**RAD sequencing output and coverage**

*Average number of reads per individual after demultiplexing:*

After demultiplexing, an average 3 032 292 reads per tagged individual were obtained. The number of reads per sample ranged from 1 558 796 to 5 686 234 (Table 1). It does not appear that samples with the lowest or highest number of reads are from a specific population (although 5 of the 12 lowest read count samples were from Kidd; Tables 2 and 3). The range of read counts per sample seems in line with other projects (see Mastretta-Yanes et al., 2015 *Molecular Ecology* they report an average of 1 632 914 reads per individual). I did not test for a lane effect because each sample was run on each lane. So a lane effect should affect each sample in a similar way.

Table 1. Summary of read counts following demultiplexing using process\_radtags in stacks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Quarter | Median | Mean | 3rd Qu. | Max |
| 1558796 | 2523975 | 2895875 | 3032292 | 3335963 | 5686234 |

Table 2. Samples where the number of reads is less than the first quantile number of reads (2523975).

|  |  |
| --- | --- |
| Read count | Sample ID |
| 1939234 | 2012HBMd01.fq |
| 1843208 | 2012HBMd02.fq |
| 2303884 | 2014KdHi2604.fq |
| 1803790 | 2014KdLo2402b.fq |
| 1618144 | 2014KdMH0405.fq |
| 2439400 | 2014KdMH2305.fq |
| 2382166 | 2014KdML2605.fq |
| 1887678 | 2014NkSHi2201.fq |
| 2256750 | 2014NkSLo2604.fq |
| 1558796 | 2015APKHi08.fq |
| 2373906 | 2015APRMd09.fq |
| 2353000 | 2015APRMd20.fq |

Table 3. Ten samples with the highest number of reads sequenced.

|  |  |
| --- | --- |
| Read count | Sample ID |
| 5649128 | 2012HBLo02.fq |
| 5243230 | 2014KdHi1402.fq |
| 4680072 | 2014KdML0401.fq |
| 4476074 | 2014NkNML0902.fq |
| 4416290 | 2015KdLo2004d.fq |
| 5686234 | 2015KdMH07A04.fq |
| 5668586 | 2015KdML0905.fq |
| 4039592 | 2015NkSMH0302.fq |
| 4181860 | 2015NkSML1301b.fq |
| 4874922 | 2015SPLo15.fq |

