BIOINFORMATICS INSTITUTE

GPS-L: Estimation of local ancestry

Ivan Dmitrievsky
Tatiana Tatarinova (Prof. at USC)

GenoChip

Specifically designed for anthropological testing and includes SNPs from

- autosomal DNA
- X-chromosome DNA
- Y-chromosome DNA
- mitochondrial DNA

ADMIXTURE

Software tool for estimation of individual ancestries

Working modes (same as for me)

— unsupervised

— supervised

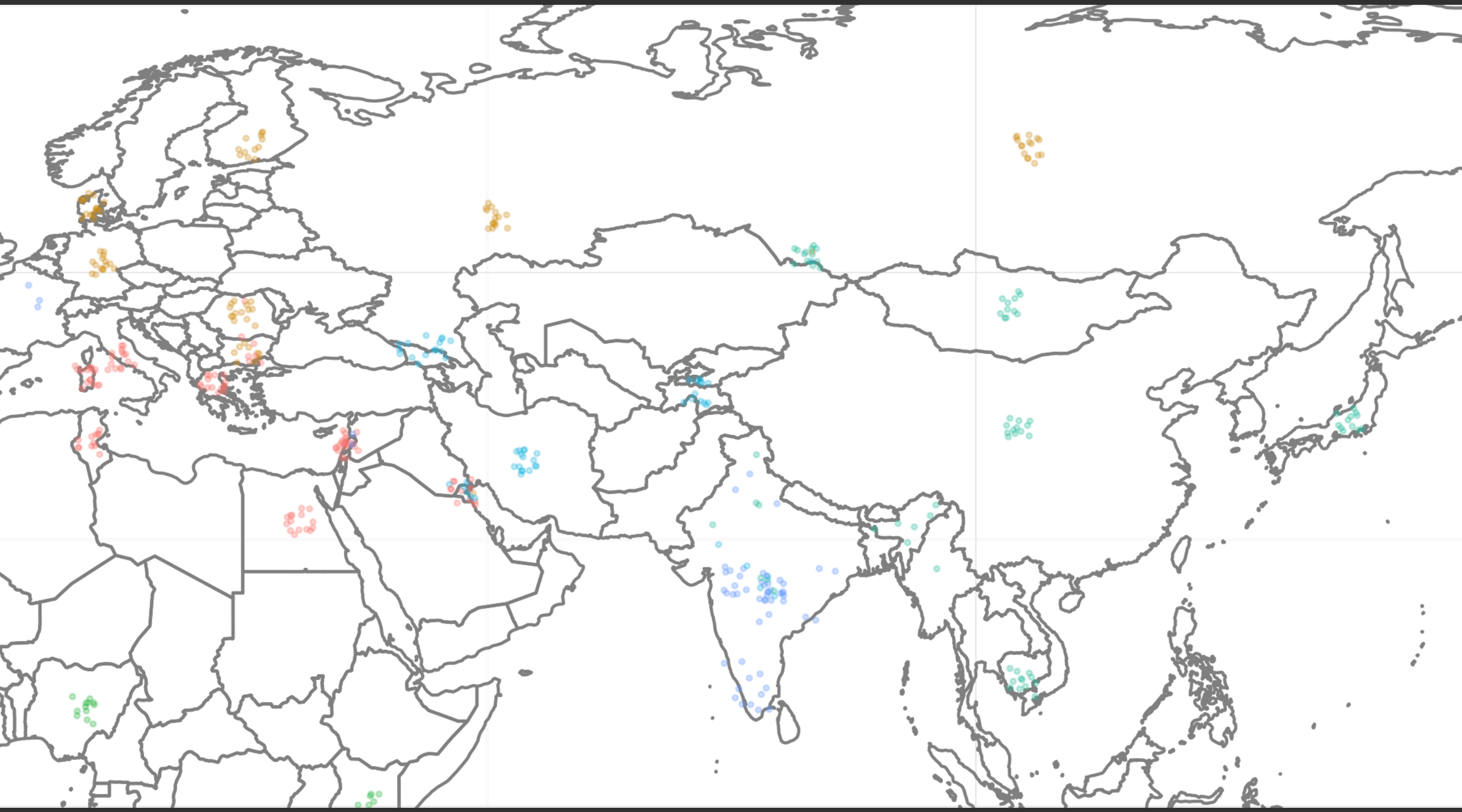
Unsupervised mode

— Determine number of ancestry populations (8-9)

— Split genome into windows

— Determine the same number inside windows (10)

But CV isn't sufficient for determining K. Common sense should also be applied.

