Population Genomics on The X Chromosome

201404379 (Carl M. Kobel)

kobel@pm.me

May 22th, 2018

Abstract

In order to get a hands on experience with some of the most used population genomics statistics, we processed SNP data from the Simons Genome Diversity Project with the tests: Fst, iHS, XPEHH and LD. In order to limit the bounds, we constrained the study to the X chromosome only.
For each of the population genomics-related tests, we found the top ten genes, and compared them across the populations as a means to seek for genes which differential selection. Most of the genes found by the statistics were found to be related to gene regulation, and some were pseudogenes.

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

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*

In order to calculate Fst, we need the allele frequencies for each SNP. We obtained those frequencies from the SNP files. Fst in defined as Fst = 1 – (Hes / Het) where Hes and Het are 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, it can be assumed that they should resemble the biggest amount of pairwise divergence, because these regions have been populated for a relatively long time. 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 applied on each population combination in order to get a moving average.

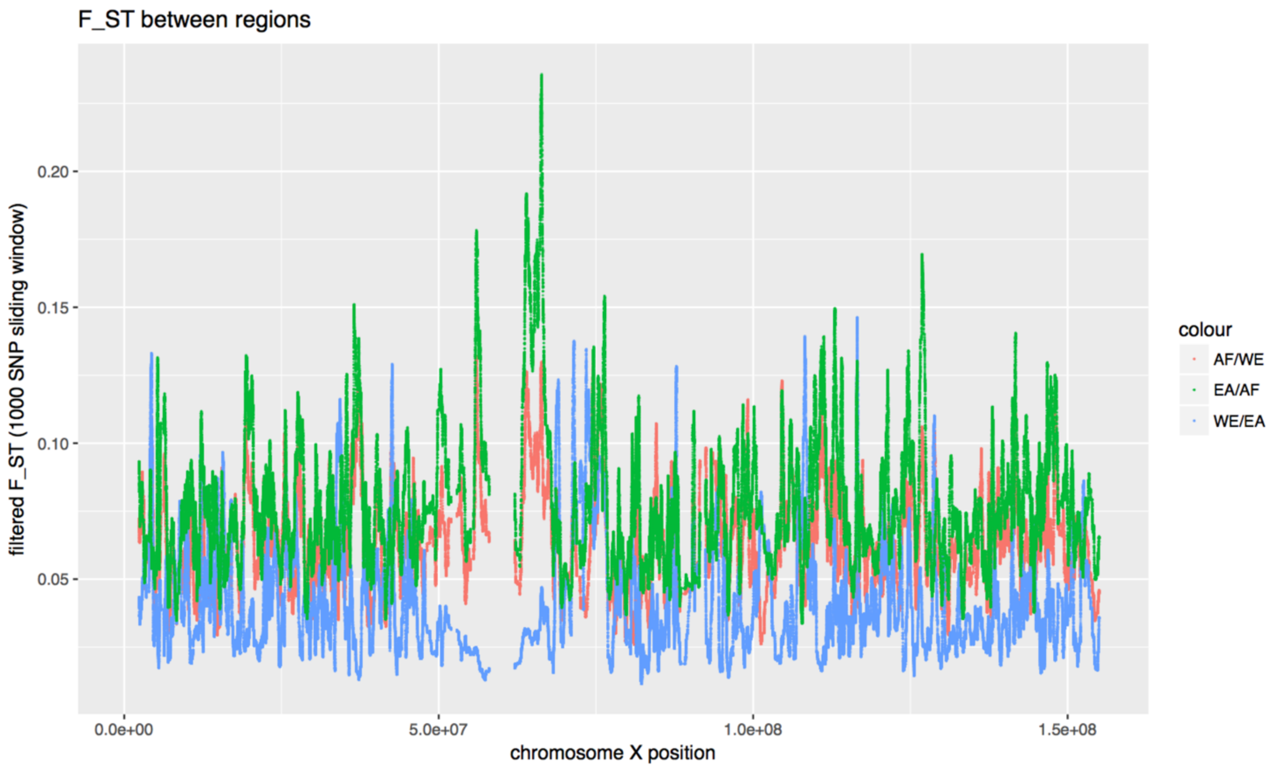

Figure 1: Fst between two populations at a time. The size of the sliding window is 1K SNPs in order to get at slowly moving average.

