**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 |

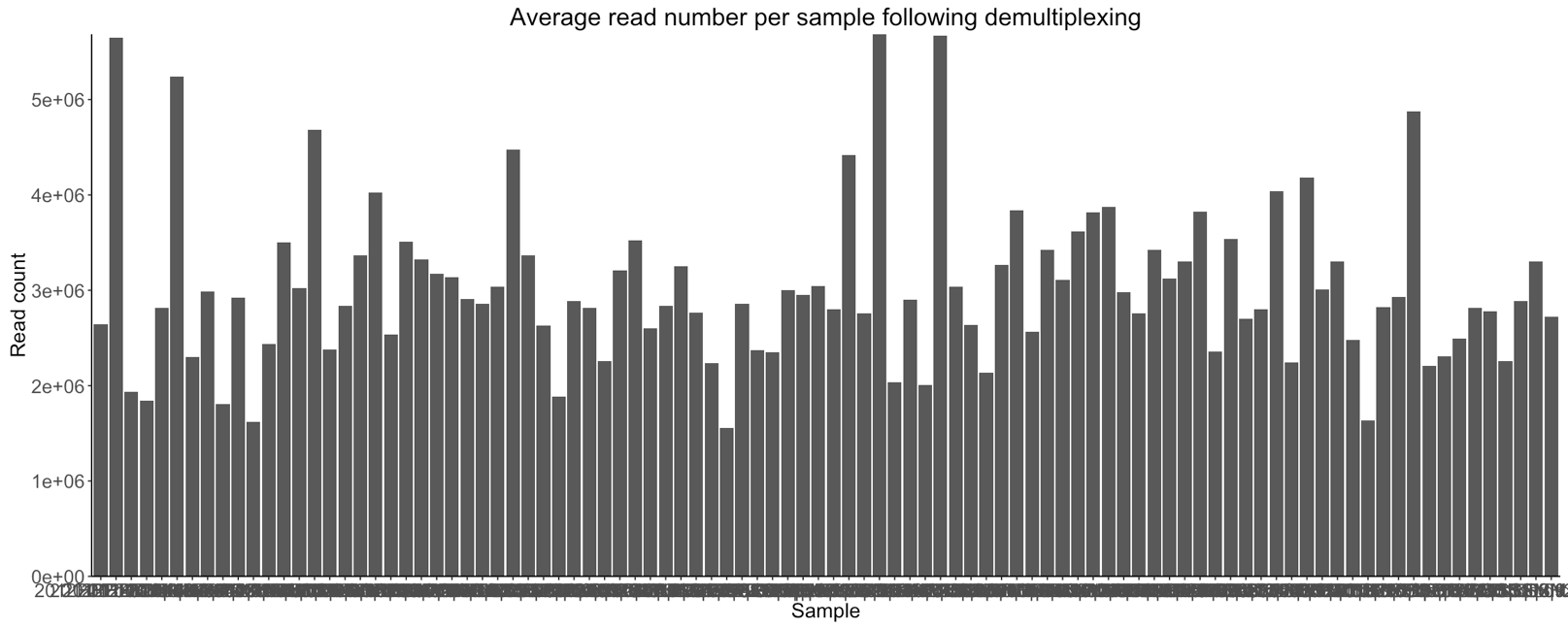


Figure 1. Number of reads per sample following demultiplexing using process\_radtags in stacks.

*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 |

Populations HB, APK and SS had the highest fraction of unique sites (0.25, 0.24, and 0. 15, respectively; Table 4 & Figure 2).

Table 4. Number of SNP sites and unique SNP sites within each population.

|  |  |  |  |
| --- | --- | --- | --- |
| pop\_name | unique\_sites | all\_sites | fraction\_unique |
| Kob | 31 | 2362 | 0.013 |
| **Kd** | **814** | **9359** | **0.087** |
| **SS** | **1306** | **8844** | **0.148** |
| APR | 134 | 3050 | 0.044 |
| SP | 182 | 5547 | 0.033 |
| **NkN** | **318** | **6819** | **0.047** |
| APK | 534 | 2227 | 0.240 |
| **NkS** | **359** | **6317** | **0.057** |
| HB | 1490 | 6074 | 0.245 |
| Ph | 9 | 2071 | 0.004 |

