Quality control report for Maternal Cardiovascular-Related Single Nucleotide Polymorphisms, Genes and Pathways Associated with Early-Onset Preeclampsia SNP chip data

Kelly Yamasato, et al.

November 14, 2018

Abstract

This report encompasses the quality control summary for the Maternal Cardiovascular-Related Single Nucleotide Polymorphisms, Genes and Pathways Associated with Early-Onset Preeclampsia SNP chip data. A total of 60 samples were genotyped for 196725 SNPs. Quality control was performed across samples, across snps and on physical location. The results for each of these and the filtering criteria used are discussed herein.

1 QC filtering results

Out of the 60 samples, 25 did not pass the filtering criteria (41.67%). From the 196725 SNPs 80630 were excluded (40.99%). Out of the total 11803500 genotypes, 7743053 were excluded (65.6%). Filtering criteria consisted of QC metrics across SNPs, across arrays and on the physical mapping as detailed in the following sections.

Table 1 summarizes the number of SNPs and samples rejected for each QC criterion. Note that many of these overlap across criteria, thus the final numbers are not simply a sum of the rejection numbers for each criterion.

The correlation criterion for samples was not used to reject samples but simply to flag potential replicates which should be checked before further analyses. Correlation includes SNPs and samples flagged as bad which makes samples less similar than they should be. The correlation matrix should be used only for QC purposes. For downstream analysis the GRM constructed after data filtering should be used.

2 SNP statistics

In this section the descriptive statistics for the dataset on a per SNP basis are discussed. Figures 1 and 2 illustrate the difference between good and bad quality genotypes.

Table 1: Summary of SNPs and samples rejected for each QC criterion.

SNP criteria	number
>10 percent genotyping fail	12934
median GC scores <0.5	18914
all GC scores 0	10926
GC <0.5 in less than 90 percent samples	22491
100 percent homozygous	46308
MAF <0.01	12693
heterozygosity 3SD	15
Hardy-Weinberg at 1e-15	0
sample criteria	**number**
call rates <0.9	25
correlation >0.98	0
heterozygosity 3SD	0
mapping criteria	**number**
Chromosome X	93
Chromosome Y	21

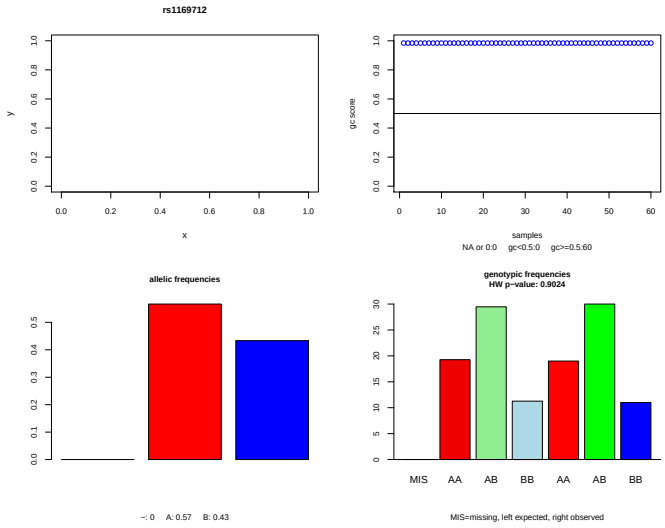

Figure 1: Example of a good quality SNP. Top left: clustering for each genotype (non calls are shown as black circles). Top right: GC scores. Bottom left: non-calls and allelic frequencies (actual counts are shown under the histogram). Bottom right: genotypic counts, on the left hand side the expected counts and on the right the observed counts; the last block shows number of non-calls.

