Practical Introduction to Population Structure

Scott Hazelhurst

2014

Population Structure

Explosion of data from complete genome sequencing, GWAS, . . . can be used to explore structure in the population

Goals:

- Understanding population histories
- Dealing with confounding effect of population structure in GWAS
- Anomalies in the data.

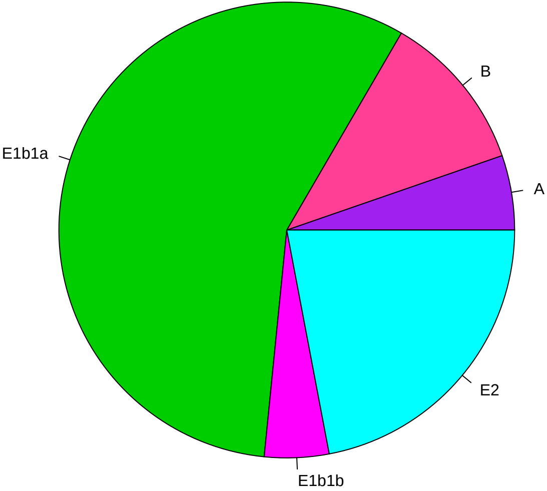
Haplotype analysis

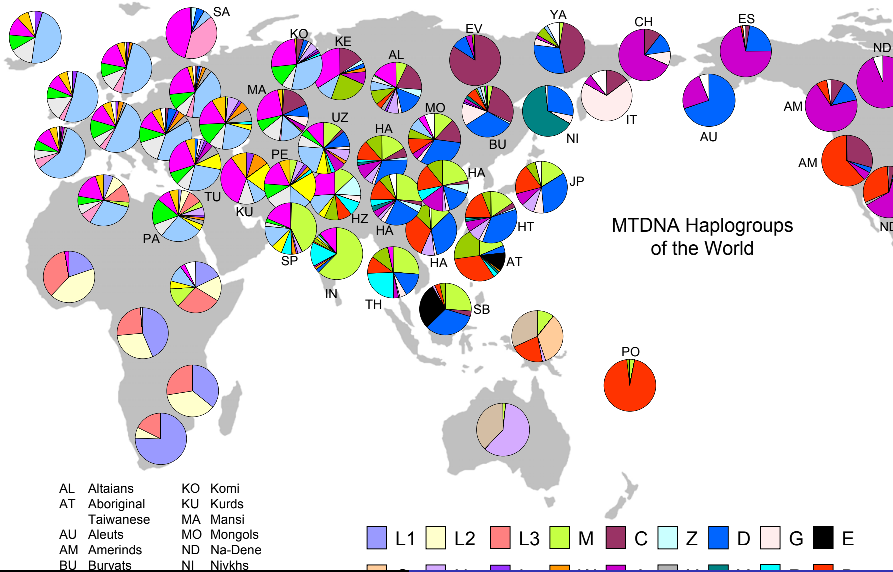
Until recently pop history done with:

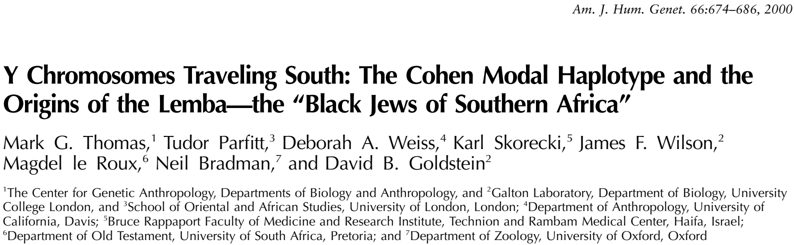
- Y-chromosome data (e.g., STR)
- mt-DNA

Done at population level:

- Resolve haplotypes
- Characterise populations with haplotype composition


B
A
E1b1a
E2
E1b1b


MTDNA Haplogroups
of the World
SA
KO
KE
AL
EV
YA
CH
ES
ND
MA
UZ
MO
BU
NI
IT
AU
AM
AM
HA
PE
TU
KU
PA
HZ
HA
HA
HT
JP
SP
IN
TH
HA
AT
SB
PO
ND
AL Altaians
AT Aboriginal Taiwanese
AU Aleuts
AM Amerinds
BU Buryats
KO Komi
KU Kurds
MA Mansi
MO Mongols
ND Na-Dene
NI Nivkhs
L1
L2
L3
M
C
Z
D
G
E