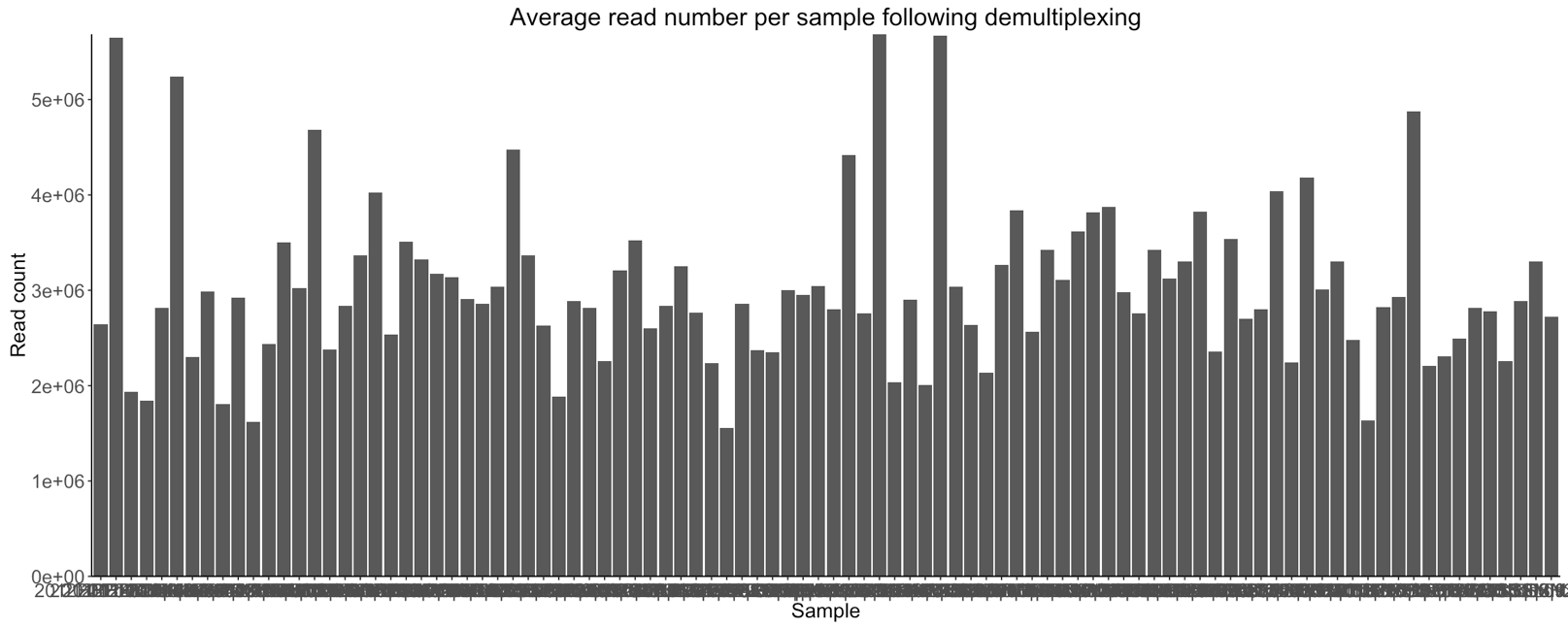


Figure 1. Number of reads per sample following demultiplexing using process\_radtags in stacks.

*Number of RAD-loci and SNPs*

When using default parameters for process\_radtags (-m 3) the number of RAD-loci is much higher than when the minimum number of reads required to build a stack is raised to 6 (Table 3).

Table 3. Summary of RAD-loci

|  |  |  |
| --- | --- | --- |
| # RAD-loci | Default parameters (-m 3) | Higher depth in process\_radtags (-m 6) |
| Total | 15421 | 8319 |
| >20 samples with a genotype | 1860 | 1270 |
| All samples have genotype | 156 | 129 |
| #SNPs | 15460 | 8358 |

*Sample depth*

The default parameter of populations did not exclude any of the samples. In Mastretta-Yanes et al., (2015) 15 of the 96 samples had too few reads to pass the filter requiring the sample to have more than 50% of the mean number of loci per individual. All *R. minor* samples had genotypes at more than 50% of the average number of loci per individual. I don’t understand why this is an appropriate filter.

There is a severe drop in the number of SNP sites when the site-level filter used by Mastretta-Yanes et al., (2015) was imposed—requiring a site to have more than 80% of samples with a genotype. After removing sites with fewer than 80% of samples that have a genotype call, we were left with 2 213 SNP sites (down from 15 460). Either the genome size of *R. minor* is much larger than anticipated (genome size= 1373 Mbp; 1.4GB Castro et al., 2011 –thesis) and thus we are not recovering sufficient number of overlapping reads. Or, the parameter setting in stacks is causing allele/site dropout.

