6.047 Problem Set 5 Writeup

Matthew Feng

November 16, 2018

1 Positive selection in the human genome

A. Long haplotypes

(a) Recent selection

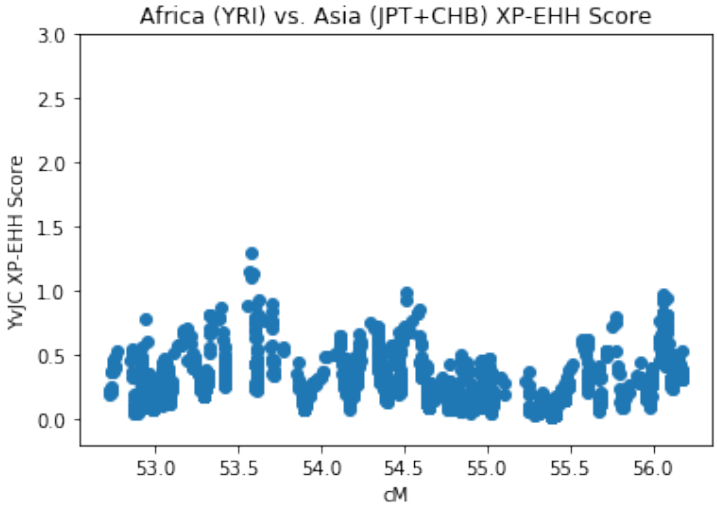
A haplotype block's length gives an indication of how recently that block appeared. Longer blocks are more recent, while shorter blocks are more ancient, since recombination has had time to decay the haplotype block. In other words, an allele may rise to high frequency rapidly enough that long-range association with nearby polymorphisms (i.e. the long-range haplotype) will not have time to be eliminated by recombination.

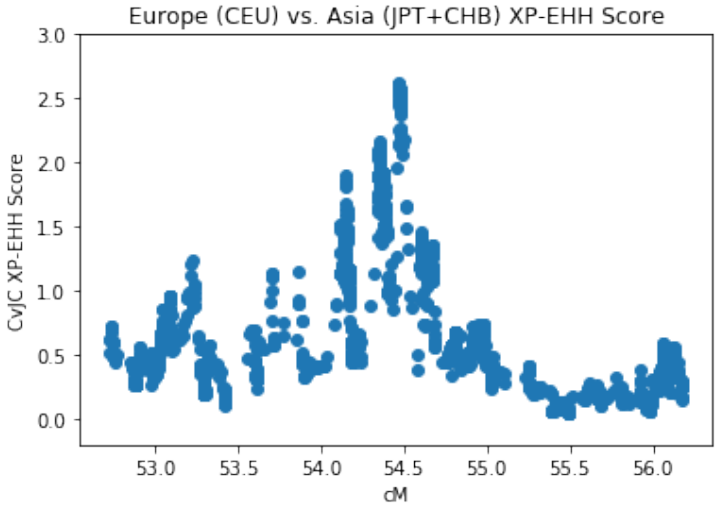
(b) Individual polymorphisms

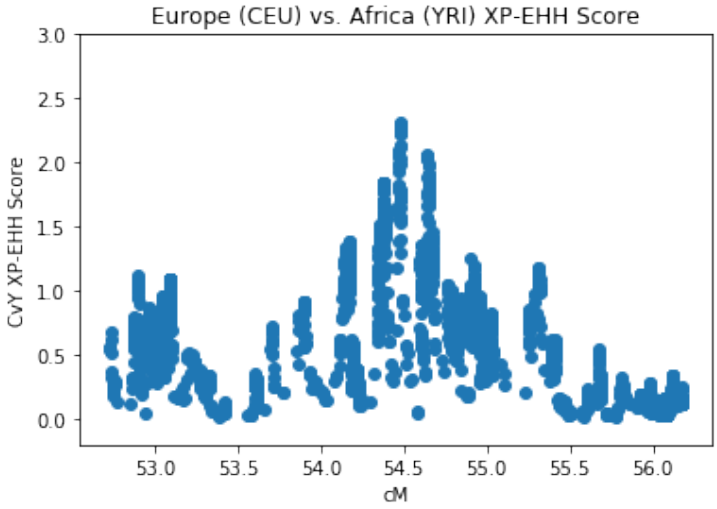
With strong selection, the number of SNPs in the candidate region will be quite large, and thus hard to identify precisely which of those SNPs are the polymorphisms under selection.

(c) XP-EHH Plots

Recombinant frequencies (cM) are more preferable because we are dealing with linkage disequilibria and recombination; thus plotting based on recombination frequencies allows us to have a better scale of long haplotypes.