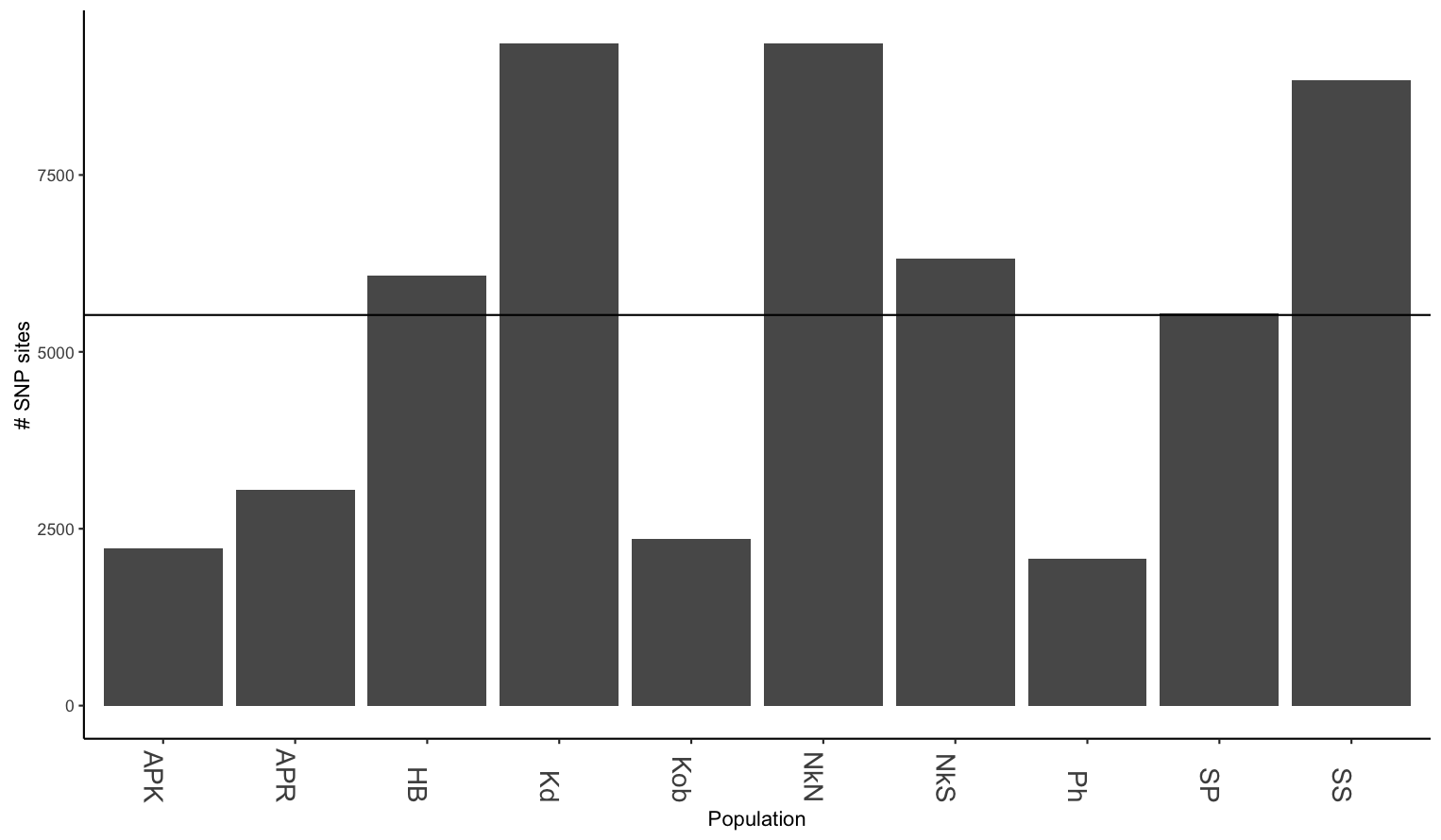


Figure 2. Number of SNP sites called within each populations.

