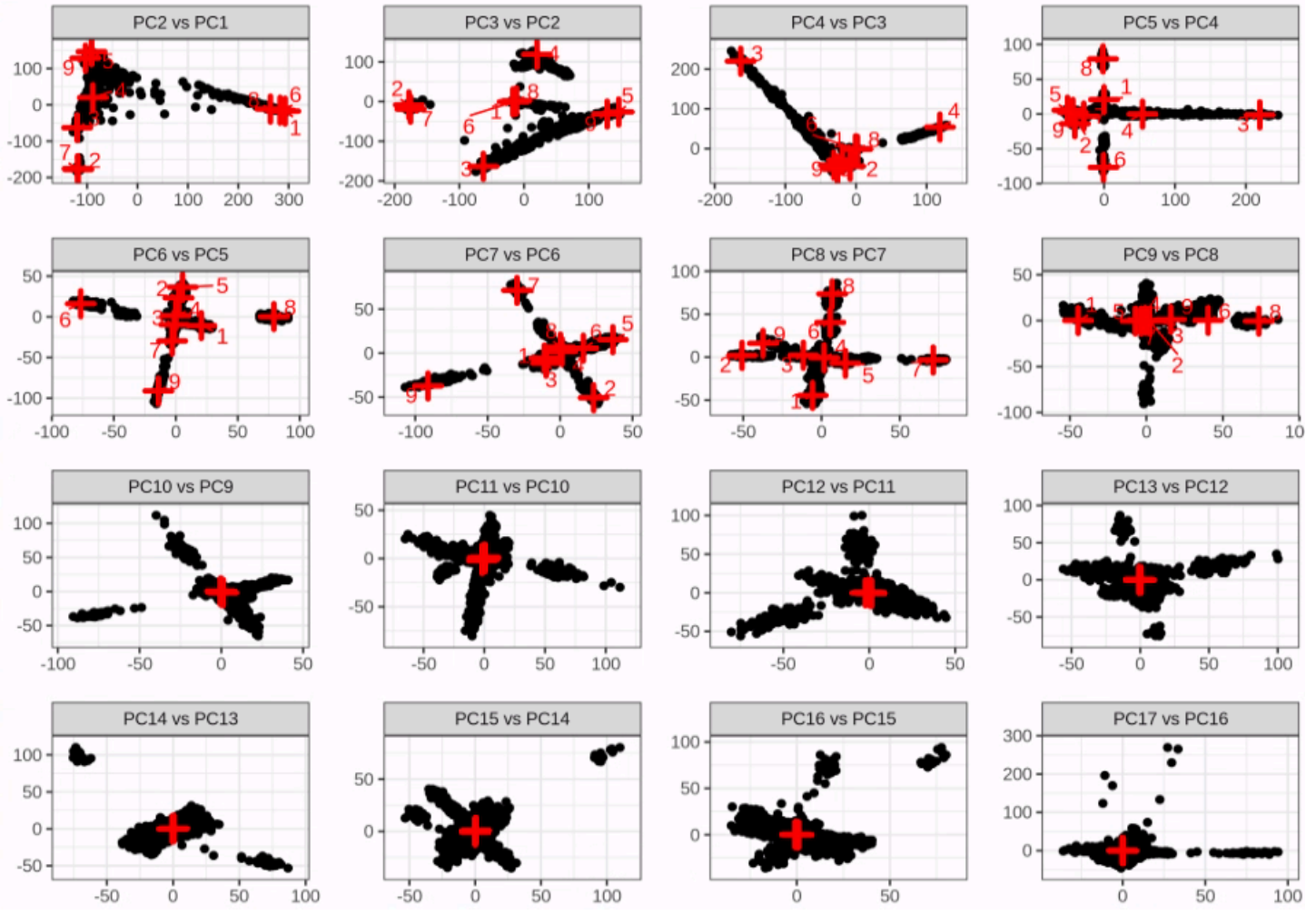
Admixture coefficients Q using L=9 reference populations