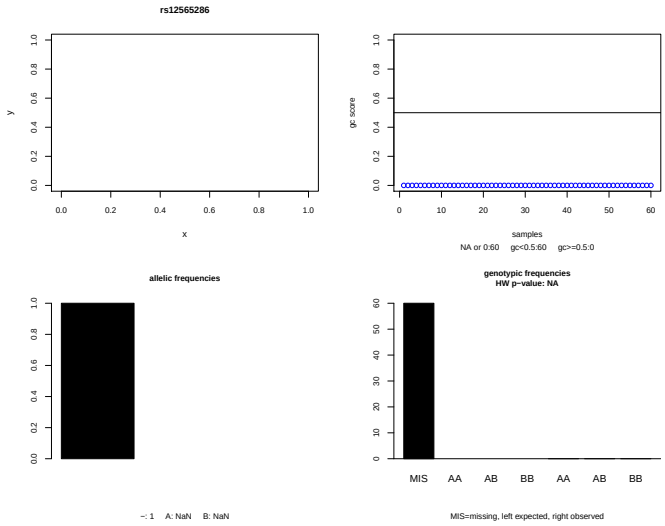

Figure 2: Example of a bad quality SNP. Top left: clustering for each genotype (non calls are shown as black circles - here all samples). Top right: GC scores. Bottom left: non-calls and allelic frequencies (actual counts are shown under the histogram). Bottom right: genotypic counts, on the left hand side the expected counts and on the right the observed counts; the last block shows number of non-calls.

Table 2: Call rates for SNPs.

rate	count	frequency
<0.9	22491	0.114
0.9-0.95	4108	0.021
0.95-0.99	22907	0.116
0.99-0.995	0	0.000
>=0.995	147219	0.748

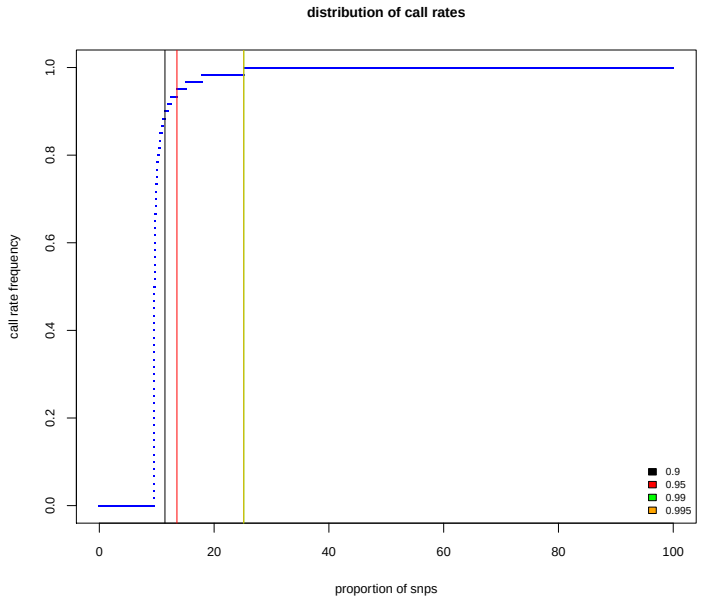

Figure 3: distribution of call rates per SNP.

2.1 SNP call rates

The number of SNPs with a call rate higher than 99.5% was 74.8% (Table 2 and Figure 3). As a rule of thumb around 90% of the snps would be expected to have a call rate above 99.5% and less than 2% would have call rates under 90%. In some cases the bulk of the data may be just below, in the 0.99-0.995 band (see breakdown of call rates in 2). Note that this will not hold well if there is ascertainment bias problems with the SNPs (i.e. SNPs selected for the chip derived from one population and the samples come from a very different one). In this dataset 22491 SNPs failed genotyping in over 10% of the samples (these were removed from the dataset). Note that the number of SNPs failed depends on the GC cutoff threshold – all SNPs below 0.5 are deemed to have failed (see further details in GC scores section).