Because the Fst varies a lot throughout the chromosome, the plotting was done with data from a bigger sliding window. The biggest peaks are around the centromere, we don't know why.
As values of high Fst show differentiation between the two subpopulations, we can investigate these peak-regions and maybe find genes that correlate with the way populations have diverged. By overlapping the SNPs with a genome annotation, we selected a threshold for each population, such that the ten genes containing the highest Fst would stand out.

Results from Fst tests

Table 1: 10 Genes with an Fst above the 99,8% percentile with the populations Africa and Westeurasia. The Fst used here is the mean of 100 adjacent SNP Fsts. Note that 'transcript_type' denotes the transcript of the specific Fst peak, and not the gene region as a whole.

gene_name	position	fst_peak	transcript_type
IL1RAPL2	104558536	0,233493218	protein_coding
SYTL5	37889821	0,219400423	protein_coding
TM4SF2	37889821	0,219400423	protein_coding
RP13-188A5.1	55982443	0,210125629	processed_transcript
UPRT	74513664	0,198155786	protein_coding
PRRG1	37300948	0,197715863	protein_coding
RP11-357K9.2	37300948	0,197715863	pseudogene
DMD	32713394	0,196936313	protein_coding
RP11-54I5.1	55988008	0,195909956	pseudogene
OCRL	128707674	0,192535606	protein_coding

Table 2: Genes with an Fst above the 99,89% percentile with the populations Westeurasia and Eastasia. The Fst used here is the mean of 100 adjacent SNP Fsts. Note that 'transcript_type' denotes the transcript of the specific Fst peak, and not the gene region as a whole.

gene_name	**position**	**fst_peak**	**transcript_type**
FTX	73308940	0,303662952	lincRNA
RP11-262D11.2	71377253	0,286265802	pseudogene
PIN4	71472853	0,269973204	protein_coding
NHSL2	71352521	0,250927669	protein_coding
RP11-262D11.1	71352521	0,250927669	pseudogene
RPS4X	71476289	0,246681874	protein_coding
RGAG4	71349753	0,246583034	protein_coding
TMEM164	109370144	0,246502551	protein_coding
BX119917.1	71372190	0,233755654	miRNA
UHRF2P1	73325745	0,228391375	pseudogene

Table 3: Genes with an Fst above 99,3% percentile with the populations Eastasia and Africa. The Fst used here is the mean of 100 adjacent SNP Fsts. Note that 'transcript_type' denotes the transcript of the specific Fst peak, and not the gene region as a whole.

gene_name	**position**	**fst_peak**	**transcript_type**
CTD-2076M15.1	126794656	0,304382696	lincRNA
RP13-188A5.1	55982877	0,293978003	processed_transcript
RP11-54I5.1	55988008	0,283576529	pseudogene
RP5-964N17.1	112907371	0,260622912	lincRNA
KRT8P27	63843830	0,25835841	pseudogene
CHRDL1	109929651	0,245930567	protein_coding
DCAF8L2	27620805	0,240543042	protein_coding
YWHAZP7	63832930	0,236082919	pseudogene
HEPH	65473327	0,22653709	protein_coding
HTR2C	114083321	0,226096665	protein_coding

By looking up the gene names on NCBI gene, we looked through the summary of each gene. No obvious adaptation genes appeared in this analysis. Many genes are associated with genes related to gene regulation and/or retinis pigmentosa. This might be because the X chromosome has many genes related to the development of sex, and that many of the genes related to population divergence might be on other chromosomes. This might be pure speculation though.

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.*

iHS (integrated haplotype score) is a test statistic developed by Voight et al. (Voight et al. 2006). iHS can be used to test for positive selection (where the frequency of an allele increases or decreases monotonically). It is based on the EHH statistic which is developed by Sabeti et al. (Sabeti et al. 2007). EHH indicates the decay of linkedness (or identity) away from each SNP. Because selection happens on regions covering several SNPs, we can use the SNP data to infer how much selection has occoured on each SNP position. Because the area under the curve of EHH is more highly correlated with selection than is EHH by itself, we integrate EHH, and get the iHH (integrated Haplotype Homozygosity). This iHH can be computed either with respect to the ancestral or derived allele, and when we take the log-ratio between them we get the unstandardized iHS. This undstandardized iHS is then standardized with the variance, and we finally get iHS, which can be used to score the amount of selection in each SNP.