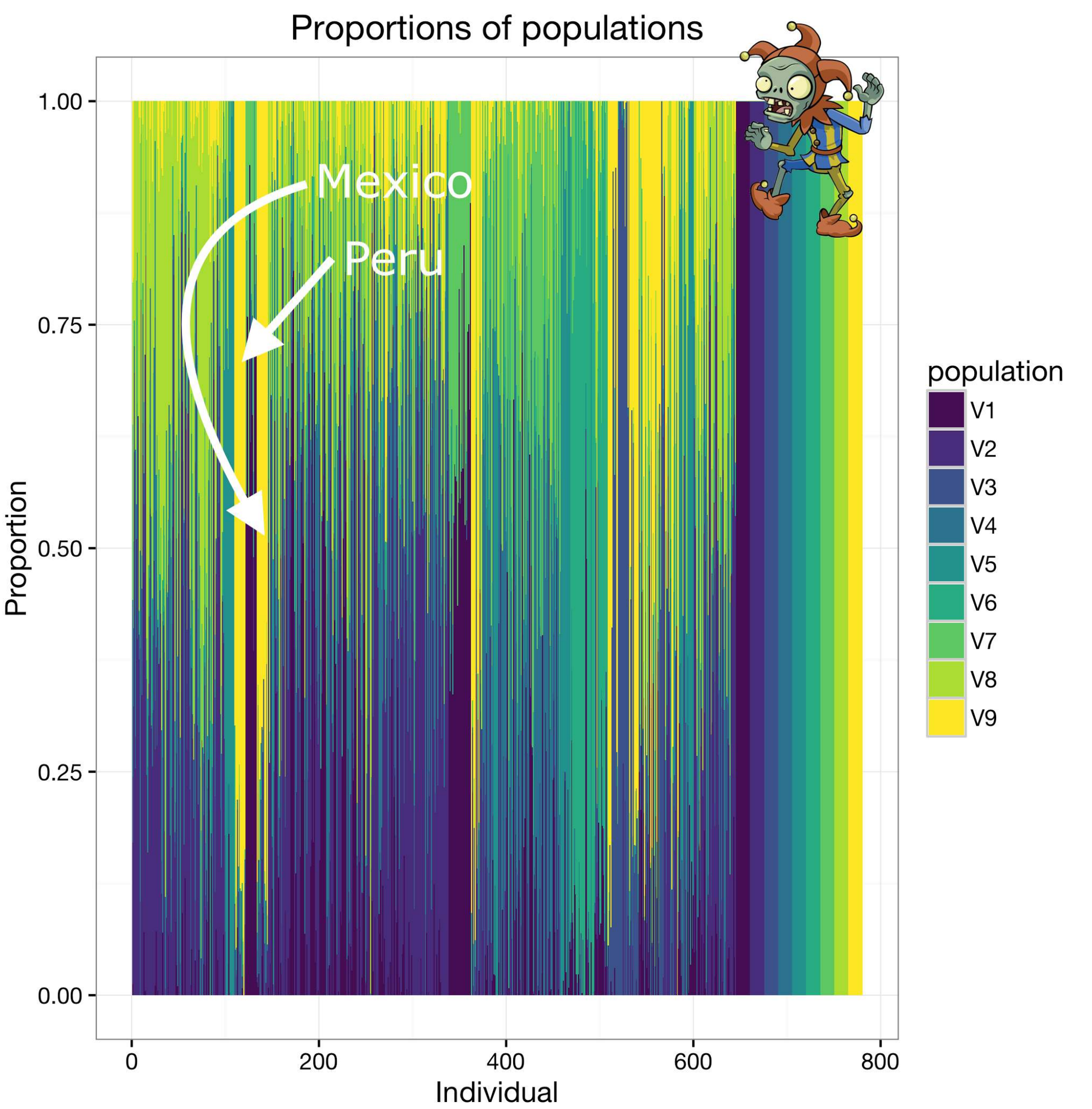


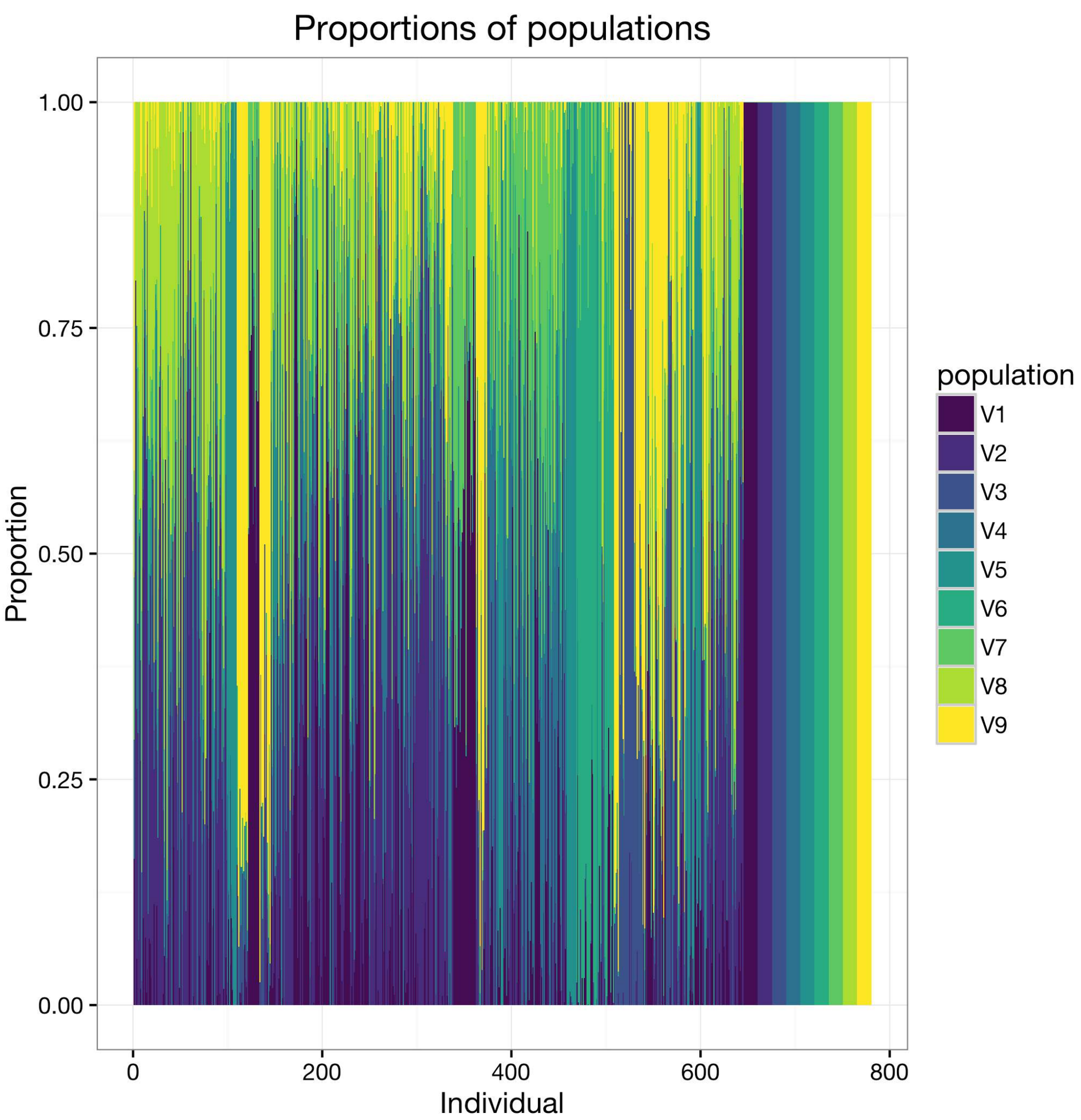
Supervised mode — Zombie

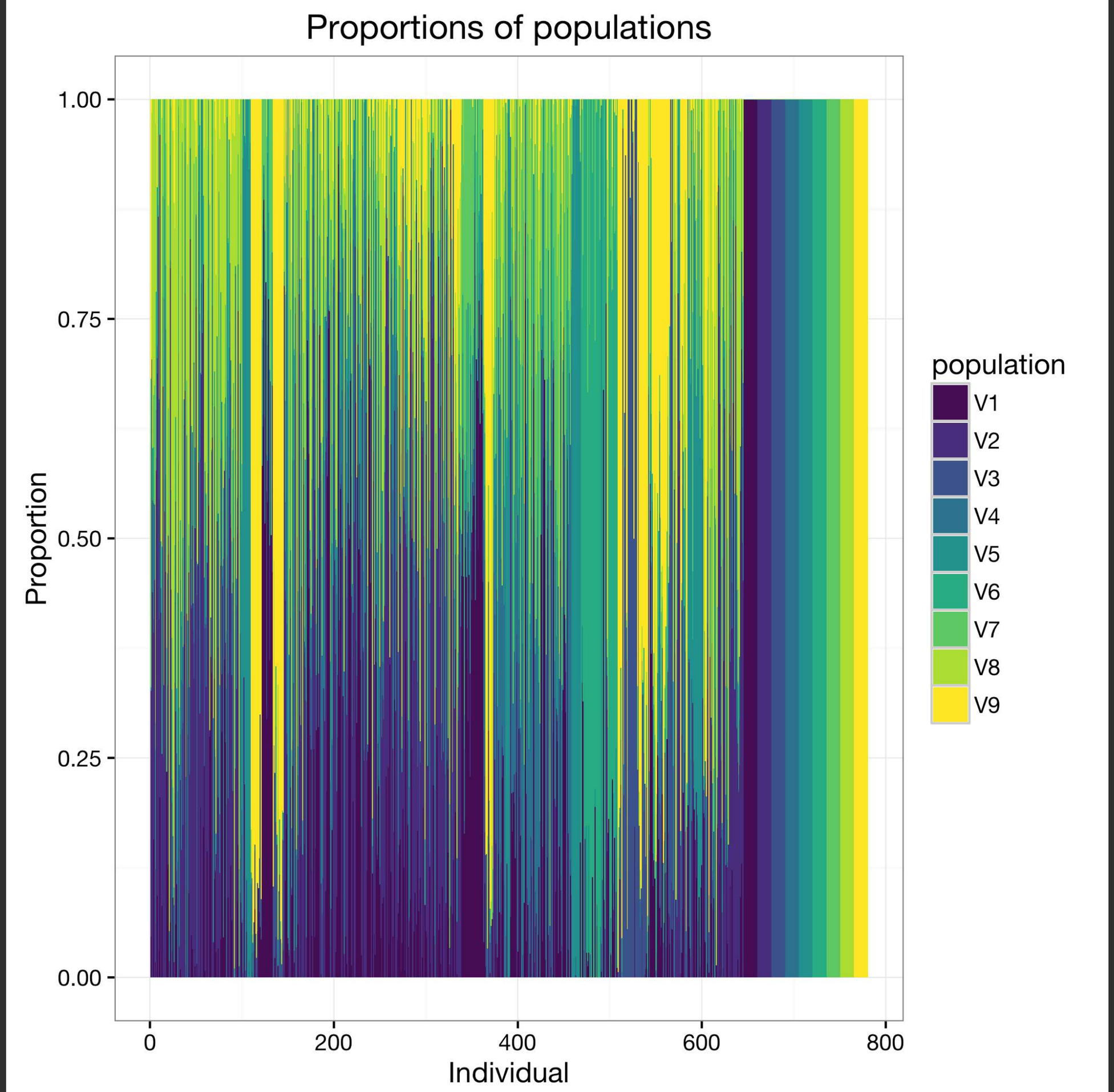
— Modeled individuals

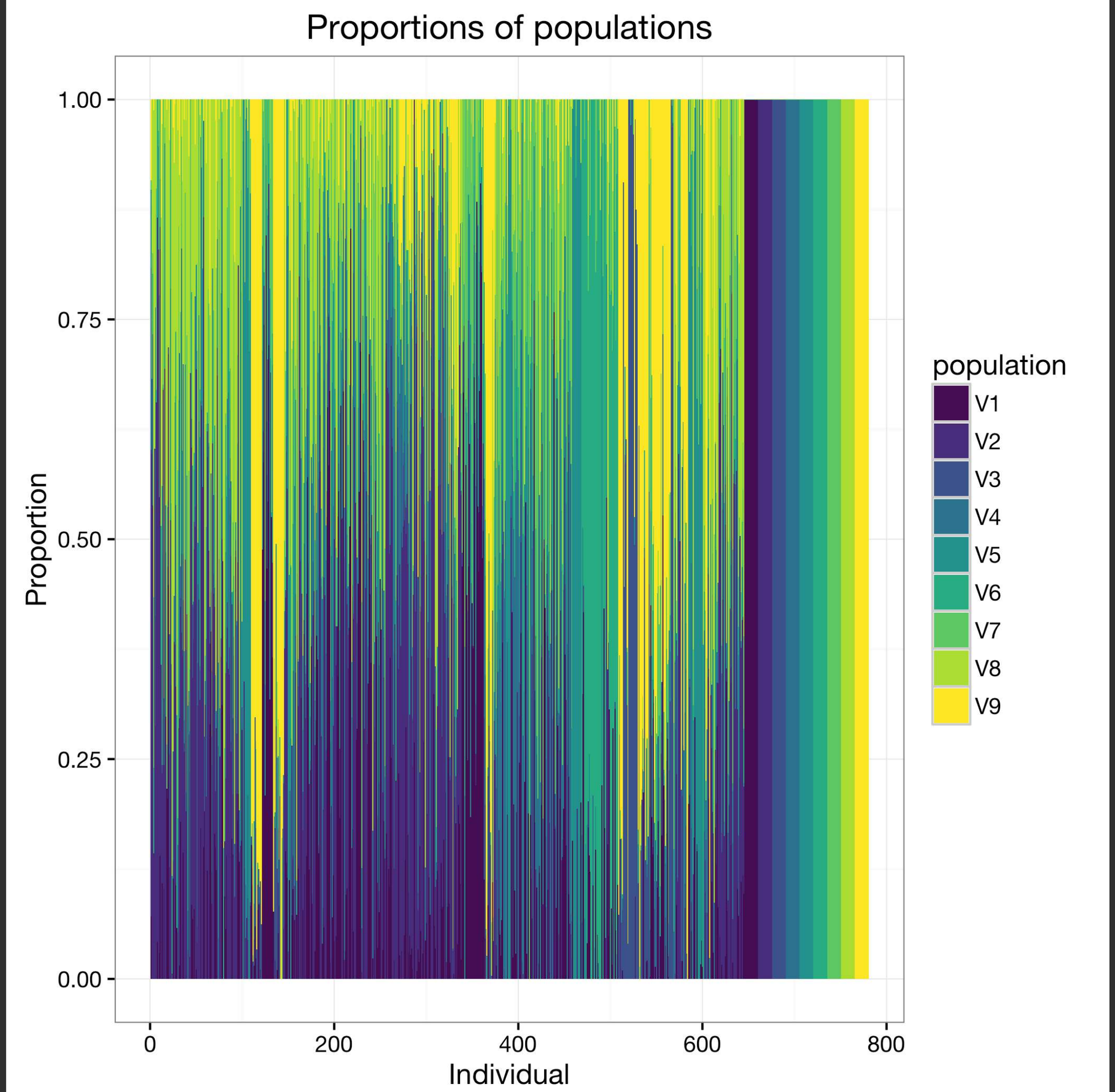
— Belong to ancestral population

