**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 website<https://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**

1. *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

These regions were chosen because it can be assumed that they should resemble the biggest amount of pairwise divergence. 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.

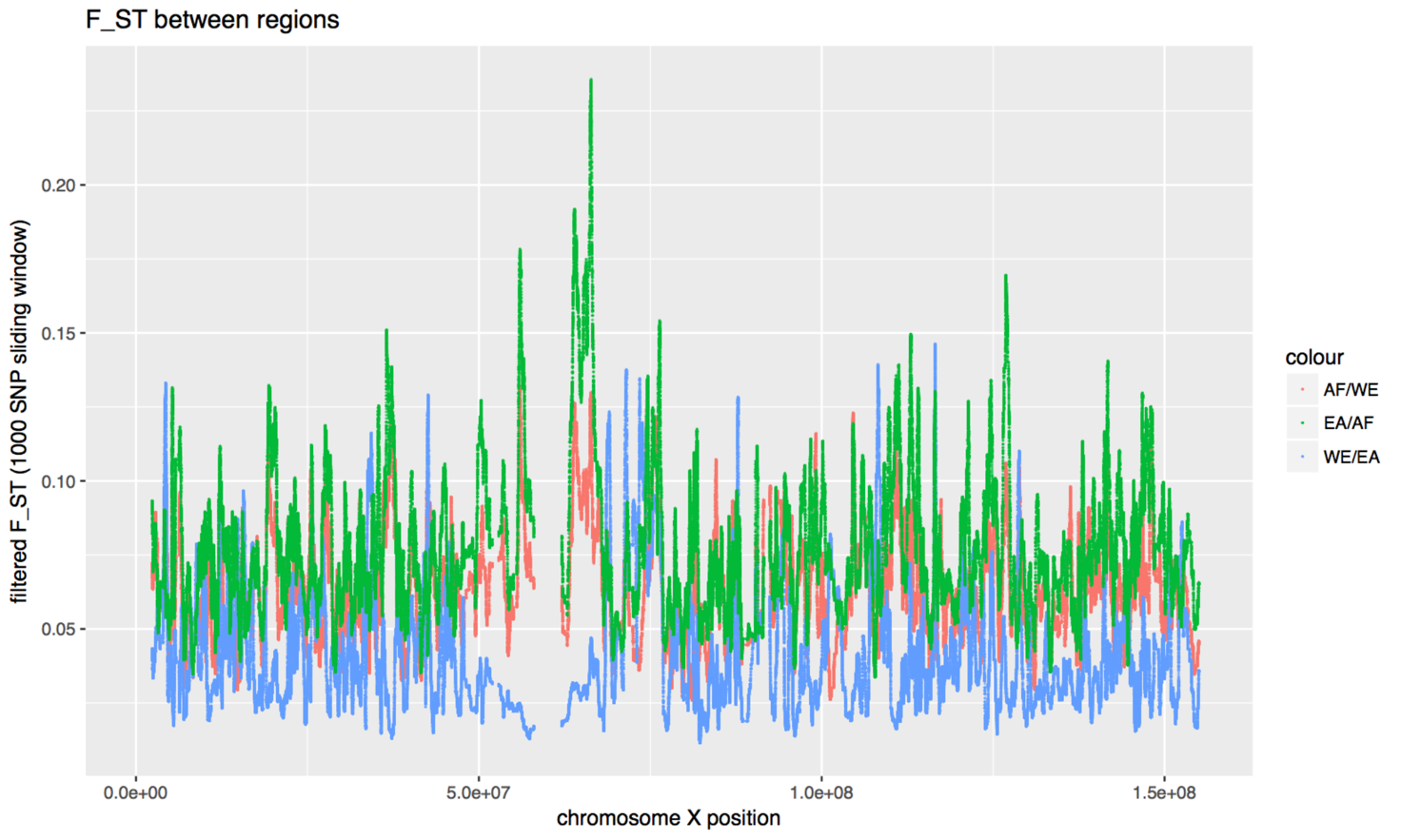


Figure 1: Fst between two populations at a time. The size of the sliding window is 1K SNPs in order to get at slowly moving average.