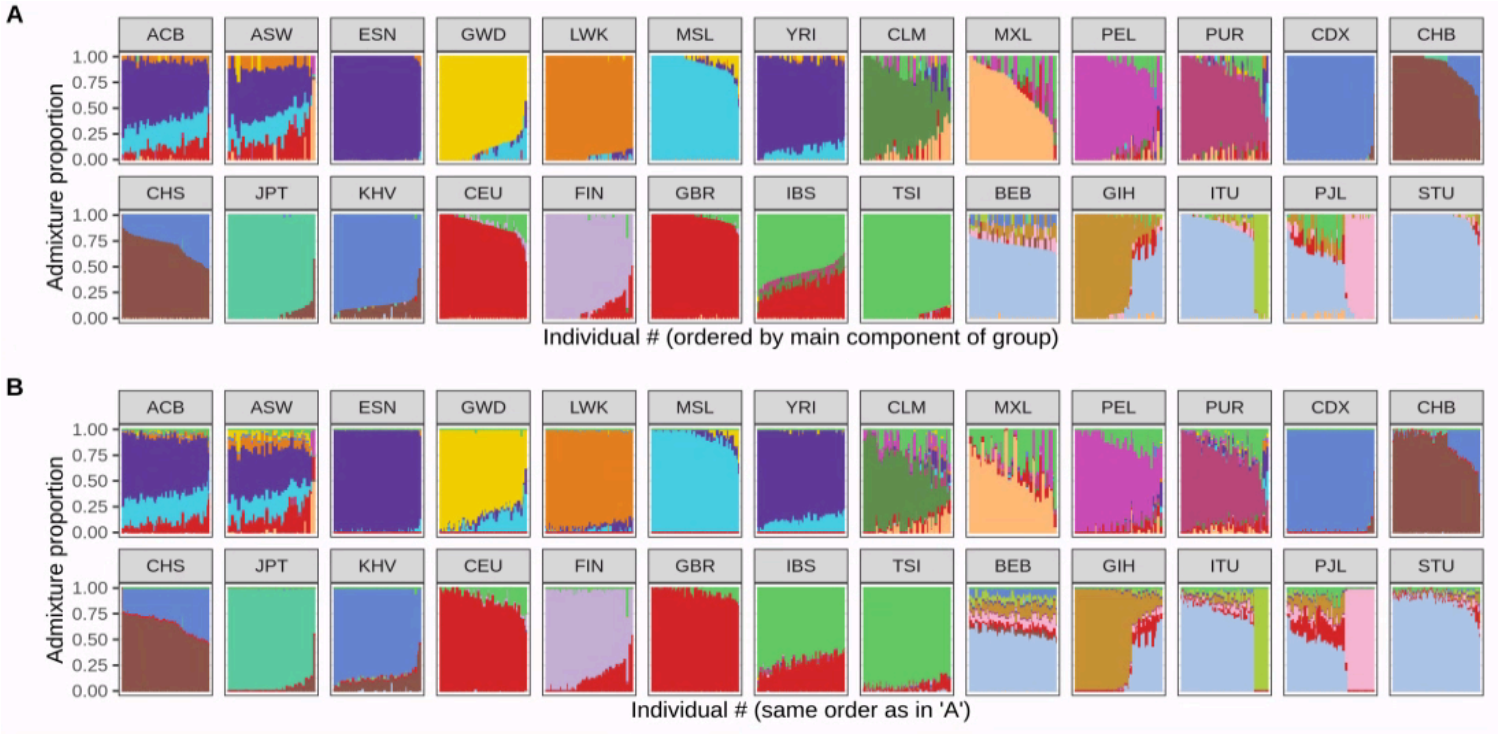
A: with my method; B: with ADMIXTURE

ACB: African Caribbean in Barbados; **ASW**: African Ancestry in Southwest US; **ESN**: Esan in Nigeria; **GWD**: Gambian in Western Division, The Gambia; **LWK**: Luhya in Webuye, Kenya; **MSL**: Mende in Sierra Leone; **YRI**: Yoruba in Ibadan, Nigeria; **CLM**: Colombian in Medellin, Colombia; **MXL**: Mexican Ancestry in Los Angeles, California; **PEL**: Peruvian in Lima, Peru; **PUR**: Puerto Rican in Puerto Rico; **CDX**: Chinese Dai in Xishuangbanna, China; **CHB**: Han Chinese in Bejing, China; **CHS**: Southern Han Chinese, China; **JPT**: Japanese in Tokyo, Japan; **KHV**: Kinh in Ho Chi Minh City, Vietnam; **CEU**: Utah residents with Northern and Western European ancestry; **FIN**: Finnish in Finland; **GBR**: British in England and Scotland; **IBS**: Iberian populations in Spain; **TSI**: Toscani in Italy; **BEB**: Bengali in Bangladesh; **GIH**: Gujarati Indian in Houston,TX; **ITU**: Indian Telugu in the UK; **PJL**: Punjabi in Lahore,Pakistan; **STU**: Sri Lankan Tamil in the UK

Projected reference allele frequencies ($2F \cdot V$) from ADMIXTURE



Admixture coefficients Q using L=18 reference populations



A: with my method; B: with ADMIXTURE

ACB: African Caribbean in Barbados; **ASW**: African Ancestry in Southwest US; **ESN**: Esan in Nigeria; **GWD**: Gambian in Western Division, The Gambia; **LWK**: Luhya in Webuye, Kenya; **MSL**: Mende in Sierra Leone; **YRI**: Yoruba in Ibadan, Nigeria; **CLM**: Colombian in Medellin, Colombia; **MXL**: Mexican Ancestry in Los Angeles, California; **PEL**: Peruvian in Lima, Peru; **PUR**: Puerto Rican in Puerto Rico; **CDX**: Chinese Dai in Xishuangbanna, China; **CHB**: Han Chinese in Bejing, China; **CHS**: Southern Han Chinese, China; **JPT**: Japanese in Tokyo, Japan; **KHV**: Kinh in Ho Chi Minh City, Vietnam; **CEU**: Utah residents with Northern and Western European ancestry; **FIN**: Finnish in Finland; **GBR**: British in England and Scotland; **IBS**: Iberian populations in Spain; **TSI**: Toscani in Italy; **BEB**: Bengali in Bangladesh; **GIH**: Gujarati Indian in Houston,TX; **ITU**: Indian Telugu in the UK; **PJL**: Punjabi in Lahore,Pakistan; **STU**: Sri Lankan Tamil in the UK