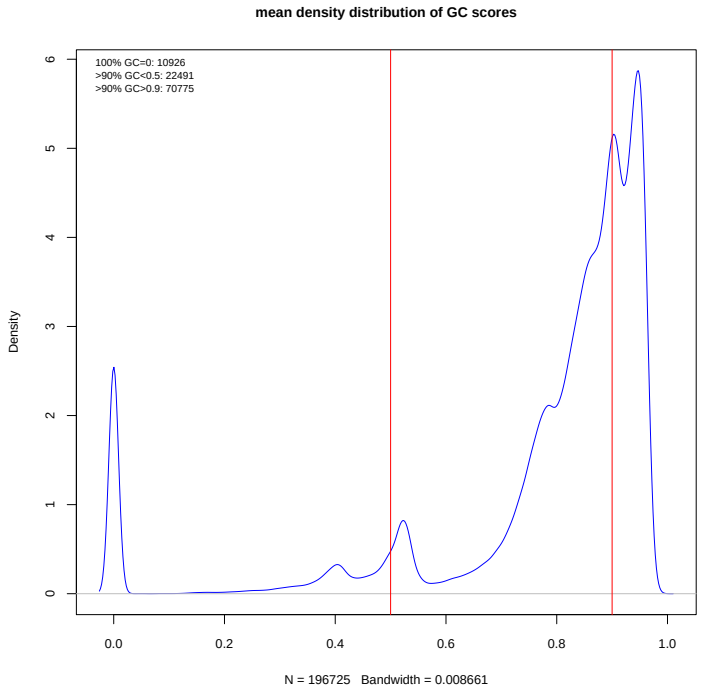

Figure 4: Histogram of GC scores.

2.2 GC scores

GC scores were filtered for a threshold value of 0.5. All calls under this value were discarded (note that this is specific for each snp on an individual sample). The dataset contained 10926 SNPs where all GC scores were 0. A further 22491 SNPs had a GC score over 0.5 in less than 90% of the samples. 70775 SNPs had a GC score of at least 0.9 for at least 90% of the genotypes. The mean GC scores for this data is 0.784 and the median is 0.791. The distribution of GC scores is shown in Figures 4 and 5.

2.3 Minor allele frequency

The minor allele frequency (MAF) was calculated for each SNP. 46308 SNPs are homozygous for the locus. A further 12693 had a MAF below 0.01 and were discarded. The distribution of MAFs is shown in figure 6. The average heterozygosity for the SNPs is 0.191 and the standard deviation is 0.188. A total 15 SNPs were detected as outliers (3SD from the mean and removed). Heterozygosity (He) and gene diversity (Ho) distributions are shown in figure 7.

2.4 Hardy-Weinberg equilibrium

Hardy-Weinberg (HW) equilibrium was calculated for each individual SNP using an exact chi-square test with continuity correction. HW equilibirum could not be determined for 65035 SNPS because these were either homozygous or

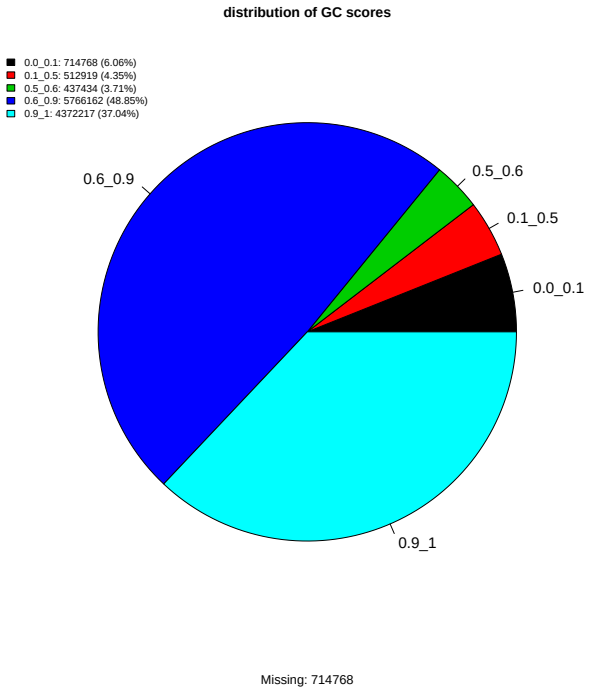

Figure 5: Pie plot of GC scores.