Kob, APK, APR and Ph had the fewest number SNP sites. Three populations (NkN, SS and Kd) shared at least one SNPs site with all other populations. All other populations had two to four populations with which they did not share a single SNP site (Table 5). Overall, population SS had the highest proportion of shared sites for 9 populations (excluding HB; Table 6). One reason may be that it simply has the second highest number of SNP sites. Alternatively, it could indicate that variation in the other pops is a subset of variation of SS. This should be followed up on. I can’t say anything about nested diversity based on SNP sites alone.

Table 5. Missing

|  |  |
| --- | --- |
| Focal pop | Missing pops |
| SS |  |
| NkN |  |
| Kd |  |
| SP | ['APR', 'HB'] |
| NkS | ['HB', 'APK', 'Ph'] |
| APR | ['Kob', 'HB', 'SP'] |
| Kob | ['APR', 'APK', 'Ph'] |
| Ph | ['NkS', 'APK', 'Kob'] |
| APK | ['NkS', 'HB', 'Kob', 'Ph'] |
| HB | ['APR', 'APK', 'SP', 'NkS'] |
|  |  |

Table 6. Frequency of co-occurrence for each of 10 populations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| pop1 | pop2 | # shared sites | Pop1 site count | Pop2 site count | Fraction shared of pop1 | Fraction shared of pop2 |
| Kob | SP | 755 | 2362 | 5547 | 0.319644369 | 0.136109609 |
| Kob | NkN | 18 | 2362 | 6819 | 0.00762066 | 0.002639683 |
| Kob | Kd | 123 | 2362 | 9359 | 0.052074513 | 0.01314243 |
| Kd | Kob | 21 | 9359 | 2362 | 0.002243829 | 0.008890771 |
| Kd | APK | 45 | 9359 | 2227 | 0.004808206 | 0.020206556 |
| Kd | SS | 3080 | 9359 | 8844 | 0.329094989 | 0.348258706 |
| Kd | SP | 1288 | 9359 | 5547 | 0.137621541 | 0.232197584 |
| Kd | NkN | 1549 | 9359 | 6819 | 0.165509136 | 0.227159408 |
| Kd | APR | 449 | 9359 | 3050 | 0.047975211 | 0.147213115 |
| Kd | NkS | 600 | 9359 | 6317 | 0.064109413 | 0.094981795 |
| Kd | HB | 1400 | 9359 | 6074 | 0.149588631 | 0.230490616 |
| Kd | Ph | 120 | 9359 | 2071 | 0.012821883 | 0.057943023 |
| SS | SP | 1072 | 8844 | 5547 | 0.121212121 | 0.193257617 |
| SS | Kd | 1259 | 8844 | 9359 | 0.1423564 | 0.134522919 |
| SS | APR | 533 | 8844 | 3050 | 0.060266848 | 0.174754098 |
| SS | Kob | 801 | 8844 | 2362 | 0.090569878 | 0.33911939 |
| SS | NkN | 2036 | 8844 | 6819 | 0.230212573 | 0.298577504 |
| **SS** | **APK** | **1108** | **8844** | **2227** | **0.125282678** | **0.49753031** |
| SS | NkS | 77 | 8844 | 6317 | 0.008706468 | 0.01218933 |
| SS | HB | 736 | 8844 | 6074 | 0.083220262 | 0.121172209 |
| **SS** | **Ph** | **892** | **8844** | **2071** | **0.10085934** | **0.430709802** |
| SP | Kd | 1075 | 5547 | 9359 | 0.19379845 | 0.114862699 |
| **SP** | **SS** | **2227** | **5547** | **8844** | **0.401478277** | 0.251809136 |