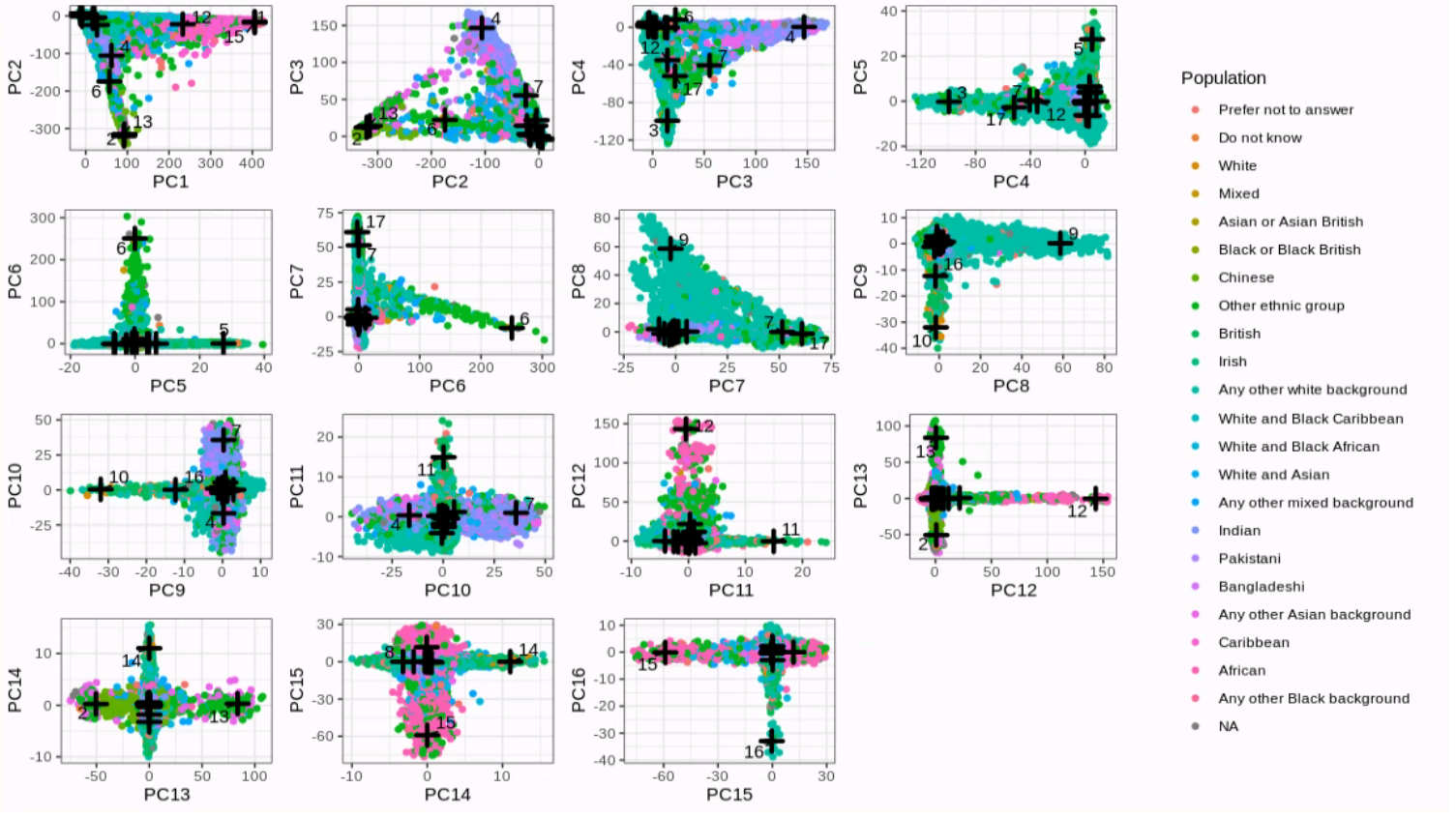
Runtimes

- for ADMIXTURE (with 15 cores), it takes
  - 1 hour for L=5
  - 4 hours for L=9
  - 13 hours for L=18

- for my method (with 1 core), it takes
  - 2 minutes to get all solutions for L=3 to L=18

In the UK Biobank data

After convergence with 17 references and 16 PCs from the UK Biobank



Country (of birth) counts with ancestry > 0.6 for each reference

- United Kingdom: 126045 – NA: 1352 – Germany: 915 – South Africa: 477 – Netherlands: 443 – USA: 400 – France: 300 – Australia: 226 – Denmark: 197 – Canada: 195 – …
- United Kingdom: 22206 – NA: 86 – Germany: 16 – Ireland: 14
- United Kingdom: 32123 – Ireland: 289 – NA: 265 – New Zealand: 75 – Canada: 58 – India: 54 – Germany: 47 – South Africa: 43 – Australia: 40 – Kenya: 36 – Malaysia: 29 – …
- United Kingdom: 10647 – Ireland: 9360 – NA: 290 – USA: 47 – Australia: 43 – …
- United Kingdom: 1347 – NA: 30
- United Kingdom: 4080 – NA: 77
- NA: 752 – Poland: 599 – United Kingdom: 415 – Russia: 131 – Finland: 105 – Germany: 87 – Lithuania: 71 – Ukraine: 55 – Czech Republic: 53 – Latvia: 52 – Slovakia: 28 – …