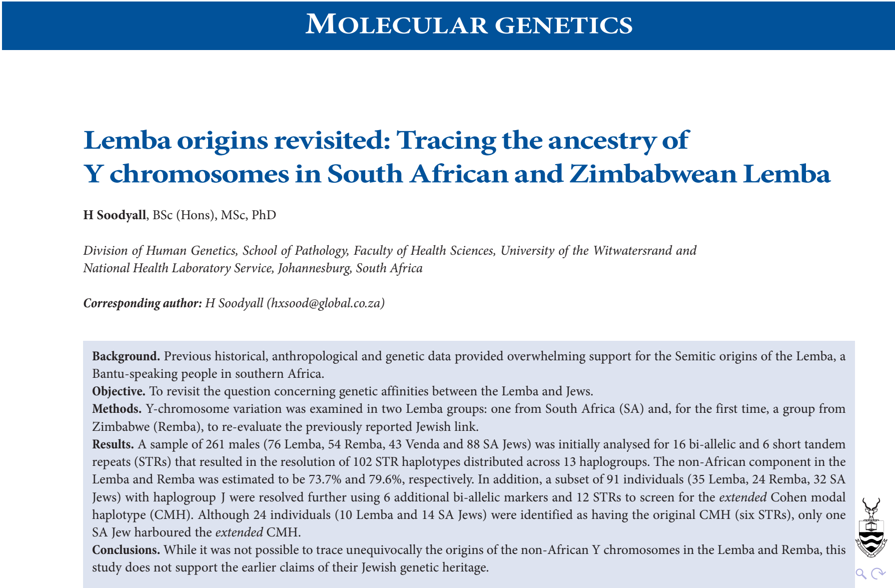


Am. J. Hum. Genet. 66:674–686, 2000

Y Chromosomes Traveling South: The Cohen Modal Haplotype and the Origins of the Lemba—the "Black Jews of Southern Africa"

Mark G. Thomas,[1] Tudor Parfitt,[3] Deborah A. Weiss,[4] Karl Skorecki,[5] James F. Wilson,[2] Magdel le Roux,[6] Neil Bradman,[7] and David B. Goldstein[2]

[1]The Center for Genetic Anthropology, Departments of Biology and Anthropology, and [2]Galton Laboratory, Department of Biology, University College London, and [3]School of Oriental and African Studies, University of London, London; [4]Department of Anthropology, University of California, Davis; [5]Bruce Rappaport Faculty of Medicine and Research Institute, Technion and Rambam Medical Center, Haifa, Israel; [6]Department of Old Testament, University of South Africa, Pretoria; and [7]Department of Zoology, University of Oxford, Oxford



MOLECULAR GENETICS

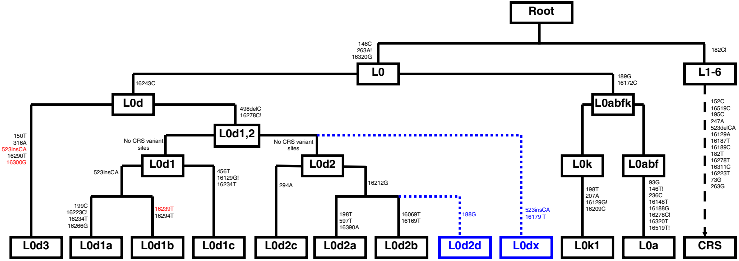
Lemba origins revisited: Tracing the ancestry of Y chromosomes in South African and Zimbabwean Lemba

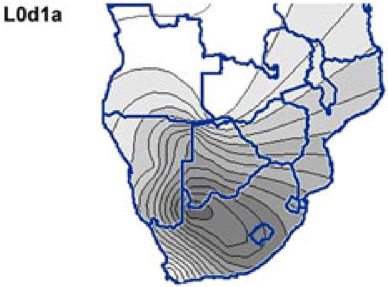
H Soodyall, BSc (Hons), MSc, PhD

Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand and National Health Laboratory Service, Johannesburg, South Africa

