

Analysis of the Exome Aggregation Data Set

Emilia Wieczorek
Lior Pachter
Math 127
UC Berkeley

December 15, 2014

Contents

1	Introduction	3
1.1	The ExAC data set	3
1.2	My interest in this particular data set	3
2	Background	4
2.1	How does the data look like?	4
3	Methods	6
3.1	Things I wanted to look at	6
3.2	Challenges	6
4	Results	6
5	Conclusions	10

1 Introduction

1.1 The ExAC data set

The ExAC data set was released in October 2014 by the Exome Aggregation Consortium in the form of a browser and a raw data file available for download. "The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community. The data set provided spans 61486 unrelated individuals sequenced as part of various disease-specific and population genetic studies" [1].

For my analysis, I used release 0.1 that contains 63352 individuals. The individuals in the data set come from different populations. The data set reports minor allele count for each population at Single Nucleotide Polymorphism (SNP) sites. The data focuses only on the exomes, i.e. regions that code for proteins, and is formatted as a vcf file.

1.2 My interest in this particular data set

My main motivation was to gain more experience parsing large data files, trying to visualize large data sets, and drawing meaningful conclusions. I wanted my project to include a significant amount of programming. I have only taken one computer science class, but learning how to program is very important to me, so I learn through attempting projects like this one.

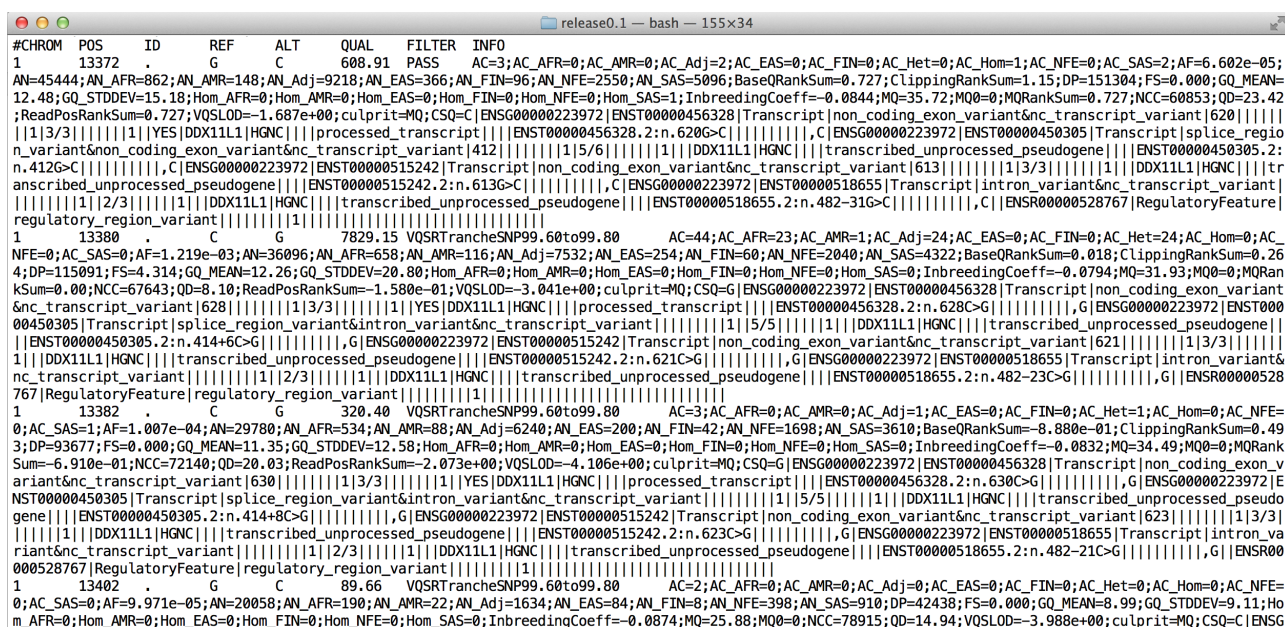
The fact that this data set has just been released, and not so many people have looked at it yet was very appealing to me.

Also, I am interested in the Genome-Wide Association Studies (GWAS). The Consortium writes, "we have removed individuals with severe pediatric diseases, making this (we believe) a reasonable comparison data set for childhood-onset Mendelian diseases" [2]. It would be interesting to me to explore ways in which I could use this data set to identify SNPs associated with common diseases. I am currently working as a student research assistant at UCSF on a GWAS study involving Multiple Sclerosis (using WTCCC2 and 23andMe data sets).