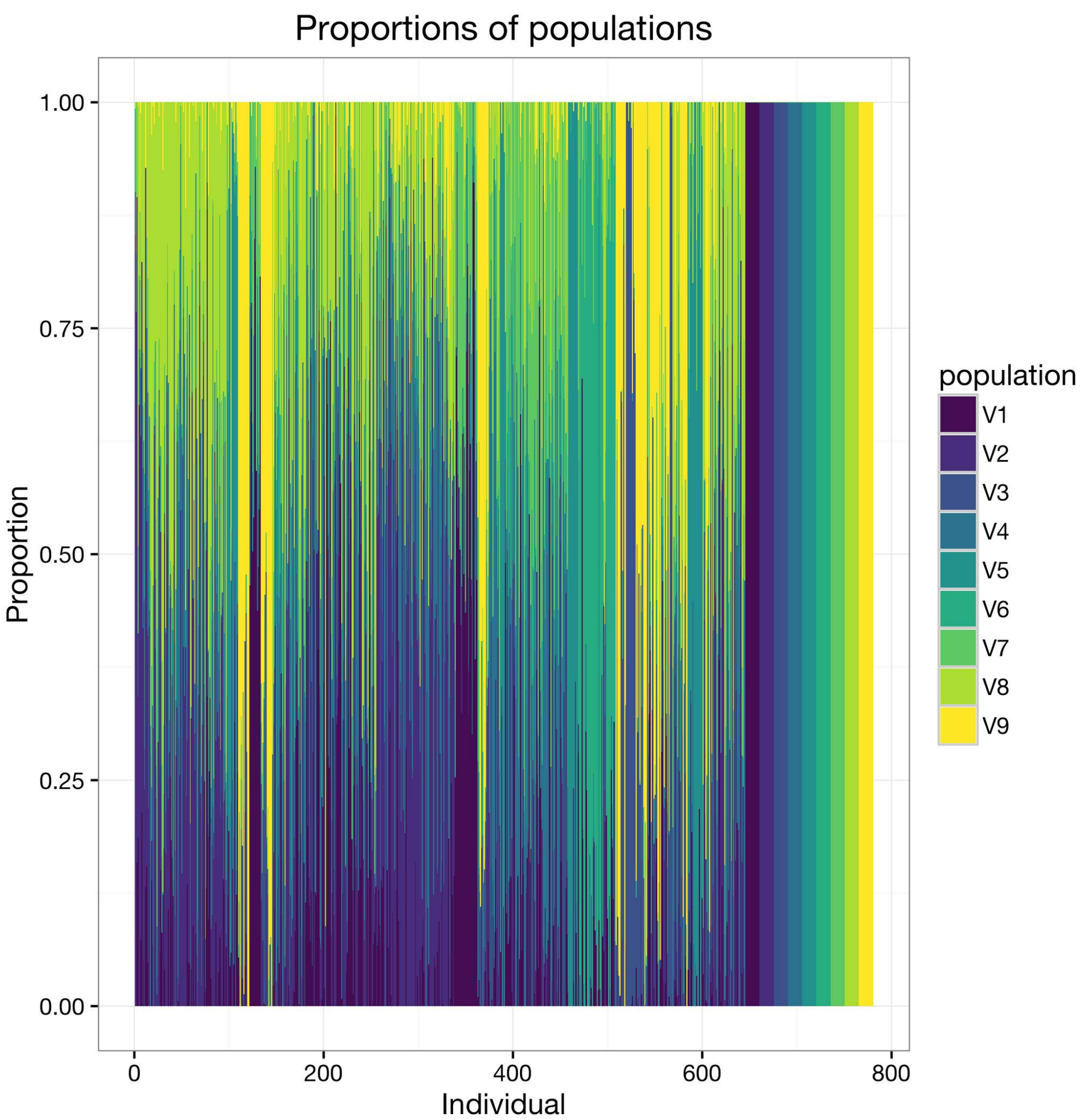
— Act as basis vectors


Proportions of populations
Mexico
Peru
1.00
0.75
0.50
0.25
0.00
Proportion
0
200
400
600
800
Individual
population
V1
V2
V3
V4
V5
V6
V7
V8
V9


Proportions of populations
1.00
0.75
0.50
0.25
0.00
Proportion
0
200
400
600
800
Individual
population
V1
V2
V3
V4
V5
V6
V7
V8
V9


Proportions of populations
Proportion
1.00
0.75
0.50
0.25
0.00
0
200
400
600
800
Individual
population
V1
V2
V3
V4
V5
V6
V7
V8
V9