Corresponding author: *H Soodyall (hxsood@global.co.za)*

Background. Previous historical, anthropological and genetic data provided overwhelming support for the Semitic origins of the Lemba, a Bantu-speaking people in southern Africa.
Objective. To revisit the question concerning genetic affinities between the Lemba and Jews.
Methods. Y-chromosome variation was examined in two Lemba groups: one from South Africa (SA) and, for the first time, a group from Zimbabwe (Remba), to re-evaluate the previously reported Jewish link.
Results. A sample of 261 males (76 Lemba, 54 Remba, 43 Venda and 88 SA Jews) was initially analysed for 16 bi-allelic and 6 short tandem repeats (STRs) that resulted in the resolution of 102 STR haplotypes distributed across 13 haplogroups. The non-African component in the Lemba and Remba was estimated to be 73.7% and 79.6%, respectively. In addition, a subset of 91 individuals (35 Lemba, 24 Remba, 32 SA Jews) with haplogroup J were resolved further using 6 additional bi-allelic markers and 12 STRs to screen for the *extended* Cohen modal haplotype (CMH). Although 24 individuals (10 Lemba and 14 SA Jews) were identified as having the original CMH (six STRs), only one SA Jew harboured the *extended* CMH.
Conclusions. While it was not possible to trace unequivocally the origins of the non-African Y chromosomes in the Lemba and Remba, this study does not support the earlier claims of their Jewish genetic heritage.


Root
146C
263A!
16320G
182C!
L0
L1-6
16243C
189G
16172C
L0d
L0abfk
498delC
16278C!
152C
16519C
195C
247A
523delCA
16129A
16187T
16189C
182T
16278T
16311C
16223T
73G
263G
L0d1,2
No CRS variant sites
No CRS variant sites
150T
316A
523insCA
16290T
16300G
L0d1
L0d2
L0k
L0abf
523insCA
456T
16129G!
16234T
294A
16212G
198T
207A
16129G!
16209C
93G
146T!
236C
16148T
16188G
16278C!
16320T
16519T!
199C
16223C!
16234T
16266G
16239T
16294T
198T
597T
16390A
16069T
16169T
188G
523insCA
16179 T
L0d3
L0d1a
L0d1b
L0d1c
L0d2c
L0d2a
L0d2b
L0d2d
L0dx
L0k1
L0a
CRS


L0d1a

FGS & SNP chip data

Availability of SNP-chip data and complete genome sequence data makes much richer data available

- Explore all lineages of individuals
- Complexities
  - Half of genetic information "lost" in each generation
  - Recombination through meiois ($\sim$ 35 breakpoints per generation)
- Statistical in nature

Complements mt and Y chromosome data

Two basic approaches

- Principal component analysis (PCA) : eigenstrat, SNPRelat
- Structure based : structure, admixture

Principal Component Analysis

PCA is general technique for dealing with high-dimensional data which is

- Difficult to visualise
- Not all dimensions in the data same importance, many correlated

PCA does dimension reduction – project data into lower-dimension space

- axes independent to each other
- can estimate importance of axes

Genotype to population structure

Each individual in study has a vector of genotype information

- $s_i = \langle g_{i,0}, g_{i,1}, \ldots, g_{i,n-1} \rangle$

Compute *distance* between two individuals, s_i, s_j

- Common: sum of differences between vectors (0, 1, 2 per position).

From pairwise distances:

- Matrix D
- $D[i, j]$ is *normalised* distance between s_1, s_j
- Implicitly embeds the individuals in a high dimensional space

Goal is to cluster individuals that are close to each other

```
P0: AA AC AT AA
P1: TT AA TT AA
P2: AA AC TT TT
```

```
P0: AA AC AT AA
P1: TT AA TT AA
P2: AA AC TT TT
```

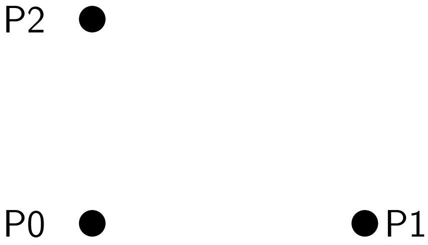
Distance:

