MLE for individual ancestries, population covariances, and selection

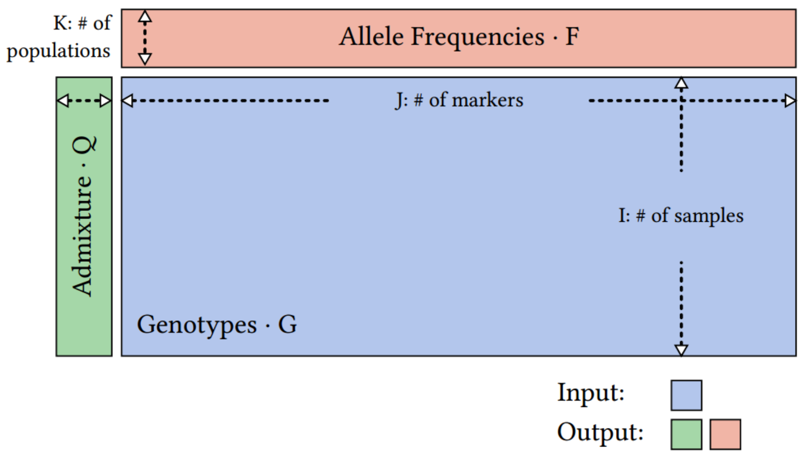
Jade Yu Cheng[1,2*], Rasmus Nielsen[1,3]

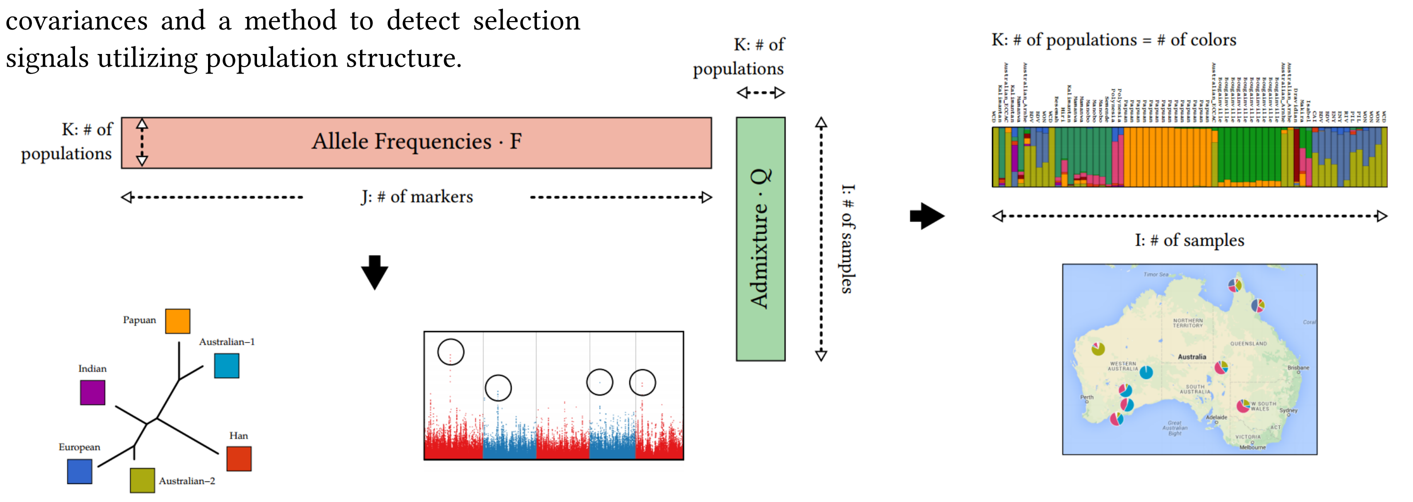
1. Departments of Integrative Biology and Statistics, University of California, Berkeley, CA, USA
2. Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
3. Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Denmark
* Email: ycheng@cs.au.dk

Overview

Questions about population structure arise in many contexts such as human evolution and disease genetics association studies. The initial focus of our project has been to study population structure by developing a model-based inference strategy using quadratic programming with the active set algorithm.