Africa (YRI) vs. Asia (JPT+CHB) XP-EHH Score
YvJC XP-EHH Score
3.0
2.5
2.0
1.5
1.0
0.5
0.0
53.0
53.5
54.0
54.5
55.0
55.5
56.0
cM


Europe (CEU) vs. Asia (JPT+CHB) XP-EHH Score
CvJC XP-EHH Score
3.0
2.5
2.0
1.5
1.0
0.5
0.0
53.0
53.5
54.0
54.5
55.0
55.5
56.0
cM



(d) Subpopulation

Subpopulation **CEU (Europe)** seems to have the strongest evidence for natural selection, as CvJC and CvY XP-EHH scores both peak above 2.0 around 54.5cM, but that peak does not appear in YvJC.

(e) XP-EHH > 2.0

```
xpehh = pd.read_csv("./XPEHH.txt", sep="\t")
xpehh[(xpehh.iloc[:, 4] > 2.0) |
      (xpehh.iloc[:, 5] > 2.0) |
      (xpehh.iloc[:, 6] > 2.0)]
```

79 SNPs (79 rows returned).

B. Derived allele frequency

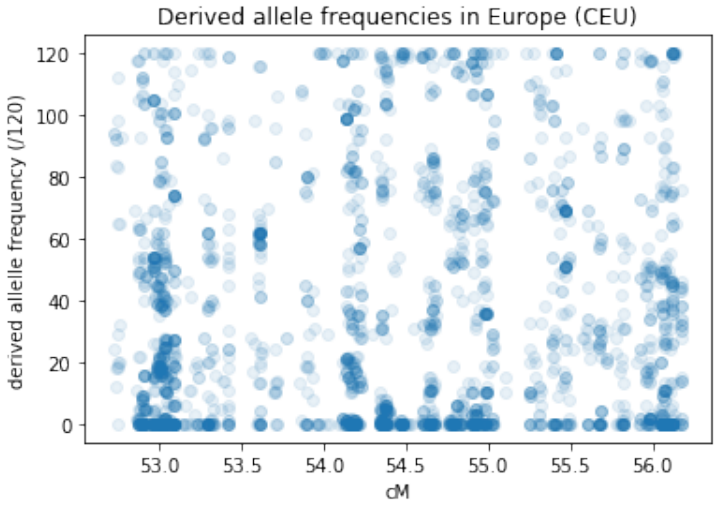
(a) Ancestral alleles

An error can occur when the chimpanzee base has mutated to the same base of the derived human allele, or when the human allele had a fixed mutation previously, and then recently a reversion mutation. The probability of error is then

$$P_e = P_c(1 - P_c)(P_{same}) + (1 - P_c)(P_c)(P_{same}) \approx 0.6\%$$

since $P_c \approx 1.23\%/2$, and P_{same}, the probability that the chimpanzee allele matches the human allele, is 0.5.

(b) Derived allele frequencies



(c) Long haplotype and derived > 0.6

```
derived = pd.read_csv("./Derived.txt", sep="\t")
hi_xpehh = (xpehh.iloc[:, 4] > 2.0) |
           (xpehh.iloc[:, 5] > 2.0) |
           (xpehh.iloc[:, 6] > 2.0)
high_daf = derived.iloc[:, 5] > 0.6 * 120
derived[high_daf & hi_xpehh]
```

22 SNPs.

C. Population differentiation

(a) F_{ST} derivation

$$H_T = 2p(1-p) \tag{1}$$

$$H_S = \frac{1}{2}(2(p+d)(1-p-d) + 2(p-d)(1-p+d)) \tag{2}$$

$$= 2p - 2p^2 - 2d^2 \tag{3}$$

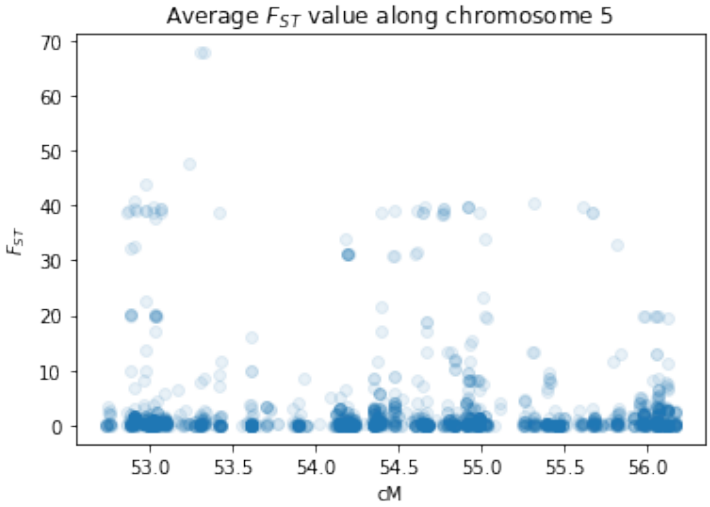
$$\frac{H_T - H_S}{H_T} = \frac{2p - 2p^2 - 2p + 2p^2 + 2d^2}{2p(1-p)} \tag{4}$$

$$= \frac{d^2}{p(1-p)}. \tag{5}$$

Therefore, $F_{ST} = \dfrac{d^2}{p(1-p)}$.