- India: 1852 – Kenya: 782 – Sri Lanka: 653 – NA: 547 – Pakistan: 410 – Mauritius: 273 – Bangladesh: 235 – Uganda: 231 – Tanzania: 175 – Caribbean: 114 – The Guianas: 83 – …
- Caribbean: 2110 – NA: 2100 – Nigeria: 1017 – Ghana: 866 – Barbados: 255 – Sierra Leone: 202 – The Guianas: 151 – Gambia: 39 – Ivory Coast: 32 – …
- Italy: 389 – NA: 353 – Cyprus: 170 – United Kingdom: 168 – Egypt: 147 – Malta: 116 – Greece: 99 – Algeria: 68 – Lebanon: 50 – Morocco: 46 – Libya: 40 – Palestine: 30 – …
- United Kingdom: 1844 – NA: 830 – USA: 169 – South Africa: 95 – Israel: 41 – …
- Iran: 476 – Iraq: 140 – NA: 59 – Turkey: 54 – India: 36 – Afghanistan: 13 – Pakistan: 10
- China: 287 – Japan: 241 – Malaysia: 185 – Hong Kong: 161 – Nepal: 123 – NA: 63 – Singapore: 56 – South Korea: 26 – Mauritius: 25 – Taiwan: 25 – Indonesia: 15 – …
- Zimbabwe: 268 – Congo: 133 – Uganda: 115 – Kenya: 73 – South Africa: 59 – Zambia: 56 – NA: 41 – Tanzania: 26 – Angola: 23 – Burundi: 17 – Rwanda: 16 – Seychelles: 14 – …
- Philippines: 315 – Malaysia: 20 – NA: 17 – Indonesia: 15 – Thailand: 13
- Peru: 33 – Ecuador: 25 – Mexico: 20 – Colombia: 17 – Bolivia: 14 – Chile: 11
- Somalia: 81 – Ethiopia: 58 – Sudan: 51 – Eritrea: 45 – NA: 20