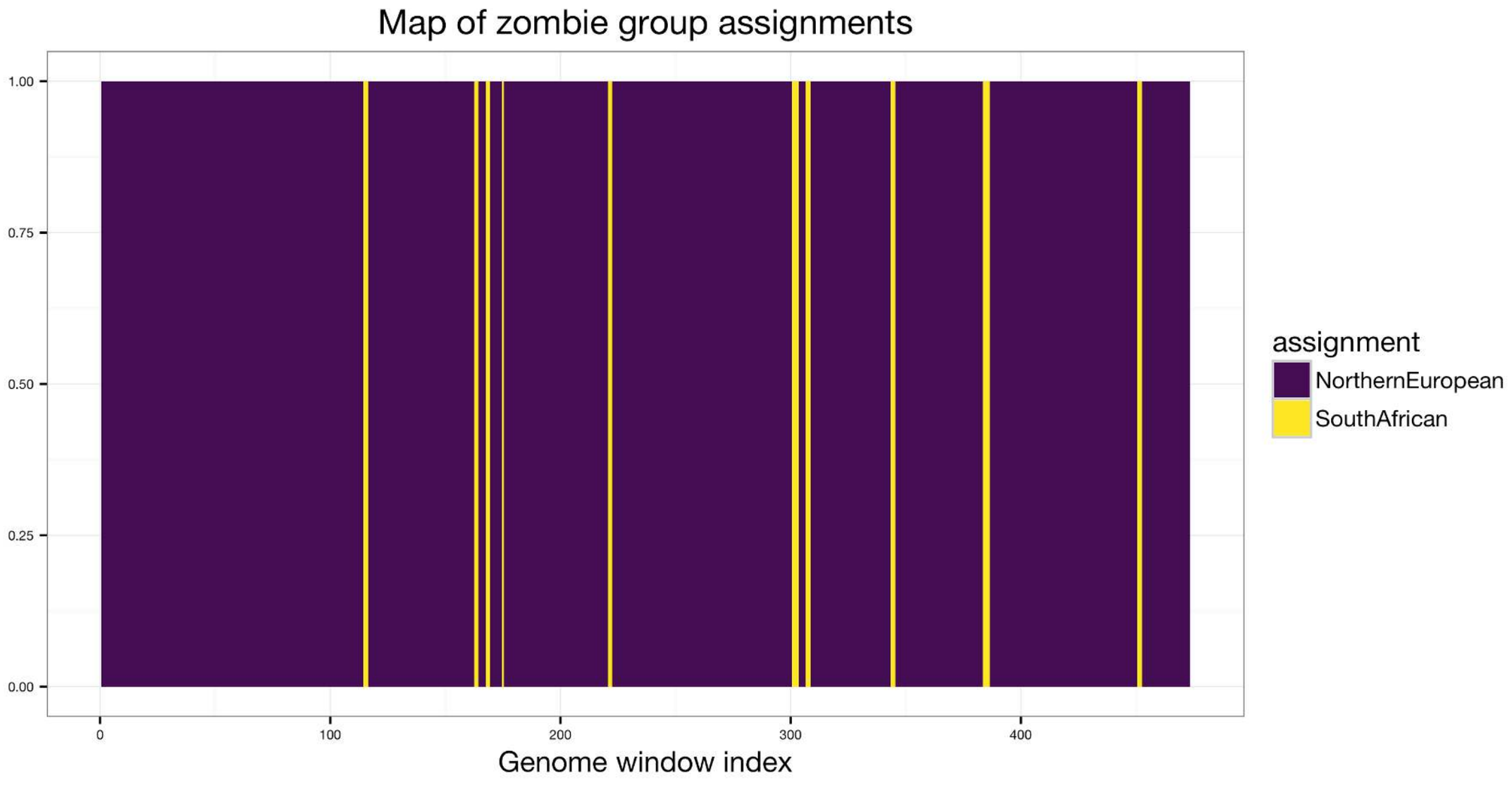

Proportions of populations
1.00
0.75
0.50
0.25
0.00
Proportion
0
200
400
600
800
Individual
population
V1
V2
V3
V4
V5
V6
V7
V8
V9

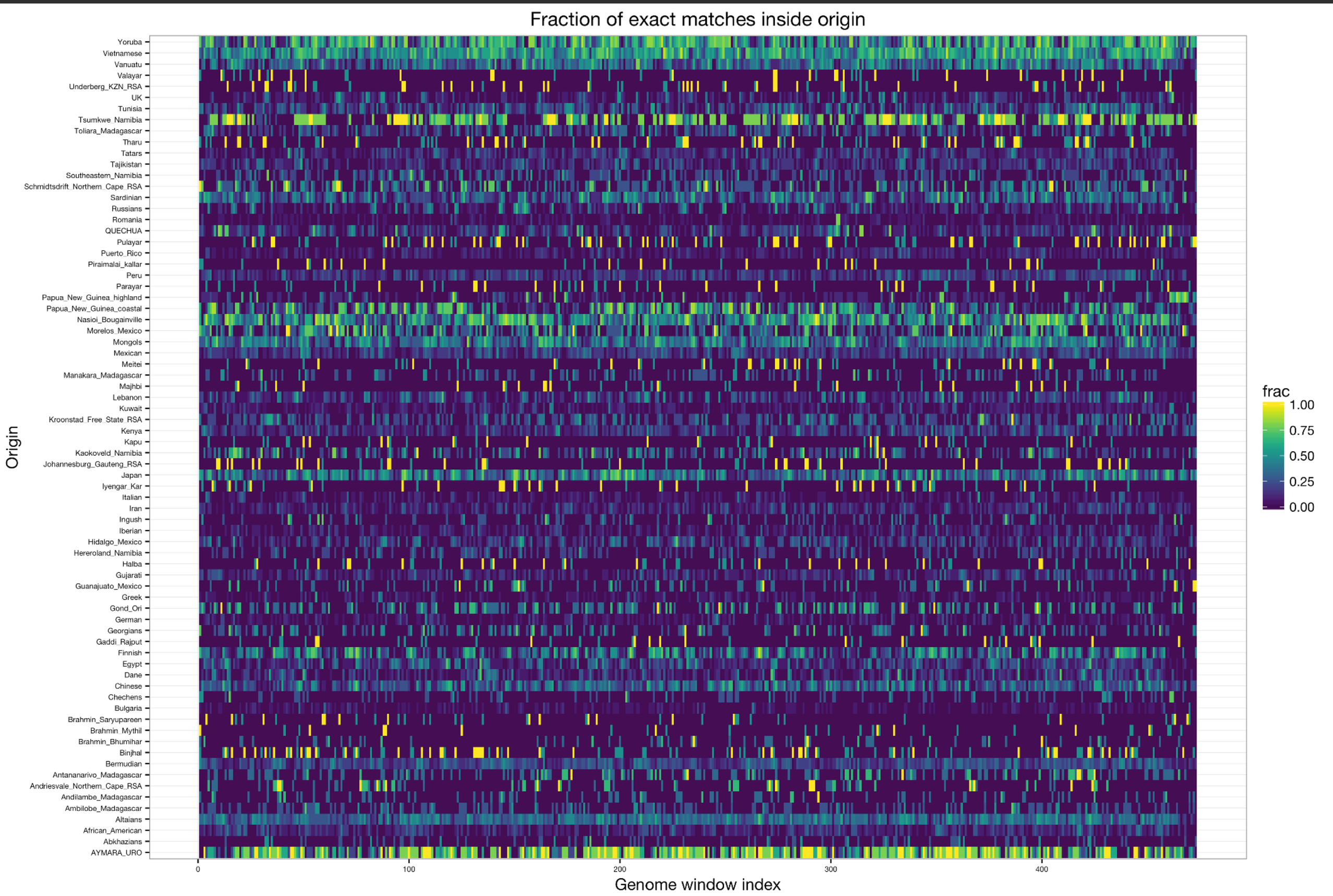

Proportions of populations
1.00
0.75
0.50
0.25
0.00
Proportion
0
200
400
600
800
Individual
population
V1
V2
V3
V4
V5
V6
V7
V8
V9

Origin assignment procedure

- For every reference origin get mean admixture vector
- Compare given vector to mean vectors
- Choose the closest one