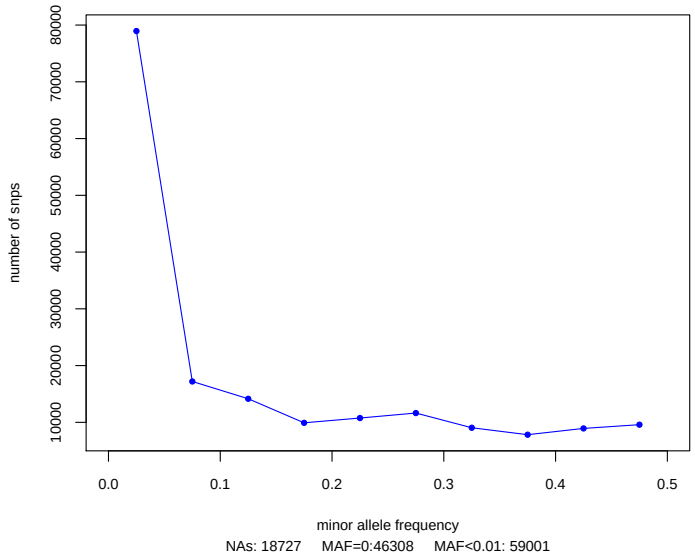

Figure 6: Minor allele frequency distribution for SNPs.

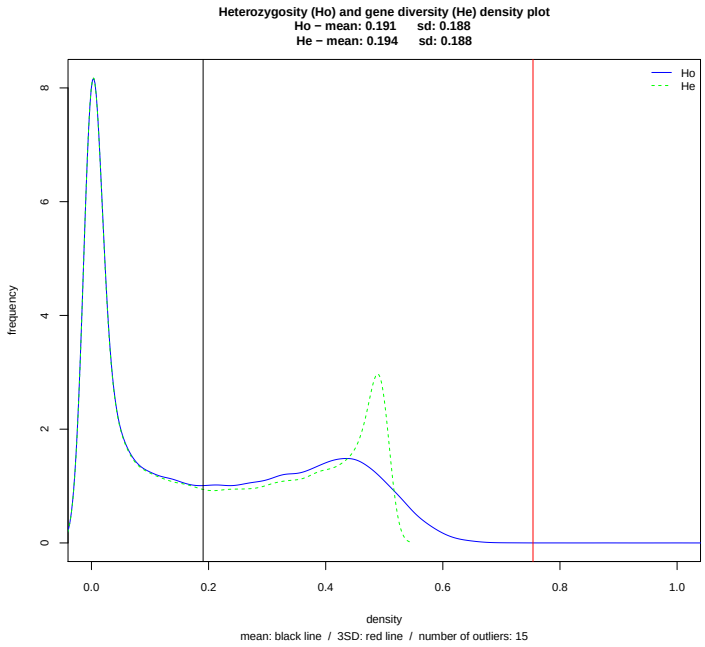

Figure 7: Heterozygosity distribution for SNPs. Note: standard deviations are biased.

had no calls assigned. 0 SNPs had a p-value of 0. A p-value cutoff of 1e-15 shows 0 SNPs out of HW equilibrium (note that this also includes SNPs that would not be expected to be in HW equilibrium such as those on sex chromosomes, mitochondria, etc). Figure 8 shows the distribution of p-values for HW equilibrium.

3 Array and sample statistics

In this section the descriptive statistics for the dataset on a per chip/sample basis are discussed.

3.1 Sample call rates

Out of the total 60 samples, 35 samples had a call rate at or above 0.9% and 0 samples had a call rate at or above 97%. The mean call rate across samples was 89.6%. An overview is given in Table 3.