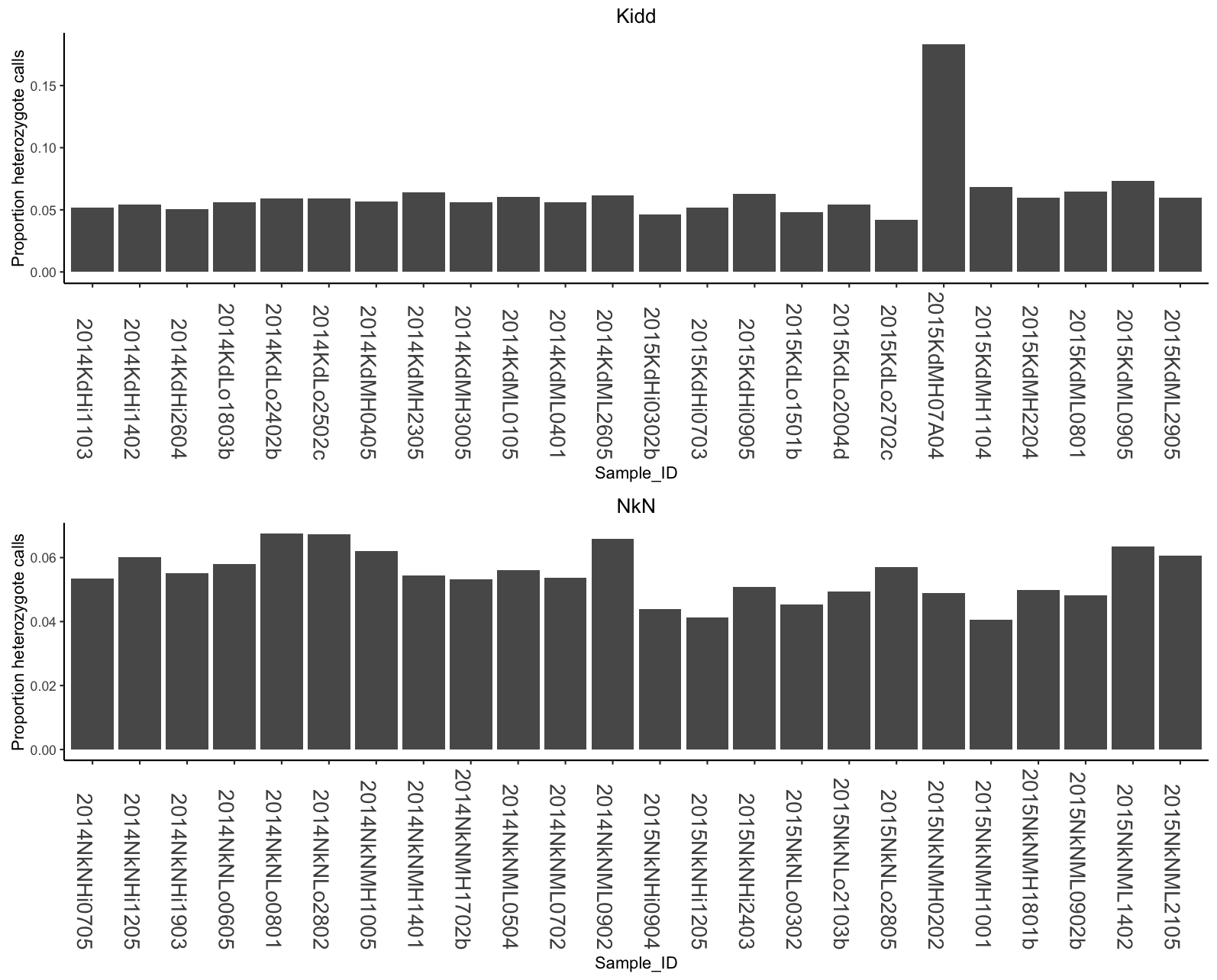
The average sample depth was not overly low. Mastretta-Yanes et al., 2015 obtained average depth per sample of 10.32 and considered that low. I don’t know what would be considered high. We achieved an average sample depth of 15.75 reads (ranging from 7.02 for 2015APKHi08 to 27.14 for 2014NkNML0902). After imposing the 80% genotype at site filter, the average sample depth was 21.51 (ranging from 9.55-52.91), suggesting that good sites have higher depth for all samples.