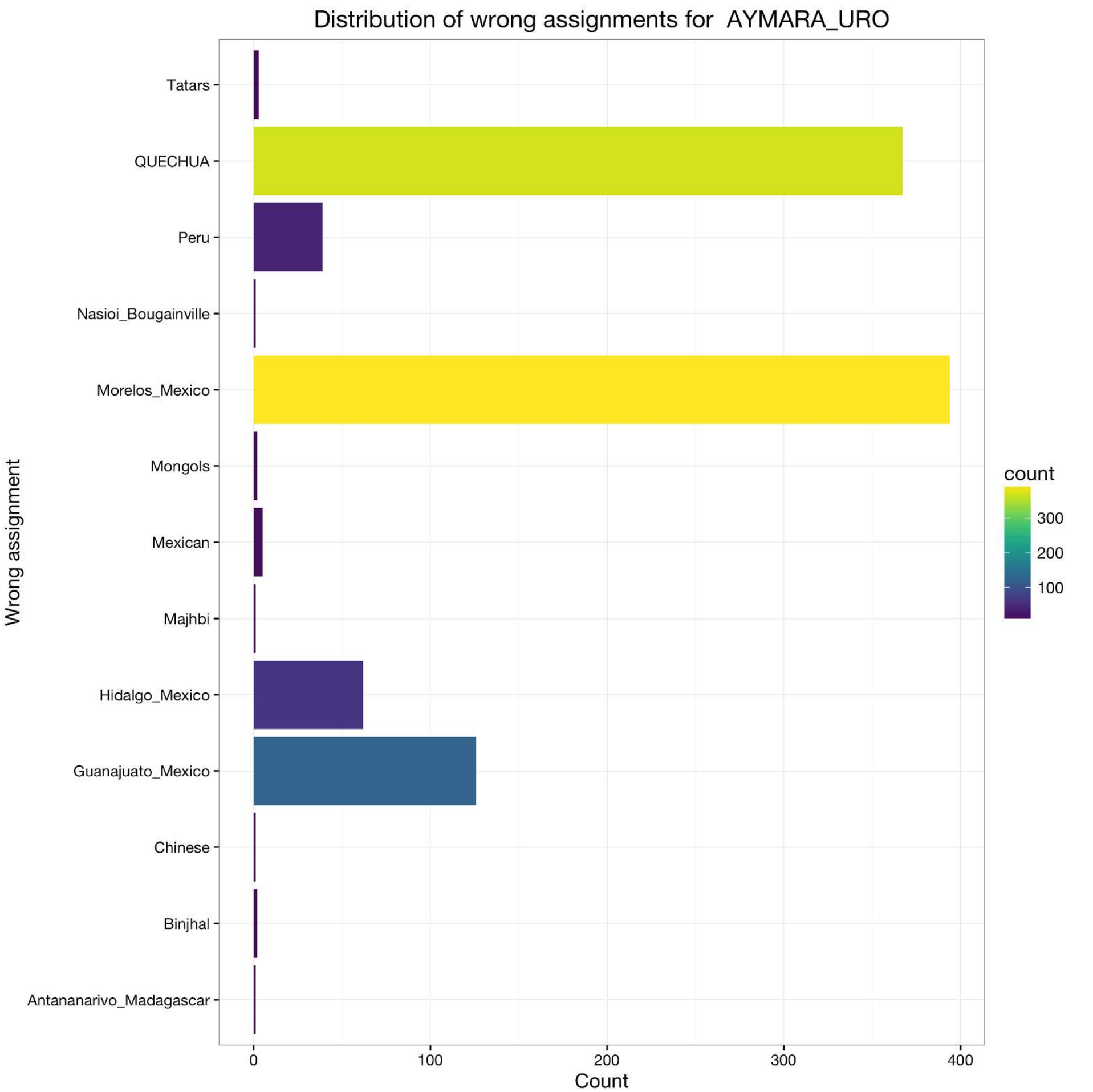
Proof of concept (zombie mix)




Fraction of exact matches inside origin
Origin
Yoruba
Vietnamese
Vanuatu
Valayar
Underberg_KZN_RSA
UK
Tunisia
Tsumkwe_Namibia
Toliara_Madagascar
Tharu
Tatars
Tajikistan
Southeastern_Namibia
Schmidtsdrift_Northern_Cape_RSA
Sardinian
Russians
Romania
QUECHUA
Pulayar
Puerto_Rico
Piraimalai_kallar
Peru
Parayar
Papua_New_Guinea_highland
Papua_New_Guinea_coastal
Nasioi_Bougainville
Morelos_Mexico
Mongols
Mexican
Meitei
Manakara_Madagascar
Majhbi
Lebanon
Kuwait
Kroonstad_Free_State_RSA
Kenya
Kapu
Kaokoveld_Namibia
Johannesburg_Gauteng_RSA
Japan
Iyengar_Kar
Italian
Iran
Ingush
Iberian
Hidalgo_Mexico
Hereroland_Namibia
Halba
Gujarati
Guanajuato_Mexico
Greek
Gond_Ori
German
Georgians
Gaddi_Rajput
Finnish
Egypt
Dane
Chinese
Chechens
Bulgaria
Brahmin_Saryupareen
Brahmin_Mythil
Brahmin_Bhumihar
Binjhal
Bermudian
Antananarivo_Madagascar
Andriesvale_Northern_Cape_RSA
Andilambe_Madagascar
Ambilobe_Madagascar
Altaians
African_American
Abkhazians
AYMARA_URO
0
100
200
300
400
Genome window index
frac
1.00
0.75
0.50
0.25
0.00