3.2 Sample correlations

The average correlation between samples is 0.907. The statistic is useful to identify replicates in the dataset and samples that show very divergent genotypes due to quality problems (Figure 9). The minimum is 0.668 and the maximum is 0.967. 0 samples have a correlation above 0.98. Figure 10 shows the distribution of correlations between samples. The sample pairs with high correlations are

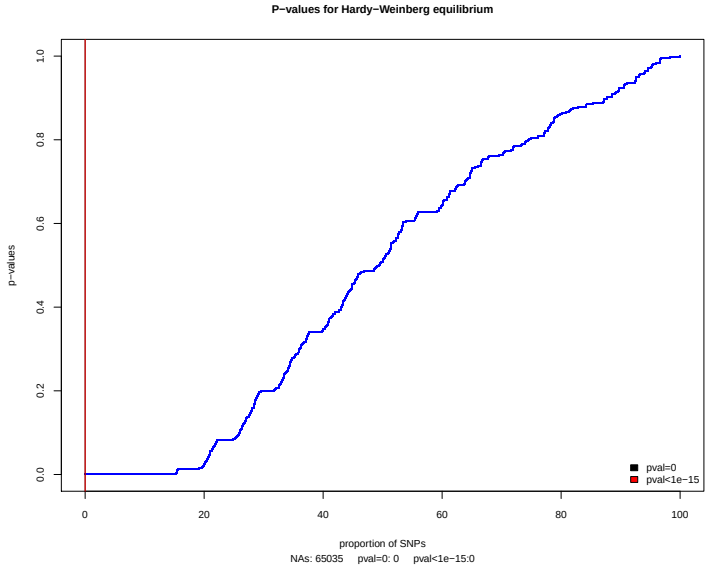

Figure 8: P-value distribution and thresholds for Hardy-Weinberg equilibrium.

Table 3: Call rates for samples.

statistic	value
num samples	60
min	0.829
max	0.904
mean	0.896
<0.97	60
<0.9	25

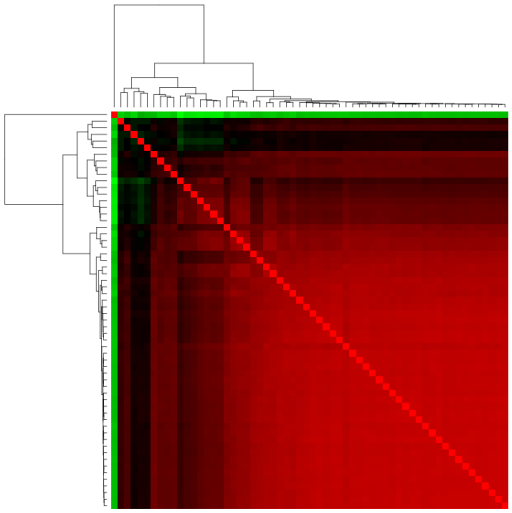

Figure 9: Heatmap of correlations between samples.

Table 4: Sample pairs with high correlations.

sample1	sample2	correlation

given in Table 4. Note: correlation herein is a simple Pearson correlation of the entire dataset without correcting for allelic frequencies or removing missing calls (use the GRM for downstream analyses). For this reason, even replicate samples will not have a perfect correlation of one (e.g. a given snp is called in one sample and missing in the replicate). A missing value of nine is used which teases genotypes with problems quite strongly apart.

3.3 Sample heterozygosity

The average heterozygosity for the samples is 0.19 and the standard deviation is 0.017. A total 0 samples were detected as outliers (3SD from the mean). Sample heterozygosity is shown in figure 11.