The same three populations as in exercise *a)* were selected for this iHS exercise.

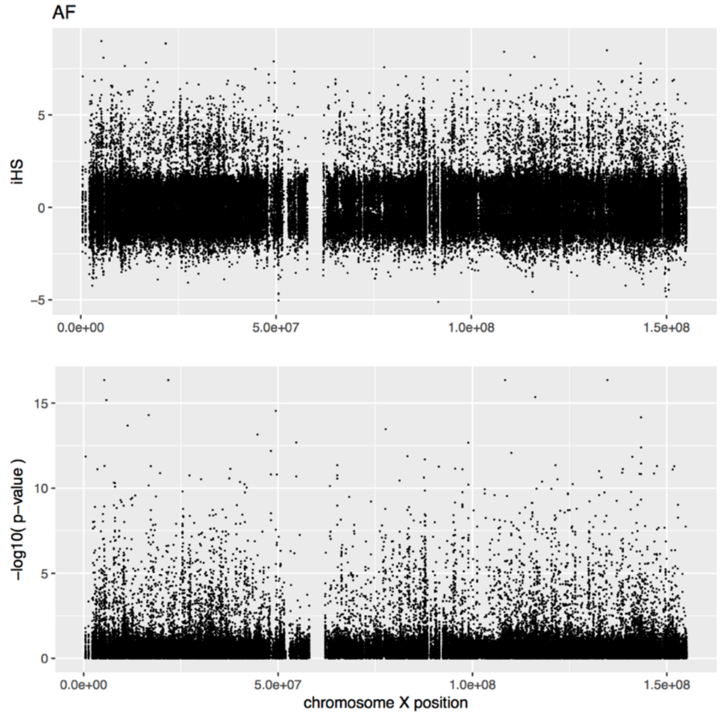

Figure 2: iHS and associated p-values for the Africa population

Table 4: The 10 genes with the highest iHS in the Africa population

gene_name	position	iHS	ppval	gene_type
DDX26B	1,35E+08	8,494167	16,35253	protein_coding
SMPX	21753631	8,866302	16,35253	protein_coding
NLGN4X	5809792	8,096836	15,17644	protein_coding
CTPS2	16700579	7,82612	14,29183	protein_coding
ARHGAP6	11287745	7,644587	13,68043	protein_coding
XRCC6P5	98924229	7,340968	12,67356	pseudogene
CHRDL1	1,1E+08	7,154159	12,07481	protein_coding
PCDH11X	91347968	6,894997	11,26869	protein_coding
DYNLT3	37698495	6,850401	11,13287	protein_coding
TM4SF2	37698495	6,850401	11,13287	protein_coding