2 Background

2.1 How does the data look like?

The data set is formatted as a vcf file. First, I had to download it from <http://exac.broadinstitute.org/downloads>. I was hoping that I would be able to use the VCFtools [3] to parse the file. However, even though it is a vcf, it is slightly different than the 1000 Genomes Project vcf format that the VCFtools are geared towards. I tried parsing it with the VCFtools, but ended up getting various errors. This forced me to write my own script. In order to do so, I first had to take a look at how the data looks like.



```

#CHROM POS ID REF ALT QUAL FILTER INFO
1 13372 . G C 608.91 PASS AC=3;AC_AFR=0;AC_AMR=0;AC_Adj=2;AC_EAS=0;AC_FIN=0;AC_Het=0;AC_Hom=1;AC_NFE=0;AC_SAS=2;AF=6.602e-05;
AN=45444;AN_AFR=862;AN_AMR=148;AN_Adj=9218;AN_EAS=366;AN_FIN=96;AN_NFE=2550;AN_SAS=5096;BaseQRankSum=0.727;ClippingRankSum=1.15;DP=151304;FS=0.000;GQ_MEAN=
12.48;GQ_STDDEV=15.18;Hom_AFR=0;Hom_AMR=0;Hom_EAS=0;Hom_FIN=0;Hom_NFE=0;Hom_SAS=1;InbreedingCoeff=-0.0844;MQ=35.72;MQ0=0;MQRankSum=0.727;NCC=60853;QD=23.42
;ReadPosRankSum=0.727;VQSLOD=-1.687e+00;culprit=MQ;CSQ=C|ENSG00000223972|ENST00000456328|Transcript|non_coding_exon_variant&nc_transcript_variant|620|
|||||1|3/3|||||1|YES|DDX11L1|HGNC||||processed_transcript||||ENST00000456328.2:n.620G>C|||||,C|ENSG00000223972|ENST00000450305|Transcript|splice_regio
n_variant&non_coding_exon_variant&nc_transcript_variant|412|
|||||1|5/6|||||1|DDX11L1|HGNC||||transcribed_unprocessed_pseudogene||||ENST00000450305.2:
n.412G>C|||||,C|ENSG00000223972|ENST00000515242|Transcript|non_coding_exon_variant&nc_transcript_variant|613|
|||||1|3/3|||||1|DDX11L1|HGNC||||tr
anscribed_unprocessed_pseudogene||||ENST00000515242.2:n.613G>C|||||,C|ENSG00000223972|ENST00000518655|Transcript|intron_variant&nc_transcript_variant|
|||||1|2/3|||||1|DDX11L1|HGNC||||transcribed_unprocessed_pseudogene||||ENST00000518655.2:n.482-31G>C|||||,C|ENSR00000528767|RegulatoryFeature|
regulatory_region_variant|
1 13380 . C G 7829.15 VQSRTTrancheSNP99.60to99.80 AC=44;AC_AFR=23;AC_AMR=1;AC_Adj=24;AC_EAS=0;AC_FIN=0;AC_Het=24;AC_Hom=0;AC_
NFE=0;AC_SAS=0;AF=1.219e-03;AN=36096;AN_AFR=658;AN_AMR=116;AN_Adj=7532;AN_EAS=254;AN_FIN=60;AN_NFE=2040;AN_SAS=4322;BaseQRankSum=0.018;ClippingRankSum=0.26
4;DP=115091;FS=4.314;GQ_MEAN=12.26;GQ_STDDEV=20.80;Hom_AFR=0;Hom_AMR=0;Hom_EAS=0;Hom_FIN=0;Hom_NFE=0;Hom_SAS=0;InbreedingCoeff=-0.0794;MQ=31.93;MQ0=0;MQRank
Sum=0.00;NCC=67643;QD=8.10;ReadPosRankSum=-1.580e-01;VQSLOD=-3.041e+00;culprit=MQ;CSQ=G|ENSG00000223972|ENST00000456328|Transcript|non_coding_exon_variant
&nc_transcript_variant|628|