Classification by results

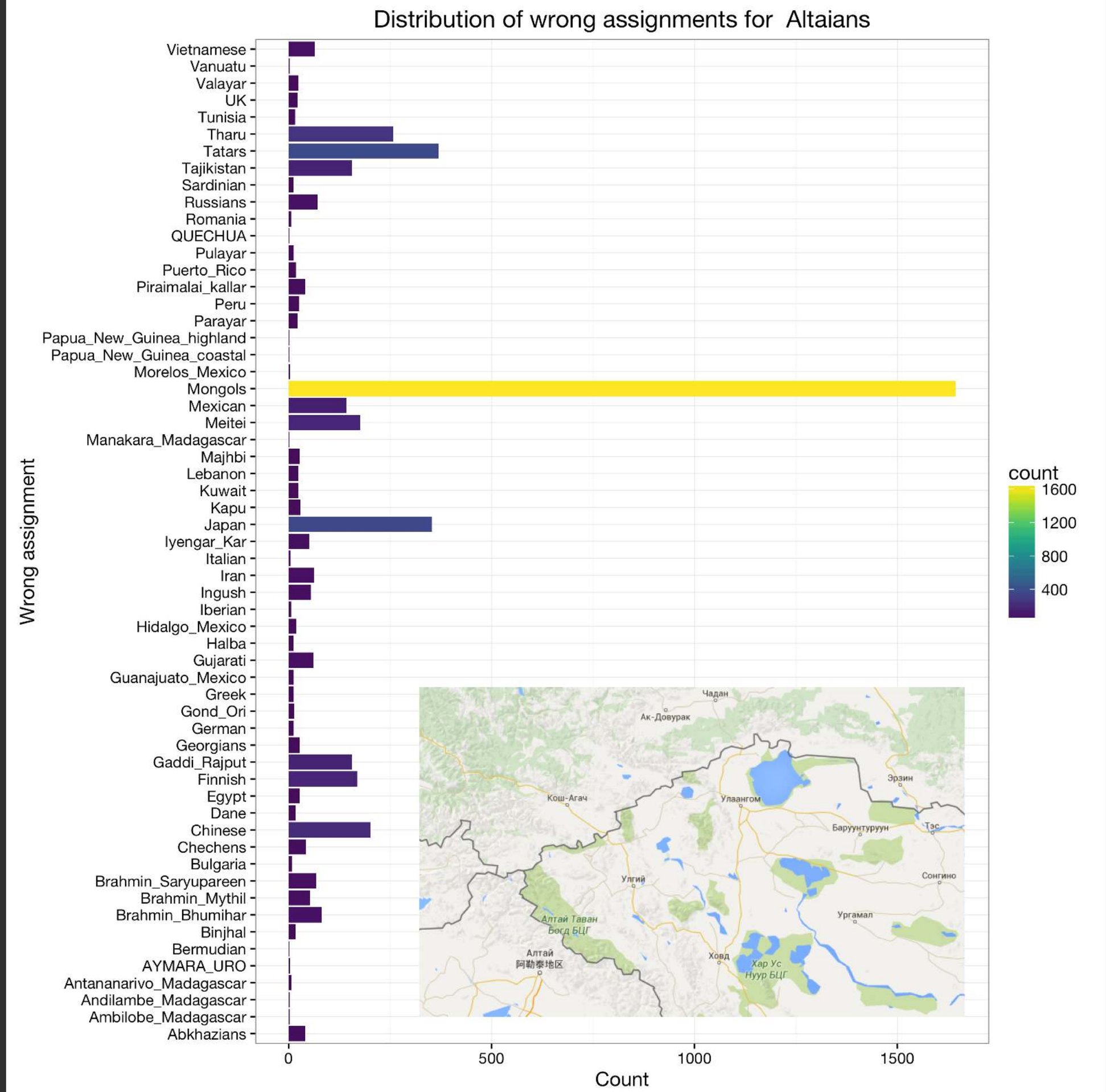
— Consistently good


Distribution of wrong assignments for AYMARA_URO
Tatars
QUECHUA
Peru
Nasioi_Bougainville
Morelos_Mexico
Mongols
Mexican
Majhbi
Hidalgo_Mexico
Guanajuato_Mexico
Chinese
Binjhal
Antananarivo_Madagascar
Wrong assignment
0
100
200
300
400
Count
count
300
200
100

Classification by results

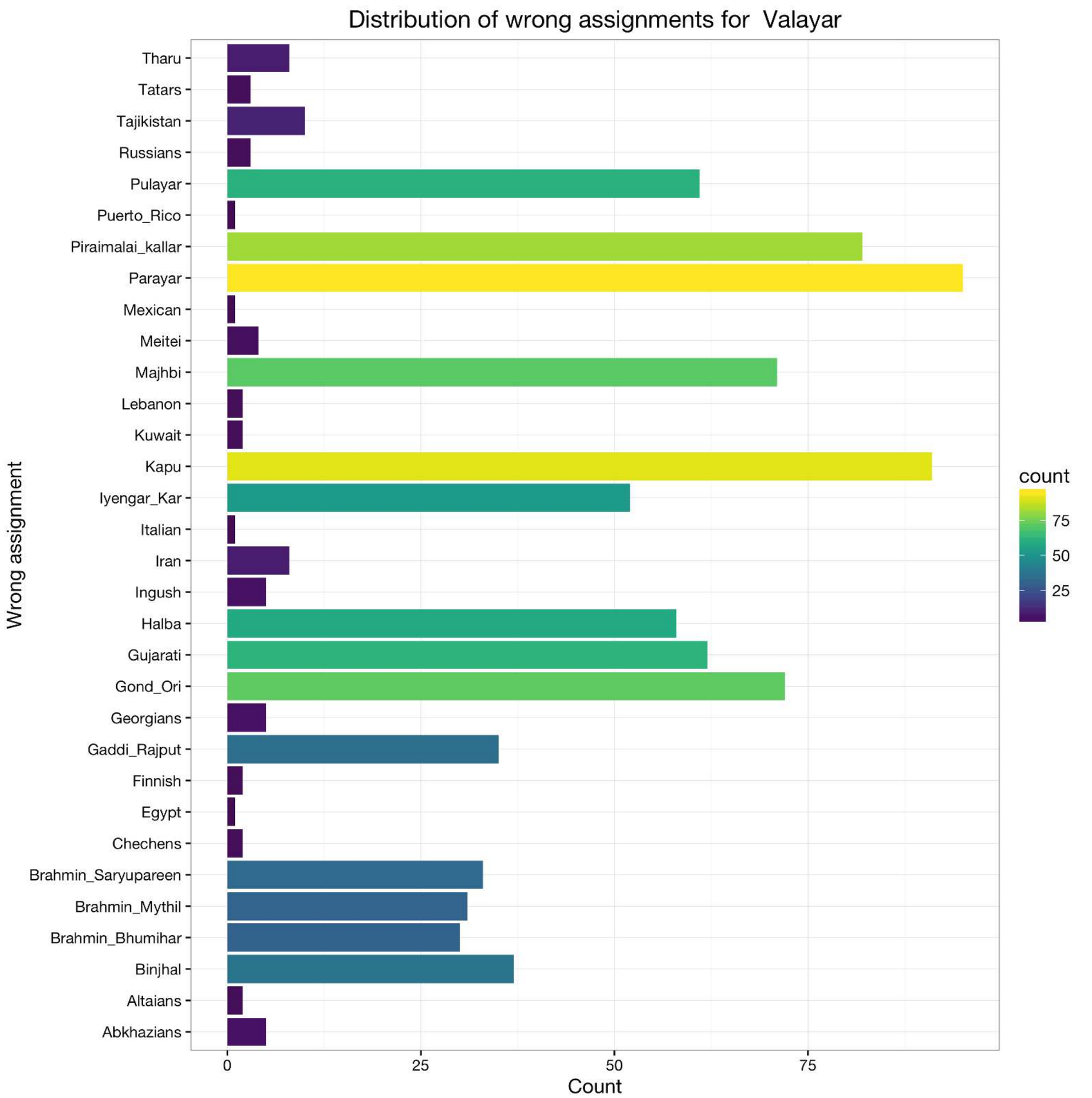
— Consistently good – relatively isolated

— Consistently average


Distribution of wrong assignments for Altaians
Wrong assignment
Vietnamese
Vanuatu
Valayar
UK
Tunisia
Tharu
Tatars
Tajikistan
Sardinian
Russians
Romania
QUECHUA
Pulayar
Puerto_Rico
Piraimalai_kallar
Peru
Parayar
Papua_New_Guinea_highland
Papua_New_Guinea_coastal
Morelos_Mexico
Mongols
Mexican
Meitei
Manakara_Madagascar
Majhbi
Lebanon
Kuwait
Kapu
Japan
Iyengar_Kar
Italian
Iran
Ingush
Iberian
Hidalgo_Mexico
Halba
Gujarati
Guanajuato_Mexico
Greek
Gond_Ori
German
Georgians
Gaddi_Rajput
Finnish
Egypt
Dane
Chinese
Chechens
Bulgaria
Brahmin_Saryupareen
Brahmin_Mythil
Brahmin_Bhumihar
Binjhal
Bermudian
AYMARA_URO
Antananarivo_Madagascar
Andilambe_Madagascar
Ambilobe_Madagascar
Abkhazians
0
500
1000
1500
Count
count
1600
1200
800
400
Чадан
Ак-Довурак
Кош-Агач
Улаангом
Эрзин
Баруунтуруун
Тэс
Улгий
Сонгино
Ургамал
Алтай Таван Богд БЦГ
Алтай
阿勒泰地区
Ховд
Хар Ус Нуур БЦГ

