**Summary**

**\*NOTE I HAD TO MOVE THE RAW\_READS TO /u1/work/hpc3461/ BECAUSE I RAN OUT OF MEMORY!!!!**

Here I present basic summary statistics on the quality of data we obtained from employing a GBS approach on 96 samples pooled into a single library and sequenced across 2 HiSeq Illumina lanes. We used stacks to obtain genotype calls and present some preliminary analysis exploring the quality of those calls and performance of stacks. I begin by exploring: 1. the coverage obtained after genotyping using stacks, 2. the number of genotype calls per sample and population and 3. The distribution of missing genotypes across sites, samples and populations. I also perform a few exploratory tests to see how altering a few of the parameters in stacks alters the number of loci and samples for which genotypes are obtained. The most striking result of this section is that although we obtained approx. 15000 loci most samples only have genotype calls for approx. 2000 loci and the overlap between samples and populations of sites with genotype calls is poor. This could be a problem with the coverage, the enzymes used or something else.

I also begin exploring genetic diversity within and among samples. These analyses are not a formal “pop gen” analysis, but instead provide a simple way of understanding what kinds of genotype differences underlie pop gen estimates of differentiation. Stacks does provide basic pop gen estimates (like Fst), but we would first need to define populations more carefully than I had done originally. I take two approaches in these analyses. First, I provide estimates of total heterozygosity for each sample. Second, I summarize the number of genotype differences (heterozygous to homozygous and homozygous to homozygous differences) that occur between samples within and between populations and I use a simple neighbor joining tree to explore how samples are related to one another. Exploring the types of genotype difference can inform us on the quality of genotype call and errors associated with those calls. For example, within highly selfing, bottlenecked populations, like those on the mountain, the main genotype differences should be het to hom. We will look at whether there are subsets of alleles and reduced diversity along the elevational gradient. That would be the real interesting analysis.

The documentation for how each of these analysis was done is presented at the end of this document.

**Methods**

*Sample preparation*

DNA was extracted from silica dried tissue using a Qiagen DNeasy plant mini kit with slight modifications to the protocol. The volume of the extraction buffer AP1 was increased to 500uL (and all other buffer volumes were increased accordingly). The initial spin at step 10 was increased to 7-9minutes. The final incubation period was increased to 7-10 min. We evaluated DNA quality based on agaraose gel visualization and the spectral properties of the sample using Nanodrop One/Onec UV-Vis spectrophotometer. Only samples with 260/280 ratios of 1.8 -1.85 and 260/230 ratios greater than 1.5 were used for library construction. DNA concentration was estimated using Qubit 3.0 flourometric quantitation. All samples were standardized to 20ng/uL +/2ng/uL.

*GBS library construction and sequencing*

We used a two-enzyme (*PstI/MspI*) GBS protocol (Poland et al 2012), which is purported to provide a greater genome complexity reduction in large complex genomes. This uses a “rare-cutter” and one “common-cutter”. The use of two enzymes amplified fragments in the two-enzyme libraries will all consist of the barcoded forward and reverse adaptor.

GBS library was constructed using double digest with *Pstl* and *Mspl* and was prepared by Brian Boyle at Laval University’s IBIS centre (Institut de Biologie Integrative et des Systemes). GBS relies on reducing genome complexity to ensure sufficient overlap in sequence coverage. In this protocol genome complexity is reduced by restriction enzymes (RE). Choosing appropriate REs repetitive regions can be avoided and lower copy regions can be targeted with two to three-fold efficiency. Barcodes are included in one of the adaptor sequences (Elshire et al 2011 *PlosOne*). The barcode is just upstream of the RE cut-site. GBS is similar to RAD but the barcode composition and length results in fewer sequence phasing errors (phase errors is the sequencing errors that usually appear more frequently towards ends of the reads -Sleep et al. 2013 *BMC Bioinformatics*).

The first GBS library comprised of 96 samples was sent to McGill by Brian Boyle around May 12, 2016. Brian sent one GBS library, which will be sequenced on 2 lanes of Hi Seq 100bp sequencing lane. Single-end genomic sequencing of GBS libraries was conducted at the Genome Quebec Innovation Centre at McGill University using the Illumina HiSeq SR 100bp platform.

*SNP calling*

We used the stacks v 1.35 (REF) pipeline to assemble loci and call genotypes.

**RAD sequencing output and *coverage***

*Average number of reads per individual after demultiplexing:*

We obtained 339,763,458 Illumina reads in total. Following demultiplexing, using process\_radtags 86% of reads (291,100,064) were assigned to an individual sample. On average 3 032 292 reads per tagged individual were obtained, with a CV of 0.28 (according to Rob Elshire at cornell GBS tutorial this is “good” enough: https://www.youtube.com/watch?v=NGqKJ0TnL9o&list=PLCLuDSotcmhL2zP1\_mUIhfw8vAWX-v1yT). 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 |

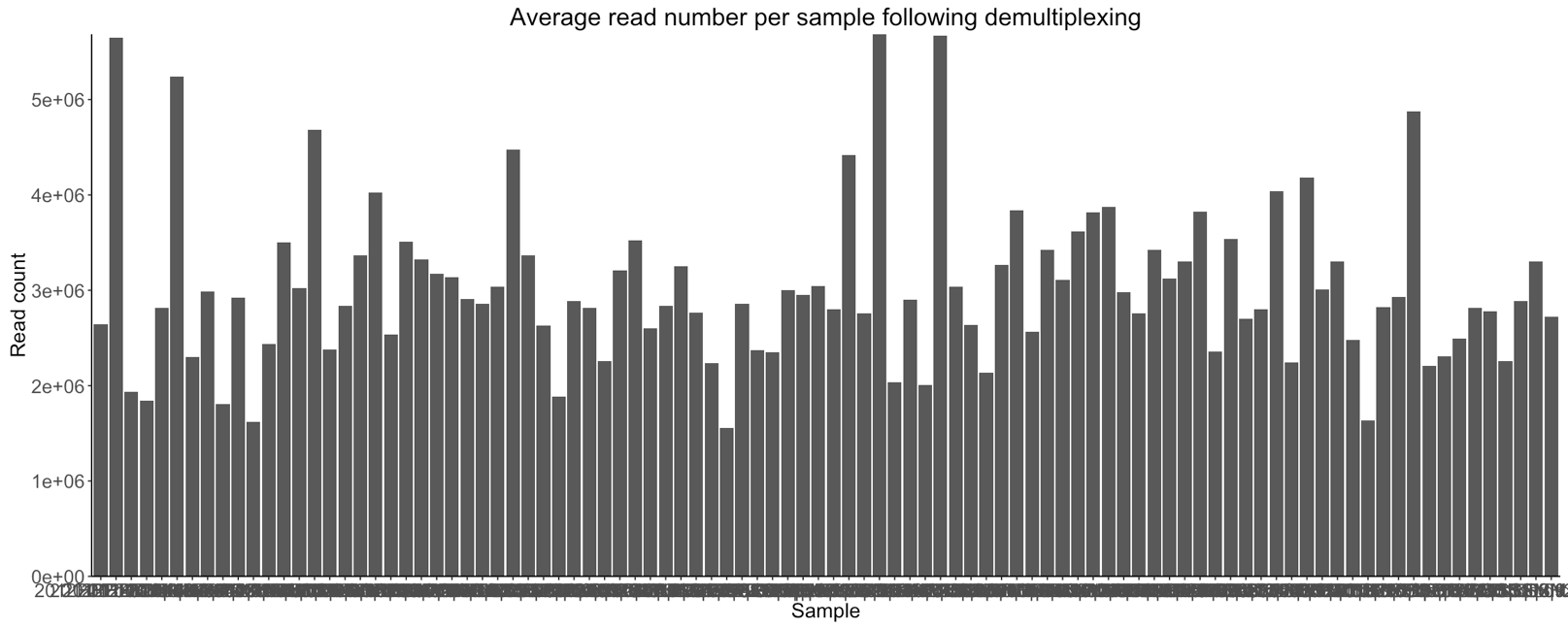


Figure 1. Number of reads per sample following demultiplexing using process\_radtags in stacks.

*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). A total of 414,049 RAD-tag loci were constructed. Of all 15421 sites called by stacks, an average sample is missing data (genotype call) at 83% of these sites. In other words, a given sample only has a genotype call at 17% of all called sites (see output from table from Het\_Hom\_Alt\_Missing\_counts\_by\_sample.py).

Not only do most samples lack a genotype call at the majority of stack sites, but there are also very few sites with genotype calls for each population (Table S4 and Fig S1). I counted the number of sites where there are no genotype calls for any samples for a given population (irrespective of year). It’s not clear if the lack of overlap is simply an issue of low coverage or if stacks is mistakenly calling separate loci, thus reducing overlap between populations. It might be worth running stacks with more permissible options on the number of differences between stacks/loci between samples/populations.

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 |

\*Note that there can be more snp’s that stacks. Stacks are the reconstructed sequences. A given reconstructed segment can have more than 1 SNP

Table S4. Number of missing sites by population. This count is the number of sites where there are no genotype calls for any sample within this population.

|  |  |  |
| --- | --- | --- |
| Population | Sites with no samples | Number of samples |
| SP | 9974 | 6 |
| Kd | 6405 | 24 |
| SS | 6975 | 6 |
| Kob | 13094 | 2 |
| NkN | 8670 | 24 |
| APR | 12442 | 2 |
| NkS | 9142 | 24 |
| APK | 13216 | 2 |
| HB | 9415 | 4 |
| Ph | 13385 | 2 |
| SP | 9974 | 6 |

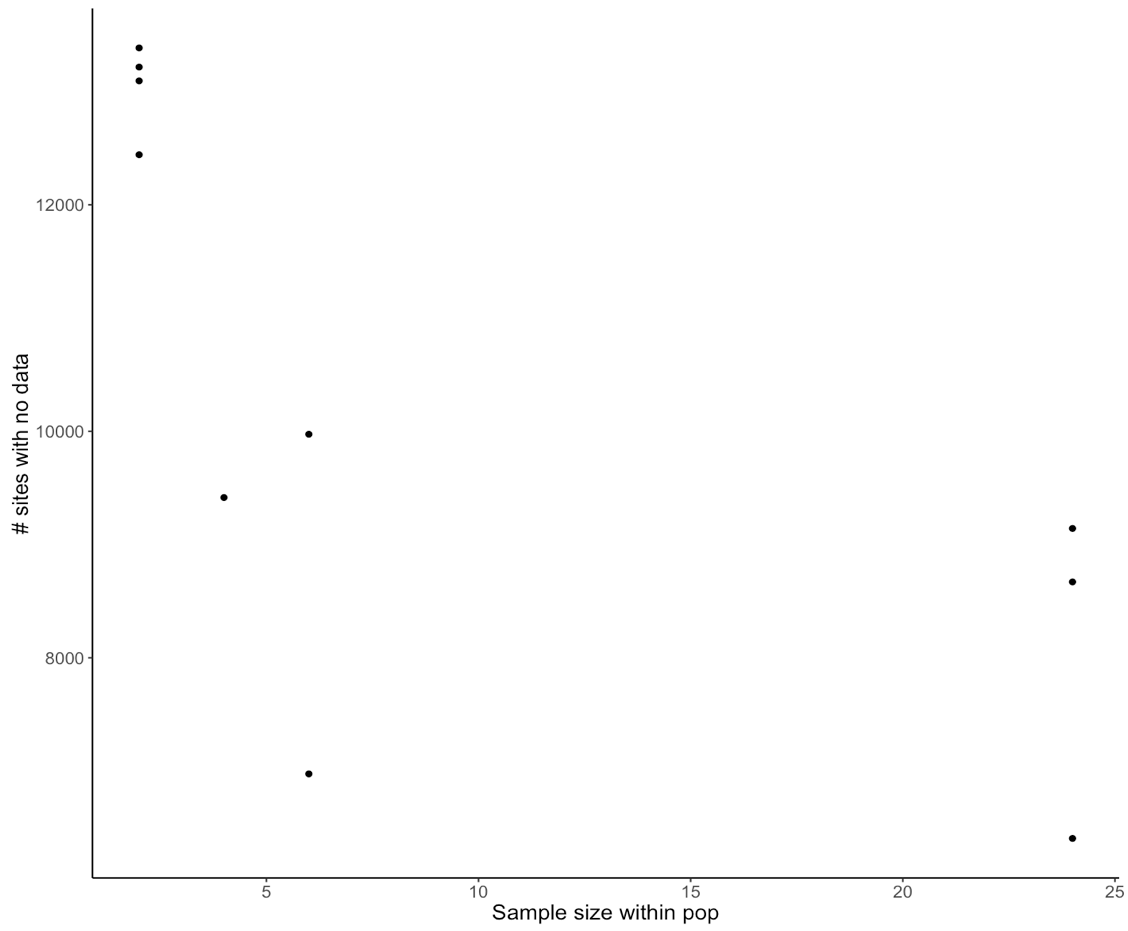


Figure S1. Association between sample size within population and number of sites with no genotype call.

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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| pop\_name | Sample size | unique\_sites | all\_sites | fraction\_unique |
| Kob | 2 | 31 | 2362 | 0.013 |
| Kd | 24 | 814 | 9359 | 0.087 |
| **SS** | **6** | **1306** | **8844** | **0.148** |
| APR | 2 | 134 | 3050 | 0.044 |
| SP | 6 | 182 | 5547 | 0.033 |
| NkN | 24 | 318 | 6819 | 0.047 |
| **APK** | **2** | **534** | **2227** | **0.240** |
| NkS | 24 | 359 | 6317 | 0.057 |
| **HB** | **4** | **1490** | **6074** | **0.245** |
| Ph | 2 | 9 | 2071 | 0.004 |

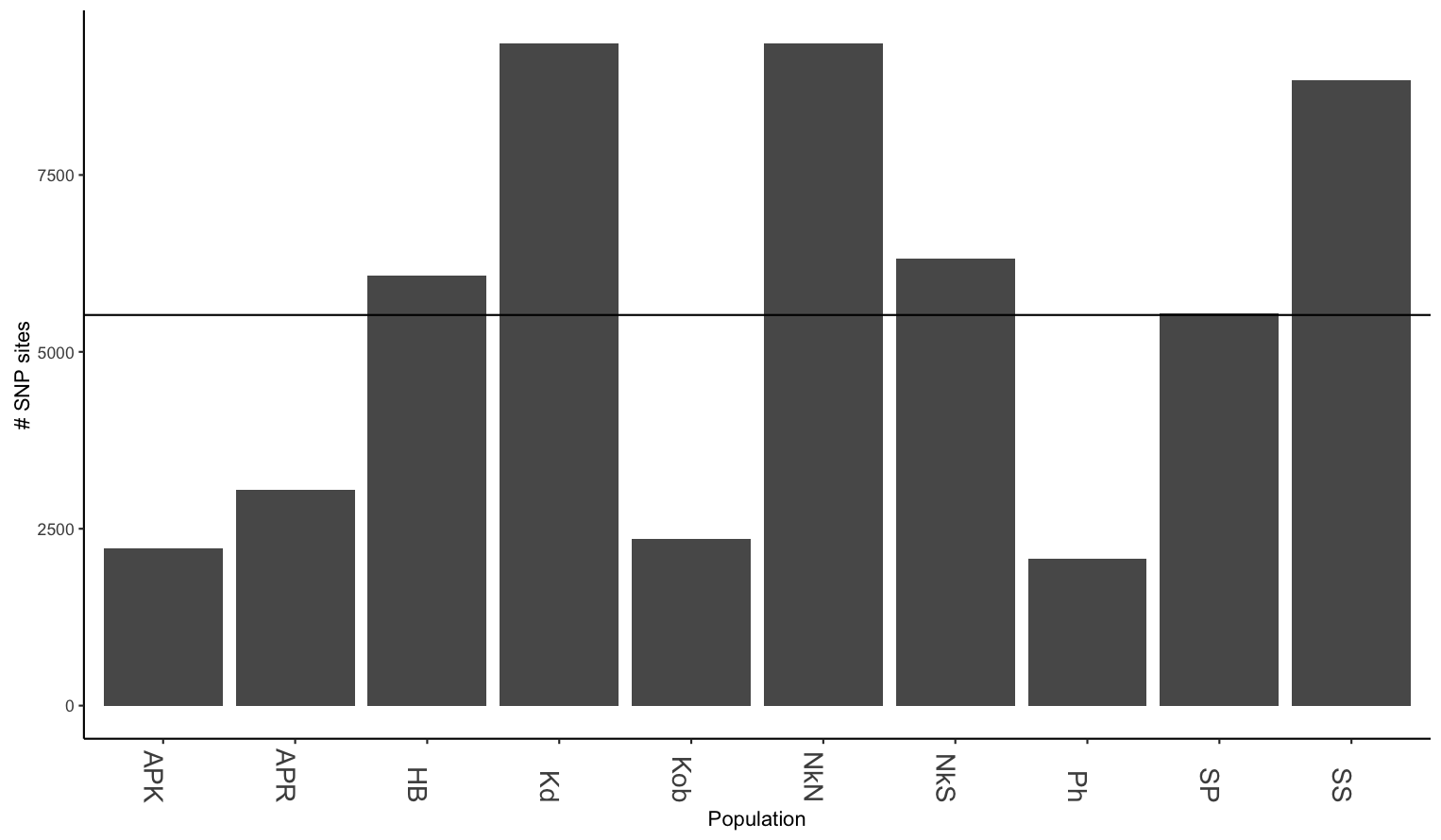


Figure 2. Number of SNP sites called within each populations.

*Overlap of sites among 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. Summary of populations which do not have any shared sites. E.g., Sp does not share any sites with APR or HB.

|  |  |
| --- | --- |
| 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.

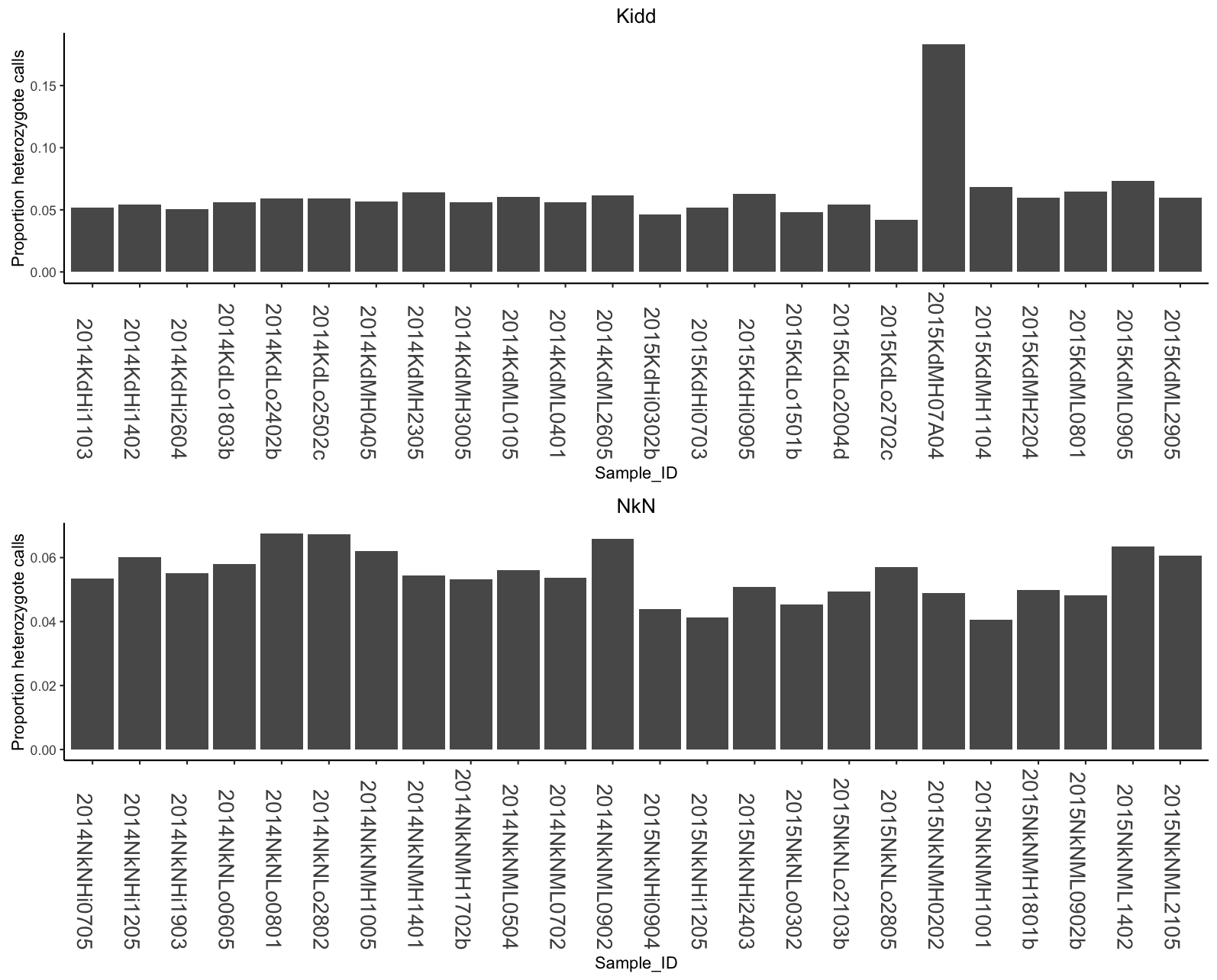
The average sample depth was not overly low. Mastretta-Yanes et al., 2015 obtained average depth per sample of 10.32 and considered that low. I don’t know what would be considered high. We achieved an average sample depth of 15.75 reads (ranging from 7.02 for 2015APKHi08 to 27.14 for 2014NkNML0902).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

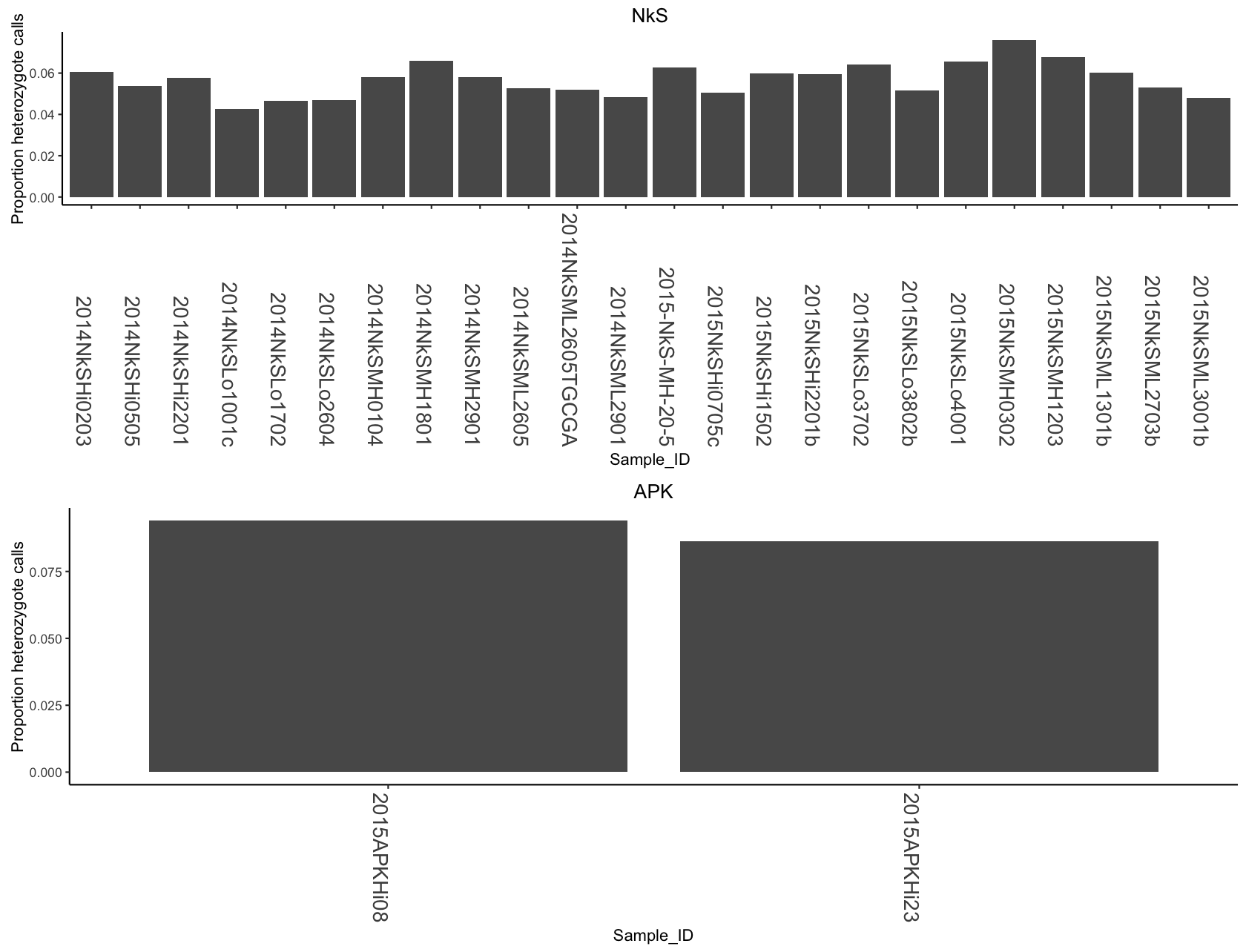
../depth/all_samples_depth.txt.pdf

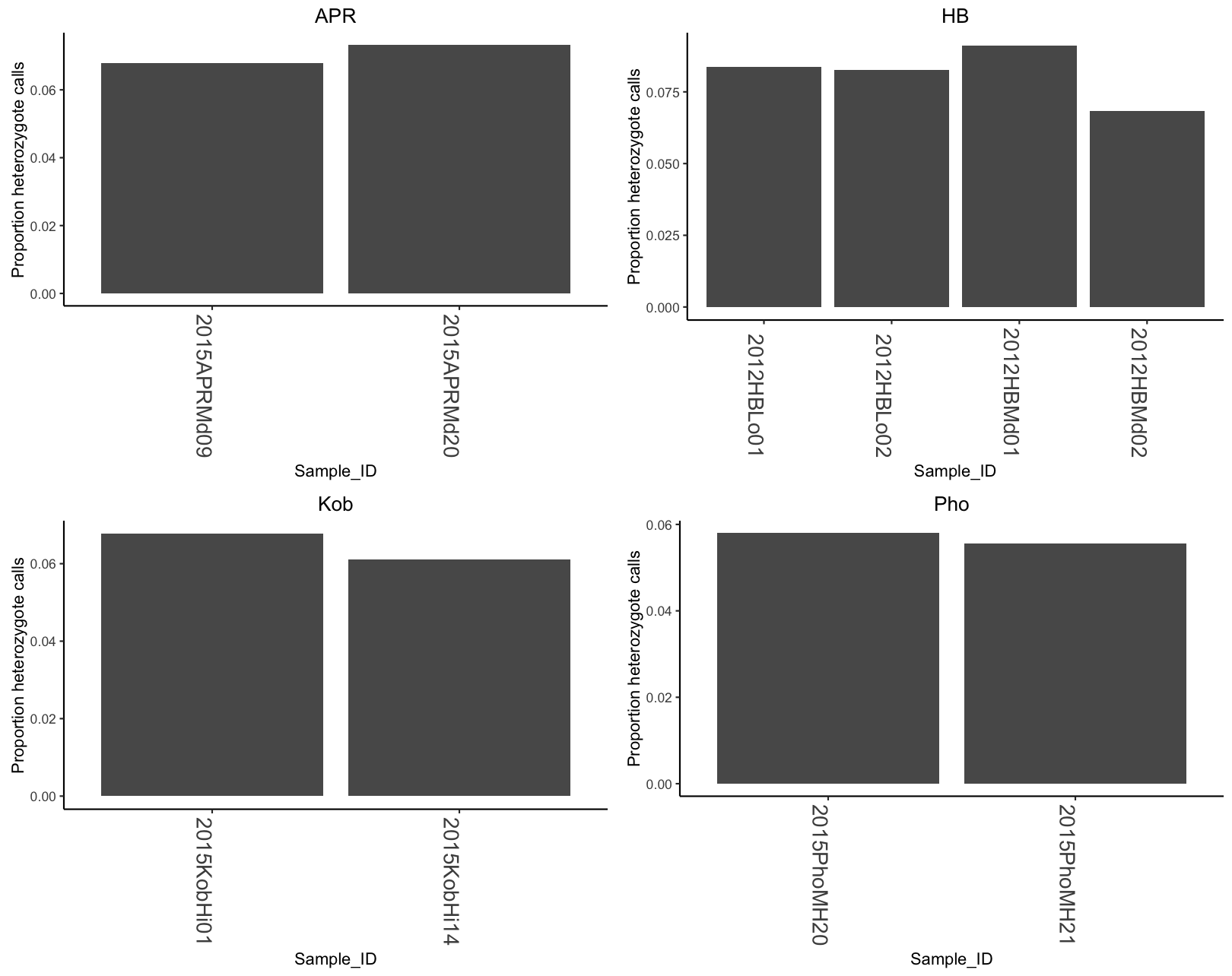
Figure 2. Distribution of sample depth across all samples after using default populations filter.

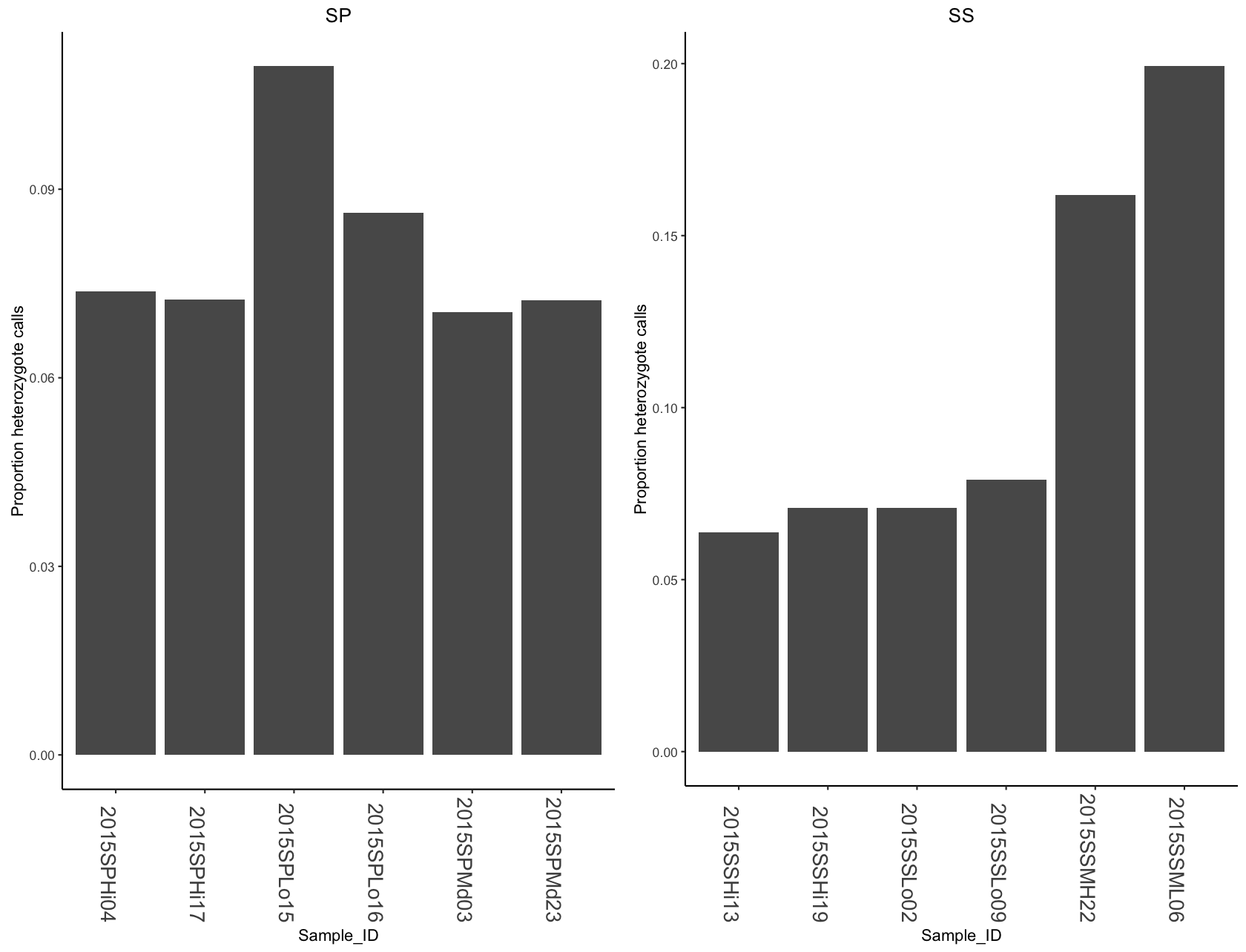
*Preliminary diversity estimates*

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 (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).







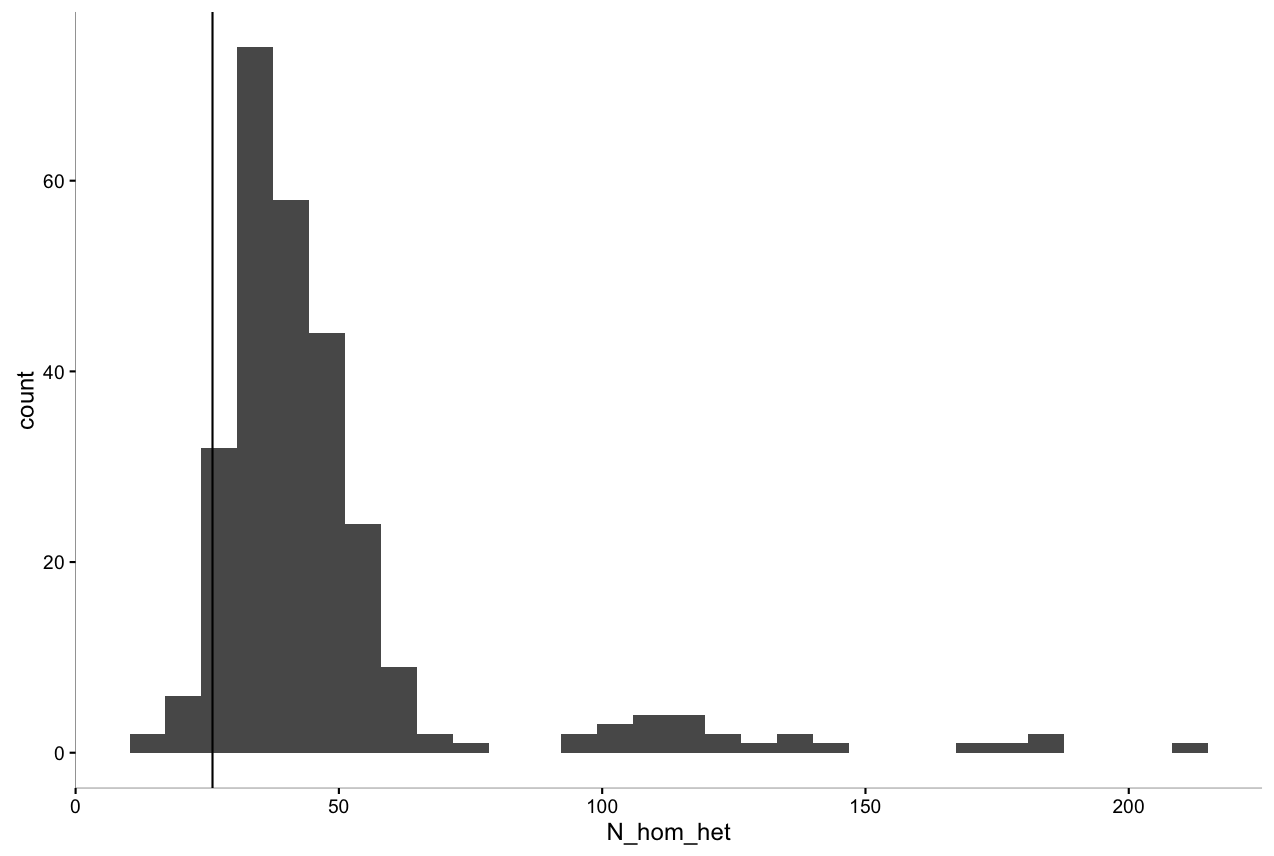


**Figure 3. Proportion of heterozygote calls for each sample.**

*Examining “duplicate” samples*

The two samples which are “duplicates” (2014NkSML2605TGCGA and 2014NkSML2605) have 107 homozygous to heterozygous differences and 26 homozygous differences. The het to hom differences could simply be an indication of genotype call error (i.e., a sample should be called as het but if the depth of coverage is too low it may be called as a homozygote). There are few homozygous to heterozygous call differences, but there are many homozygous differences (Figure 4 b). To me this suggests that these two samples are not duplicates. If they were duplicates there should be very few homozygous differences. If these are in fact duplicates the high number of homozygous differences indicates that stacks is doing a very poor job of calling genotypes. I could see if the depth or quality of these duplicated samples is different at the homozygous differences. My suspicion is that these samples are not duplicates, but do come from the same population.

a)



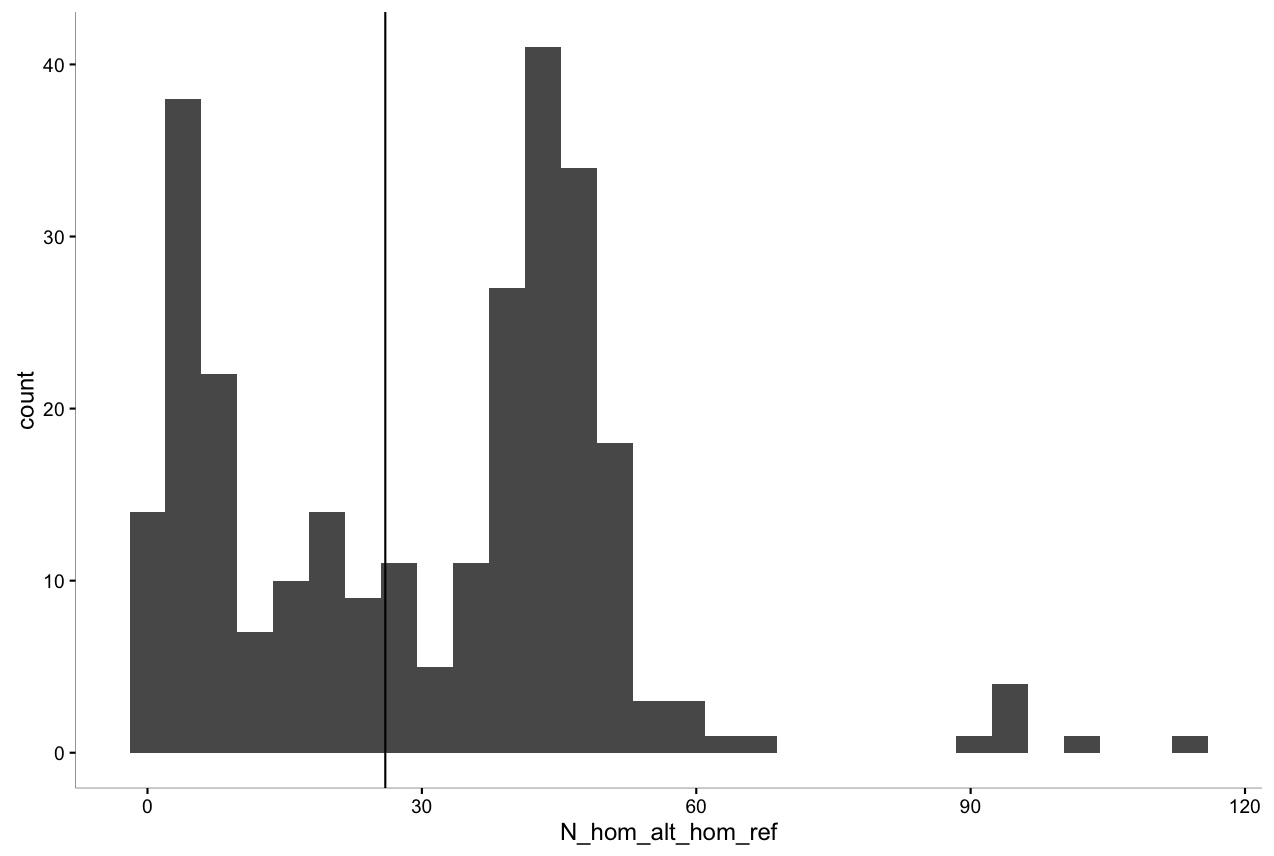
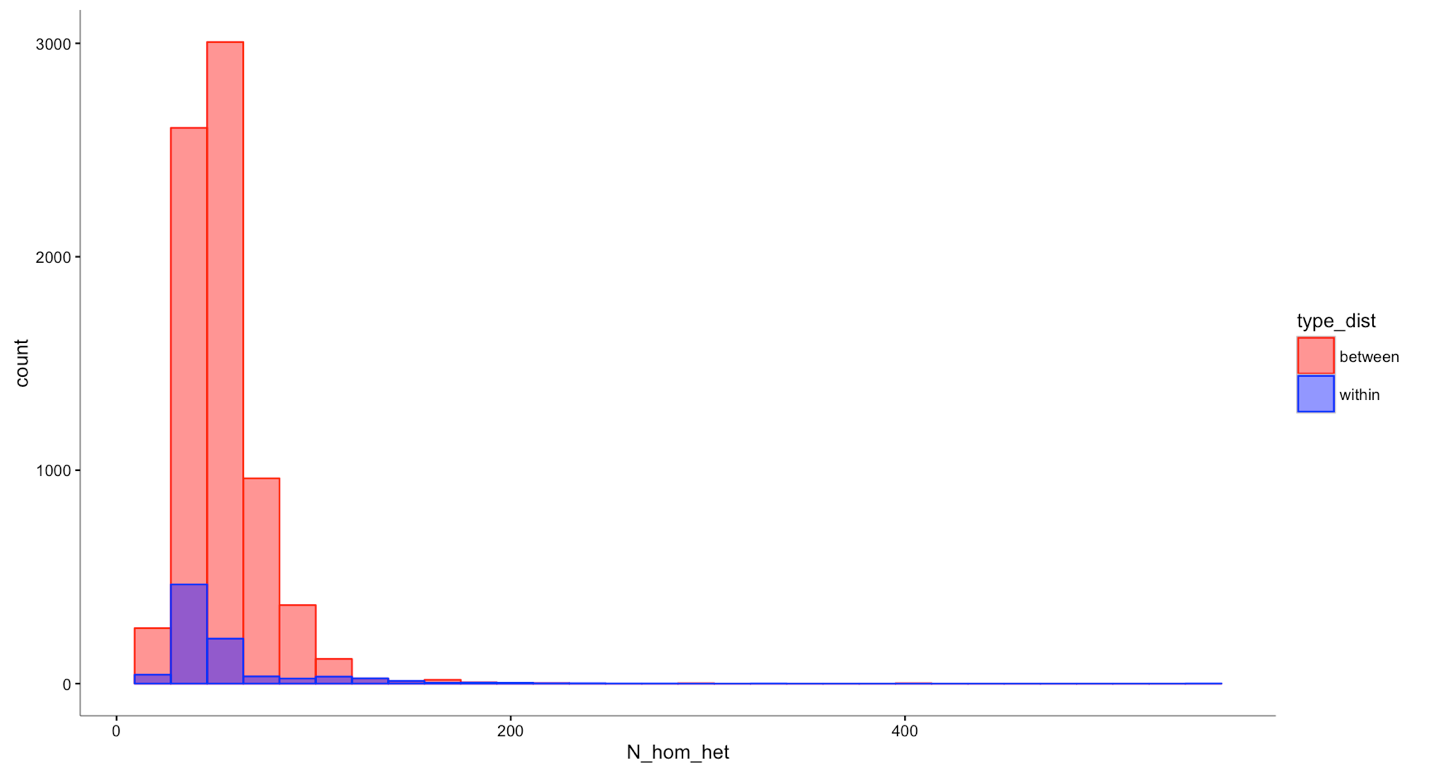
b) 

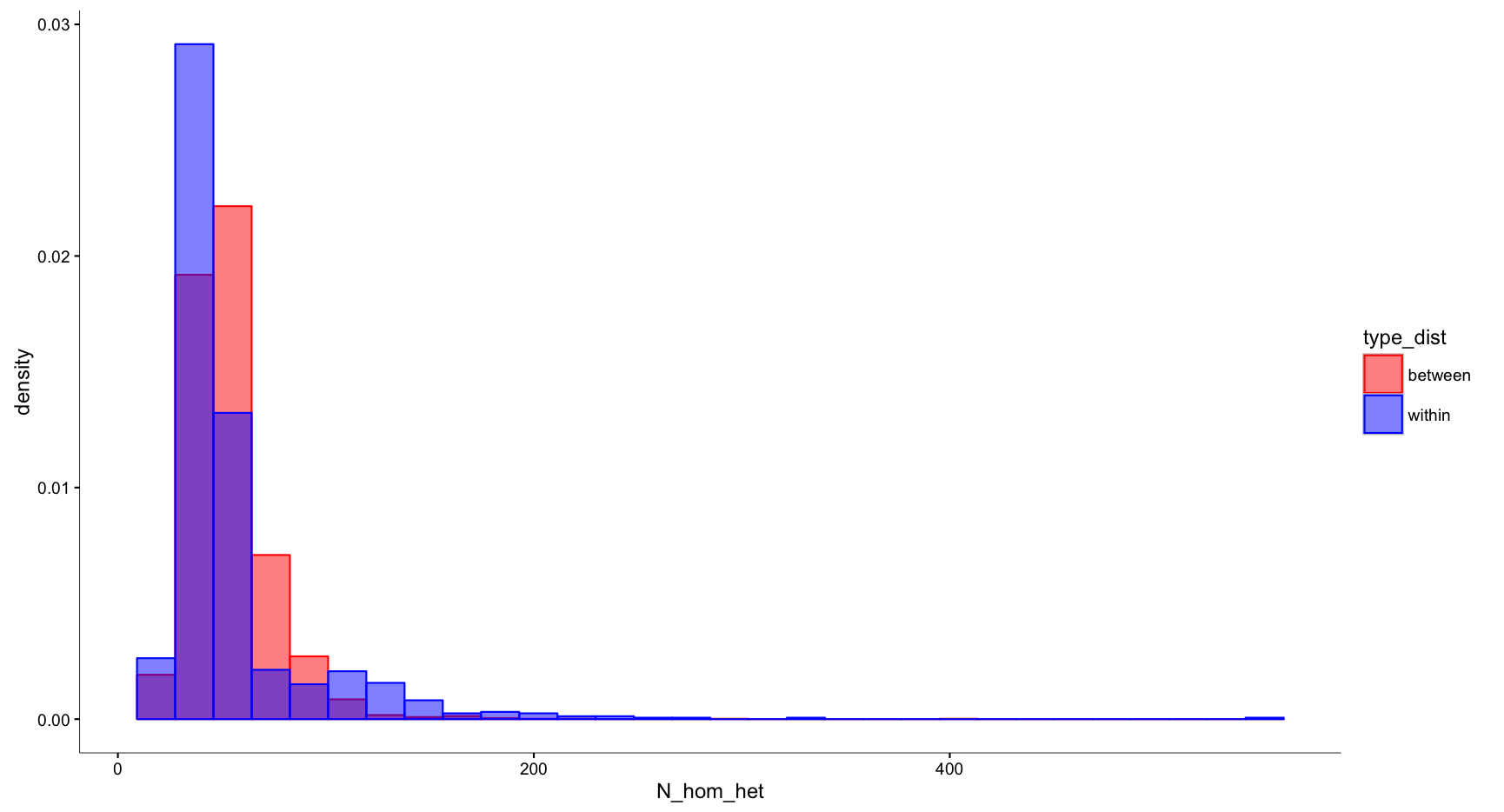
Figure 4. Distribution of a) homozygous to heterozygous differences and b) homozygous difference among samples in population NkS. The vertical line shows where the “duplicated” sample lies.

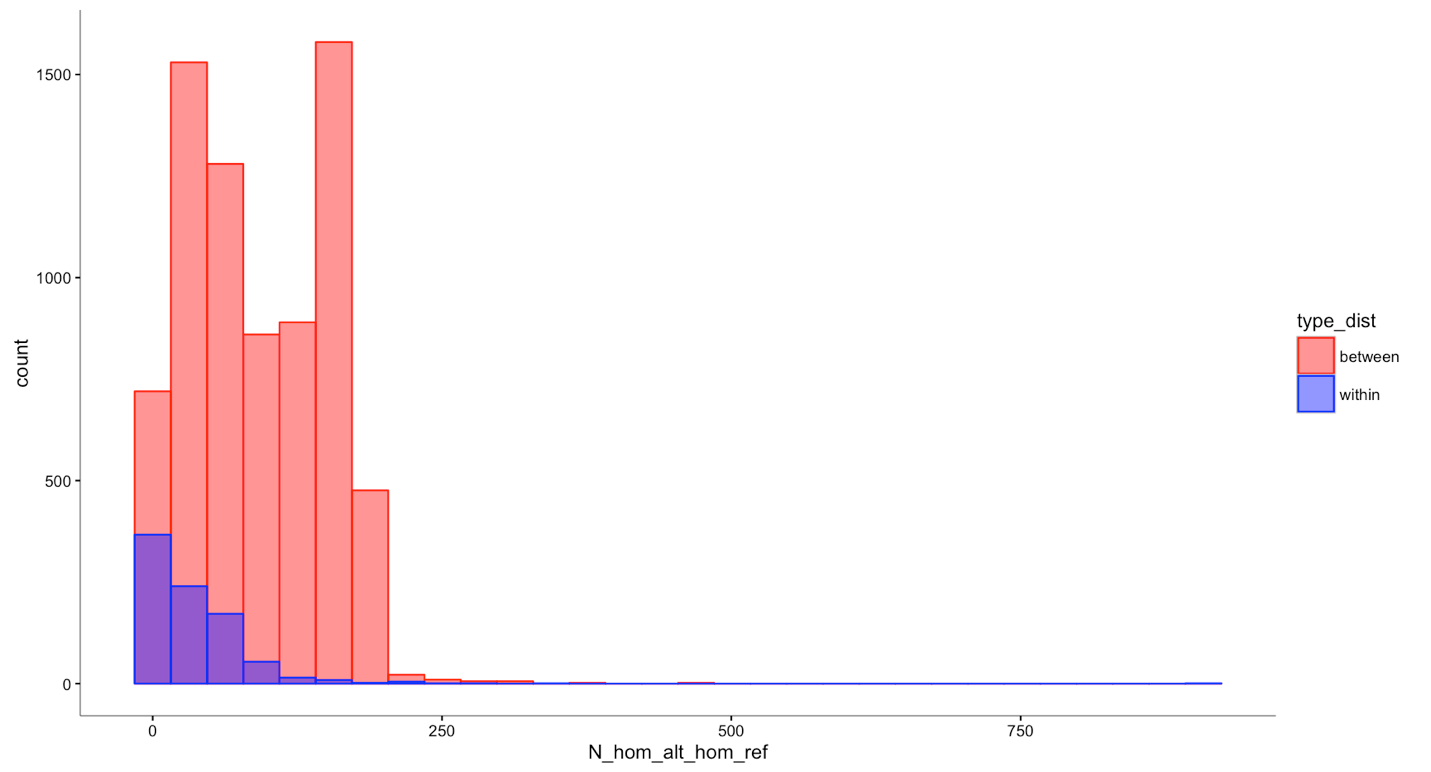
*Relationship among samples*

On average, there are 53.40 heterozygous to homozygous differences and an average of 81.59 homozygous differences between all samples (Table 7). Population SP is the most different (het to hom) from the other populations. There are fewer genotypic differences within populations (54-171 for het-hom and 22-188 hom differences in Kd and HB) than between population. Overall there are far fewer heterozygous to homozygous and homozygous differences within populations relative to between. This is exactly what we’d expect. I ran this little exploration to asses how many differences were there in genotypes between samples collected within versus among populations.

a) 

b)

 Figure 5. Distribution of het to hom differences within and between samples among populations.

1. 

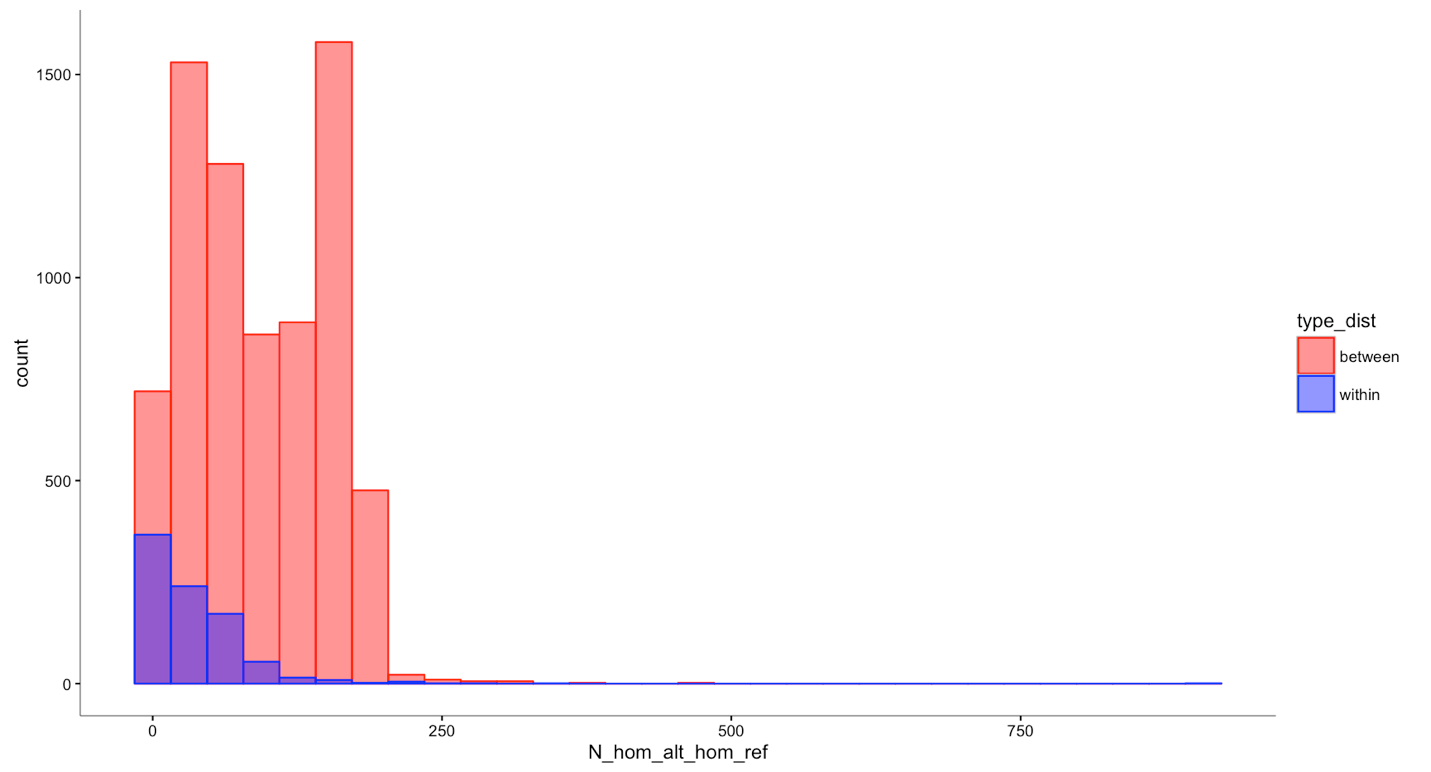


Figure 6. Distribution of homozygous differences within and between samples among populations.

Table 7. Mean number of genotypic differences among samples from different populations.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| population\_name | population\_name2 | N\_hom\_het | N\_hom\_alt\_hom\_ref | N\_same | N\_different | n |
| Kob | Kd | 54.6 | 143.1 | 1099.1 | 197.7 | 48 |
| Kob | SS | 61.2 | 101.0 | 1320.3 | 162.2 | 12 |
| Kob | APR | 49.5 | 67.5 | 1291.8 | 117.0 | 4 |
| Kob | SP | 61.7 | 108.8 | 1280.3 | 170.4 | 12 |
| Kob | NkN | 51.2 | 146.6 | 1067.9 | 197.8 | 48 |
| Kob | APK | 37.5 | 47.8 | 826.3 | 85.3 | 4 |
| Kob | NkS | 51.1 | 144.3 | 1066.7 | 195.4 | 48 |
| Kob | HB | 66.3 | 138.1 | 1074.9 | 204.4 | 8 |
| Kob | Ph | 46.5 | 91.8 | 1033.8 | 138.3 | 4 |
| Kd | Kob | 54.6 | 143.1 | 1099.1 | 197.7 | 48 |
| **Kd** | **SS** | **73.6** | **157.2** | **1275.2** | **230.8** | **144** |
| Kd | APR | 67.2 | 161.8 | 1164.0 | 229.0 | 48 |
| **Kd** | **SP** | **71.4** | **170.4** | **1243.0** | **241.8** | **144** |
| Kd | NkN | 46.3 | 52.6 | 1290.9 | 99.0 | 576 |
| Kd | APK | 46.6 | 102.9 | 683.6 | 149.5 | 48 |
| Kd | NkS | 44.2 | 30.9 | 1294.2 | 75.1 | 576 |
| Kd | HB | 61.9 | 74.9 | 1246.3 | 136.7 | 96 |
| Kd | Ph | 48.5 | 127.8 | 961.7 | 176.3 | 48 |
| SS | Kob | 61.2 | 101.0 | 1320.3 | 162.2 | 12 |
| SS | Kd | 73.6 | 157.2 | 1275.2 | 230.8 | 144 |
| SS | APR | 84.3 | 125.5 | 1534.8 | 209.8 | 12 |
| **SS** | **SP** | **100.1** | **100.3** | **1657.6** | **200.4** | 36 |
| SS | NkN | 66.0 | 155.8 | 1256.2 | 221.7 | 144 |
| SS | APK | 52.7 | 85.3 | 872.3 | 137.9 | 12 |
| SS | NkS | 63.1 | 150.7 | 1201.2 | 213.8 | 144 |
| SS | HB | 81.4 | 156.8 | 1290.4 | 238.3 | 24 |
| SS | Ph | 55.1 | 94.9 | 1114.8 | 150.0 | 12 |
| APR | Kob | 49.5 | 67.5 | 1291.8 | 117.0 | 4 |
| APR | Kd | 67.2 | 161.8 | 1164.0 | 229.0 | 48 |
| APR | SS | 84.3 | 125.5 | 1534.8 | 209.8 | 12 |
| **APR** | **SP** | **78.2** | **137.1** | **1425.8** | **215.3** | 12 |
| APR | NkN | 58.8 | 164.6 | 1133.0 | 223.4 | 48 |
| APR | APK | 67.0 | 61.5 | 938.0 | 128.5 | 4 |
| APR | NkS | 55.5 | 159.2 | 1080.7 | 214.6 | 48 |
| APR | HB | 68.3 | 156.3 | 1131.3 | 224.5 | 8 |
| APR | Ph | 52.5 | 105.5 | 1109.5 | 158.0 | 4 |
| SP | Kob | 61.7 | 108.8 | 1280.3 | 170.4 | 12 |
| **SP** | **Kd** | **71.4** | **170.4** | **1243.0** | **241.8** | **144** |
| **SP** | **SS** | **100.1** | **100.3** | **1657.6** | **200.4** | **36** |
| **SP** | **APR** | **78.2** | **137.1** | **1425.8** | **215.3** | **12** |
| SP | NkN | 64.4 | 171.9 | 1200.6 | 236.4 | 144 |
| SP | APK | 49.0 | 88.1 | 825.0 | 137.1 | 12 |
| SP | NkS | 63.3 | 164.0 | 1161.7 | 227.3 | 144 |
| SP | HB | 75.5 | 166.0 | 1199.1 | 241.5 | 24 |
| SP | Ph | 57.2 | 111.4 | 1078.6 | 168.6 | 12 |
| NkN | Kob | 51.2 | 146.6 | 1067.9 | 197.8 | 48 |
| NkN | Kd | 46.3 | 52.6 | 1290.9 | 99.0 | 576 |
| NkN | SS | 66.0 | 155.8 | 1256.2 | 221.7 | 144 |
| NkN | APR | 58.8 | 164.6 | 1133.0 | 223.4 | 48 |
| NkN | SP | 64.4 | 171.9 | 1200.6 | 236.4 | 144 |
| NkN | APK | 38.0 | 106.2 | 660.9 | 144.2 | 48 |
| NkN | NkS | 42.8 | 41.3 | 1255.1 | 84.1 | 576 |
| NkN | HB | 55.5 | 92.4 | 1188.1 | 147.9 | 96 |
| NkN | Ph | 45.9 | 126.1 | 931.1 | 172.0 | 48 |
| APK | Kob | 37.5 | 47.8 | 826.3 | 85.3 | 4 |
| APK | Kd | 46.6 | 102.9 | 683.6 | 149.5 | 48 |
| APK | SS | 52.7 | 85.3 | 872.3 | 137.9 | 12 |
| APK | APR | 67.0 | 61.5 | 938.0 | 128.5 | 4 |
| APK | SP | 49.0 | 88.1 | 825.0 | 137.1 | 12 |
| APK | NkN | 38.0 | 106.2 | 660.9 | 144.2 | 48 |
| APK | NkS | 37.6 | 103.5 | 654.5 | 141.1 | 48 |
| APK | HB | 46.5 | 108.9 | 708.3 | 155.4 | 8 |
| APK | Ph | 33.5 | 63.0 | 711.5 | 96.5 | 4 |
| NkS | Kob | 51.1 | 144.3 | 1066.7 | 195.4 | 48 |
| NkS | Kd | 44.2 | 30.9 | 1294.2 | 75.1 | 576 |
| NkS | SS | 63.1 | 150.7 | 1201.2 | 213.8 | 144 |
| NkS | APR | 55.5 | 159.2 | 1080.7 | 214.6 | 48 |
| NkS | SP | 63.3 | 164.0 | 1161.7 | 227.3 | 144 |
| NkS | NkN | 42.8 | 41.3 | 1255.1 | 84.1 | 576 |
| NkS | APK | 37.6 | 103.5 | 654.5 | 141.1 | 48 |
| NkS | HB | 55.4 | 81.2 | 1156.7 | 136.6 | 96 |
| NkS | Ph | 41.9 | 122.8 | 912.8 | 164.6 | 48 |
| HB | Kob | 66.3 | 138.1 | 1074.9 | 204.4 | 8 |
| HB | Kd | 61.9 | 74.9 | 1246.3 | 136.7 | 96 |
| HB | SS | 81.4 | 156.8 | 1290.4 | 238.3 | 24 |
| HB | APR | 68.3 | 156.3 | 1131.3 | 224.5 | 8 |
| **HB** | **SP** | **75.5** | **166.0** | **1199.1** | **241.5** | **24** |
| HB | NkN | 55.5 | 92.4 | 1188.1 | 147.9 | 96 |
| HB | APK | 46.5 | 108.9 | 708.3 | 155.4 | 8 |
| HB | NkS | 55.4 | 81.2 | 1156.7 | 136.6 | 96 |
| HB | Ph | 46.5 | 128.5 | 894.9 | 175.0 | 8 |
| Ph | Kob | 46.5 | 91.8 | 1033.8 | 138.3 | 4 |
| Ph | Kd | 48.5 | 127.8 | 961.7 | 176.3 | 48 |
| Ph | SS | 55.1 | 94.9 | 1114.8 | 150.0 | 12 |
| Ph | APR | 52.5 | 105.5 | 1109.5 | 158.0 | 4 |
| Ph | SP | 57.2 | 111.4 | 1078.6 | 168.6 | 12 |
| Ph | NkN | 45.9 | 126.1 | 931.1 | 172.0 | 48 |
| Ph | APK | 33.5 | 63.0 | 711.5 | 96.5 | 4 |
| Ph | NkS | 41.9 | 122.8 | 912.8 | 164.6 | 48 |
| Ph | HB | 46.5 | 128.5 | 894.9 | 175.0 | 8 |

Table 8. Comparison of mean number of genotypic difference among samples within population groups

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Population name | Num het to hom | Num homozygous differences | Total | Num of Samples |
| Kob | 119 | 37 | 156 | 2 |
| **Kd** | **54.40942** | **22.42** | **76.83** | **24** |
| SS | 122.26667 | 48.53 | 170.8 | 6 |
| APR | 219 | 136 | 355 | 2 |
| SP | 124.13333 | 87 | 211.13 | 6 |
| **NkN** | **49.86594** | **48.82** | **98.69** | **24** |
| APK | 250 | 343 | 593 | 2 |
| **NkS** | **47.73188** | **30.43478** | **78.17** | **24** |
| HB | 171.33333 | 188.33 | 359.67 | 4 |
| Ph | 118 | 59 | 177 | 2 |

Kidd and Nikkiska South cluster highly together. SS and APK have greater number of differences to the other populations. This adegenet package in R uses the average allele frequency to fill in data. (see <http://lists.r-forge.r-project.org/pipermail/adegenet-forum/2014-March/000815.html)>

This approach strikes could be problematic here (given the high number of sites where there is no data for a given population) and could be why the two duplicated samples do not cluster perfectly together. If sites are missing in one of the duplicated samples and vice versa but that are not the average allele, then the algorithm will consider this to be a difference between the two samples.

I’m going to find how many genotypic differences there are between samples scaled by the number of shared sites. If sequencing error is low then the two duplicates should have identical genotype calls for shared sites.

One approach to deal with this issue is to simply downsample the data. Rather than using all samples within a pop we can use average genotype or some representation of genotype within each pop when we do between pop comparisons. At least for a NJ tree using population mean rather than total mean allele frequency at a site would be better for sites where there is no data for any sample within the population then the average sample wide allele frequency is a fine substitute.

Also because doing Fst type stats for a few samples for each pop instead I might just do number of genotype and allele differences between all pairs of samples and then do average difference within groups versus between groups.

****

**Notes on how to recreate analyses/plots etc.**

*Read counts*

I used a custom Unix script to count number of reads and length of each read for each sample following demultiplexing. Tables 1-3 and Figure 1. The summary tables and Figure were made using the Rscript “sample\_depth.R”

|  |
| --- |
| 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

*Counting genotype calls per sample and population*

I converted the vcf file output by stacks “population” to a summary file format using my script called “convert\_vcf\_to summary.py”.The summary file format simplified the output of a vcf file. Each row is information about a SNP site. Each column is the genotype call for a specific sample. The summary file format 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 the summary file called “batch\_1.summary”. The output of my script is called (this is housed in the output folder on the github project directory and is housed among the stacks output on the server):

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

The output is also presented in Table S5

I imported the file "output/Het\_Hom\_alt\_Missing\_counts\_by\_sample.txt"

into R and create basic summaries of the genotype calls per sample using a script called “preliminary\_diversity.R”. I use this R script “preliminary\_diversity.R” to make Table 4 and Figure 2.

I made another python script called “Missing\_sites\_by\_pop.py” to count the number of missing sites for a population. This script counts missing sites for population if there is no sample within that population that has a genotype call. This script takes in the summary file and outputs a short text file that I called “missing\_sites.by\_pop.txt” which is housed in the output/summary\_stats/ folder. This is used in the rscript called “preliminary\_diversity.R” to make Figure S1 and Table S4.

*Overlap of sites among populations*

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”.

This script pulls in the stacks output file called “summarized\_sumstats\_tsv.txt” and does the following tasks:

1. finds which populations have a representative sample at each site.
2. prints out for each population how many sites are unique to this population and how many sites are called for each population
3. for each population it assesses which populations do not co occur at any site.

This is used to make Tables 5 and 6

*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 also updated this script so it can run on the vcf file output.

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

*Preliminary diversity estimates*

I basically used the output of heterozygous calls from the python script described aboe that counts genotype calls for each sample. The output it called “Het\_Hom\_alt\_Missing\_counts\_by\_sample.txt”

I use the Rscript called “preliminary\_diverstiy.R” to plot heterozygosity per sample and cluster the plots by population.

*Examining “duplicate” samples*

*Relationship among samples*