|||||1|3/3|||||1|YES|DDX11L1|HGNC||||processed_transcript||||ENST00000456328.2:n.628G>G|||||,G|ENSG00000223972|ENST000
00450305|Transcript|splice_region_variant&intron_variant&nc_transcript_variant|
|||||1|5/5|||||1|DDX11L1|HGNC||||transcribed_unprocessed_pseudogene|
|ENST00000450305.2:n.414+6C>G|
|||||,G|ENSG00000223972|ENST00000515242|Transcript|non_coding_exon_variant&nc_transcript_variant|621|
|||||1|3/3|||||1|DDX11L1|HGNC||||transcribed_unprocessed_pseudogene|
|ENST00000515242.2:n.621C>G|
|||||,G|ENSG00000223972|ENST00000518655|Transcript|intron_varia
nt&nc_transcript_variant|
|||||1|2/3|||||1|DDX11L1|HGNC||||transcribed_unprocessed_pseudogene|
|ENST00000518655.2:n.482-23C>G|
|||||,G|ENSR00000528
767|RegulatoryFeature|regulatory_region_variant|
1 13382 . C G 320.40 VQSRTTrancheSNP99.60to99.80 AC=3;AC_AFR=0;AC_AMR=0;AC_Adj=1;AC_EAS=0;AC_FIN=0;AC_Het=1;AC_Hom=0;AC_NFE=
0;AC_SAS=1;AF=1.007e-04;AN=29780;AN_AFR=534;AN_AMR=88;AN_Adj=6240;AN_EAS=200;AN_FIN=42;AN_NFE=1698;AN_SAS=3610;BaseQRankSum=-8.880e-01;ClippingRankSum=0.49
3;DP=93677;FS=0.000;GQ_MEAN=11.35;GQ_STDDEV=12.58;Hom_AFR=0;Hom_AMR=0;Hom_EAS=0;Hom_FIN=0;Hom_NFE=0;Hom_SAS=0;InbreedingCoeff=-0.0832;MQ=34.49;MQ0=0;MQRank
Sum=-6.910e-01;NCC=72140;QD=20.03;ReadPosRankSum=-2.073e+00;VQSLOD=-4.106e+00;culprit=MQ;CSQ=G|ENSG00000223972|ENST00000456328|Transcript|non_coding_exon_v
ariant&nc_transcript_variant|630|
|||||1|3/3|||||1|YES|DDX11L1|HGNC||||processed_transcript||||ENST00000456328.2:n.630C>G|||||,G|ENSG00000223972|E
NST00000450305|Transcript|splice_region_variant&intron_variant&nc_transcript_variant|
|||||1|5/5|||||1|DDX11L1|HGNC||||transcribed_unprocessed_pseudo
gene|
|ENST00000450305.2:n.414+8C>G|
|||||,G|ENSG00000223972|ENST00000515242|Transcript|non_coding_exon_variant&nc_transcript_variant|623| |
|||||1|3/3|
|||||1|DDX11L1|HGNC||||transcribed_unprocessed_pseudogene|
|ENST00000515242.2:n.623C>G|
|||||,G|ENSG00000223972|ENST00000518655|Transcript|intron_va
riant&nc_transcript_variant|
|||||1|2/3|||||1|DDX11L1|HGNC||||transcribed_unprocessed_pseudogene|
|ENST00000518655.2:n.482-21C>G|
|||||,G|ENSR00
00528767|RegulatoryFeature|regulatory_region_variant|
1 13402 . G C 89.66 VQSRTTrancheSNP99.60to99.80 AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=0;AC_EAS=0;AC_FIN=0;AC_Het=0;AC_Hom=0;AC_NFE=
0;AC_SAS=0;AF=9.971e-05;AN=20058;AN_AFR=190;AN_AMR=22;AN_Adj=1634;AN_EAS=84;AN_FIN=8;AN_NFE=398;AN_SAS=910;DP=42438;FS=0.000;GQ_MEAN=8.99;GQ_STDDEV=9.11;Ho
m_AFR=0;Hom_AMR=0;Hom_EAS=0;Hom_FIN=0;Hom_NFE=0;Hom_SAS=0;InbreedingCoeff=-0.0874;MQ=25.88;MQ0=0;NCC=78915;QD=14.94;VQSLOD=-3.988e+00;culprit=MQ;CSQ=C|ENSG