Inspired by a number of recent methods that use Gaussian distributions to approximate the distribution of allele frequencies among populations, our project has also focused on developing a method to infer population covariances and a method to detect selection signals utilizing population structure.





Population Covariances Inference

$$P\left(f_j \mid \Omega,\ \mu_j\right) \sim \mathcal{N}\left(\mu_j,\ \mu_j\left(1-\mu_j\right)\Omega\right).$$

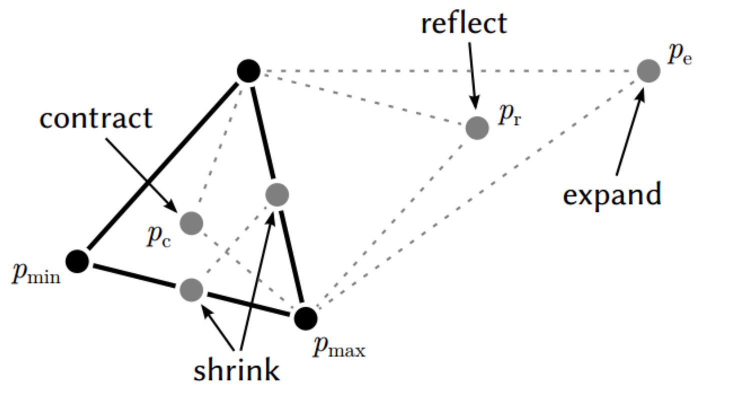
Likelihood Model

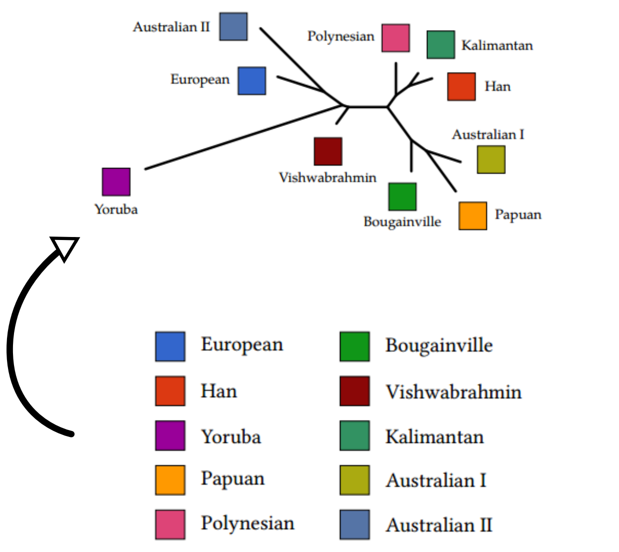
$$\begin{aligned}
&\ln\left[P_2\left(F\right)\right]\\
&= \ln\left\{\prod_j^J\left[\frac{1}{\sqrt{|2\pi c_j\Omega'|}}\right.\right.\\
&\qquad\left.\left.\exp\left(-\frac{1}{2}\cdot f_j'^T\cdot\left(c_j\Omega'\right)^{-1}\cdot f_j'\right)\right]\right\}\\
&= -\frac{1}{2}\cdot\sum_j^J\left\{\left(K-1\right)\cdot\ln\left(2\pi c_j\right)+\right.\\
&\qquad\left.\ln\left[\det\left(\Omega'\right)\right]+\frac{1}{c_j}\cdot f_j'^T\cdot\Omega'^{-1}\cdot f_j'\right\}
\end{aligned}$$