- P0, P1: 4
- P0, P2: 3
- P1, P2: 5

```
P0: AA AC AT AA
P1: TT AA TT AA
P2: AA AC TT TT
```

Distance:

- P0, P1: 4
- P0, P2: 3
- P1, P2: 5

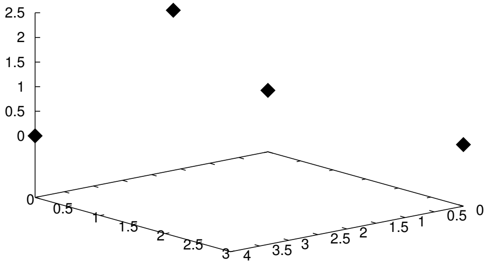


```
P0: AA AC AT AA
P1: TT AA TT AA
P2: AA AC TT TT
P3: AT CC TT AA
```

Distance:

	P0	P1	P2	P3
P0	0	4	3	3
P1		0	5	3
P2			0	4

Can only be embedded in 3D space

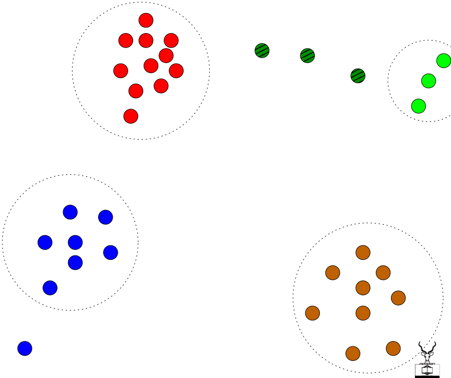


P0: (0,0,0) P1: (4,0,0)
P2: (0,3,0) P3=(2,0.33,2.21)

Cluster analysis

From distance matrix draw individuals in *k*-dimensional space.

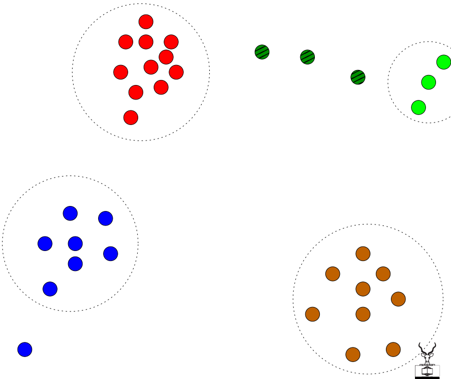
Clusters: individuals closer to each other than others.



Cluster analysis

From distance matrix draw individuals in *k*-dimensional space.

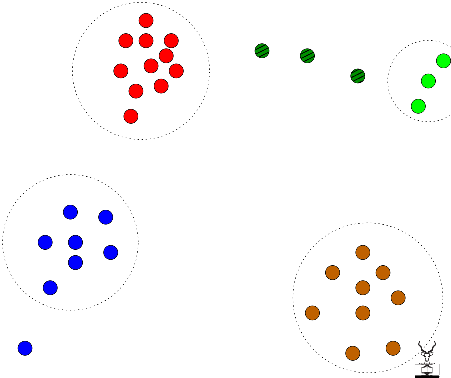
May see outliers, admixed individuals



Cluster analysis

From distance matrix draw individuals in *k*-dimensional space.

Position of individuals from genotype information.

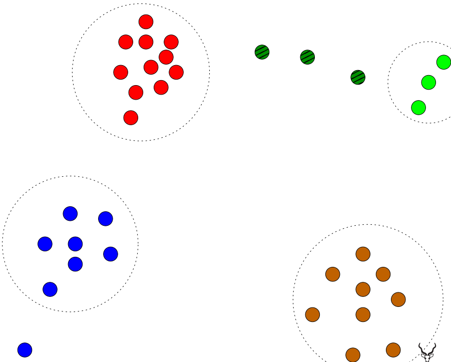


Cluster analysis

From distance matrix draw individuals in k-dimensional space.

May have external info – e.g., case/control, population group – use colours.