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??.

As values of high Fst show differentiation between the two subpopulations, we can investigate these regions and maybe find genes that correlate with the way people of differing populations have diverged. This was done by sorting the Fst values in each population combination, and overlapping them with gencode v17 for Hg19.

Africa - Westeurasia

I order to get the 10 highest peaks, only the Fst values above the 0.998th percentile were considered. With gencode v17, the following results were obtained:

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 |

No obvious adaptation genes appeared in this analysis. Many genes are associated with genes related to gene regulatio and 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. iHS will show the validity of this approach.

1. *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.*

The same three populations as in exercise *a)* were selected for this iHS exercise, as it makes a comparison of the methods possible.

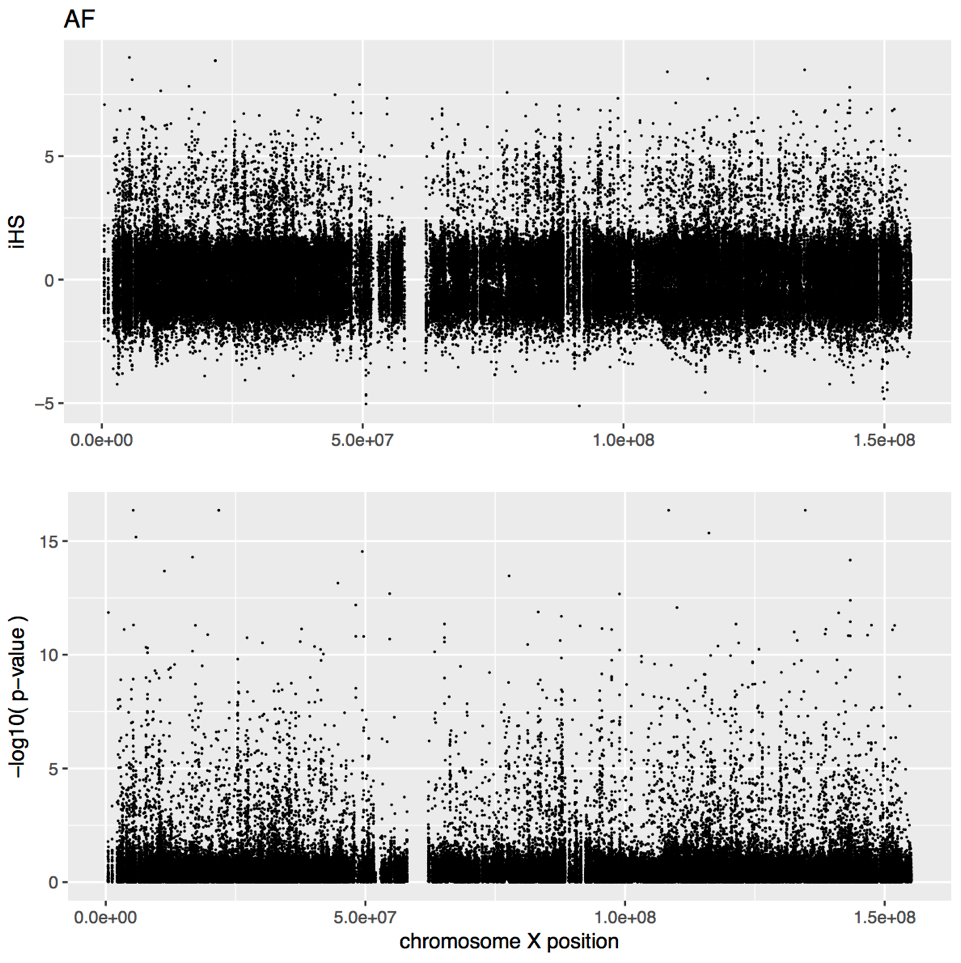


Figure 2: iHS and associated p-values for the Africa population

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 |