Capturing more population structure (with less individuals)

Efficient toolkit implementing best practices for principal component analysis of population genetic data

Florian Privé ✉, Keurcien Luu, Michael G B Blum, John J McGrath, Bjarni J Vilhjálmsson ✉

Bioinformatics, Volume 36, Issue 16, August 2020, Pages 4449–4457, https://doi.org/10.1093/bioinformatics/btaa520

Efficient toolkit implementing best practices for principal component analysis of population genetic data

Florian Privé ✉, Keurcien Luu, Michael G B Blum, John J McGrath, Bjarni J Vilhjálmsson ✉

Bioinformatics, Volume 36, Issue 16, August 2020, Pages 4449–4457, https://doi.org/10.1093/bioinformatics/btaa520

In the UK Biobank data,

- only the first 16 PCs actually capture population structure
  (PC 19–40 capture LD only; never use them!)

Efficient toolkit implementing best practices for principal component analysis of population genetic data

Florian Privé ✉, Keurcien Luu, Michael G B Blum, John J McGrath, Bjarni J Vilhjálmsson ✉

Bioinformatics, Volume 36, Issue 16, August 2020, Pages 4449–4457, https://doi.org/10.1093/bioinformatics/btaa520

In the UK Biobank data,

- only the first 16 PCs actually capture population structure (PC 19–40 capture LD only; never use them!)

When subsampling British and Irish individuals (self-reported ancestry)

- can obtain 40 PCs that capture some population structure
- using the best practices for PCA of genetic data

Efficient toolkit implementing best practices for principal component analysis of population genetic data

Florian Privé ✉, Keurcien Luu, Michael G B Blum, John J McGrath, Bjarni J Vilhjálmsson ✉

Bioinformatics, Volume 36, Issue 16, August 2020, Pages 4449–4457, https://doi.org/10.1093/bioinformatics/btaa520

In the UK Biobank data,

- only the first 16 PCs actually capture population structure (PC 19–40 capture LD only; never use them!)

When subsampling British and Irish individuals (self-reported ancestry)

- can obtain 40 PCs that capture some population structure
- using the best practices for PCA of genetic data

In my current work:

- I've also looked at using Q to do the subsampling
- I've run my deconvolution algorithm on K=41 PCs to get L=42 references

Rerun the algorithm with new PCs (K=41, L=42)