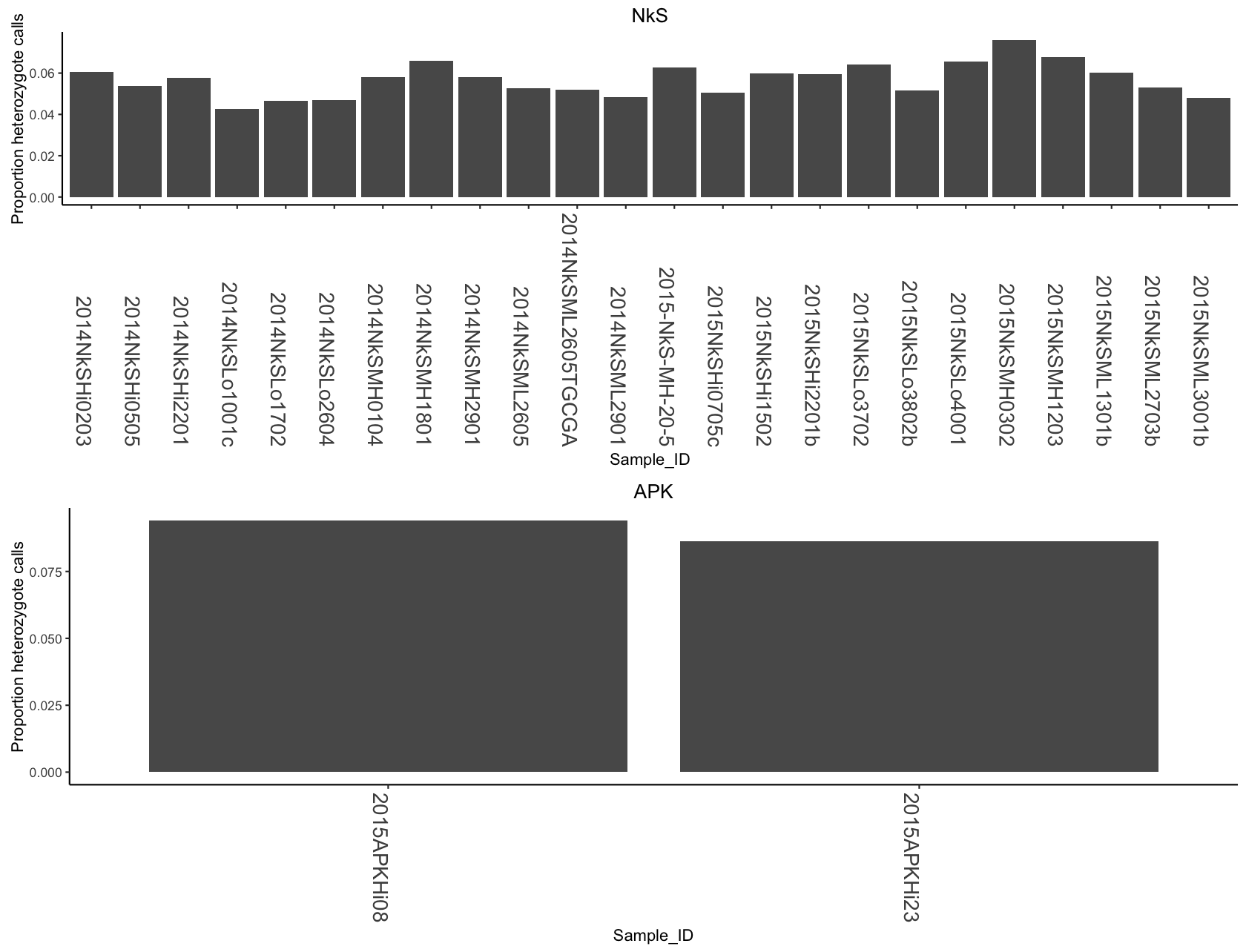
../depth/all_samples_depth.txt.pdf

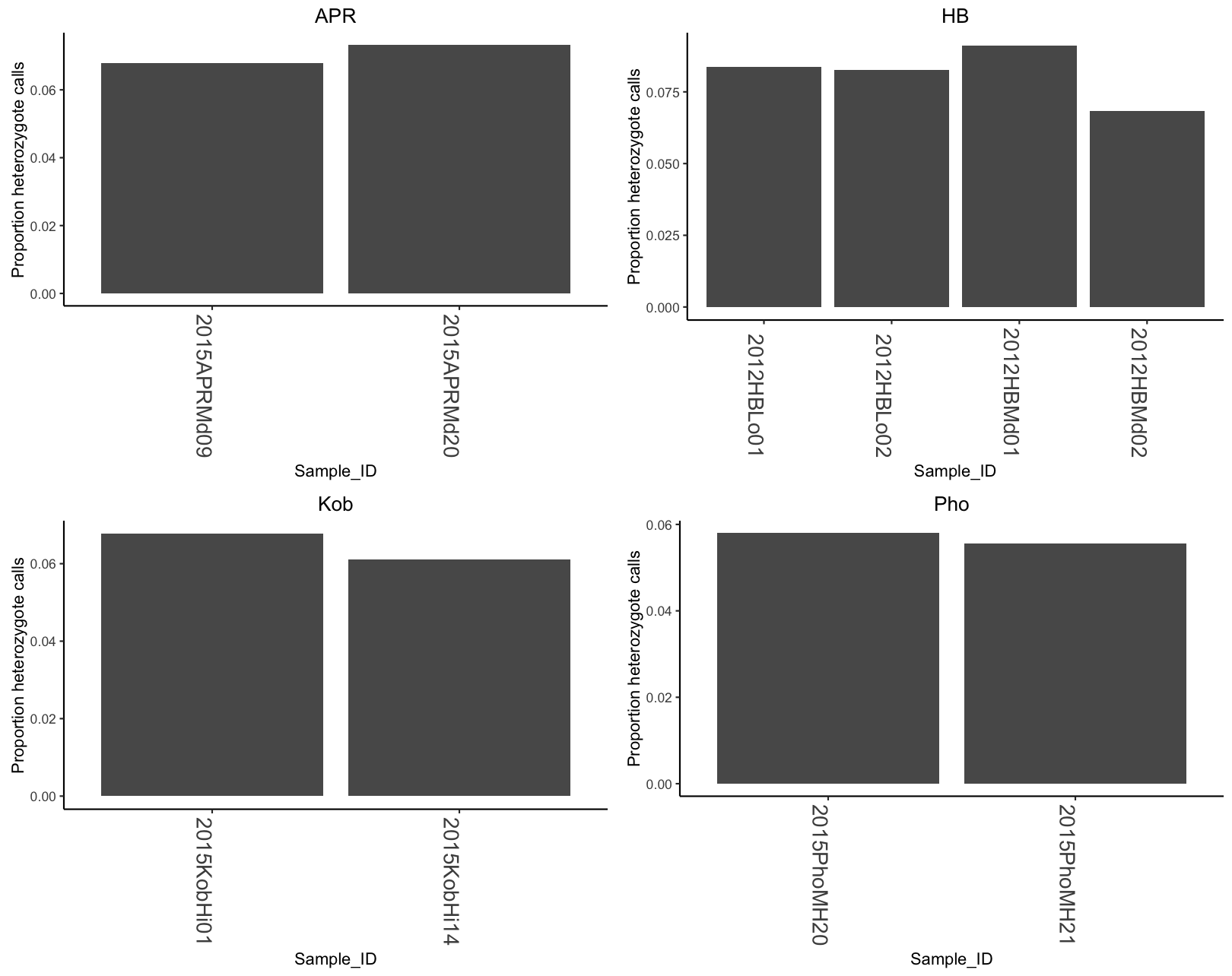
Figure 2. Distribution of sample depth across all samples after using default populations filter.