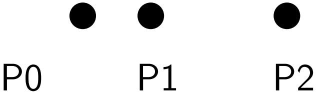
- Interested in interplay between internal, external evidence



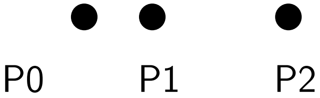
Problems. . .

- *Extremely* high dimensional
- Some dimensions correlated, not useful

	P0	P1	P2
P0	0	2	4
P1		0	2

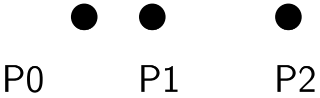


	P0	P1	P2
P0	0	2	4
P1		0	2

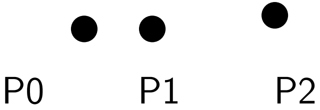


	P0	P1	P2
P0	0	2	4.1
P1		0	2

	P0	P1	P2
P0	0	2	4
P1		0	2



	P0	P1	P2
P0	0	2	4.1
P1		0	2



Principal Component Analysis – PCA

Transform data so that it is embedded in another space.

- Preserve relative distance between individuals.
- Number of dimensions/components are reduced.
- Components independent-ish of each other.
- Ordered by importance.

Common method is *eigendecomposition* – takes distance matrix and produces:

- Eigenvalues: λ_i is relative importance of dimension i
- Eigenvectors: $\mathbf{v}_i$ coordinates of each individual in the i-th dimension

Coordinate of indvidual j is

$$(v_1[j], v_2[j], v_3[j], \ldots)$$

NB: although fewer dimensions, still many dimensions.

How many dimensions?

Use 2 or 3 at a time for display – but may need to look at/analyse several

Top PCs are most important, lower ones are noise. Where cut-off?

- If use too few, miss real signal structure, and may get false signals in GWAS.
- if use too many, may reduce power because of noise in data

Choice of number of PCs by eyeball method, or using Tracy-Widom statistic, Velicer's MAP test, ANOVA test.

Guidelines for PCA

- Most studies – autosomal SNPs
- The more SNPs the better (Eigenstrat says $\geq$100k), but YMMV.
- SNPs should not be in LD with each other – prune first.
- Skew group sizes may skew results – try to have balanced groups sizes.
- Interpret distances with caution.
- Can you explain significant PCs: true biology, or indication of problem

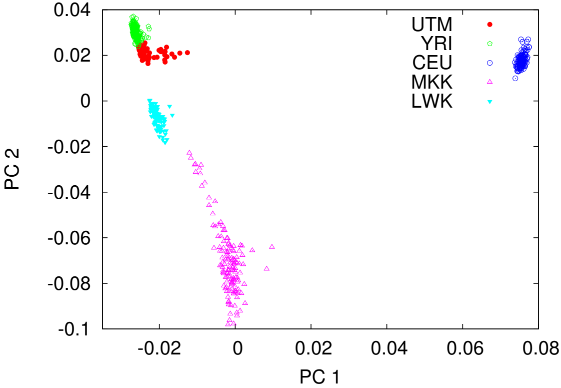
SNPRelate – alternative to eigenstrat

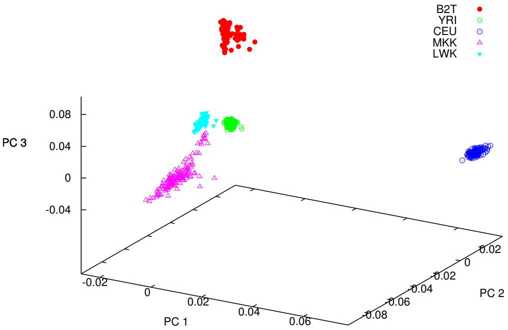
R package

- Faster than eigenstrat – at least version ≤ 4
- Suggestion: if you know R or have a very big set, use SNPRelate
  Otherwise eigenstrat

Eigenstrat

Powerful program for doing PCA – lots of features.




