Population Genetics on X-chromosome

The data consists of a vcf file of 150 male full X chromosomes, a bed file with callable regions, a gif gene annotation file, a metafile with information about the samples and a set of files for use with REHH.

Gene annotation:

Gene annotation (gtf format) for Hg19 can be found in the following websitehttps://www.gencodegenes.org/releases/17.html It was also uploaded to the dropbox as **gencode.v17.annotation.gtf**

Fst Calculation:

You will do the analysis from scratch by reading the genotype file of each population into different tables (remember rows are SNP positions and columns are individuals), the information about the snps are in the .snp file (ancestral and derived alleles). The data is haploid (n) therefore calculating Fst consists of estimating the allele frequencies for each position and calculating the expected heterozygosity within population Hs and contrasting Expected Heterozygosity across populations Ht.

Fst = (Ht - Hs)/Ht.

This can be done by averaging Fst values for a set of consecutive markers in a given window size (100 SNPs).

All my code for this project can be found on the github repo:
https://github.com/cmkobel/population-genomics-X-chromosome

Investigate the following

a) *Perform an Fst scan between sets of populations in a sliding window of 100 SNP positions, including at least the contrast between Africa and Europe, between Europe and East Asia, and between East Asia and Africa. Identify the 10 strongest Fst outlier regions in each case. Identify their genomic position and the genes covered by these Fst peaks. Discuss potential adaptive explanations.*

Parts of exercise *b)* were completed in order to extract the allele frequencies from the SNP files from the different regions.

Fst in defined as Fst = 1 – (Hes / Het) where Hes and Het is the expected frequency of heterozygotes when two populations are considered either as two subpopulations (Hes) or as one total (Het). It measures the lack of heterozygotes in the subpopulations in relation to the total population. Fst was calculated for each individual SNP in the following region combinations:

Africa – Westeurasia
Westeurasia – Eastasia
Eastasia – Africa

The regions in this exercise are considered to be separate populations though they may not be practically speaking.

A rolling window of 100 SNPs was then applied on each region in order to get a moving average.

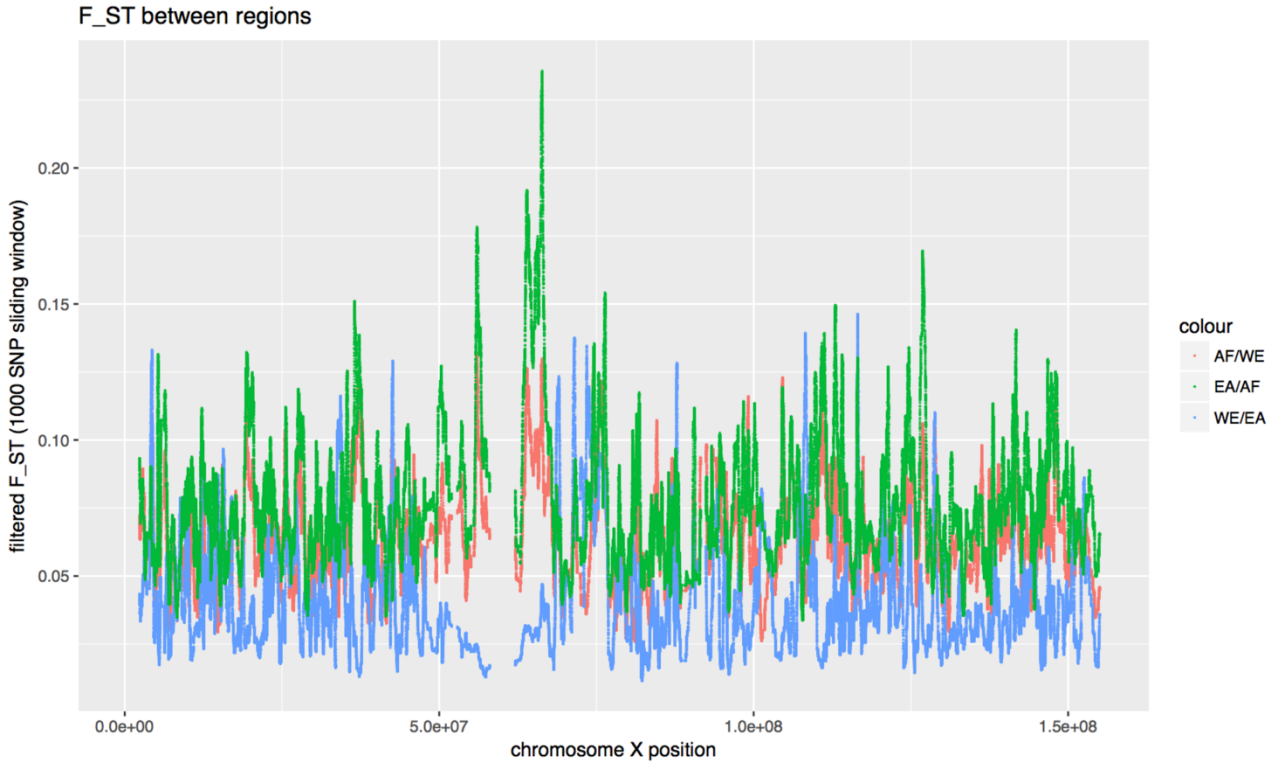

Figure 1: Fst between two populations at a time. The size of the sliding window is 1K SNPs in order to get at slowly moving average.

Because the Fst is varying a lot throughout the chromosome, the plotting was done with data from a bigger sliding window.
As values of high Fst show differentiation between the two subpopulations, we can investigate the

b) *Perform an iHS scan of the whole X chromosome for at least three populations. Identify the 10 most significant regions and associated with genes as in A.*

c) *Perform an XP-EHH scan of the whole X chromosome for at least three populations. Identify the 10 most significant regions and associated with genes as in A.*

d) *Intersect the analysis of Fst and XP-EHH*

e) *Perform any additional analysis of your own choice, such as (diversity along the C X chromosome)*