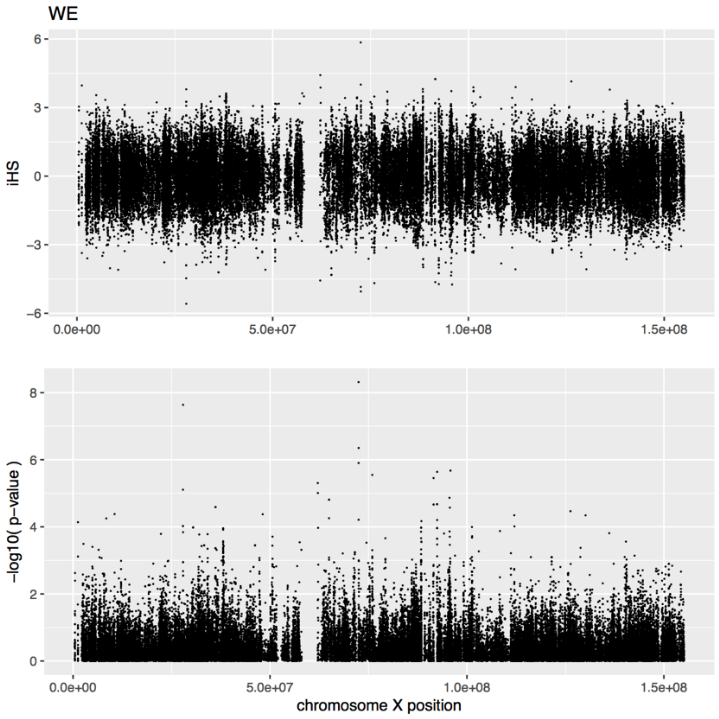

Figure 3: iHS and associated p-values for the Westeurasia population

Table 5: The 10 genes with the highest iHS in the Westeurasia population

gene_name	position	iHS	ppval	gene_type
PCDH11X	91522531	4,24694	4,664121	protein_coding
MID1	10543062	-4,09752	4,37924	protein_coding
AMOT	1,12E+08	3,89799	4,343454	protein_coding
RP1-23K20.2	1,3E+08	-4,07714	4,341094	antisense
GPR101	1,36E+08	3,78353	3,810735	protein_coding
DMD	32629538	-3,55711	3,770822	protein_coding
CD99	2648871	-3,59655	3,49151	protein_coding
OTC	38257658	-3,55653	3,425053	protein_coding
TM4SF2	38105192	3,618469	3,382885	protein_coding
CHDC2	36153918	-3,51765	3,361119	protein_coding

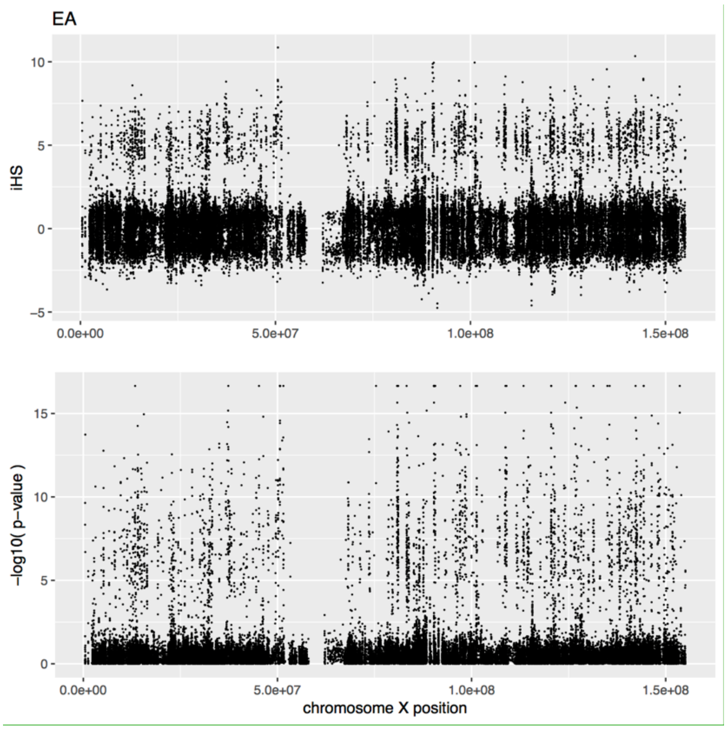

Figure 4: iHS and associated p-values for the East asia population

Table 6: The 10 genes with the highest iHS in the Eastasia population

gene_name	position	iHS	ppval	gene_type
BEX5	1,01E+08	8,533277	16,65356	protein_coding
GS1-600G8.3	13334199	8,583642	16,65356	antisense
RP1-192P9.1	45240847	8,298389	16,65356	pseudogene
RP5-842K24.2	1,31E+08	8,93403	16,65356	antisense
RPL10	1,54E+08	8,512827	16,65356	protein_coding
SAGE1	1,35E+08	9,550098	16,65356	protein_coding
SPANXN1	1,44E+08	8,982066	16,65356	protein_coding
PRRG1	37296759	8,0713	15,17644	protein_coding
TM4SF2	37344835	8,80894	15,17644	protein_coding
GUCY2F	1,09E+08	8,05655	15,0515	protein_coding
RPS6KA6	83428893	8,044866	15,0515	protein_coding

Ppvalues for Westeurasia are all very low

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.