| SP | Kob | 376 | 5547 | 2362 | 0.067784388 | 0.15918713 |
| SP | NkN | 1136 | 5547 | 6819 | 0.204795385 | 0.166593342 |
| SP | APR | 290 | 5547 | 3050 | 0.052280512 | 0.095081967 |
| SP | NkS | 132 | 5547 | 6317 | 0.023796647 | 0.020895995 |
| SP | APK | 28 | 5547 | 2227 | 0.005047774 | 0.012572968 |
| SP | HB | 292 | 5547 | 6074 | 0.052641067 | 0.048073757 |
| SP | Ph | 147 | 5547 | 2071 | 0.026500811 | 0.070980203 |
| APR | Kob | 59 | 3050 | 2362 | 0.019344262 | 0.024978831 |
| APR | Kd | 656 | 3050 | 9359 | 0.215081967 | 0.070092959 |
| APR | SS | 894 | 3050 | 8844 | 0.293114754 | 0.101085482 |
| APR | SP | 626 | 3050 | 5547 | 0.205245902 | 0.112853795 |
| APR | NkN | 202 | 3050 | 6819 | 0.066229508 | 0.029623112 |
| APR | APK | 150 | 3050 | 2227 | 0.049180328 | 0.067355186 |
| APR | NkS | 192 | 3050 | 6317 | 0.06295082 | 0.030394174 |
| APR | HB | 90 | 3050 | 6074 | 0.029508197 | 0.014817254 |
| APR | Ph | 71 | 3050 | 2071 | 0.023278689 | 0.034282955 |
| NkS | SP | 1034 | 6317 | 5547 | 0.163685294 | 0.186407067 |
| NkS | Kd | 1235 | 6317 | 9359 | 0.195504195 | 0.131958543 |
| NkS | SS | 1764 | 6317 | 8844 | 0.279246478 | 0.199457259 |
| NkS | Kob | 93 | 6317 | 2362 | 0.014722178 | 0.039373412 |
| NkS | NkN | 468 | 6317 | 6819 | 0.0740858 | 0.068631764 |
| NkS | APK | 2 | 6317 | 2227 | 0.000316606 | 0.000898069 |
| NkS | Ph | 31 | 6317 | 2071 | 0.004907393 | 0.014968614 |
| APK | SP | 141 | 2227 | 5547 | 0.063313875 | 0.025419145 |
| APK | Kd | 180 | 2227 | 9359 | 0.080826224 | 0.019232824 |
| APK | Kob | 9 | 2227 | 2362 | 0.004041311 | 0.00381033 |
| APK | NkN | 30 | 2227 | 6819 | 0.013471037 | 0.004399472 |
| APK | NkS | 7 | 2227 | 6317 | 0.003143242 | 0.001108121 |
| APK | Ph | 2 | 2227 | 2071 | 0.000898069 | 0.000965717 |
| HB | SP | 844 | 6074 | 5547 | 0.138952914 | 0.152154318 |
| HB | Kd | 1038 | 6074 | 9359 | 0.170892328 | 0.110909285 |
| HB | SS | 791 | 6074 | 8844 | 0.130227198 | 0.089439168 |
| HB | Kob | 82 | 6074 | 2362 | 0.013500165 | 0.034716342 |
| HB | NkN | 1056 | 6074 | 6819 | 0.173855779 | 0.154861417 |
| HB | APK | 35 | 6074 | 2227 | 0.005762265 | 0.01571621 |
| HB | NkS | 948 | 6074 | 6317 | 0.156075074 | 0.150071236 |
| HB | Ph | 41 | 6074 | 2071 | 0.006750082 | 0.019797199 |
| Ph | NkS | 1 | 2071 | 6317 | 0.000482859 | 0.000158303 |
| Ph | SP | 406 | 2071 | 5547 | 0.19604056 | 0.073192717 |
| Ph | NkN | 5 | 2071 | 6819 | 0.002414293 | 0.000733245 |
| Ph | Kd | 392 | 2071 | 9359 | 0.189280541 | 0.041884817 |

*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. Alternatively, there just is very little variation within and among these samples.