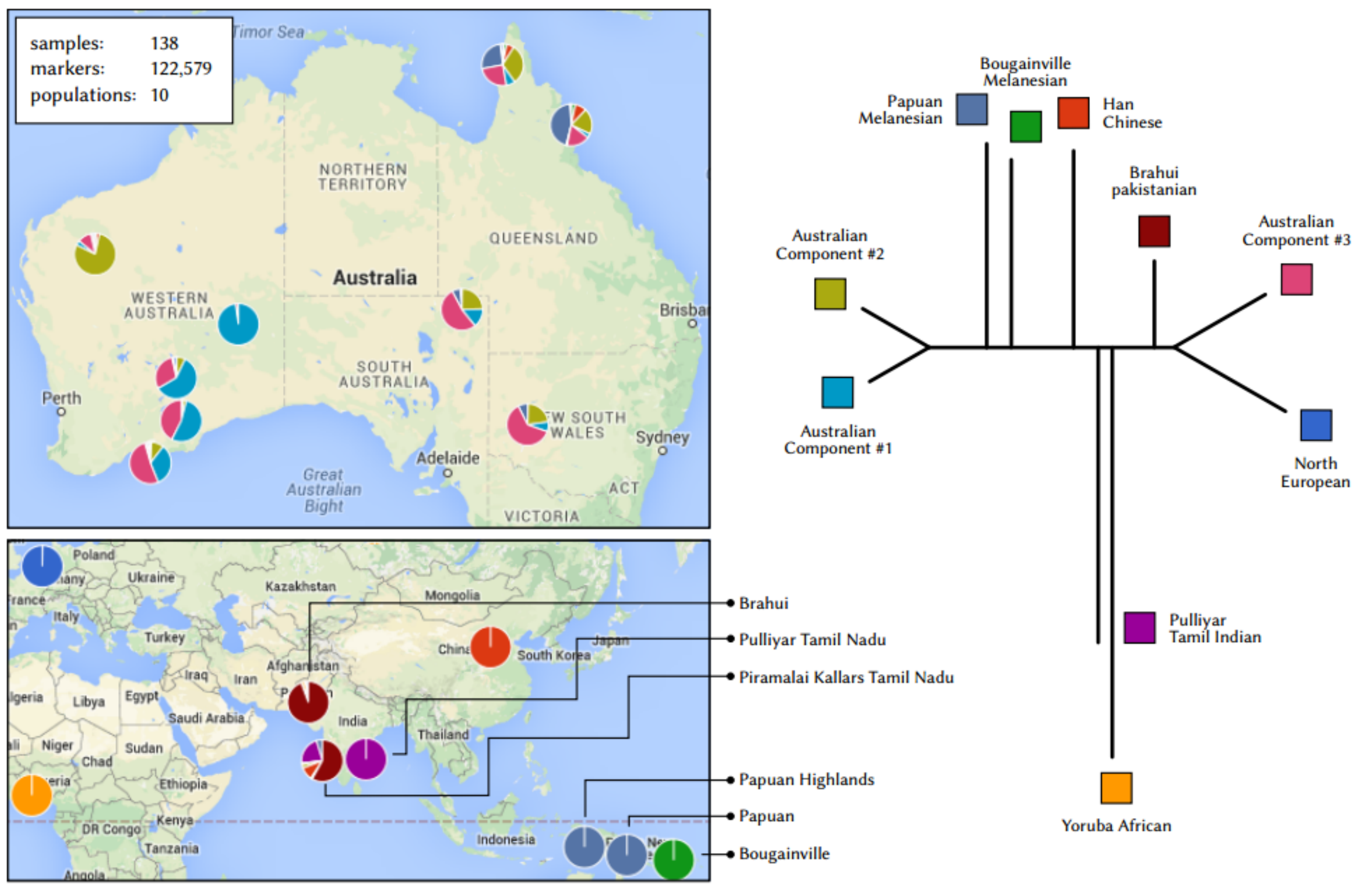
$$\text{where } c_j = \mu_j\left(1-\mu_j\right)$$
$$f_j' = f_j - f_{j0}.$$