```

Figure 1: Snapshot of the terminal showing the data set.

Figure 1 is a snapshot of my terminal showing the first few lines of the original data file. The header of the file (omitted in Figure 1) contains various descriptions of the data itself. Following the comments, each line of the file contains information about a single SNP. The first column of each line is the chromosome number, the second column is the position of the SNP, fourth column gives the major allele, fifth column the minor allele, seventh column specifies whether or not the particular site passed the filtering. For the purpose of all analysis, I only used the sites that passed the filtering. The eight column contains overall allele count, and the next columns break the allele count according to a population. The abbreviations used are as follows:

AC_AFR: African/African American Allele Counts

AC_AMR: American Allele Counts

AC_EAS: East Asian Allele Counts

AC_FIN: Finnish Allele Counts

AC_NFE: Non-Finnish European Allele Counts

AC_SAS: South Asian Allele Counts

AC_OTH: Other Allele Counts

The information following the allele counts is the allele frequency. The allele frequency is the allele count divided by the total number of alleles in called genotypes denoted by AN. The AN is broken down according to a population as well. The columns are as follows:

AN_AFR: African/African American Chromosome Count

AN_AMR: American Chromosome Count

AN_EAS: East Asian Chromosome Count

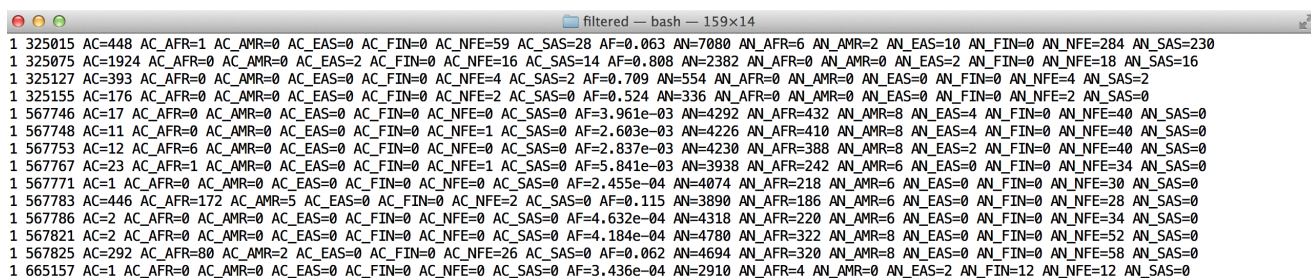
AN_FIN: Finnish Chromosome Count

AN_NFE: Non-Finnish European Chromosome Count

AN_SAS: South Asian Chromosome Count

AN_OTH: Other Chromosome Count

I first wrote a simple python script to filter the data set, producing 23 smaller files, one for each chromosome. I only kept the sites that passed the filtering, and I only kept the following columns for each site: chromosome number, allele count, allele count for each population, allele frequency, chromosome count, and chromosome count for each population. A snapshot of a filtered file is depicted in Figure 2.



```

1 325015 AC=448 AC_AFR=1 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=59 AC_SAS=28 AF=0.063 AN=7080 AN_AFR=6 AN_AMR=2 AN_EAS=10 AN_FIN=0 AN_NFE=284 AN_SAS=230
1 325075 AC=1924 AC_AFR=0 AC_AMR=0 AC_EAS=2 AC_FIN=0 AC_NFE=16 AC_SAS=14 AF=0.808 AN=2382 AN_AFR=0 AN_AMR=0 AN_EAS=4 AN_FIN=0 AN_NFE=18 AN_SAS=16
1 325127 AC=393 AC_AFR=0 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=4 AC_SAS=2 AF=0.709 AN=554 AN_AFR=0 AN_AMR=0 AN_EAS=0 AN_FIN=0 AN_NFE=4 AN_SAS=2
1 325155 AC=176 AC_AFR=0 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=2 AC_SAS=0 AF=0.524 AN=336 AN_AFR=0 AN_AMR=0 AN_EAS=0 AN_FIN=0 AN_NFE=2 AN_SAS=0
1 567746 AC=17 AC_AFR=0 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=0 AC_SAS=0 AF=3.961e-03 AN=4292 AN_AFR=432 AN_AMR=8 AN_EAS=4 AN_FIN=0 AN_NFE=40 AN_SAS=0
1 567748 AC=11 AC_AFR=0 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=1 AC_SAS=0 AF=2.603e-03 AN=4226 AN_AFR=410 AN_AMR=8 AN_EAS=4 AN_FIN=0 AN_NFE=40 AN_SAS=0
1 567753 AC=12 AC_AFR=6 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=0 AC_SAS=0 AF=2.837e-03 AN=4230 AN_AFR=388 AN_AMR=8 AN_EAS=2 AN_FIN=0 AN_NFE=40 AN_SAS=0
1 567767 AC=23 AC_AFR=1 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=1 AC_SAS=0 AF=5.841e-03 AN=3938 AN_AFR=242 AN_AMR=6 AN_EAS=0 AN_FIN=0 AN_NFE=34 AN_SAS=0
1 567771 AC=1 AC_AFR=0 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=0 AC_SAS=0 AF=2.455e-04 AN=4074 AN_AFR=218 AN_AMR=6 AN_EAS=0 AN_FIN=0 AN_NFE=30 AN_SAS=0
1 567783 AC=446 AC_AFR=172 AC_AMR=5 AC_EAS=0 AC_FIN=0 AC_NFE=2 AC_SAS=0 AF=0.115 AN=3890 AN_AFR=186 AN_AMR=6 AN_EAS=0 AN_FIN=0 AN_NFE=28 AN_SAS=0
1 567786 AC=2 AC_AFR=0 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=0 AC_SAS=0 AF=4.632e-04 AN=4318 AN_AFR=220 AN_AMR=6 AN_EAS=0 AN_FIN=0 AN_NFE=34 AN_SAS=0
1 567821 AC=2 AC_AFR=0 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=0 AC_SAS=0 AF=4.184e-04 AN=4780 AN_AFR=322 AN_AMR=8 AN_EAS=0 AN_FIN=0 AN_NFE=52 AN_SAS=0
1 567825 AC=292 AC_AFR=80 AC_AMR=2 AC_EAS=0 AC_FIN=0 AC_NFE=26 AC_SAS=0 AF=0.062 AN=4694 AN_AFR=320 AN_AMR=8 AN_EAS=0 AN_FIN=0 AN_NFE=58 AN_SAS=0
1 665157 AC=1 AC_AFR=0 AC_AMR=0 AC_EAS=0 AC_FIN=0 AC_NFE=0 AC_SAS=0 AF=3.436e-04 AN=2910 AN_AFR=4 AN_AMR=0 AN_EAS=2 AN_FIN=12 AN_NFE=12 AN_SAS=0