Classification by results

— Consistently good – relatively isolated
— Consistently average – artificial border
— Mostly bad


Distribution of wrong assignments for Valayar
Wrong assignment
Tharu
Tatars
Tajikistan
Russians
Pulayar
Puerto_Rico
Piraimalai_kallar
Parayar
Mexican
Meitei
Majhbi
Lebanon
Kuwait
Kapu
Iyengar_Kar
Italian
Iran
Ingush
Halba
Gujarati
Gond_Ori
Georgians
Gaddi_Rajput
Finnish
Egypt
Chechens
Brahmin_Saryupareen
Brahmin_Mythil
Brahmin_Bhumihar
Binjhal
Altaians
Abkhazians
0
25
50
75
Count
count
75
50
25

Classification by results

— Consistently good – relatively isolated
— Consistently average – artificial border
— Mostly bad — a lot of slightly diverged populations

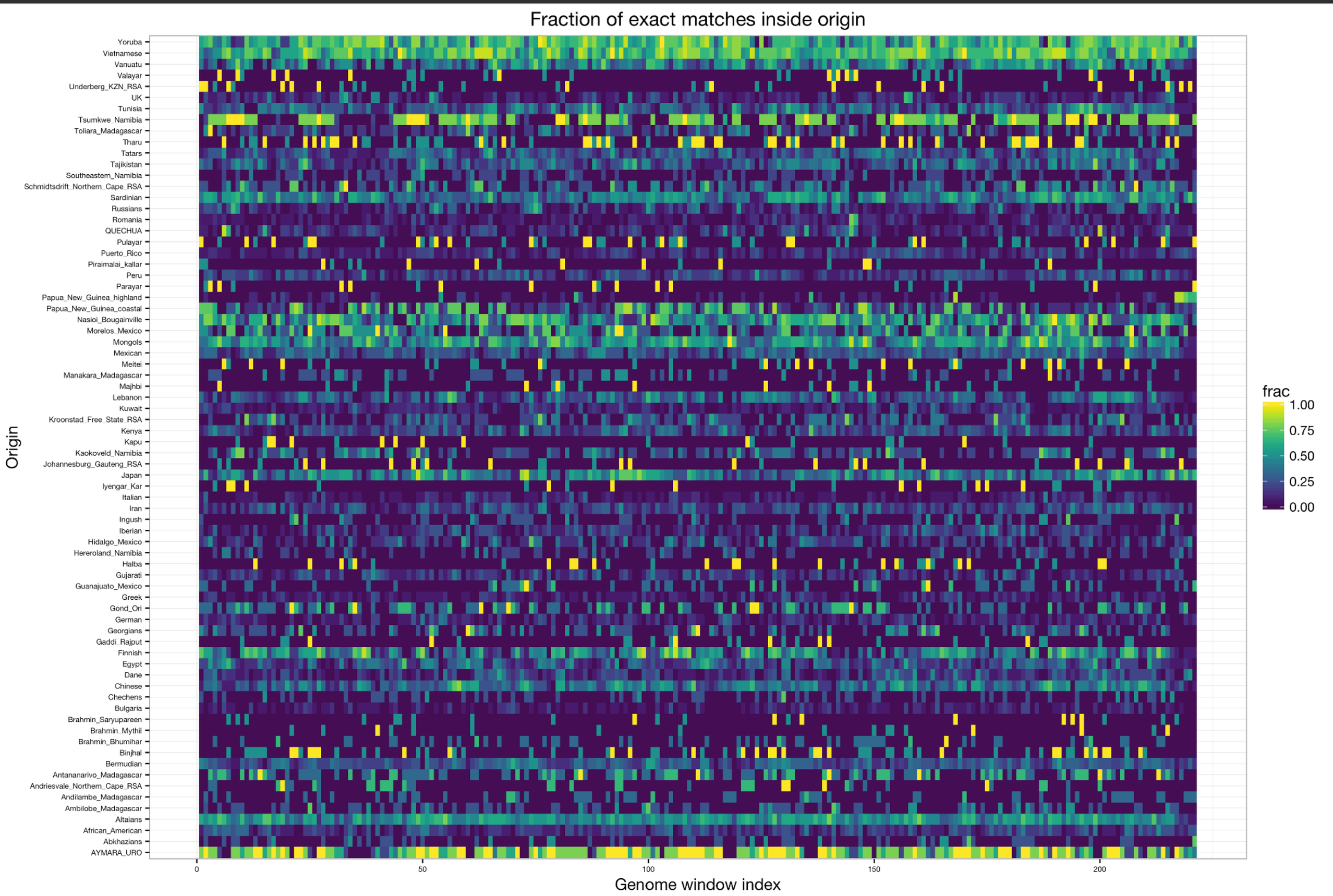
As for today

- Working classifier
- Occuring missassignments are meaningful
- «Mostly bad» ones have bright spots – highly specific windows

Next

- Check specific windows for co-localized medical markers
- See if some obvious differences in phenotype match with results
- Try other classifing approaches
- Refine pipeline for public use

Thank you


Fraction of exact matches inside origin
Origin
Yoruba
Vietnamese
Vanuatu
Valayar
Underberg_KZN_RSA
UK
Tunisia
Tsumkwe_Namibia
Toliara_Madagascar
Tharu
Tatars
Tajikistan
Southeastern_Namibia
Schmidtsdrift_Northern_Cape_RSA
Sardinian
Russians
Romania
QUECHUA
Pulayar
Puerto_Rico
Piraimalai_kallar
Peru
Parayar
Papua_New_Guinea_highland
Papua_New_Guinea_coastal
Nasioi_Bougainville
Morelos_Mexico
Mongols
Mexican
Meitei
Manakara_Madagascar
Majhbi
Lebanon
Kuwait
Kroonstad_Free_State_RSA
Kenya
Kapu
Kaokoveld_Namibia
Johannesburg_Gauteng_RSA
Japan
Iyengar_Kar
Italian
Iran
Ingush
Iberian
Hidalgo_Mexico
Hereroland_Namibia
Halba
Gujarati
Guanajuato_Mexico
Greek
Gond_Ori
German
Georgians
Gaddi_Rajput
Finnish
Egypt
Dane
Chinese
Chechens
Bulgaria
Brahmin_Saryupareen
Brahmin_Mythil
Brahmin_Bhumihar
Binjhal
Bermudian
Antananarivo_Madagascar
Andriesvale_Northern_Cape_RSA
Andilambe_Madagascar
Ambilobe_Madagascar
Altaians
African_American
Abkhazians
AYMARA_URO
0
50
100
150
200
Genome window index
frac
1.00
0.75
0.50
0.25
0.00