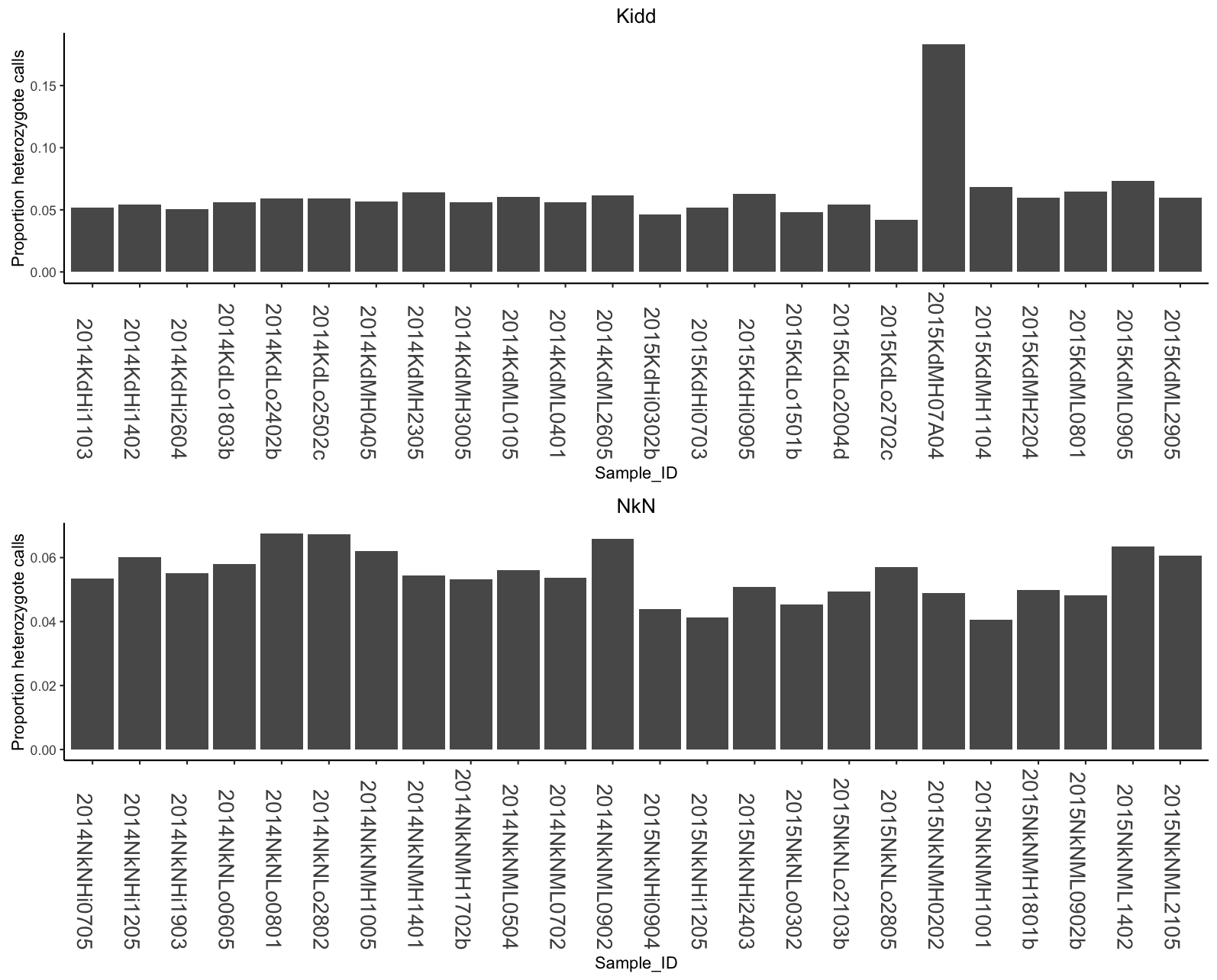
The average sample depth was not overly low. Mastretta-Yanes et al., 2015 obtained average depth per sample of 10.32 and considered that low. I don’t know what would be considered high. We achieved an average sample depth of 15.75 reads (ranging from 7.02 for 2015APKHi08 to 27.14 for 2014NkNML0902). 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, or that stacks merged duplicated regions.