4 Physical mapping summary

A summary of the mapping information for the chip is given in table 5. Physical mapping plots for Hardy-Weinberg, MAF, GC scores and heterozygosity statistics are respectively shown in Figures 12, 13, 14 and 15. 114 SNPs are on excluded chromosomes and were removed. Many SNPs on e.g. the X chromosome are, as would be expected, out of HW equilibrium. The key point is to observe if any of the other chromosomes show a clear pattern of disequilibrium

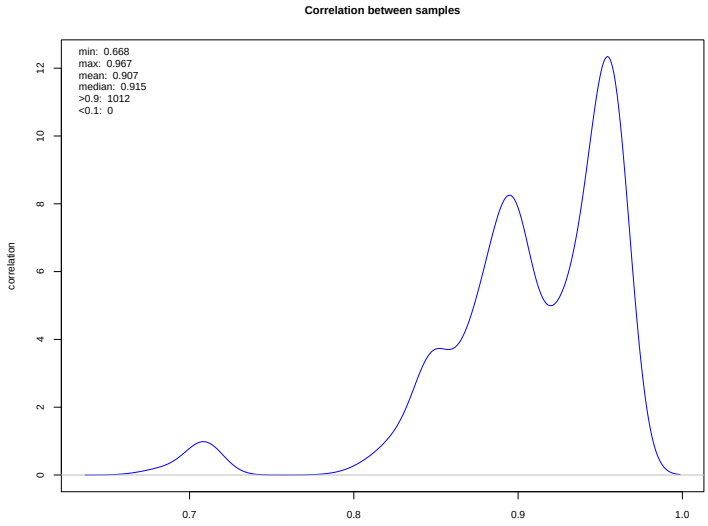

Figure 10: Correlations between samples.

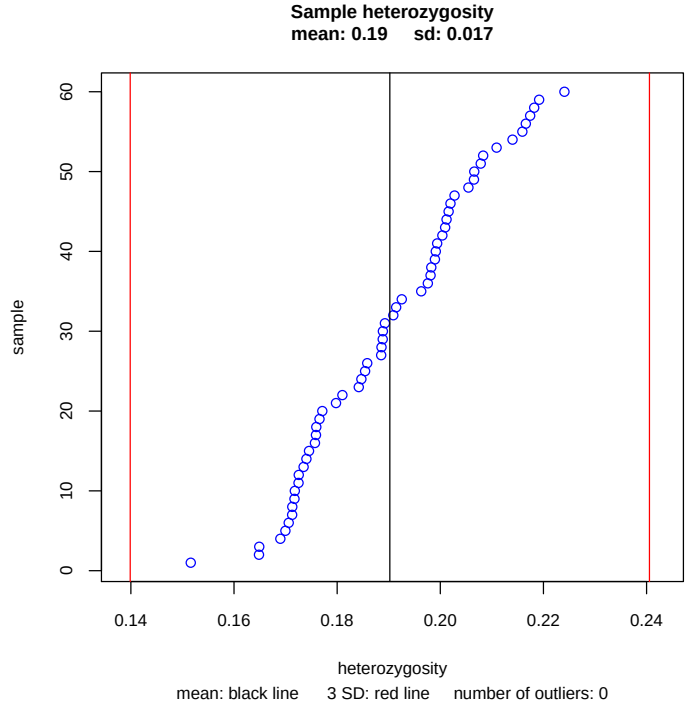

Figure 11: Heterozygosity for samples. Note: standard deviations are biased.

Table 5: Summary of mapping information per chromosome. Second column is the number of SNPs per chromosome. Columns min, max and mean are respectively the minimum distance between adjacent SNPs, the maximum distance and the average distance.

chrom	num	min	max	mean
1	18720	0	22699493	13163.6325
10	10506	1	2712749	12860.9368
11	16108	0	3105922	8334.6289
12	13791	0	1622815	9589.5287
13	3772	1	558894	25497.0438
14	2418	1	693058	36018.1713
15	9946	1	714123	8181.3318
16	8029	1	10903015	11034.0512
17	7464	1	647787	10530.8803
18	3550	1	1949829	21446.1877
19	3754	1	8345227	16929.2361
2	15750	0	3252257	15406.0056
20	3628	1	1851436	17182.9766
21	1398	1	394790	23910.6306
22	1182	0	1157483	29715.6088
3	13539	0	4594726	14714.6624
4	6169	1	3407830	30947.5717
5	7744	1	3327240	23298.3438
6	21966	0	3270910	7769.3697
7	12802	0	3784222	12393.0942
8	7596	0	3281727	19233.7002
9	6643	1	21788444	21068.7356
Mt	135	2	1326	113.9328
X	93	2211	11838402	1629453.5761
XY	1	NA	NA	NA
Y	21	350310	4277019	983913.2000