B2T
YRI
CEU
MKK
LWK
PC 3
0.08
0.04
0
-0.04
-0.02
0
0.02
0.04
0.06
PC 1
0.02
0
-0.02
-0.04
-0.06
-0.08
PC 2

```
    #eigvals: 1.499 1.107 1.096 1.090 ....  081 1.079 1.075 1.07
CH18526:NA18526 0.109 -0.198 -0.012 ...   0.1007 -0.1290 -0.187
CH18524:NA18524 0.107  0.091  0.228 ...  -0.0330 -0.2377  0.104
CH18529:NA18529 0.074  0.033  0.016 ...  -0.1024 -0.0767 -0.019
CH18558:NA18558 0.104  0.155  0.244 ...  -0.0530 -0.1045  0.290
CH18532:NA18532 0.106 -0.098 -0.017 ...   0.0290 -0.1142  0.072
CH18561:NA18561 0.106  0.268  0.072 ...  -0.3031 -0.0435 -0.236
CH18562:NA18562 0.099 -0.004 -0.055 ...  -0.1120  0.0058 -0.075
CH18537:NA18537 0.113 -0.037  0.163 ...  -0.1631 -0.2544  0.143
...
```

Calling EIGENSTRAT

Call smartpca directly.

- Most powerful. Takes a parameter file as input. Complex as eigenstrat takes different formats.

Calling EIGENSTRAT

Call smartpca.perl: much easier, but also lots of options. Creates *par* file and calls smartpca.

```
smartpca.perl -i hm3-prune.bed
    -a hm3-prune.bim -b hm3-prune.fam
    -p hm3-prune.pca
    -e hm3-prune.eval
    -o hm3-prune.pca
    -q NO -l hm3-prune.log
```

Script *runpca*

- runpca hm3-prune

```
#! /bin/bash

smartpca.perl
     -i $1.bed -a $1.bim -b $1.fam
     -p $1.pca -e $1.eval -o $1.pca -q NO
     -l $1.log $2
```

Interpreting results

Look at the log and standard out/error . . . here are some

```
Average divergence between populations:
      CEU    YRI MbutiPygmies    Han    San po
CEU  1.165  1.485  1.499  1.316  1.531  17
YRI  1.485  1.093  1.160  1.478  1.208   9
MbutiPygmies  1.499  1.160  0.875  1.483  1.05
Han  1.316  1.478  1.483  0.987  1.510  19
San  1.531  1.208  1.059  1.510  0.880       2
```

```
eigenvector 1:means
       MbutiPygmies     -0.179
                San     -0.150
                YRI     -0.121
                CEU      0.078
                Han      0.126
```

```
## Anova statistics for population
differences along each eigenvector:
                                       p-value
           eigenvector_1_overall_             0 +++
  MbutiPygmies minv:    -0.179                  Han maxv:     0.
           eigenvector_1_CEU_YRI_             0 +++
  eigenvector_1_CEU_MbutiPygmies_             0 +++
           eigenvector_1_CEU_Han_   1.11022e-16 +++
           eigenvector_1_CEU_San_             0 +++
  eigenvector_1_YRI_MbutiPygmies_             0 +++
           eigenvector_1_YRI_Han_             0 +++
           eigenvector_1_YRI_San_   8.98579e-05 ***
  eigenvector_1_MbutiPygmies_Han_   1.11022e-16 +++
  eigenvector_1_MbutiPygmies_San_   1.75826e-06 ***
           eigenvector_1_Han_San_             0 +++
eigenvector 5:means
           San     -0.004
           CEU     -0.001
           Han      0.000
  MbutiPygmies      0.001
           YRI      0.002
            eigenvector_5_overall_      0.999997
            San minv:    -0.004    YRI maxv:     0.002
            eigenvector_5_CEU_YRI_      0.973486
   eigenvector_5_CEU_MbutiPygmies_      0.979764
            eigenvector_5_CEU_Han_      0.978691
            eigenvector_5_CEU_San_      0.987923
```

Tracy-Widom statistics

Used to assess if there is significant structure in each PC

Use the `twstats` program (also need twtable file)

```
twstats -t twtable -i sample.eval -o sample.tw
```

yields

```
#N  eigenval difference  twstat p-value effect
1   8.977       NA   22.25 2.58e-32  46.592
2   4.006  -4.971 37.21 1.07e-67  206.977
3   2.057  -1.948 42.68 9.94e-83  894.681
4   1.118  -0.939  7.11 1.00e-07 2330.993
5   1.013  -0.104 -1.07 0.438162 2551.689
```

smartpca notes

- Can support quantitative traits.
- smartpca does outlier detection: by default remove individuals >6 standard deviations from 0 on any of the top 10 PCs.
- algorithm is quadratic in number of samples $\times$ number of SNPs (in space and time)
- lots of different options