```

Figure 2: Snapshot of the terminal showing the filtered data set.

3 Methods

3.1 Things I wanted to look at

When writing my research proposal, and before even looking at the data set, I had a few questions in my mind that I thought interesting to explore:

1. Are the minor alleles more common in any particular population?
2. Are the rare variants more common in any particular population?
3. Are the minor alleles uniformly distributed along the genome, or do they occur in clusters?
4. Is there one individual in some population that is distinct from everybody else?

3.2 Challenges

In order to try to answer the questions, I needed to visualize the data somehow. The amount of data is quite overwhelming, so it's been hard to produce meaningful and clear graphs. I attempted to plot the allele count as a function of the position. Since there are six distinct populations in the data set, and I wanted to graph two populations in one plot, in order to compare all of the populations with each other, I had to make 15 plots per chromosome for a total of 345 plots. I only included a few of them in this report, but made all of the plots available for viewing on my GitHub: <https://github.com/emiliawk/math127>.

4 Results

My initial analysis (discussed during the class presentation on December 3) revealed significant disparities in allele counts between the populations. In particular, as shown in Figure 3, the allele count for the Non-Finnish European population was significantly higher than for the other populations.

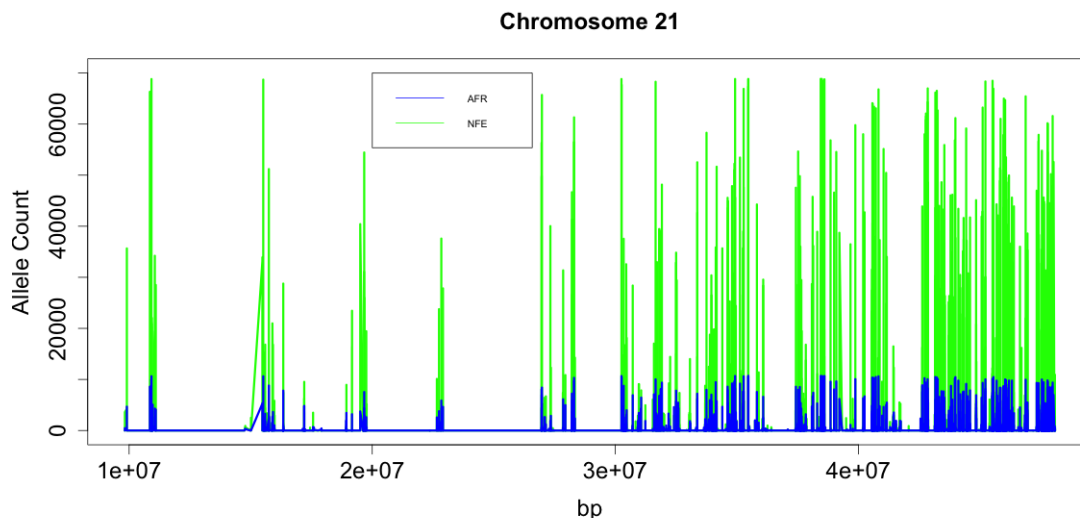


Figure 3: Allele count as a function of position (in base pairs) for chromosome 21 African and Non-Finnish European populations