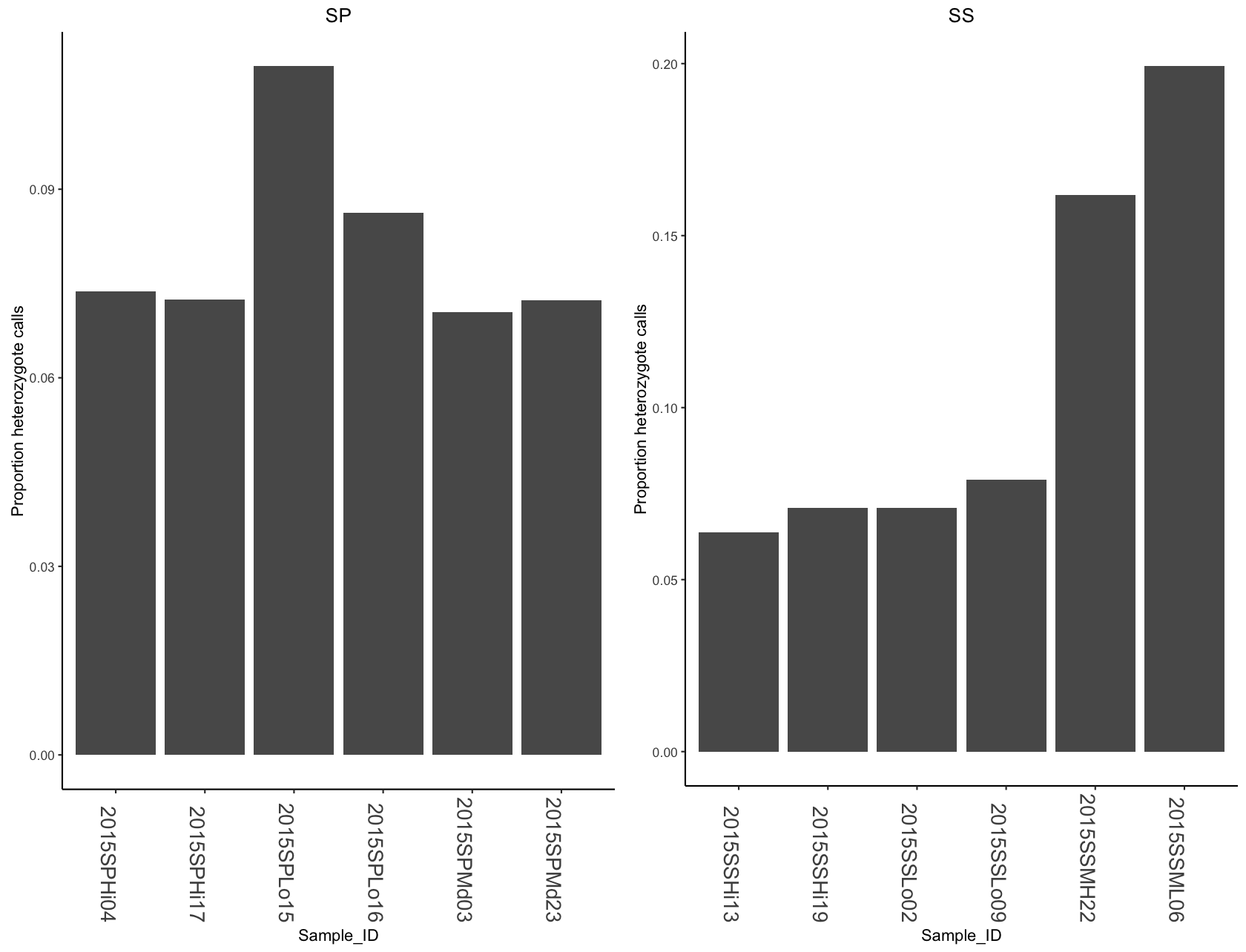
*Preliminary diversity estimates*

As a preliminary estimate of diversity, I estimated the proportion of heterozygous genotypes of all sites where a sample had a genotype call (min=0.04, mean=0.064 max=0.20). I clustered samples together into populations ignoring year and elevation (Figures below). Most samples have similar levels of heterozygosity of around 6%. APK, SS, SP and HB all have slightly higher than average proportion of heterozygote calls. Kidd at mid elevation also seems to have slightly higher heterozygote genotype calls (but not much).









**Notes on how I generated tables and plots:**

*Read counts*

I used a custom Unix script to count number of reads and length of each read for each sample following demultiplexing.

|  |
| --- |
| for i in \*.fq; do OUTPUT="$(cat "$i"| grep '^[ACTG]' | awk '{print length}'| sort -nr | uniq -c )" ; echo "$OUTPUT $i" >> read\_length\_count\_by\_sample2.txt; done& |

*Number of RAD-loci*

1. how many catalogs-sites are there in the entire dataset?

awk '$1!~"#"' batch\_1.sumstats.tsv | cut -f 2,5 |uniq -c | wc –l

15421

2. how many unique catalogs-sites are there in the entire dataset where more than 20 populations have a call?

awk '$1!~"#"' batch\_1.sumstats.tsv | cut -f 2,5 | uniq –c | awk '$1>19' |wc –l

3. How many catalog-sites are all populations represented?

awk '$1!~"#"' batch\_1.sumstats.tsv | cut -f 2,5 | uniq -c|awk '$1==38' |wc -l

4.How many SNPs?

*Sample depth*

Coverage pull out distribution from matches file