- United Kingdom: 414629 – Ireland: 12416 – NA: 4731 – Germany: 1498 – South Africa: 970 – USA: 956 – Australia: 853 – New Zealand: 656 – Canada: 644 – …
- NA: 709 – Poland: 592 – United Kingdom: 389 – Russia: 123 – Germany: 71 – Lithuania: 71 – Ukraine: 53 – Latvia: 52– ..
- Italy: 34
- Spain: 30
- United Kingdom: 1838 – NA: 831 – USA: 170 – South Africa: 95 – Israel: 40 – Canada: 18 – Hungary: 18 – France: 12
- Finland: 125
- Nigeria: 975 – NA: 292 – Caribbean: 155 – Sierra Leone: 42 – Ghana: 13
- Sri Lanka: 635 – India: 493 – Mauritius: 190 – NA: 156 – Kenya: 90 – Caribbean: 71 – Malaysia: 67 – The Guianas: 52 – …
- Malta: 114 – United Kingdom: 15 – NA: 12 – Egypt: 10
- Iran: 494 – Iraq: 247 – Turkey: 114 – NA: 58 – Syria: 11 – United Kingdom: 10
- Ghana: 817 – NA: 68 – Ivory Coast: 27
- India: 571 – NA: 207 – Kenya: 40 – Pakistan: 28 – Malaysia: 23 – Singapore: 13
- India: 28
- Yemen: 26 – Egypt: 18 – NA: 12
- Congo: 129 – Angola: 30 – Zambia: 30 – NA: 25 – Cameroon: 24
- India: 224 – Kenya: 179 – NA: 55 – Uganda: 28 – Pakistan: 21 – Tanzania: 21
- Japan: 241 – South Korea: 26
- Thailand: 61 – Vietnam: 40 – Malaysia: 10
- Algeria: 69 – Morocco: 66 – Libya: 27 – NA: 10
- Kenya: 18 – India: 13
- Philippines: 310 – NA: 16
- Pakistan: 76 – NA: 20
- Kenya: 36 – India: 25
- India: 70 – Afghanistan: 25 – NA: 19
- India: 17 – NA: 11 – Malawi: 10
- Colombia: 115
- Sierra Leone: 38 – Gambia: 33
- India: 90 – NA: 32
- Tanzania: 24
- Pakistan: 146 – NA: 42 – India: 22 – Kenya: 18
- India: 135 – Kenya: 120 – Uganda: 80 – NA: 37 – Tanzania: 24
- Nepal: 125 – NA: 14
- Peru: 31 – Ecuador: 20 – Bolivia: 14 – Mexico: 13
- Uganda: 69 – Tanzania: 43 – Kenya: 40 – India: 24
- Kenya: 42 – India: 39 – NA: 16 – Tanzania: 14
- Uganda: 101 – Kenya: 28 – Tanzania: 11
- Kenya: 114
- India: 43 – Kenya: 38 – NA: 19
- South Africa: 48 – Zimbabwe: 25
- Sudan: 17
-
- Somalia: 78

Conclusion

- A very efficient admixture deconvolution algorithm

Conclusion

- A very efficient admixture deconvolution algorithm
- Also very powerful; it can identify many reference groups (subsampling before PCA is beneficial to capture more structure)

Conclusion

- A very efficient admixture deconvolution algorithm
- Also very powerful; it can identify many reference groups (subsampling before PCA is beneficial to capture more structure)
- One can (should) check the results visually (also, L=K+1)

Conclusion

- A very efficient admixture deconvolution algorithm
- Also very powerful; it can identify many reference groups
  (subsampling before PCA is beneficial to capture more structure)
- One can (should) check the results visually (also, L=K+1)
- The algorithm is not specific to genetic data
  (merely a deconvolution algorithm based on PCA)
  $\Rightarrow$ may be used for cell type deconvolution of methylation data

Conclusion

- A very efficient admixture deconvolution algorithm
- Also very powerful; it can identify many reference groups
  (subsampling before PCA is beneficial to capture more structure)
- One can (should) check the results visually (also, L=K+1)
- The algorithm is not specific to genetic data
  (merely a deconvolution algorithm based on PCA)
  $\Rightarrow$ may be used for cell type deconvolution of methylation data
- I will provide a new set of reference populations for people to use directly

Conclusion

- A very efficient admixture deconvolution algorithm
- Also very powerful; it can identify many reference groups
  (subsampling before PCA is beneficial to capture more structure)
- One can (should) check the results visually (also, L=K+1)
- The algorithm is not specific to genetic data
  (merely a deconvolution algorithm based on PCA)
  $\Rightarrow$ may be used for cell type deconvolution of methylation data
- I will provide a new set of reference populations for people to use directly

Thank you for your attention

Presentation available at bit.ly/privefl_ISHG2025