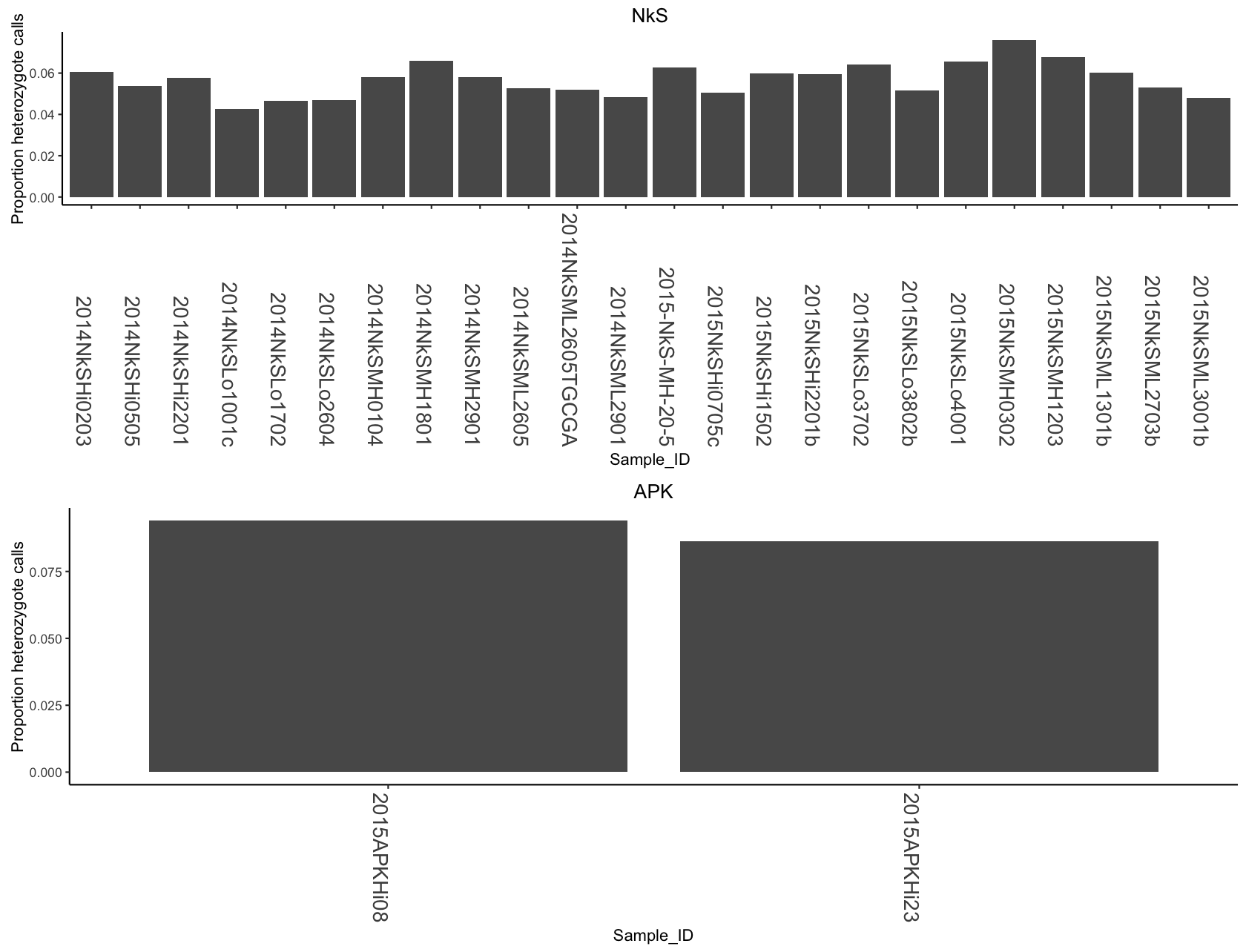
../depth/all_samples_depth.txt.pdf

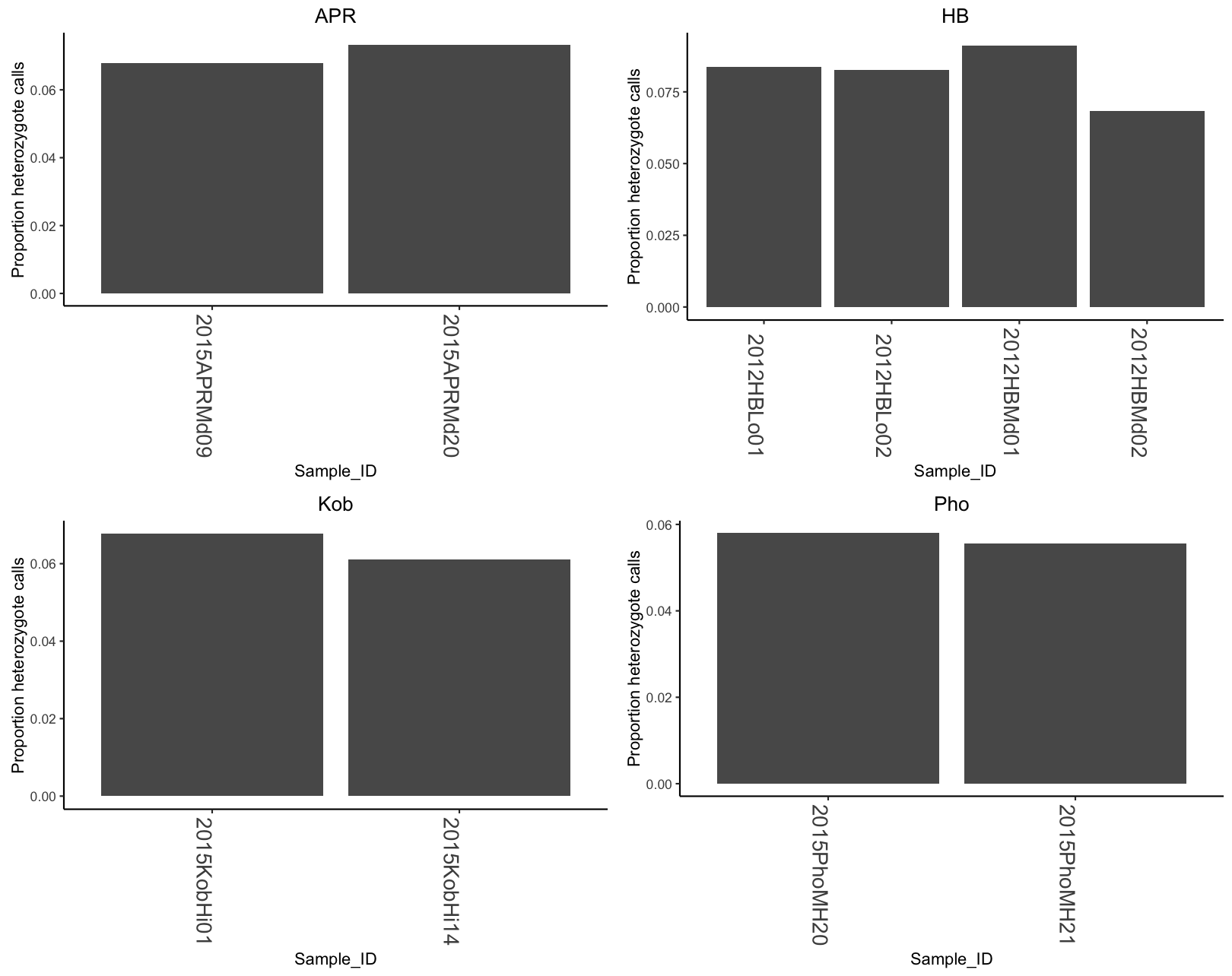
Figure 2. Distribution of sample depth across all samples after using default populations filter.