I quickly realized that plotting the allele count as a function of the position does not make for meaningful comparisons because the differences in allele counts that I was seeing between the populations could be proportional to the number of alleles called from each population. Therefore, I modified the plots by dividing the allele count for each population by the number of alleles called from each population and plotting the allele frequency as a function of the position instead. Figure 4 shows the same comparison as Figure 3 but takes into account the number of chromosomes called from each population.

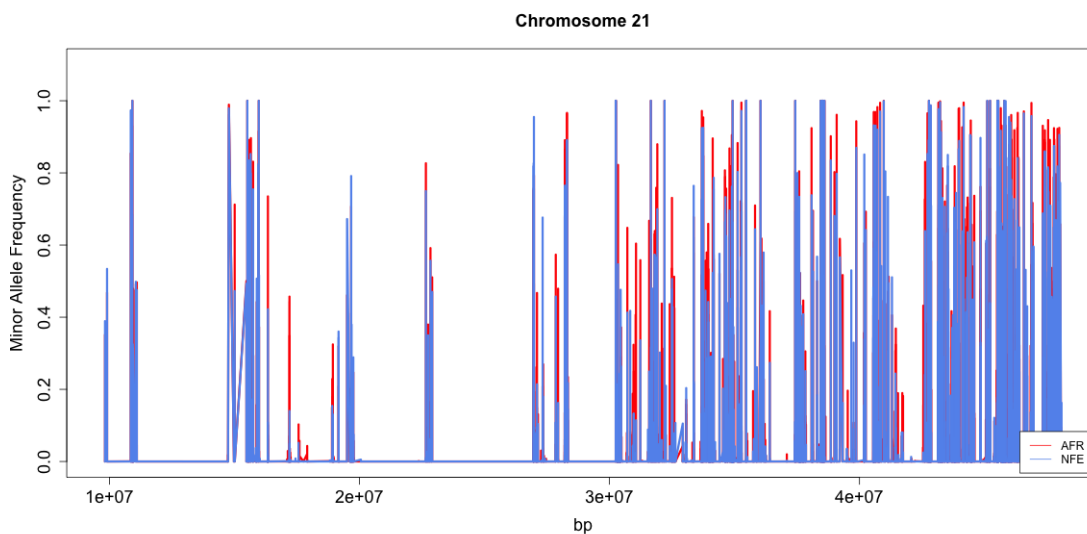


Figure 4: Allele frequency as a function of position (in base pairs) for chromosome 21 African and Non-Finnish European populations

As can be seen, now the two populations look more alike, and the significant difference between the two populations is gone.

However, after looking at the plots, I was surprised to see so many peaks where the frequency is equal to one. Based on that observation, I wanted to look at the distribution of rare variants. Therefore I took advantage of the overall allele frequency (it is given in the data set, but the allele frequency for each population I had to compute myself). I filtered my data further and kept only the sites with the overall allele frequency less than or equal to 1 percent, and remade the plots again. Figure 5 depicts the same comparison as in Figure 4, but only for sites with overall minor allele frequency $\leq 1\%$.