CLI programs

Can use gnuplot or R directly.
Wrapper programs

- ploteig
  Comes with the smartpca program
- evec2gp
  `http://www.bioinf.wits.ac.za/software/poputils`

Genesis

Interactive program for producing PCA plots

- Can choose which PCs (2D or 3D)
- Change colours
- Identify, hide individuals
- Control display

Structure based approaches

Statistical models.

- Assume K underlying groups
- Unknown: Population k contributes a fraction q_{ik} of individual i's genome.
- Unknown: Allele 1 at SNP j has frequency f_{kj} in population k.

Start off knowing only knowing *Obs*: overall frequencies of alleles and each individuals' state

For person i at SNP j, consider probability for that person that they have alleles 1/1, 1/2 or 2/2

- $\Pr(1/1) = [\sum_k q_{ik} f_{kj}]^2$
- $\Pr(1/2) = 2[\sum_k q_{ik} f_{kj}][\sum_k q_{ik}(1 - f_{kj})]$
- $\Pr(2/2) = [\sum_k q_{ik}(1 - f_{kj})]^2$

Derive probabilistic objective function:

$$\mathcal{L}(\langle q_{ik}, f_{kj}\rangle | Obs)$$

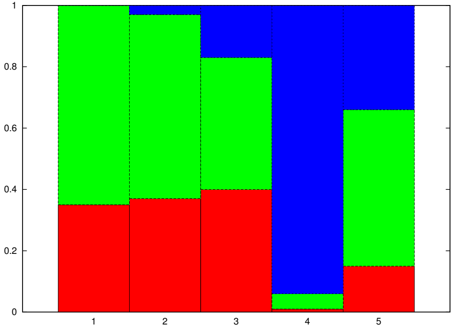
Goal is to maximise:

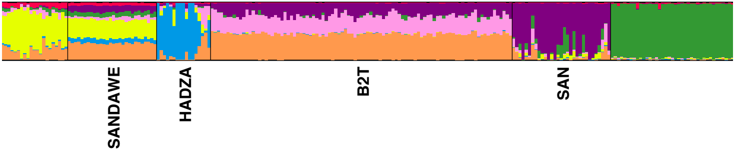
- probabilistic
- structure, admixture are programs that compute this.

We use *admixture*:

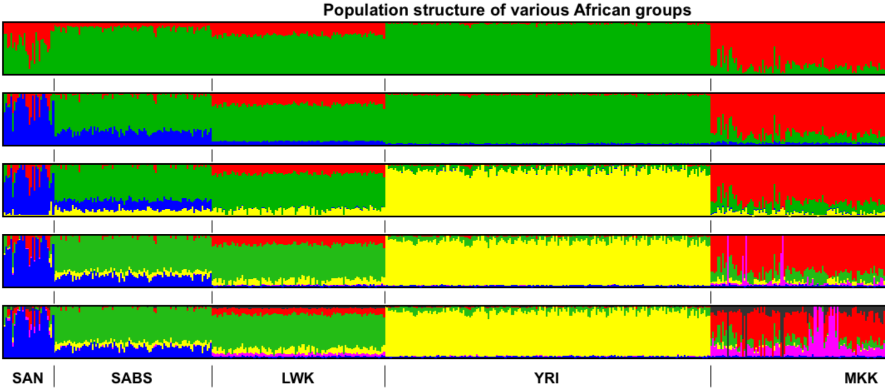
- Given value K, probabilistically compute populations tructure

```
0.35 0.65 0.00
0.37 0.60 0.03
0.4  0.43 0.17
0.01 0.05 0.94
0.15 0.51 0.34
```




SANDAWE
HADZA
B2T
SAN

- Try different values of K.



Choice of K

Important to have reasonable choice of K – population structure, history important to understand.

- Eyeball method
- Use PCA – but number ancestral pops not same as number clusters.
- Cross-validation.