XP-EHH (Cross Population Extended Haplotype Homozygosity) is developed by Sabeti et al. (Sabeti et al. 2007). It detects selective sweeps between two populations where the allele is fixed in one and polymorphic in the other. It is based on the EHH test, which measures the decay of identity from a SNP. XP-EHH compares the length of these EHH regions between populations. A positive value of XP-EHH suggests selection in the first population, and a negative in the second population.

Table 7: The ten genes with the highest xpehh values form the comparison of populations Africa and Westeurasia

gene_name	position	xpehh	ppval	transcript_type
DHRSX	2180731	-7,42314	12,94175	protein_coding
RP11-104D21.3	50909542	-5,43121	7,252011	lincRNA
PCDH11X	91523596	-5,07198	6,40484	protein_coding
ALAS2	55038885	-5,0175	6,281109	protein_coding
KAL1	8680990	-4,8634	5,937848	protein_coding
ATP6AP1	1,54E+08	-4,65318	5,48564	protein_coding
CCNB1IP1P3	94032285	-4,62797	5,432668	pseudogene
APEX2	55034419	-4,4867	5,140661	protein_coding
PRKX	3532267	-4,44306	5,052127	protein_coding
GDI1	1,54E+08	-4,23953	4,649781	protein_coding

Table 8: The ten genes with the highest xpehh values form the comparison of populations Westeurasia and Eastasia

gene_name	position	xpehh	ppval	transcript_type
PCDH11X	91517416	5,341297	7,034873405	protein_coding
FOXN3P1	101802186	-5,07722	6,416815261	pseudogene
PASD1	150837984	-4,87148	5,955593461	protein_coding
RP11-45D17.1	150837984	-4,87148	5,955593461	pseudogene
NXF4	101809168	-4,65913	5,498187928	processed_transcript
GABRA3	151369634	4,487602	5,142490811	protein_coding
SMARCA1	128611215	4,280981	4,730321307	protein_coding
RP11-258C19.5	53191756	4,192733	4,559699127	lincRNA
RP11-40F8.2	22850109	-4,14499	4,468730689	processed_transcript
MTMR1	149932703	-4,06234	4,31351351	protein_coding

Table 9: The ten genes with the highest xpehh values form the comparison of populations Eastasia and Africa

gene_name	position	xpehh	ppval	transcript_type
DHRSX	2180731	7,817278	14,27334853	protein_coding
ATP6AP1	153664698	5,344227	7,041897407	protein_coding
GDI1	153666533	5,022854	6,293204142	protein_coding
KAL1	8680990	4,691439	5,566565913	protein_coding
ALAS2	55038885	4,517497	5,203600027	protein_coding
RP11-104D21.3	50909542	4,147207	4,472942981	lincRNA
PCDH11X	91523596	3,760786	3,771136357	protein_coding
MAGED1	51600709	3,661409	3,600617079	protein_coding

RP11-22B10.3	51600709	3,661409	3,600617079	pseudogene
DMD	33113844	3,616343	3,524628171	protein_coding

DHRSX shows up both in the analysis on Africa-Westeurasia and Eastasia-Africa. According to the NCBI Gene database, it is a dehydrogenase present in most tissues. It does not have any obvious relation to population differentiation. Paralogs of RP11 show up in all the populations. ALAS2 shows up in Africa-Westeurasia and Eastasia-Africa, it is an enzyme located in the mitochondria that and is a part of the heme pathway. KAL1 shows up in the Africa-related analyses and is related the development of puberty.

It doesn’t appear to us that any of genes that come out of the XPEHH analysis has any relation to the divergence between the populations.

d) Intersect the analysis of Fst and XP-EHH

Fst can be used to indicate the divergence between two populations as it measures the relative difference in the frequency of heterozygotes. If there have been a selective sweep in the population, the frequency of heterozygotes might be significantly changed in one of the populations.

XPEHH on the other hand is similar to the Rsb statistic, but instead of using the Tang et al. iES, it uses the iES defined by Sabeti et al.(Sabeti et al. 2007). It detects selective sweeps where a SNP is fixed in one population, and polymorphic in the total population.

Fst is discards fixed SNP (as the frequency of heterozygosity is not possible to calculate from fixed SNP data.