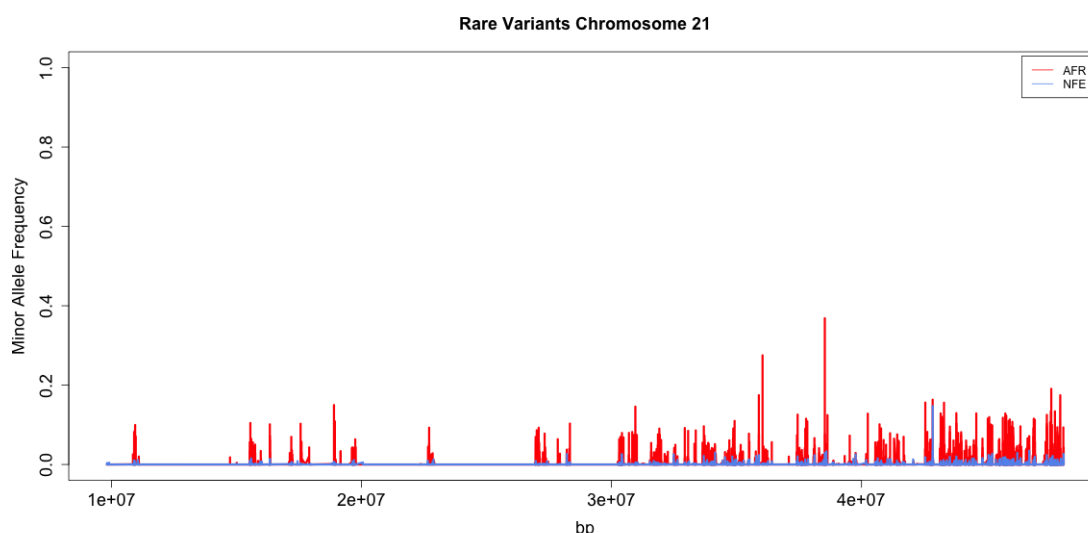


Figure 5: Allele frequency for rare variants as a function of position (in base pairs) for chromosome 21 African and Non-Finnish European populations

As can be seen, even though the overall allele frequency is lower than one percent, the frequencies for particular populations may be higher than that, which makes sense. After looking carefully at all plots, it is evident that rare variants are more common in the African population. All 345 plots of the rare variants are available for viewing at

https://github.com/emiliawk/math127/tree/master/plots/rare_variants.

Here are a few more:

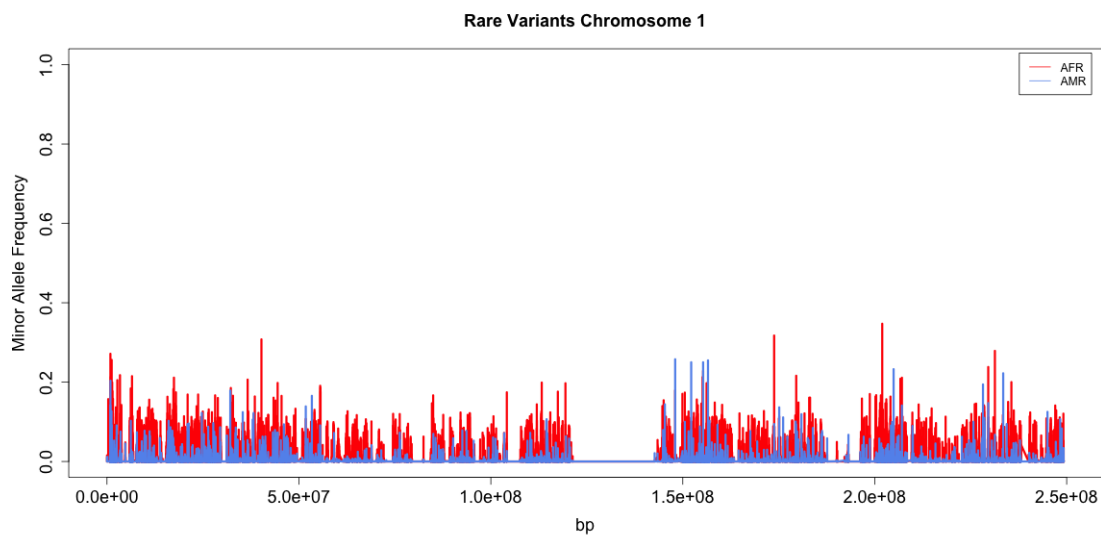


Figure 6: Allele frequency for rare variants as a function of position (in base pairs) for chromosome 1 African and American populations