NM

```
Repeat until a stopping criteria is reached
  Evaluate each point in the simplex using the objective function
  Determine the point $p_{min}$ with the lowest score
  Reflect $p_{min}$ through the centroid of the remaining vertices to $p_r$
  If the score at $p_r$ is the highest score in the simplex Then
    Expand $p_r$ away from the centroid to $p_e$
    Use $p_e$ in place of $p_{min}$
  Else If the score at $p_r$ is still the lowest score Then
    Contract $p_r$ toward the centroid to point $p_c$
    If the score at $p_c$ is no longer the lowest score Then
      Use $p_c$ to replace $p_{min}$
    Else
      Determine the point $p_{max}$ with the highest score
      Shrink all points in the simplex around $p_{max}$
    End If
  Else
    Use $p_r$ in place of $p_{min}$
  End If
End Repeat
```







Admixture Inference

$$\begin{aligned}
\ln\left[P_1\left(Q,\ F\right)\right] = \sum_i^I\sum_j^J\left\{g_{ij}\cdot\ln\left[\sum_k^K q_{ik}\cdot f_{kj}\right]+\right.\\
\left.\left(2-g_{ij}\right)\cdot\ln\left[\sum_k^K q_{ik}\cdot\left(1-f_{kj}\right)\right]\right\}.
\end{aligned}$$

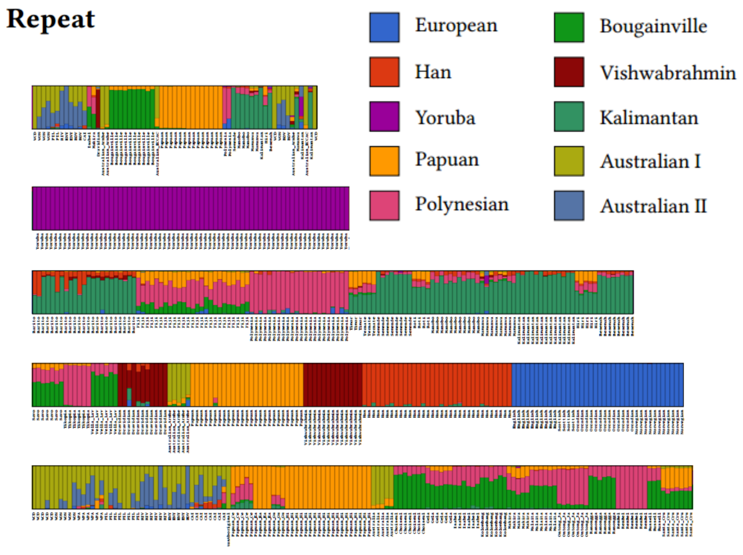
Likelihood Model

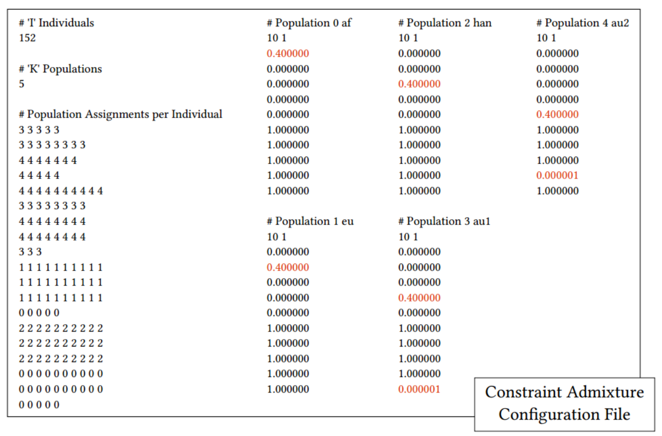
$$\max_{\Delta_{Q_i}}\left\{\frac{1}{2}\Delta_{Q_i}^T H_{Q_i}\Delta_{Q_i} + D_{Q_i}^T\Delta_{Q_i}\right\}$$
$$\text{s.t.}\quad A\Delta_{Q_i} \le a$$
$$B\Delta_{Q_i} = b$$

$$\max_{\Delta_{F_j}}\left\{\frac{1}{2}\Delta_{F_j}^T H_{F_j}\Delta_{F_j} + D_{F_j}^T\Delta_{F_j}\right\}$$
$$\text{s.t.}\quad A\Delta_{F_j} \le a$$