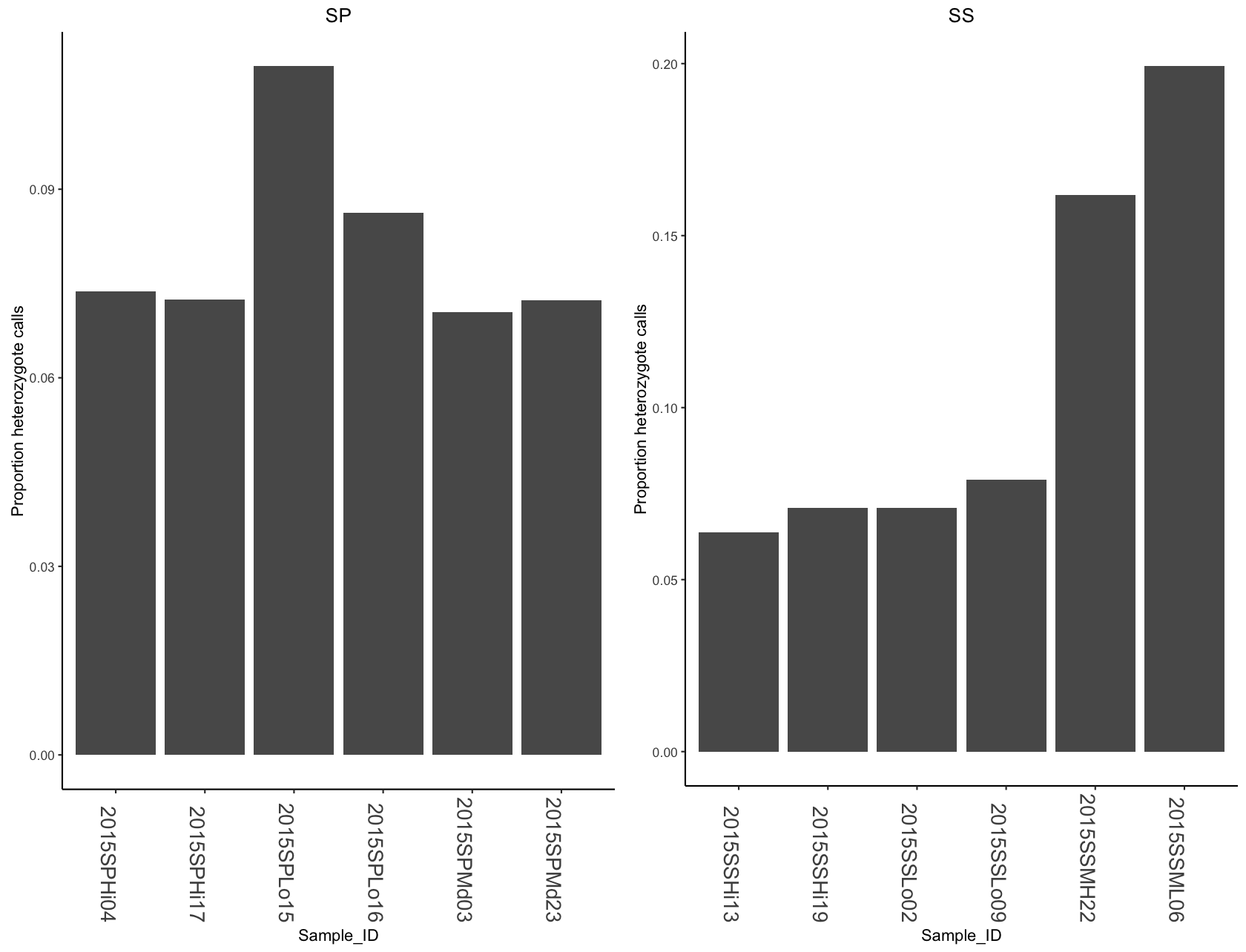
*Preliminary diversity estimates*

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).









**Figure 3. Proportion of heterozygote calls for each sample.**

**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*

Created my own custom script called “sample\_depth.py”. It pulls in all the \*.matches.tsv files that are output by the stacks script “sstacks”. There is a \*.matches.tsv file for each population. The 7th column in this file contains the number of reads for that sample at a given catalog-site.

The output from this script was downloaded to my computer and housed:

"output/read\_lenght\_counts\_post\_ustack/read\_length\_count\_by\_sample2.txt"

I then read that file into an Rscript called “sample\_depth.R” and made the plots

*Proportion of heterozygote calls per sample*

I converted the vcf file output by stakcs “population” to a summary file format using my script called “convert\_vcf\_to summary.py”. It simplifies outpu by SNP site. It reports the site, the ref allele alt allele and simple genotype for each samples (R-reference homozyogote, H –heterozygote, A-alternate homozygote and N for missing genotype)/

I then pass this summary file to another script to get counts.

Created my own script called “Het\_Hom\_Alt\_Missing\_counts\_by\_sample.py” . This takes in a batch\_1.summary file output by stacks “population” script. I downloaded the ouput of my script to the following:

"output/Het\_Hom\_alt\_Missing\_counts\_by\_sample.txt"

I imported that file into Rscript “preliminary\_diversity.R” and made plots . I also used this script to make a distribution of SNP sites for each sample. That information was obtained from my python script called “populations\_at\_each\_site.py”.