ADMIXTURE program: Alexander *et al.*

Fast, flexible program.

Basic form of operation:

```
    admixture data.bed 5
```

Find admixture on data set for $K = 5$.

Produces:

- data.5.Q
  Estimate of ancestry of each individual – one row per individual
- data.5.P
  Estimate of ancestry of each SNP – one per SNP

`-j`: multi-core option

```
admixture -j3 data.bed 5
```

Random seed

```
admixture -s 871 data.bed 5
admixture -s time data.bed 5
admixture -s $$ data.bed 5
```

Cross-validation – `--cv` for a given K, compute Q, P.

- Then, repeat analysis several times, each time removing a proportion of the SNPs.
- Compute difference between results from partial and full analysis
- Get error estimate

Then repeat for different values of K. Good K are those with lowest error estimate.

CLUMPP

Typically should run many times to get range of possibilities

- CLUMPP [Jakobsson and Rosenberg, 2006] averages, label switches

```
0.35 0.65 0.00          0.60 0.07 0.33
0.37 0.60 0.03          0.60 0.05 0.35
0.4  0.43 0.17          0.47 0.13 0.40
0.01 0.05 0.94          0.15 0.80 0.05
0.15 0.51 0.34          0.52 0.32 0.16
```

Solving label switching

- Given Q_1, Q_2 and a mapping/alignment p between columns of Q_i, can compute H – error estimate.
- Generalise to r matrices
- Want to find alignment that minimise the error.

Three basic algorithms:

- FullSearch
- Greedy: Iterative on runs
- LargeKGreedy: Simplified – iterative on runs, columns

Format of *indfile*

Concatenation of multiple Q files with some prefix

```
1 1 (1) 1 : 0.916977 0.038781 0.044243
2 2 (2) 1 : 0.930008 0.049758 0.020234
3 3 (3) 1 : 0.982630 0.000010 0.017360
4 4 (4) 1 : 0.893398 0.070601 0.036001
5 5 (5) 1 : 0.958184 0.030509 0.011307
6 6 (6) 1 : 0.972452 0.021890 0.005658
```

Format of *paramfile*

CLUMPP expects a file called *paramfile* as input in directory

```
DATATYPE 0    # 0 for Q file 1 for P file
INDFILE allh.indfile    # data file
OUTFILE allh.outfile    # output
MISCFILE allh.miscfile  # log
K 3   # num clusters
C 639 # num people
R 100 # num runs
M 2   # 1=FullSearch,2=Greedy,3=LargeKGreedy)
W 0   # weight by size of pop (0=no, 1=tes)
```

Use existing *paramfile* as template

Practical suggestion for using CLUMPP

Create directory for your plink data, e.g. *projectA* contains: uab.bed, uab.bim, uab.fam

Decide number of runs you want R, and which K.

- Make directories 1,. . . ,R
- Inside each directory, run admixture for each K value
- e.g., projectA/12 will contain *uab.Q.3*, *uab.Q.4*, ...
- Run *cdg.py* script to combine results.

runadmix.sh script

```
#!/bin/bash
DATA=hapmap1.bed
KMIN=3
KMAX=5
R=10
for i in ‘seq $R‘; do
    mkdir -p $i
    cd $i
    for k in ‘seq $KMIN $KMAX‘; do
       admixture ../$DATA $k -j3
    done
    cd ..
done
```

```
python cdg.py          \
   -K 3                \
   --glob "[0-9]*/"    \
   --output result --par_clumpp allh
```

- *glob*: a Un*x glob which specifies the directories where the Q files can be found (NB: the $/$).
- *output*: name of output *indfile*
- NB: file *paramfile* is created, over-writing existing file

Edit the paramfile as needed (e.g., change method)

Genesis

Genesis program can

- display single, multiple structure charts
- interactive
- change colours, orders, headings

Found at

http://www.bioinf.wits.ac.za/software/

Distruct

Rosenberg [2004]

- CLUMPP/Admixture/Structure provide "objective" evidence of admixture of each individual from K unknown ancestral groups.
- We know ascribed membership of groups

Distruct displays this appropriately.

`http://www.stanford.edu/group/rosenberglab/distructDownload.html`