QPAS

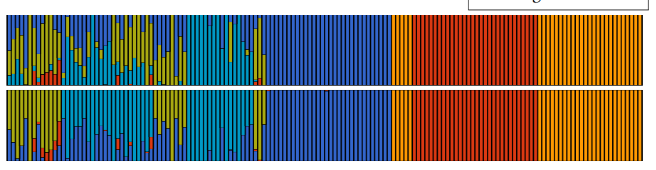
```
Find a feasible starting point
Initialize the corresponding active set
Repeat
  Solve the equality problem defined by the active set
  Compute the Lagrange multipliers of the active set
  If the solved approximation is within the feasible region
    If all Lagrange multipliers are negative
      Return the solved approximation
    Else
      Remove the constraint with the largest Lagrange multiplier
    End If
  Else
    Take the shortest step back into the feasible region
    Insert the corresponding constraint into the active set
  End If
End Repeat
```

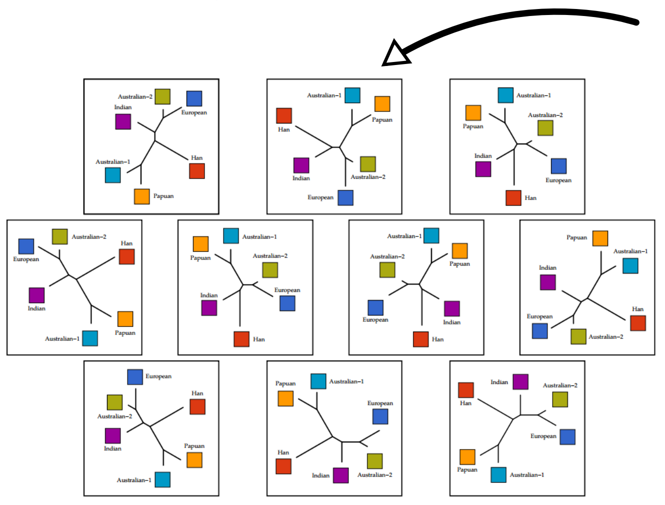




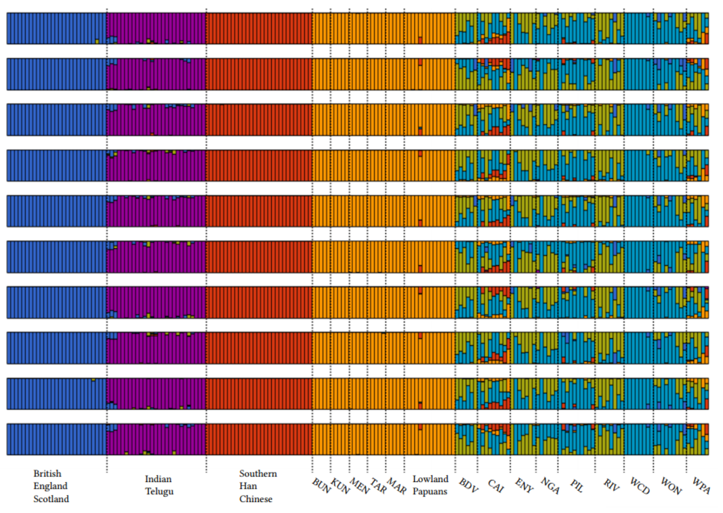
```
# 'I' Individuals                 # Population 0 af   # Population 2 han   # Population 4 au2
152                               10 1                10 1                 10 1
                                  0.400000            0.000000             0.000000
# 'K' Populations                 0.000000            0.000000             0.000000
5                                 0.000000            0.400000             0.000000
                                  0.000000            0.000000             0.000000
# Population Assignments per Individual  0.000000     0.000000             0.400000
3 3 3 3 3                         1.000000            1.000000             1.000000
3 3 3 3 3 3 3 3                   1.000000            1.000000             1.000000
4 4 4 4 4 4 4                     1.000000            1.000000             1.000000
4 4 4 4 4                         1.000000            1.000000             0.000001
4 4 4 4 4 4 4 4 4 4               1.000000            1.000000             1.000000
3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4                   # Population 1 eu   # Population 3 au1
4 4 4 4 4 4 4 4                   10 1                10 1
3 3 3                             0.000000            0.000000
1 1 1 1 1 1 1 1 1 1               0.400000            0.000000
1 1 1 1 1 1 1 1 1 1               0.000000            0.000000
1 1 1 1 1 1 1 1 1 1               0.000000            0.400000
0 0 0 0 0                         0.000000            0.000000
2 2 2 2 2 2 2 2 2 2               1.000000            1.000000
2 2 2 2 2 2 2 2 2 2               1.000000            1.000000
2 2 2 2 2 2 2 2 2 2               1.000000            1.000000
0 0 0 0 0 0 0 0 0 0               1.000000            1.000000
0 0 0 0 0 0 0 0 0 0               1.000000            0.000001
0 0 0 0 0
```