We compared the top ten peaks from the Fst test, and XPEHH test. In Africa vs. Westeurasia there wasn't any union between the results, other than paralogs of the RP11 gene (often as pseudogenes) kept on showing up across the two tests. RP11 is a pre-mRNA processing factor that is a part of the spliceosome complex. The coding paralogs are associated with retinitis pigmentosa. It doesn't appear to have any connection with adaptation.

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

We decided to calculate the linkage disequilibrium (LD) throughout the X chromosome. LD can be calculated in different ways. The method we implemented here consists of calculating the pairwise correlation of a window of (in this case) 100 SNPs. The mean was calculated from each of these 100 SNP windows, and plotted against the median chromosomal positition (see figure Figure 5: Linkage disequilibrium throughout the X chromosome from the African population.)

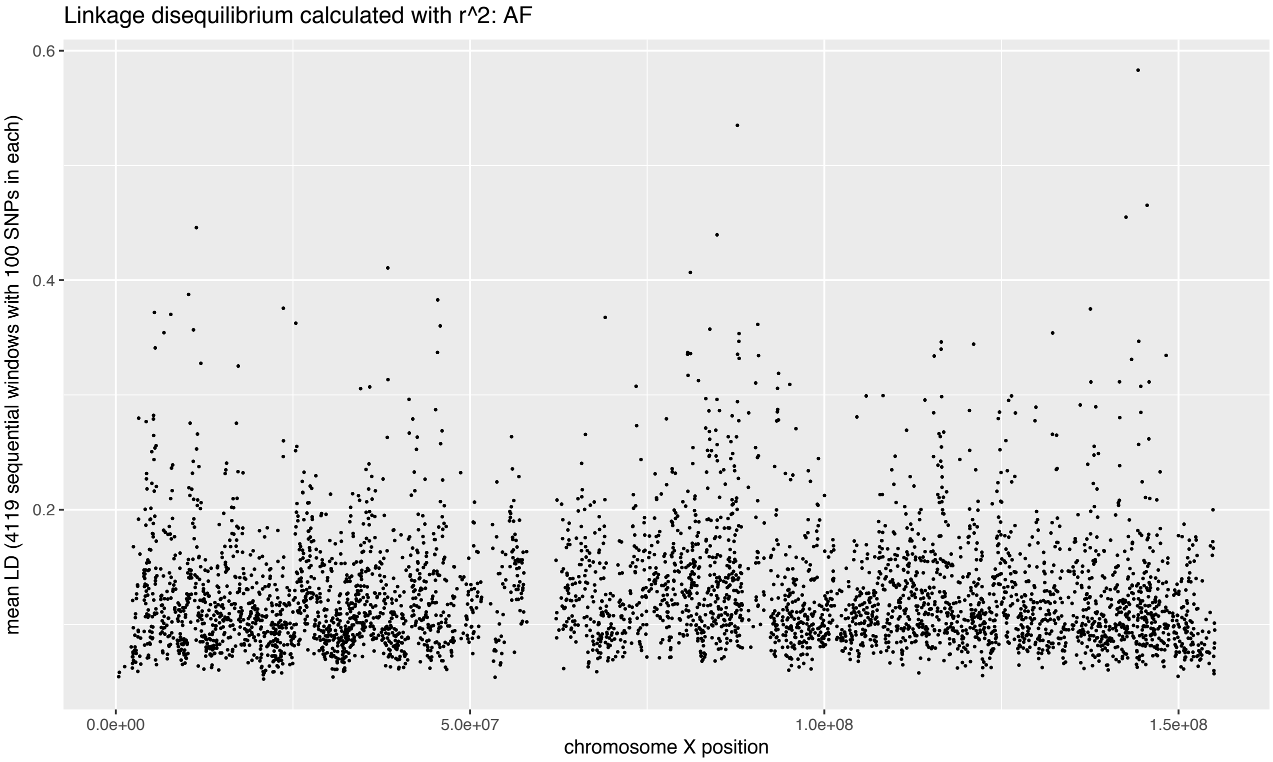

Figure 5: Linkage disequilibrium throughout the X chromosome from the African population.

We have done linkage disequilibrium calculations on many of the populations. Unfortunately, we haven’t had the time to compare them with the other tests. Our speculation is, that the absolute slope of iHS will be correlated with LD, as the iHS cannot change in a region where the correlation between positions (LD) is high.