(b) F_{ST} plots

We can estimate p by dividing the number of derived alleles by the number of samples we have from the subpopulations; since we have three subpopulations, we compute two p values: one for Europe and Africa (total $240 = 120 + 120$), and Europe and Asia (total $300 = 120 + 180$). We can estimate d by finding the absolute difference in derived allele frequencies, dividing by 2 (since there is both an addition and a subtraction), and then normalizing over the number of samples (240 and 300, respectively).



(c) Adding $F_{st} > 0.6$

3 SNPs.

D. Function

(a) Conserved elements

```
s = 0
for x in xpehh.iloc[:, 2]:
    s += 1 if (len(cons[(cons.iloc[:, 1] <= x) &
                        (x <= cons.iloc[:, 2])])) > 0 else 0
```

There are 105 SNPs that like within conserved elements.

(b) Final constraint

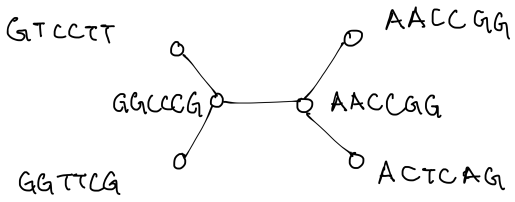
There is only a single SNP that passes the above thresholds and lies within a known or likely functional element.

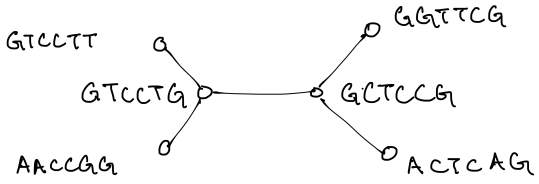
E. SNP

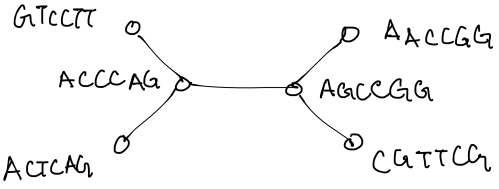
SNP ID **rs16891982**, part of the **SLC45A2** gene which is known to play a role in skin pigmentation.

2 Maximum parsimony phylogeny

A. Tree costs







Cost of tree 1 (top left edge, bottom left, top right, bottom right, middle):

$$\begin{aligned}
&0+2+0+0+1+2\\
+\ &0+0+1+1+0+0\\
+\ &0+0+0+0+0+0\\
+\ &0+2+1+0+1+0\\
+\ &1+1+0+0+2+0\\
=\ &\mathbf{15}
\end{aligned}$$

Cost of tree 2 (top left edge, bottom left, top right, bottom right, middle):

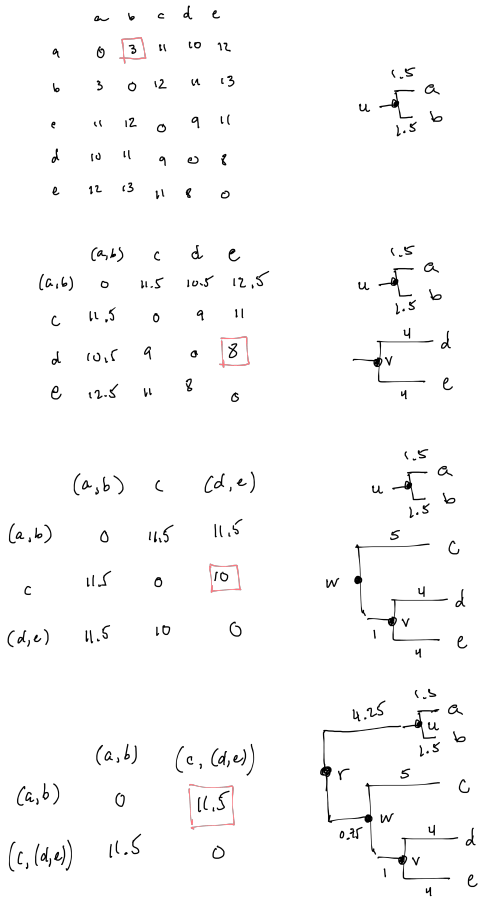
$$\begin{aligned}
&0+0+0+0+0+2\\
+\ &1+2+0+0+2+0\\
+\ &0+2+0+1+0+0\\
+\ &1+0+0+0+2+0\\
+\ &0+1+1+0+1+0\\
=\ &\mathbf{16}
\end{aligned}$$

Cost of tree 3 (top left edge, bottom left, top right, bottom right, middle):

$$\begin{aligned}
&1+1+0+0+2+2\\
+\ &0+0+1+0+0+0\\
+\ &0+1+0+0+0+0\\
+\ &2+0+1+1+2+0\\
+\ &0+2+0+0+1+0\\
=\ &\mathbf{17}
\end{aligned}$$

Tree 1 is the minimum cost tree.

B. UPGMA Clustering



No, it is not possible because the given distance matrix doesn't satisfy the ultrametric constraint, where for any three nodes (a, b, c), $d_{ab} \leq d_{bc} + d_{ac}$.