in any particular region. The same applies to MAF, GC scores and heterozygosity chromosomal plots - an indication of problems is a pattern in any given region.

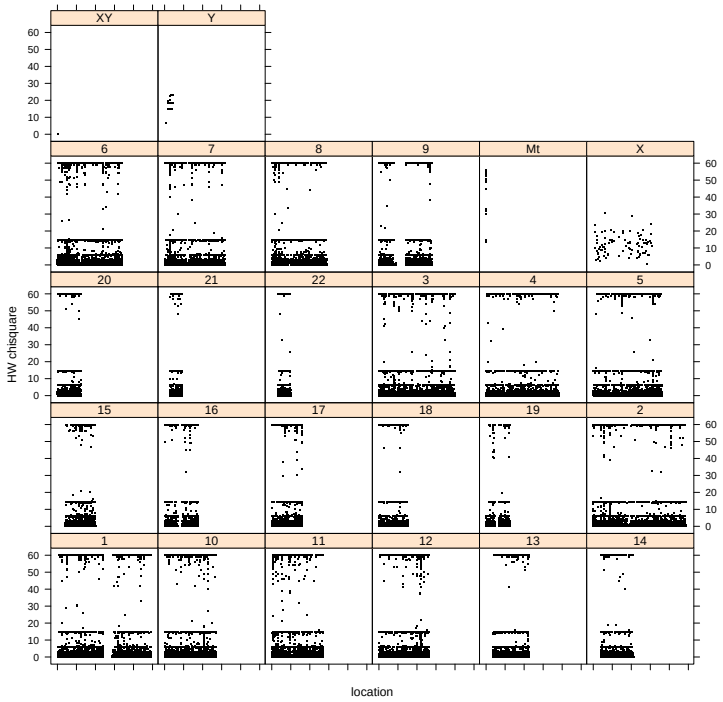

Figure 12: Hardy-Weinberg plotted against physical location for each chromosome (unmapped SNPs also included).

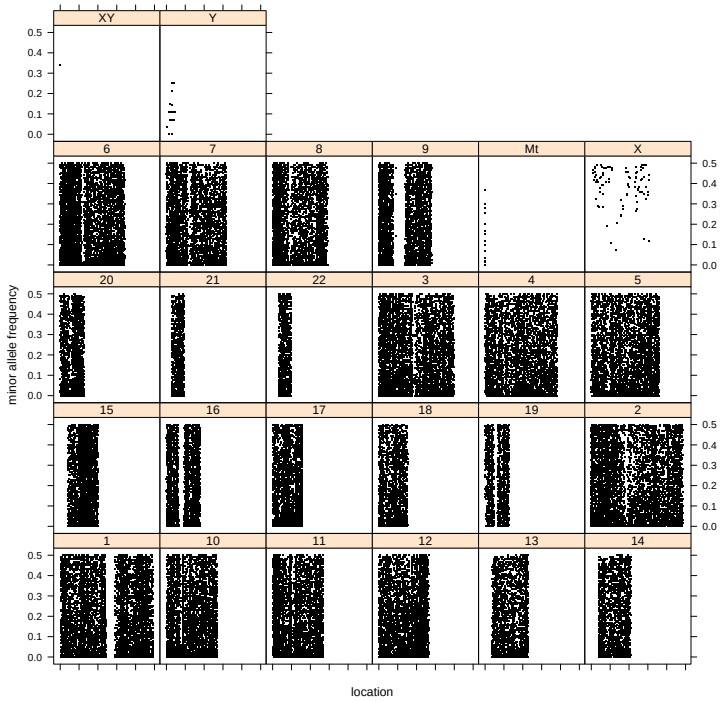

Figure 13: Minor allele frequencies plotted against physical location for each chromosome (unmapped SNPs also included).