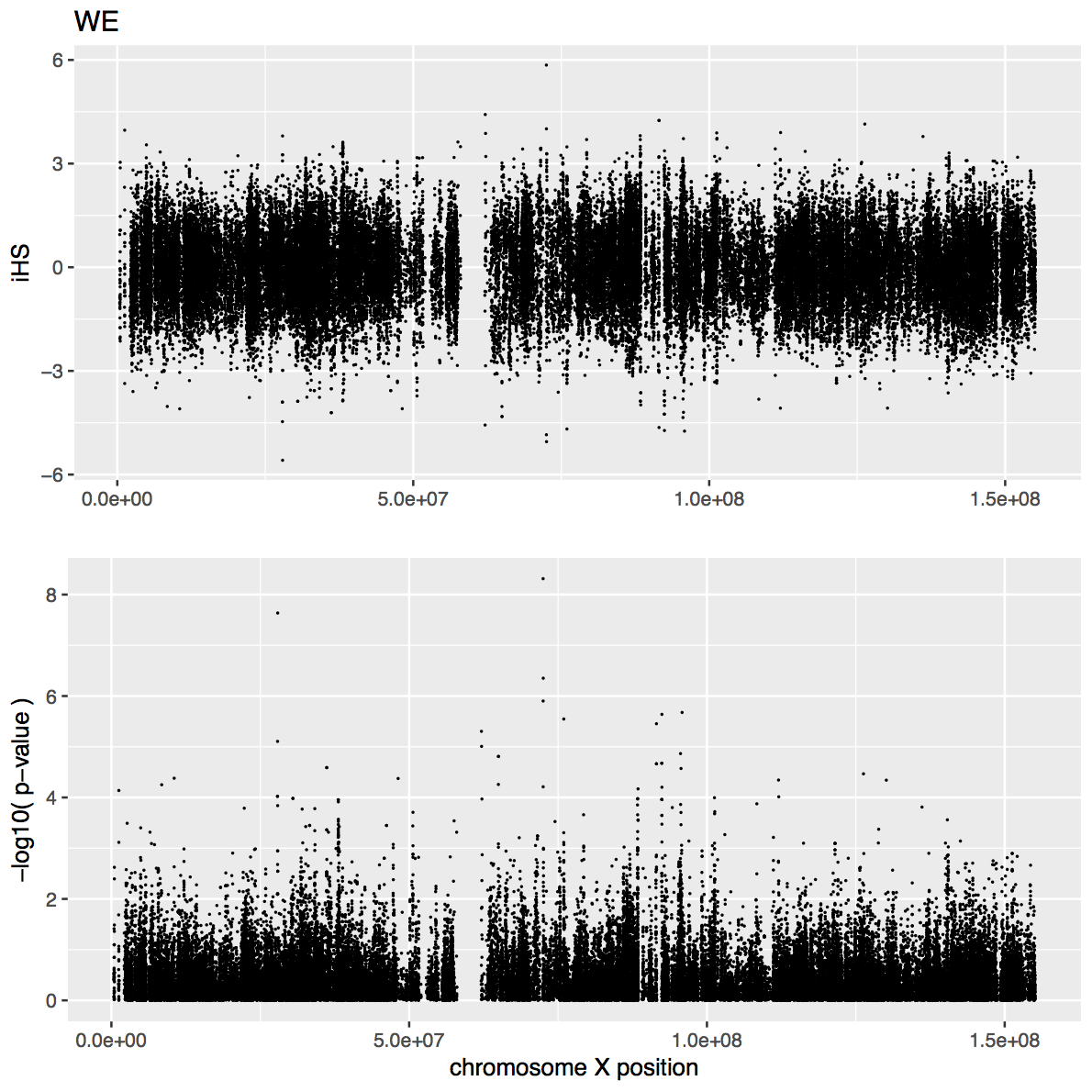


Figure 3: iHS and associated p-values for the Westeurasia population

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 |

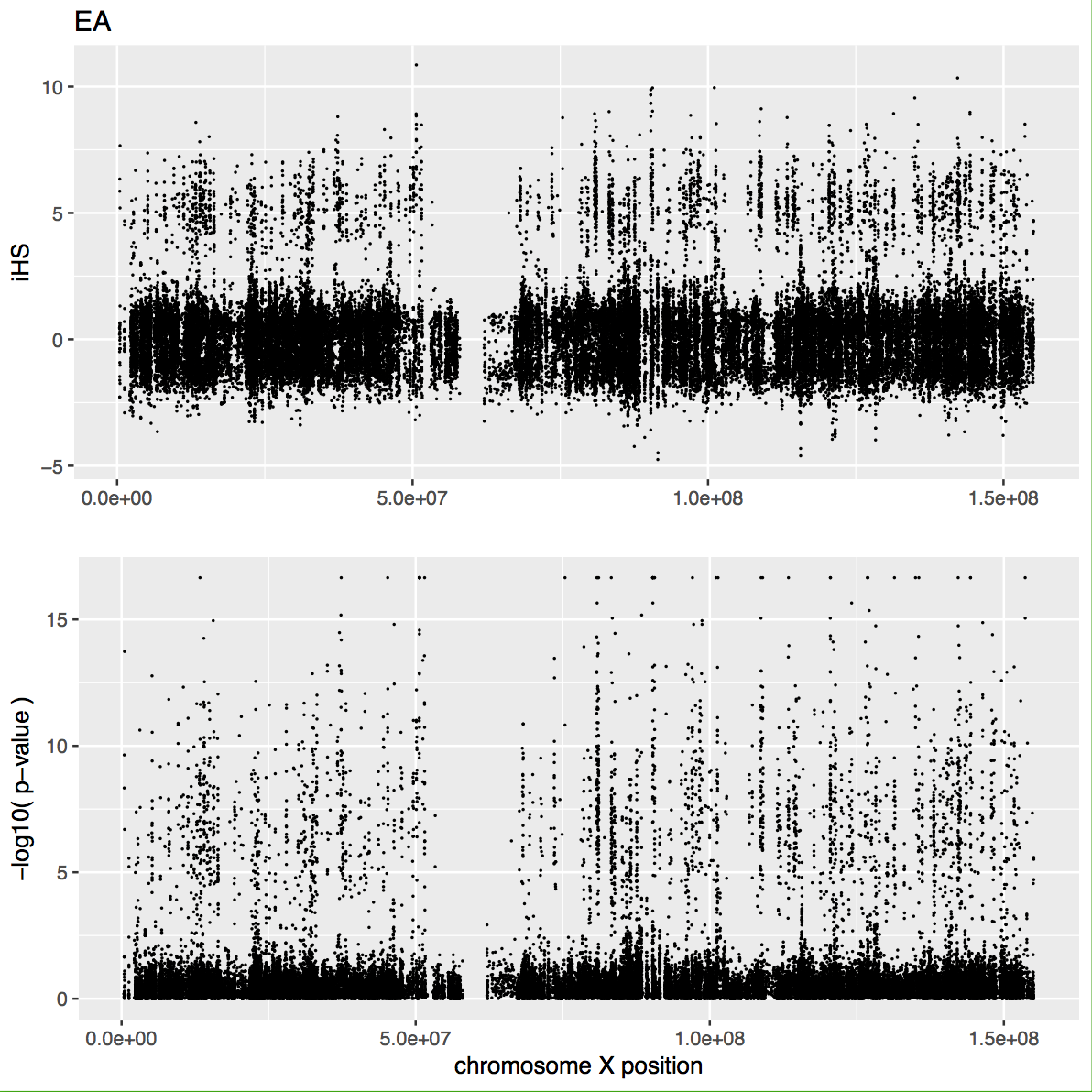


Figure 4: iHS and associated p-values for the East asia population

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

1. *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.*
2. *Intersect the analysis of Fst and XP-EHH*

Make tables that show union of the two methods

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

Document LD ??

Put in some plots