Constraint Admixture Configuration File

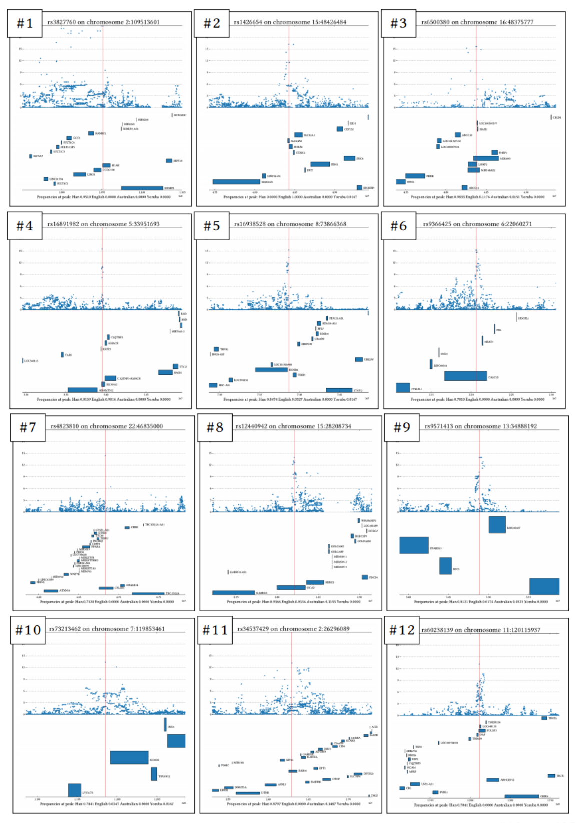




Block Bootstrap

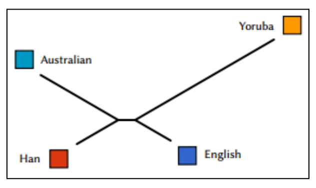


Selection Detection



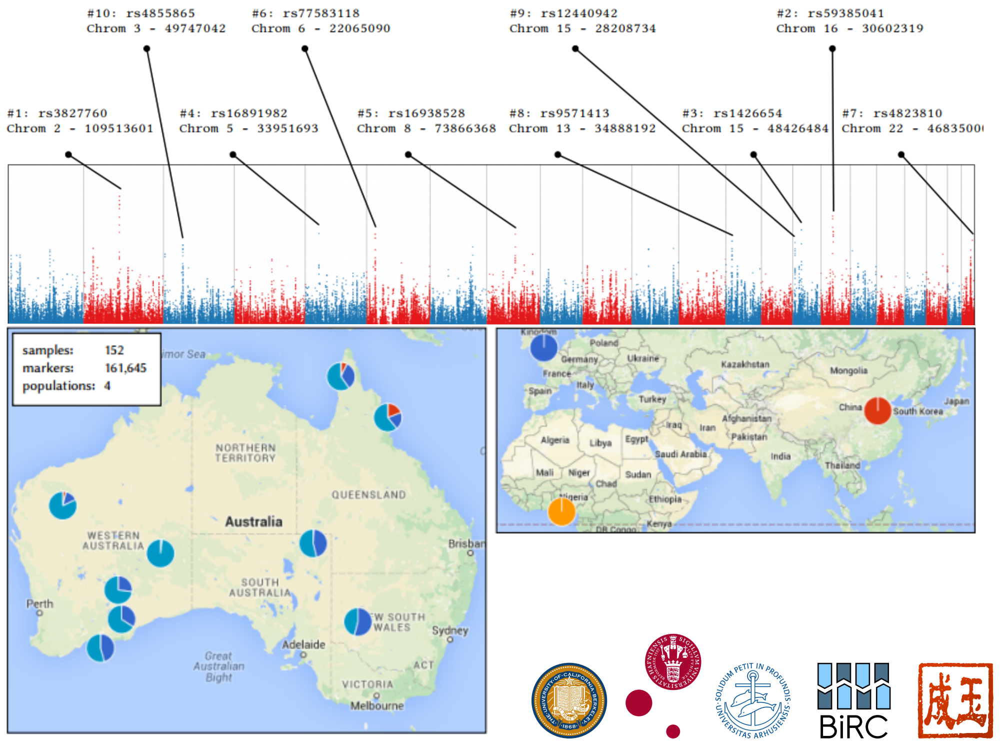
SELECTION

```
Obtain the full genotype dataset G with N markers and M samples
Sample N' markers with respect to LD (N' > 100,000) to form G'
QPAS over ln($P_1$) using G'
  Produce admixture proportions Q' of size M by K
  Produce allele frequencies F' of size K by N'
Nelder-Mead over ln($P_2$) using F'
  Produce variance covariance matrix Ω'
QPAS over ln($P_1$) using G while fixing Q'
  Produce allele frequencies F of size K by N
Repeat for each marker in F
  Set $l_{ratio}$ to zero
  Repeat for each $\alpha$ in a range of an even interval starting from 1.0
    Set $l_{new}$ to ln($P_2$) calculated for this marker using $\alpha$ × Ω'
    Set $l_{old}$ to ln($P_2$) calculated for this marker using Ω'
    If 2 × ($l_{new}$ - $l_{old}$) is greater than $l_{ratio}$
      Set $l_{ratio}$ to 2 × ($l_{new}$ - $l_{old}$)
  End Repeat
  Emit $l_{ratio}$
End Repeat
```



Top Twelve Peaks