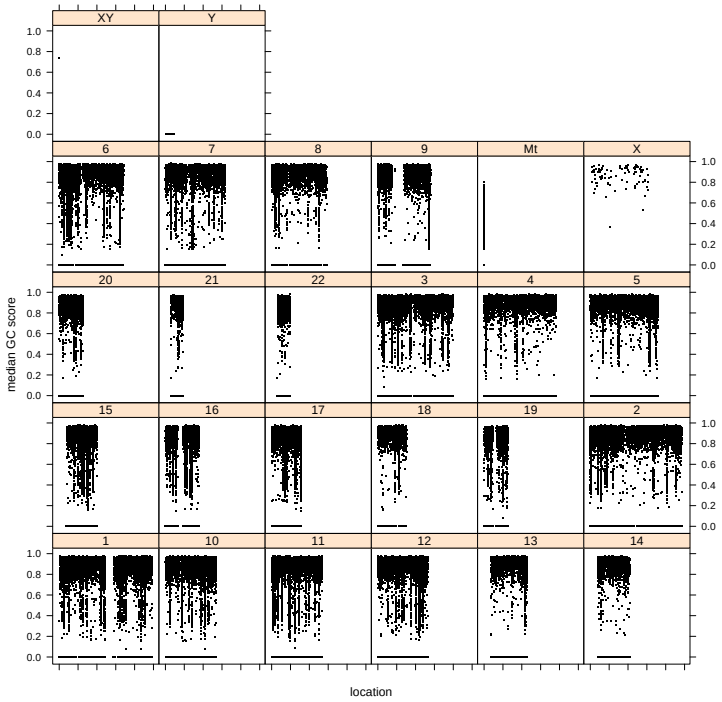

Figure 14: GC scores plotted against physical location for each chromosome (unmapped SNPs also included).

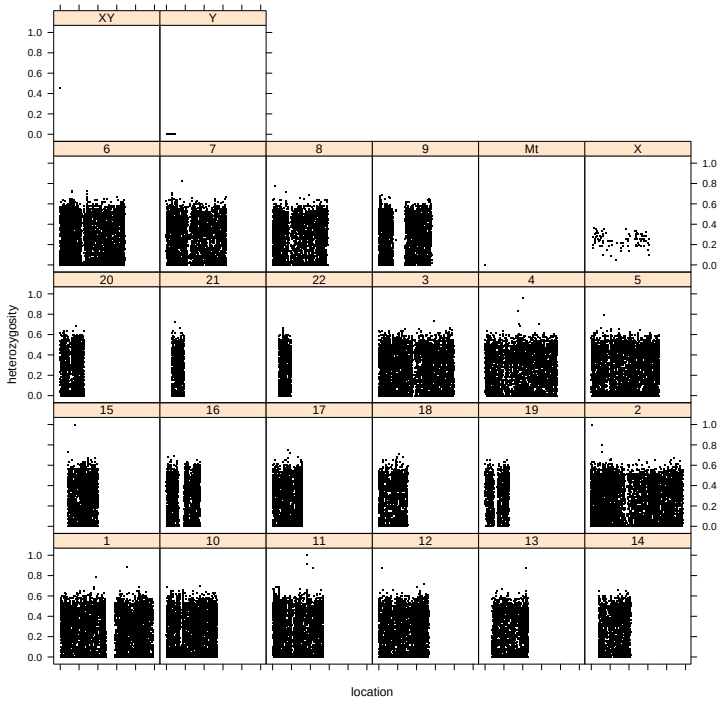

Figure 15: Heterozygosity plotted against physical location for each chromosome (unmapped SNPs also included).