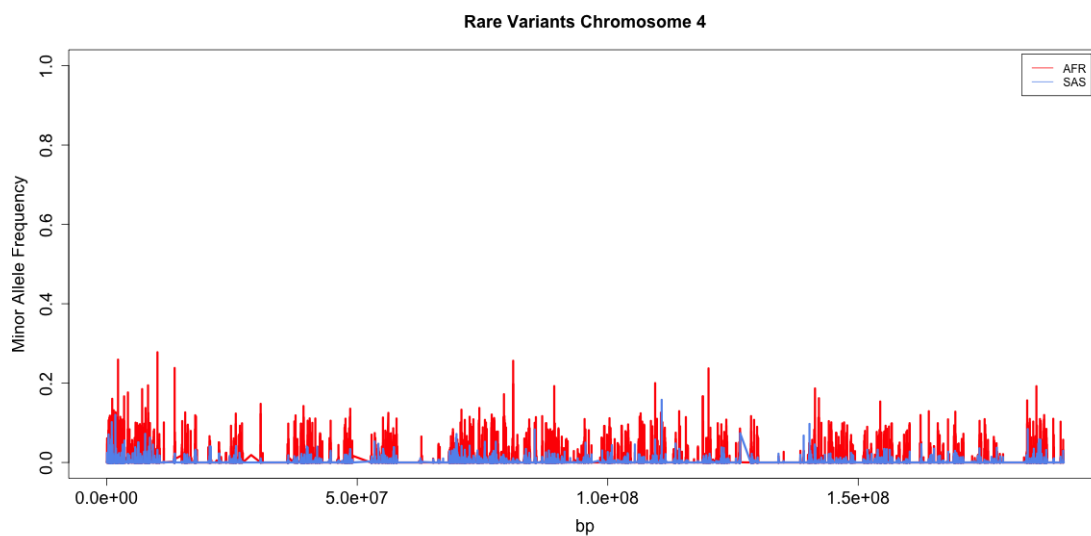


Figure 7: Allele frequency for rare variants as a function of position (in base pairs) for chromosome 4 African and South Asian populations

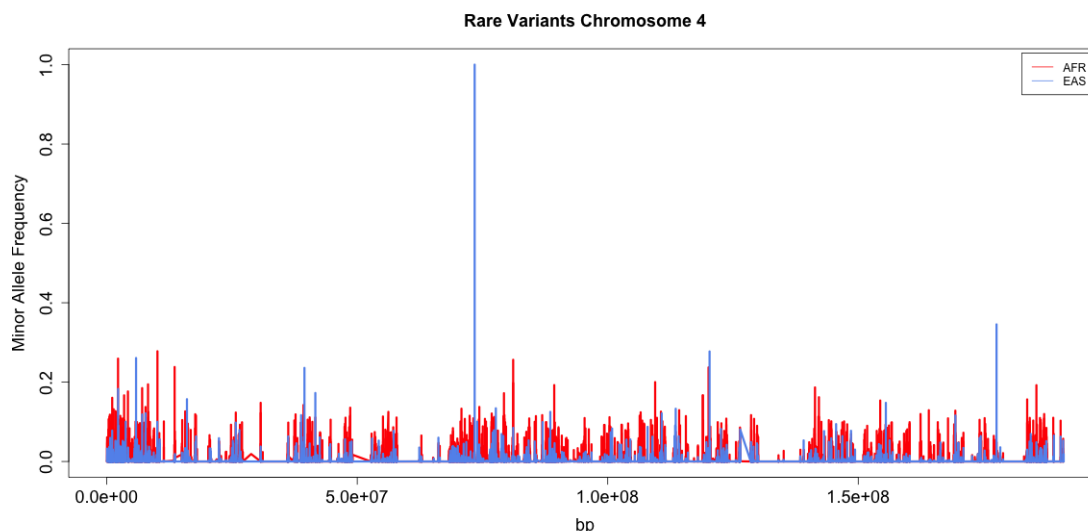


Figure 8: Allele frequency for rare variants as a function of position (in base pairs) for chromosome 4 African and East Asian populations

There were a few interesting sites with the overall allele frequency $\leq 1\%$, but with population allele frequency equal to 100%. For example, Figure 8 shows a site on chromosome 4 with a peak for the East Asian population. I investigated this particular locus further. The site is at position 73434380 with the overall allele frequency $AF=1.960e-03$, where $AC_EAS=2$, and $AF_EAS=2$. Therefore, it makes sense that the minor allele frequency for the East Asian population at this site is 100%.

5 Conclusions

It is hard to draw conclusions when taking all the available variants into account because the plots tend to be very dense, owing to the huge amount of data. After filtering for rare variants only, and looking carefully at all plots, I conclude that the rare variants seem to be more common in the African population than in the other five populations. The frequencies of rare variants seem to be very similar in the remaining five populations.

Rare variants are not uniformly distributed along the genome, but tend to occur in clusters. This of course depends on the chromosome that we look at, and is the most evident in chromosome 21, but less so in the other chromosomes, which can be attributed to the fact that chromosome 21 is the shortest and easiest to visualize.

It was hard to find a single individual that would be different from everybody else in the

ExAC data set, therefore based on the work that I've done so far, I was unable to address question 4 mentioned in section 3.1 just yet.

References

- [1] <http://exac.broadinstitute.org>
- [2] <http://macarthurlab.org/2014/11/18/a-guide-to-the-exome-aggregation-consortium-exac-data-set/>
- [3] <http://vcftools.sourceforge.net>