1. Hair thickness and curliness
2. Earwax moisture and underarm odor
3. Skin pigmentation
4. Skin pigmentation
5. 'maxDrinks' and alcohol related
6. A melanoma tumor repressor
7. Neural tube defect and hair follicles
8. Skin pigmnetation
9. Intergenic
10. Intergenic
11. Insulin dependent regulation of glucose
12. Taste cells in the mouth


#10: rs4855865
Chrom 3 - 49747042
#6: rs77583118
Chrom 6 - 22065090
#9: rs12440942
Chrom 15 - 28208734
#2: rs59385041
Chrom 16 - 30602319
#1: rs3827760
Chrom 2 - 109513601
#4: rs16891982
Chrom 5 - 33951693
#5: rs16938528
Chrom 8 - 73866368
#8: rs9571413
Chrom 13 - 34888192
#3: rs1426654
Chrom 15 - 48426484
#7: rs4823810
Chrom 22 - 4683500
samples: 152
markers: 161,645
populations: 4
NORTHERN TERRITORY
QUEENSLAND
Australia
WESTERN AUSTRALIA
SOUTH AUSTRALIA
Perth
Adelaide
Sydney
Brisbane
Great Australian Bight
VICTORIA
Melbourne
Poland
Germany
Ukraine
France
Spain
Italy
Kazakhstan
Mongolia
Turkey
China
South Korea
Japan
Iraq
Iran
Afghanistan
Pakistan
India
Thailand
Algeria
Libya
Egypt
Saudi Arabia
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya