The power and promise of genetic mapping from *Plasmodium falciparum* crosses utilizing human liver-chimeric mice

Katrina A. Button-Simons1\*, Sudhir Kumar2, Nelly Carmago2, Meseret T. Haile2, Catherine Jett3, Lisa A. Checkley1, Spencer Y. Kennedy2, Richard S. Pinapati4, Douglas A. Shoue1, Marina McDew-White5, Xue Li5, François H. Nosten6,7, Stefan H. Kappe2, Timothy J. C. Anderson5, Jeanne Romero-Severson8, Michael T. Ferdig1, Scott J. Emrich9, Ashley M. Vaughan2, Ian H. Cheeseman3\*

**Abstract**

Genetic crosses are most powerful for linkage analysis when progeny numbers are high, when parental alleles segregate evenly and, for hermaphroditic organisms, when numbers of inbred progeny are minimized. We previously developed a novel genetic crossing platform for the human malaria parasite *Plasmodium falciparum*, an obligately sexual, hermaphroditic protozoan, using mice carrying human hepatocytes (the human liver-chimeric FRG NOD huHep mouse) as the vertebrate host for the liver stage-to-blood stage transition. Here we examine the statistical power of two different genetic crosses – (1) an allopatric cross between a laboratory-adapted parasite (NF54) of African origin and a recently patient-derived Asian parasite, and (2) a sympatric cross between two recently patient-derived Asian parasites. We generated 144 unique recombinant clones over a 12-month period from the two crosses, doubling the number of unique recombinant progeny generated in the previous 30 years. Each cross shows distinct segregation patterns. The allopatric African/Asian cross has minimal levels of inbreeding (2% of clonal progeny are inbred) and extreme skews in marker segregation, while in the sympatric Asian cross, inbred progeny predominate (66% of clonal progeny are inbred) and parental alleles segregate evenly. Both new crosses show bi-parental inheritance of plastid markers amongst recombinant progeny. Using simulations, we demonstrate that the progeny provide the power to map large-effect mutations to a 31 kb interval and can capture complex, epistatic interactions. The segregation distortion in the allopatric African/Asian cross erodes power to detect linkage in several genome regions, but the repeatable distortions observed offer promising alternative approaches to identifying genes underlying traits of interest. We greatly increase the power and precision to map biomedically important traits with these new large progeny panels.

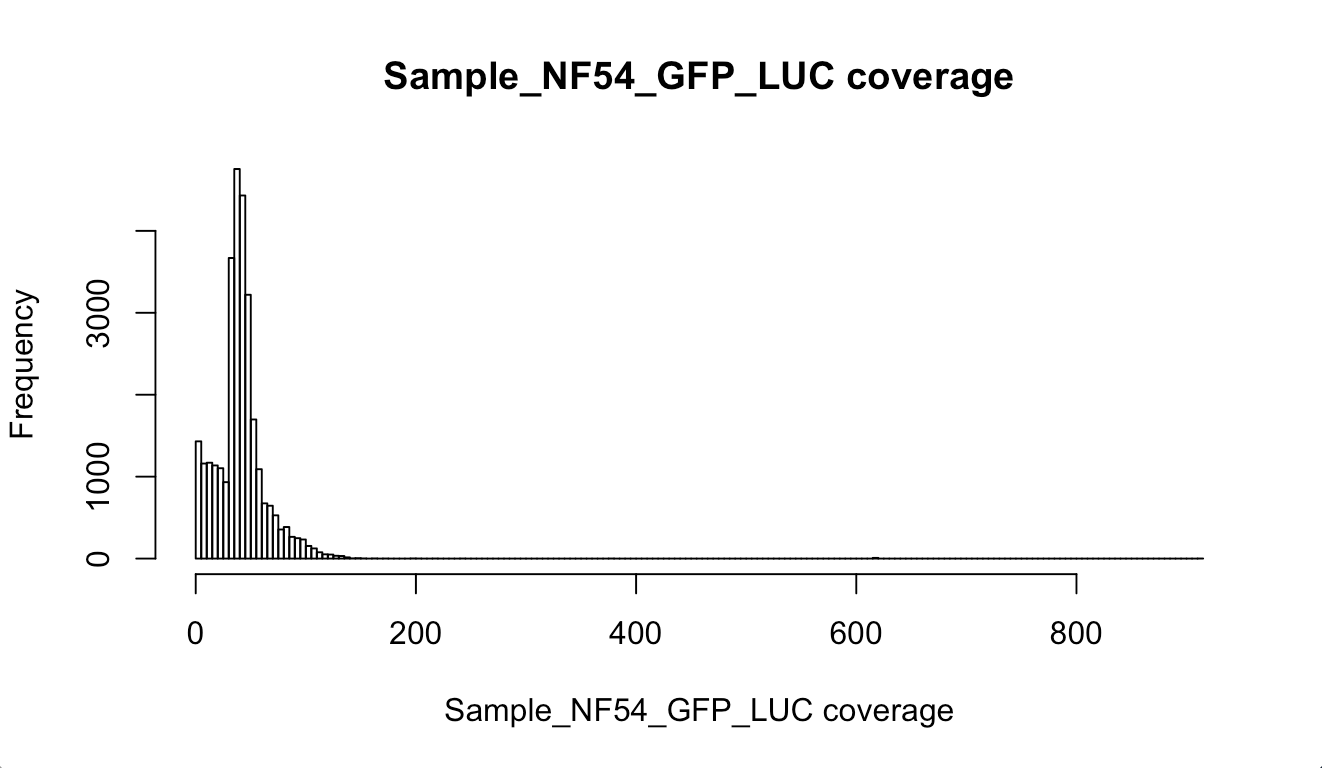
Filtering SNPs for NF54xNHP4026 cross

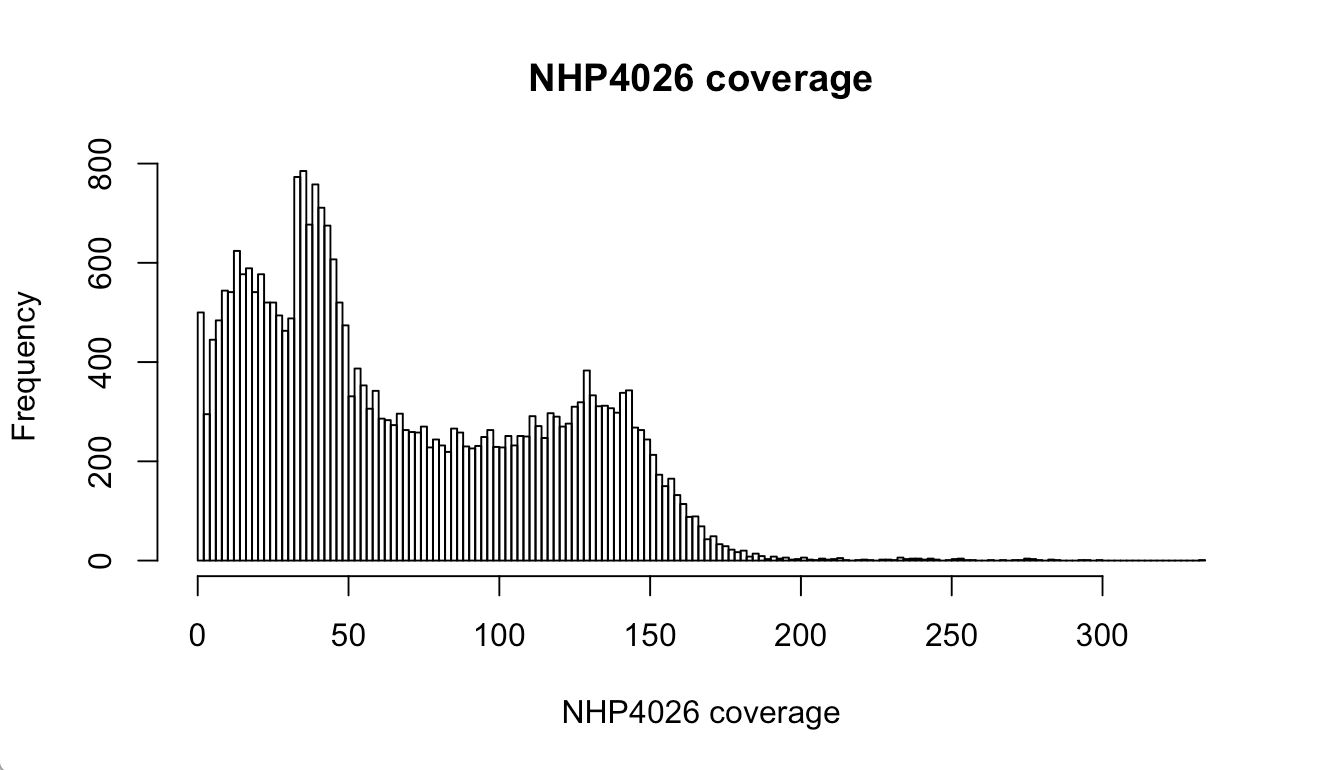
Run file Progeny\_Characterization\_Pipeline\_NF54xNHP4026.R code

Read in vcf file NF54\_NHP4026\_combo.snps.recal.sel.vcf

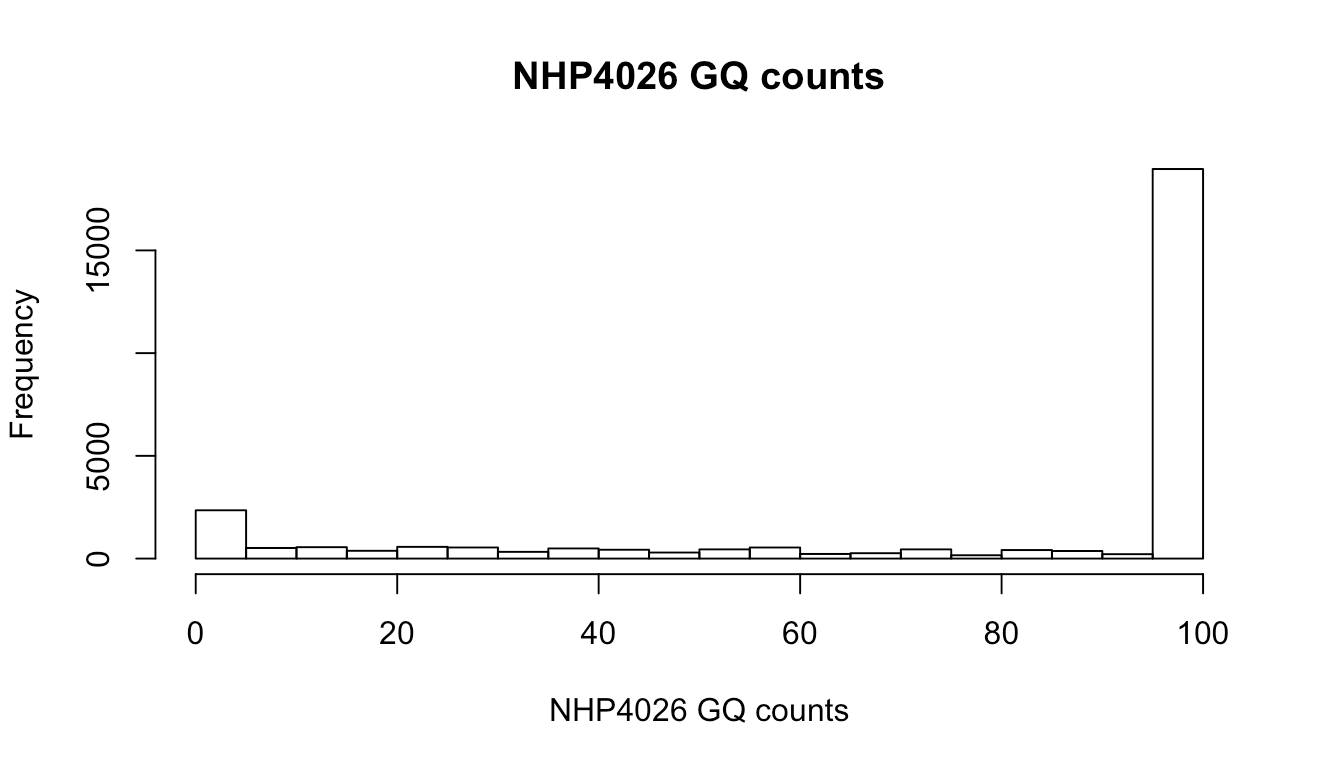
Filtering SNPs to get analysis ready set for NF54xNHP4026

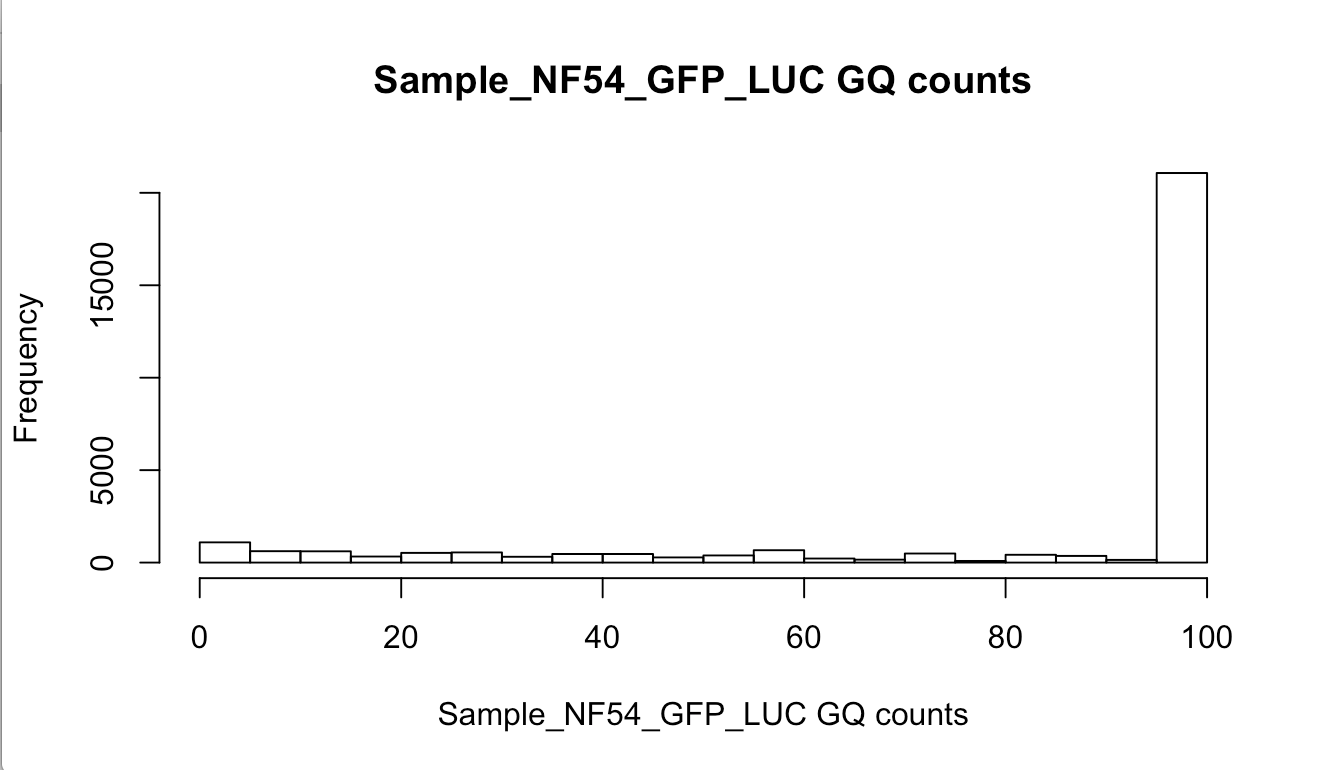
1. Filtering parental SNPs, include only if
   1. Homozygous and bi-allelic in parents
   2. Coverage ≥ 10



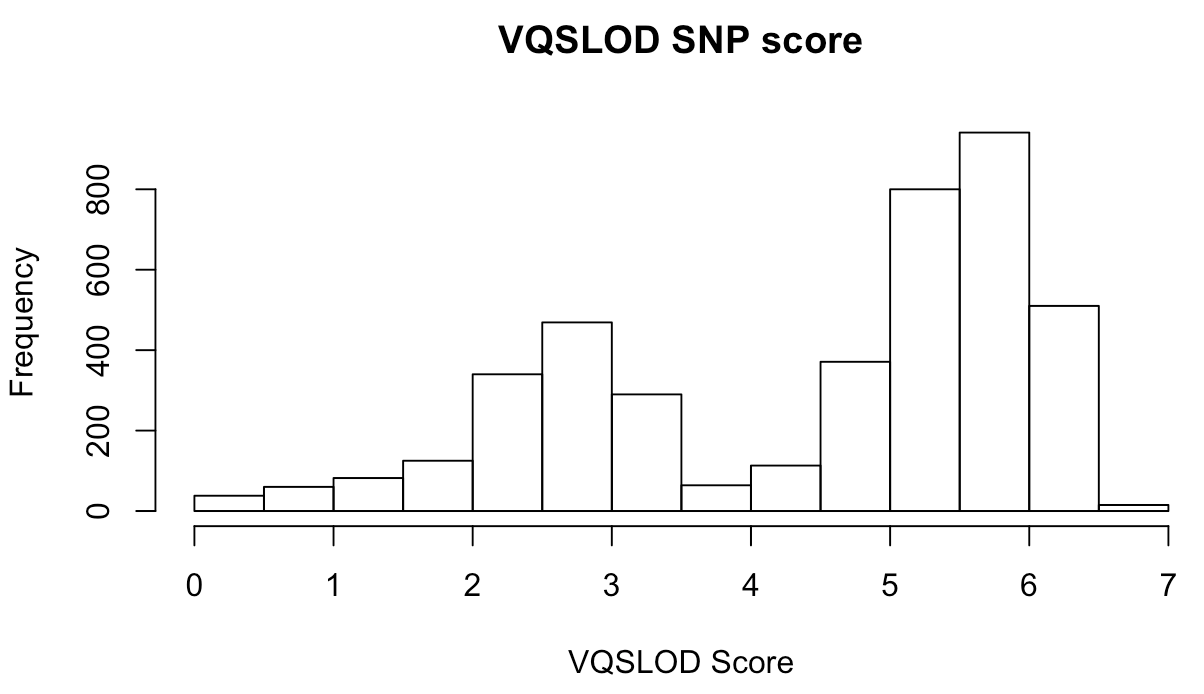


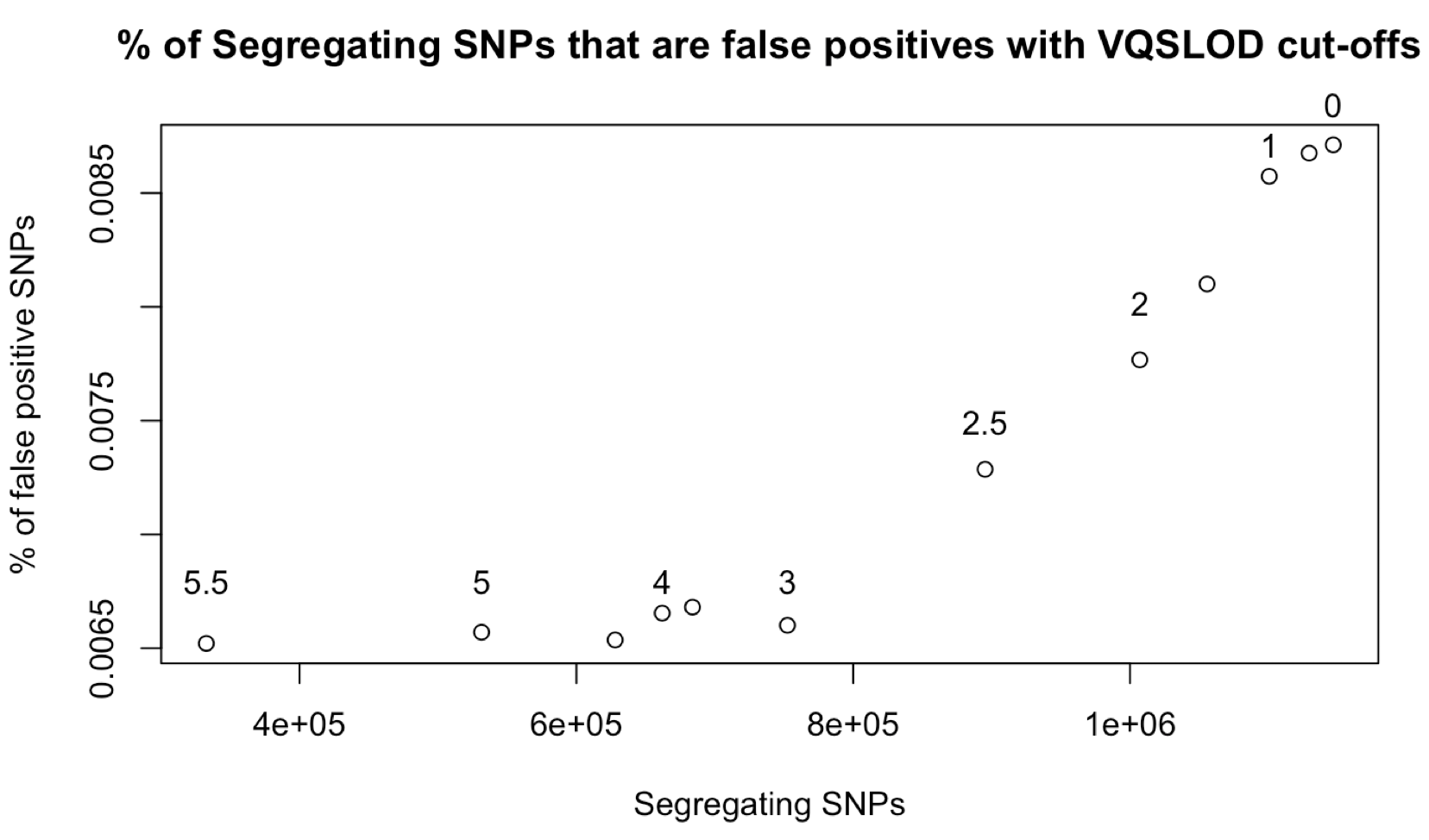
* 1. GQ ≥ 99





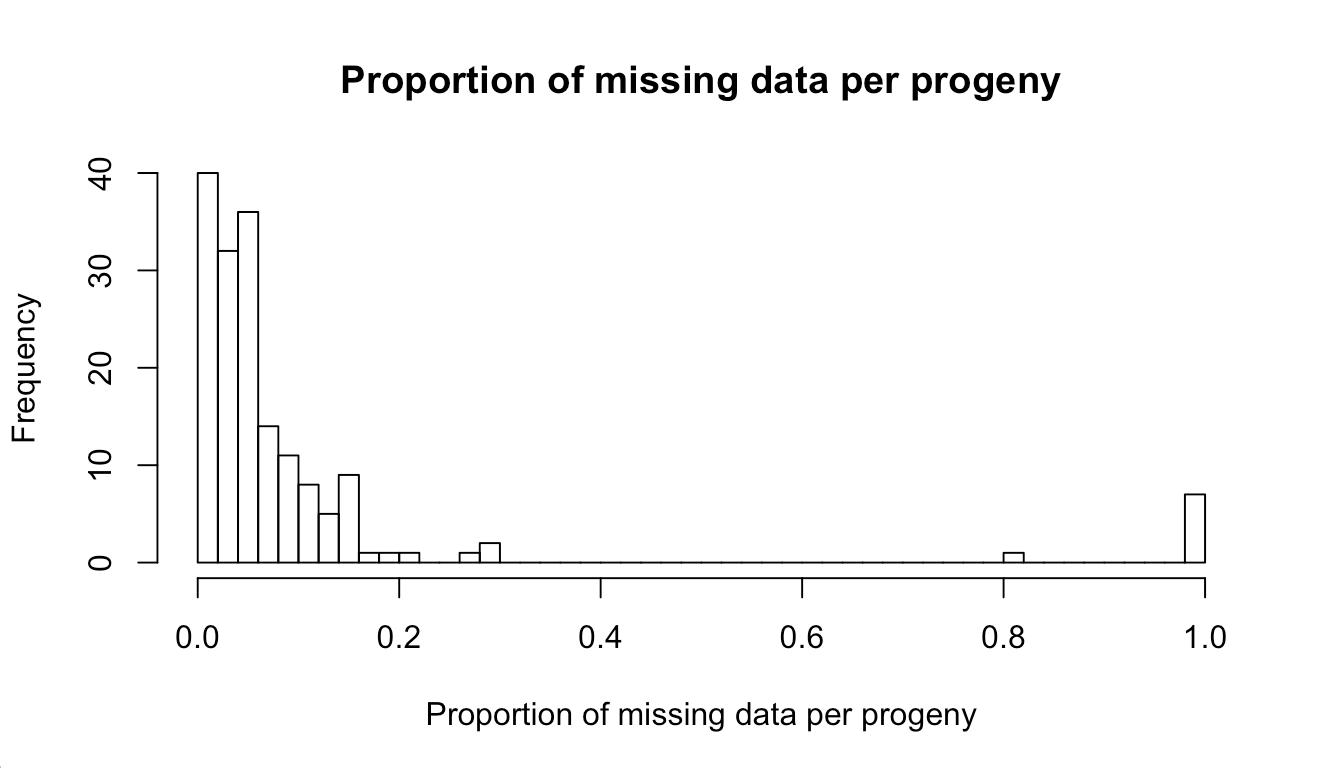
1. Progeny SNP filtering based on VQSLOD
   1. Choose VQSLOD cut-off similar to Miles et al. but using heterozygous SNP calls in clonal progeny as measure of sequencing error – cut-off is inflection point in curve of % of SNP calls that have sequencing error for all segregating markers in a set defined by different VQSLOD cut-offs (can update this later to base it on non-mendelian markers)
   2. Filter SNPs with VQSLOD < 2.5



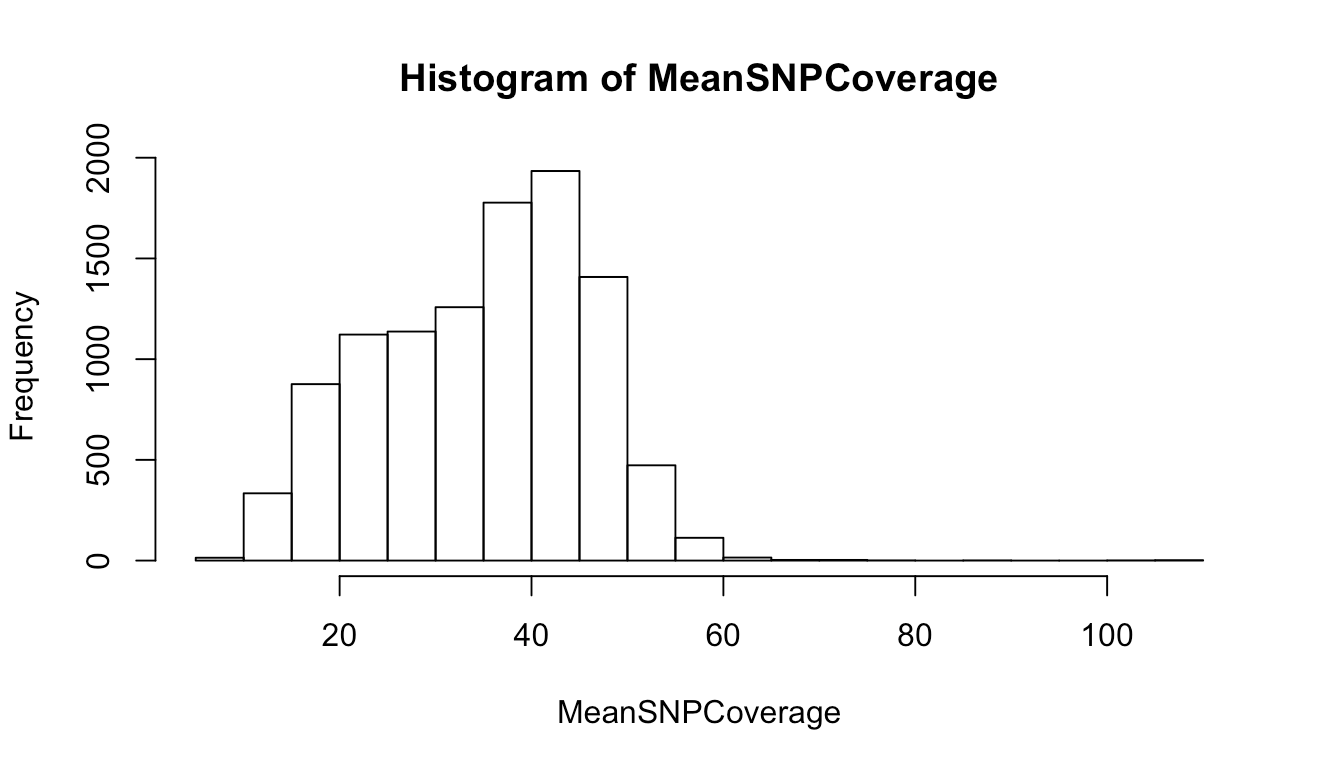


Filter Progeny

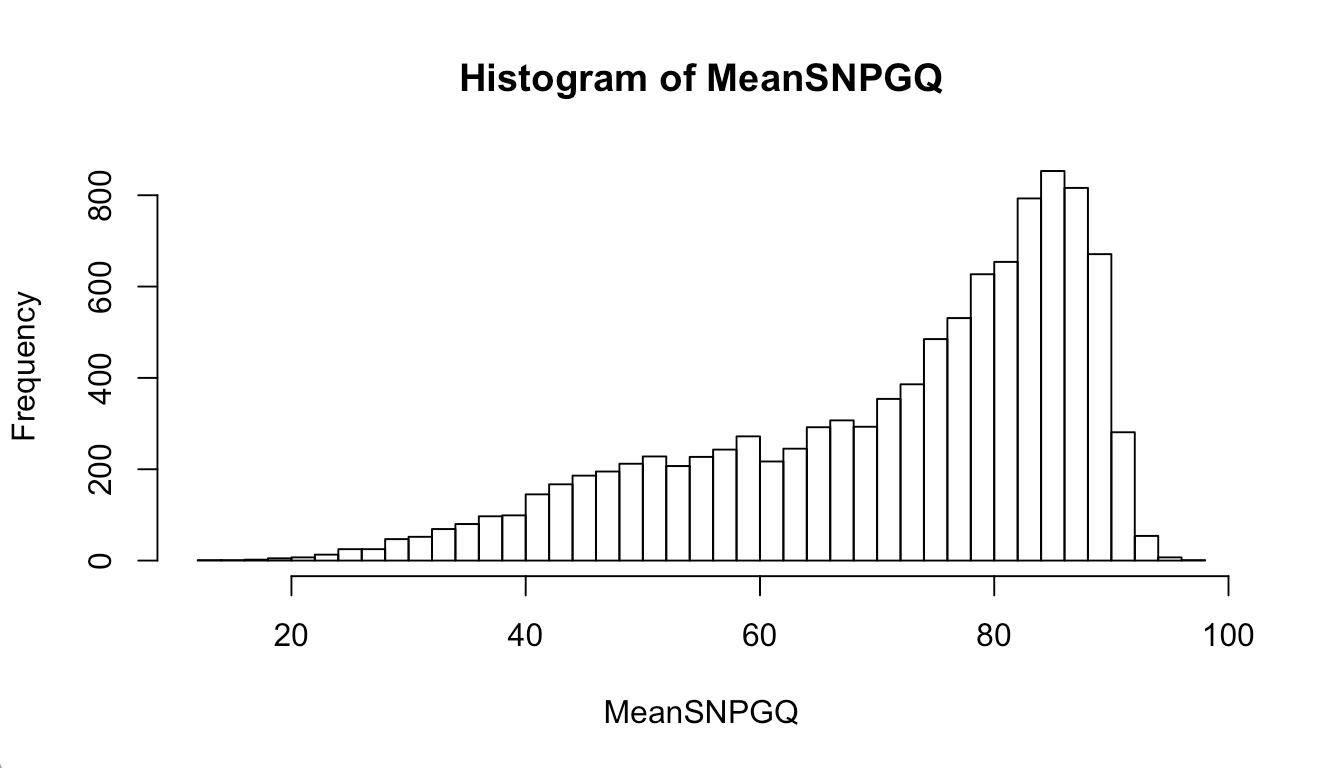
1. Filter progeny based on % missing data, coverage and GQ score (this ends of filtering out the same progeny, ie. those with high missing data have low coverage and low GQ scores)
   1. Keep progeny with Missing Data % < 80%



* 1. Mean coverage > 3



* 1. Mean GQ > 10



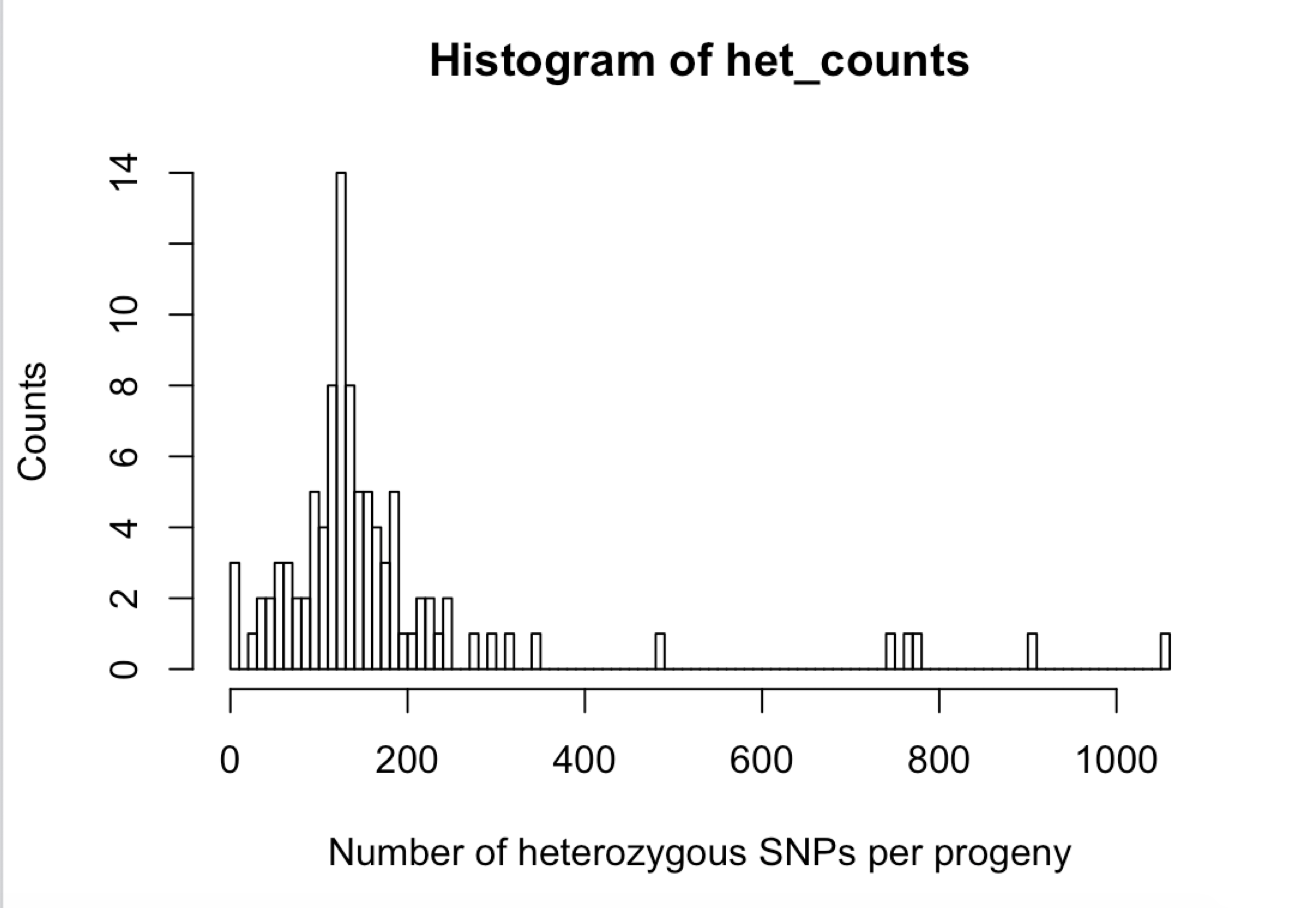
1. Filter non-clonal progeny defined as having any windows with above background heterozygosity

Coding SNP calls for final cleaned data

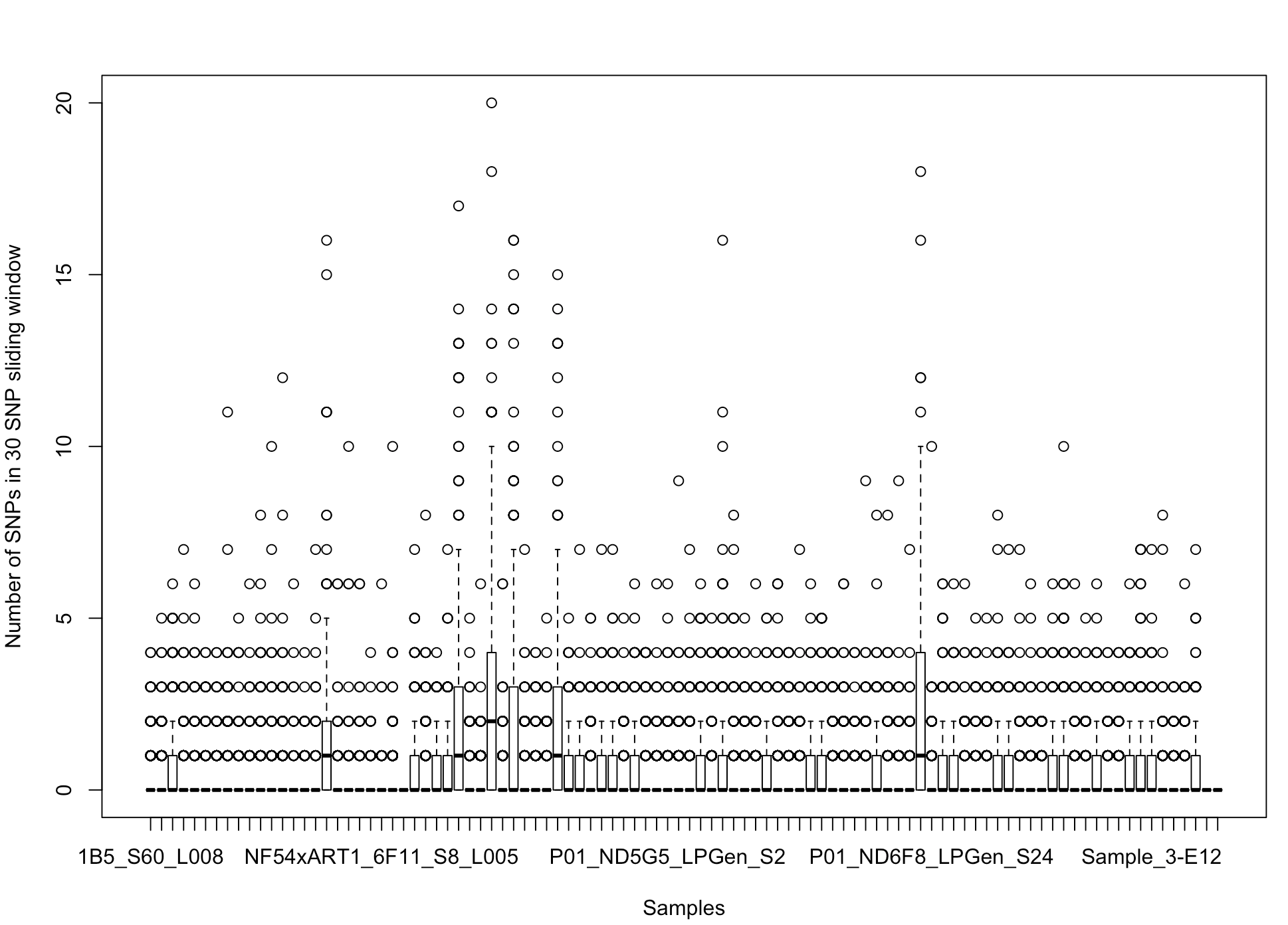
1. SNPs in progeny matching NF54 or NF54GFPLuc parent are coded as 0
2. SNPs in progeny matching NHP4026 parent are coded as 1
3. SNPs with “0/1”,”1/0”, “./1”, “./0”, “1/.”, “0/.” and “./.” are coded as NA

Identify unique recombinant progeny

1. Set cut-off for relatedness amoung progeny – cut off set at 0.9% similarity
   1. Choose single progeny and generate dataset repeating this progenies sequence with same total number of potential progeny
   2. For each simulated progeny, choose a percent heterozygous sequencing error rate distribution for clonal progeny and recode a random sampling of SNPs as other genotype
   3. Calculate % sequence similarity across all simulated progeny
2. Cluster progeny and include 1 from each cluster in final call set (this is done on coded data, not raw SNP calls)
3. For parents, identify homozygous calls, which are biallelic
   1. Subset of SNPs that are homozygous and biallelic in a vector of row numbers from vcf file
4. VCF file
   1. For each of the progeny and each row in the vector of row numbers for homozygous, biallelic SNPs we want to know if the SNP call is a heterozygous save in a matrix with columns as progeny and rows
   2. Compare counts of heterozygous SNPs for each progeny and identify outliers



1. Sliding window analysis
   1. Define window size (by number of SNPs) and count heterozygous SNP calls in a window, identify regions with high number of heterozygous SNPs



* 1. Define window based on actual location of SNPs and average cross-over length from previous crosses

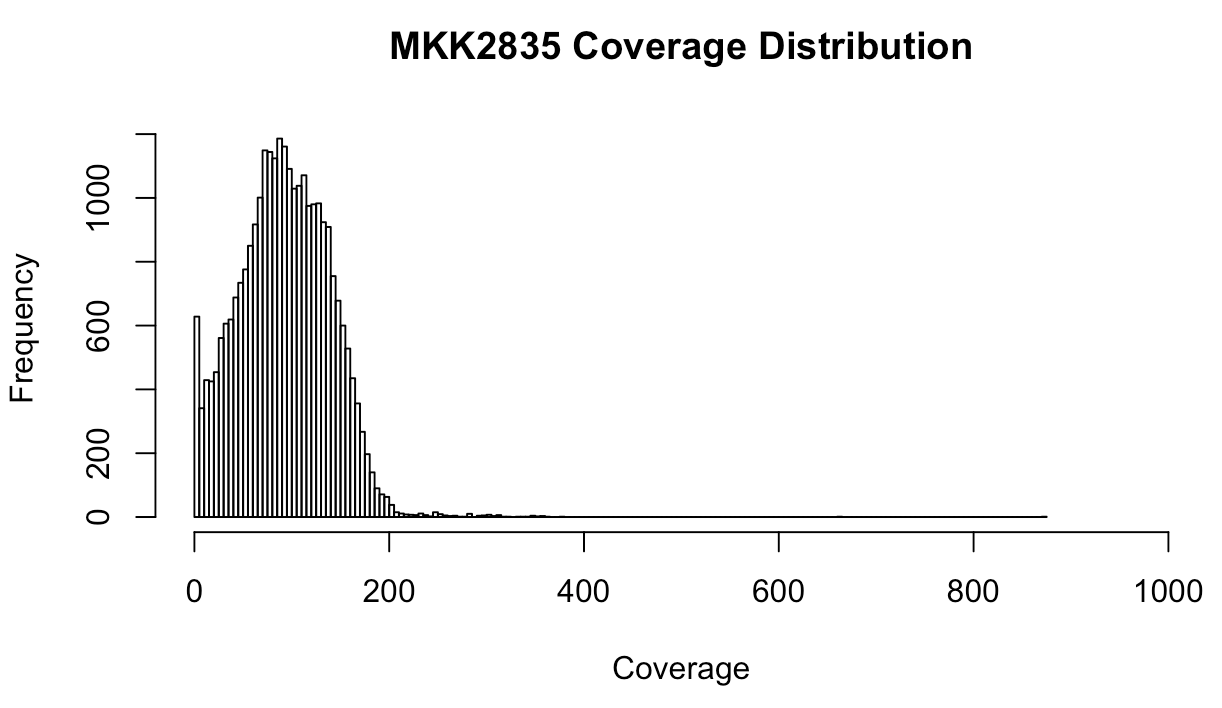
Filtering SNPs for MKK2835xNHP1337 cross

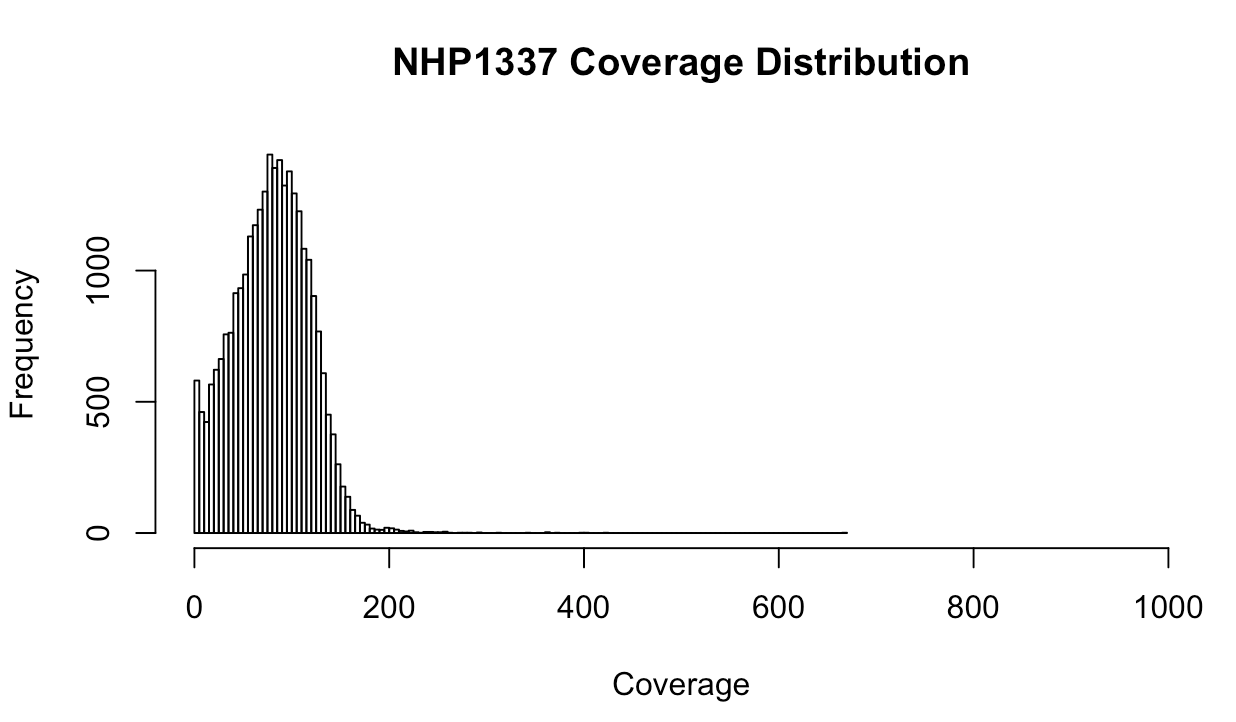
Run file Progeny\_Characterization\_Pipeline\_MKK2835xNHP1337.R code

Starting with vcf file MKK2835\_NHP1337\_combo.snps.recal.sel.vcf

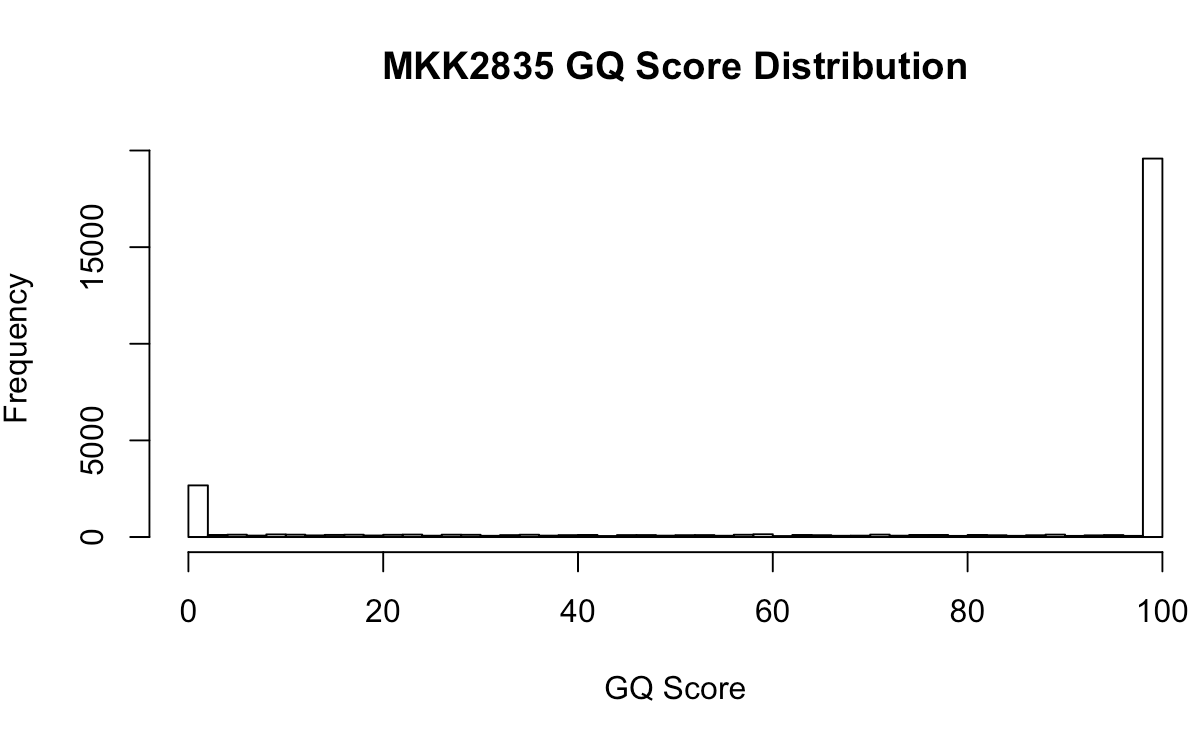
Filtering SNPs to get analysis ready set for MKK2835xNHP1337

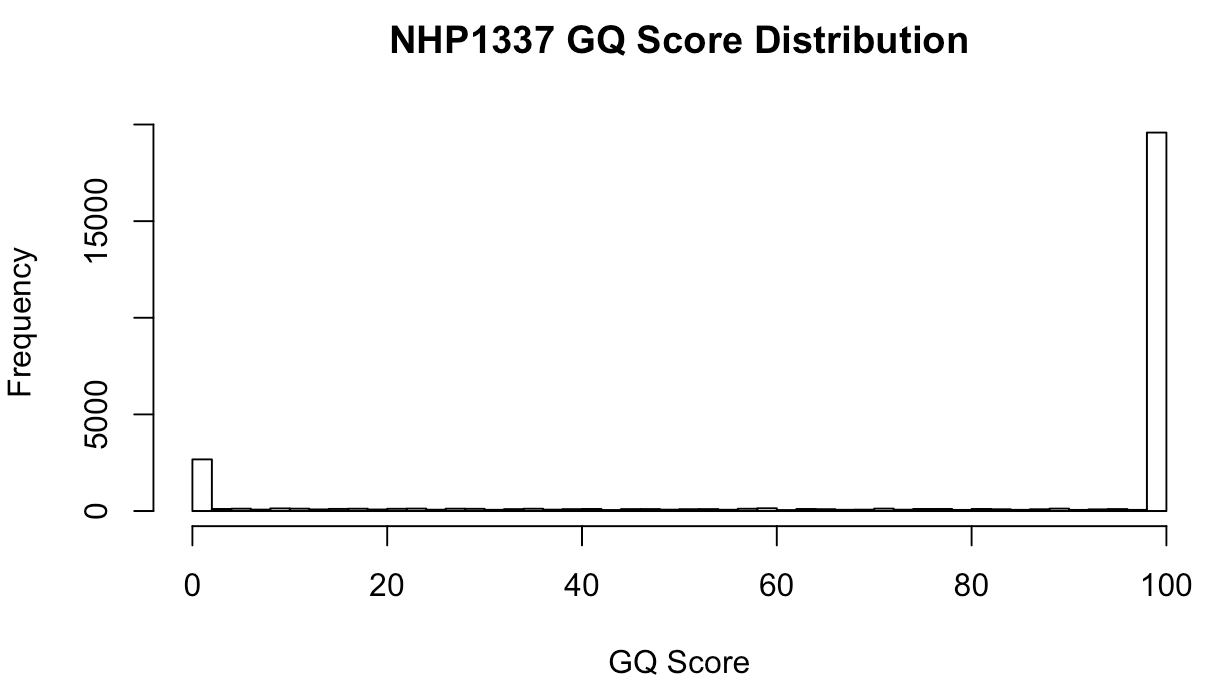
1. Filtering parental SNPs, include only if
   1. Homozygous and bi-allelic in parents
   2. Coverage ≥ 10



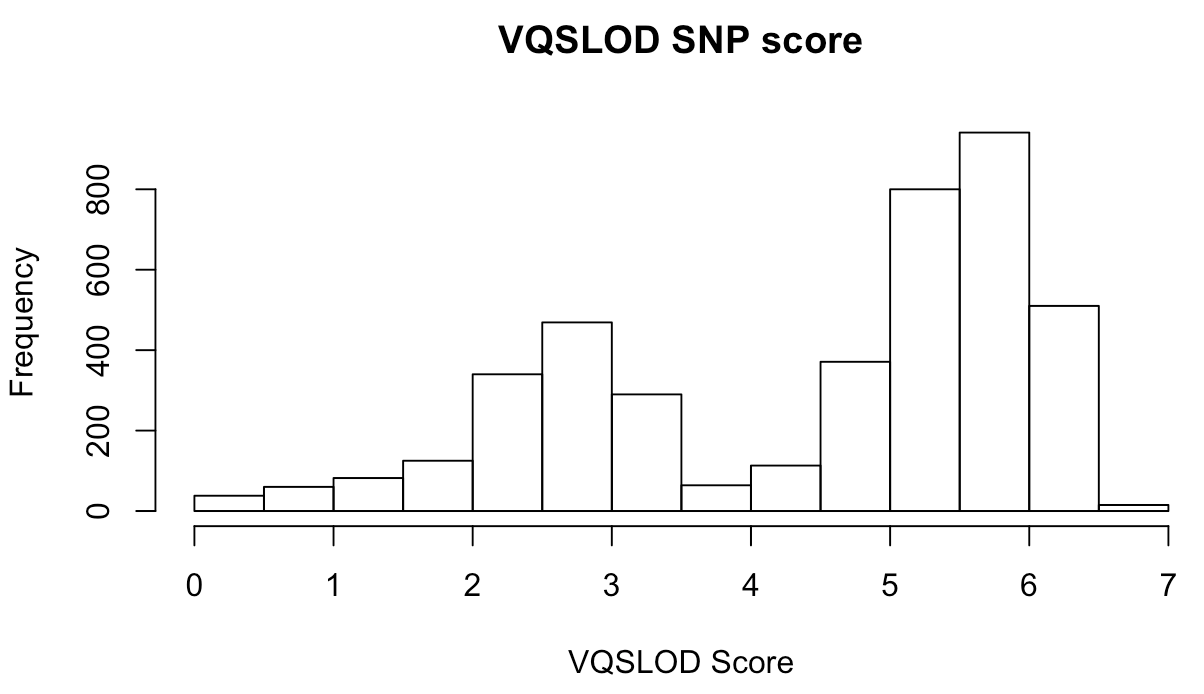


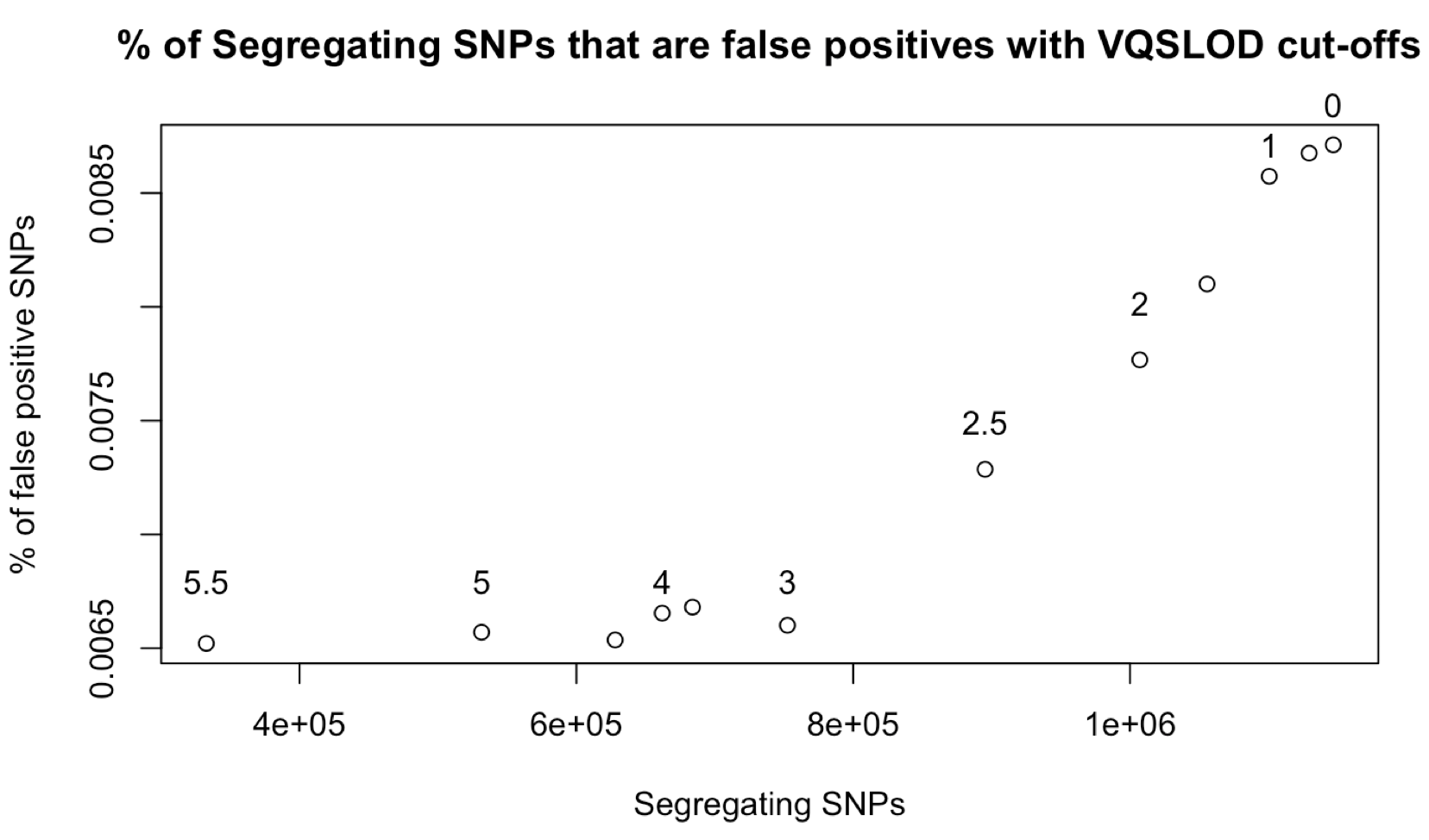
* 1. GQ ≥ 99





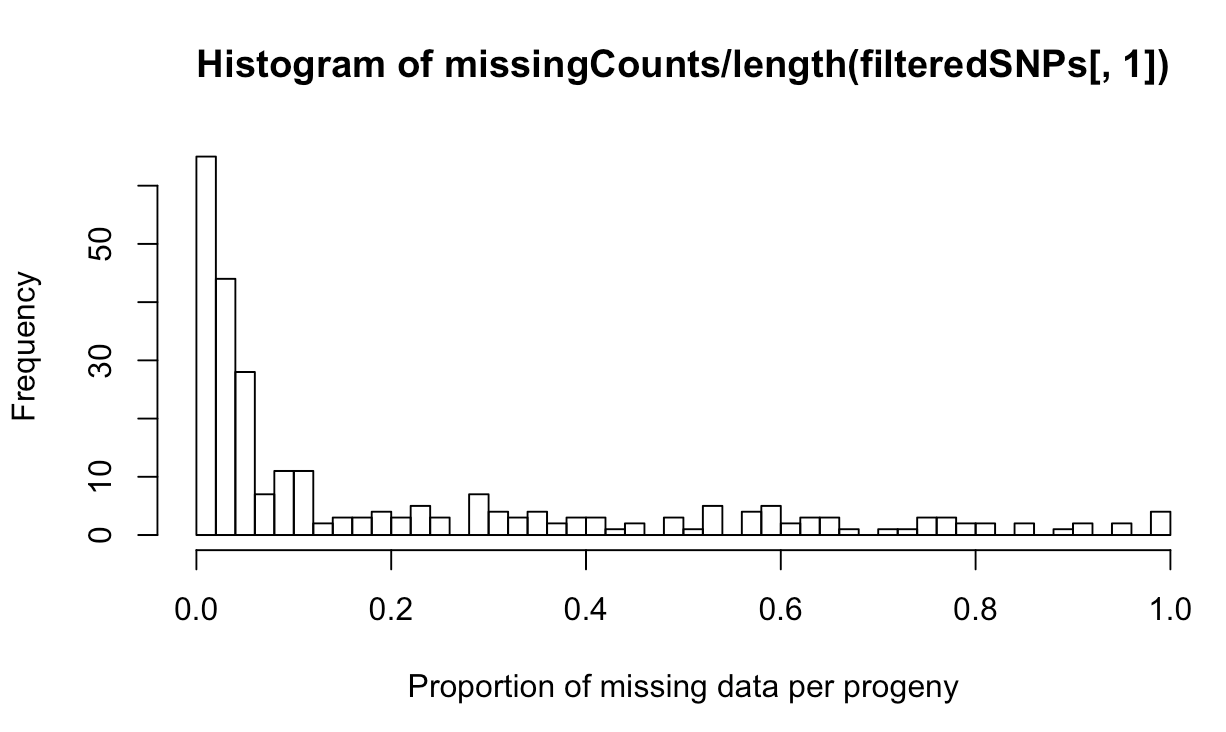
1. Progeny SNP filtering based on VQSLOD
   1. Choose VQSLOD cut-off similar to Miles et al. but using heterozygous SNP calls in clonal progeny as measure of sequencing error – cut-off is inflection point in curve of % of SNP calls that have sequencing error for all segregating markers in a set defined by different VQSLOD cut-offs (can update this later to base it on non-mendelian markers)
   2. Filter SNPs with VQSLOD < 2.5



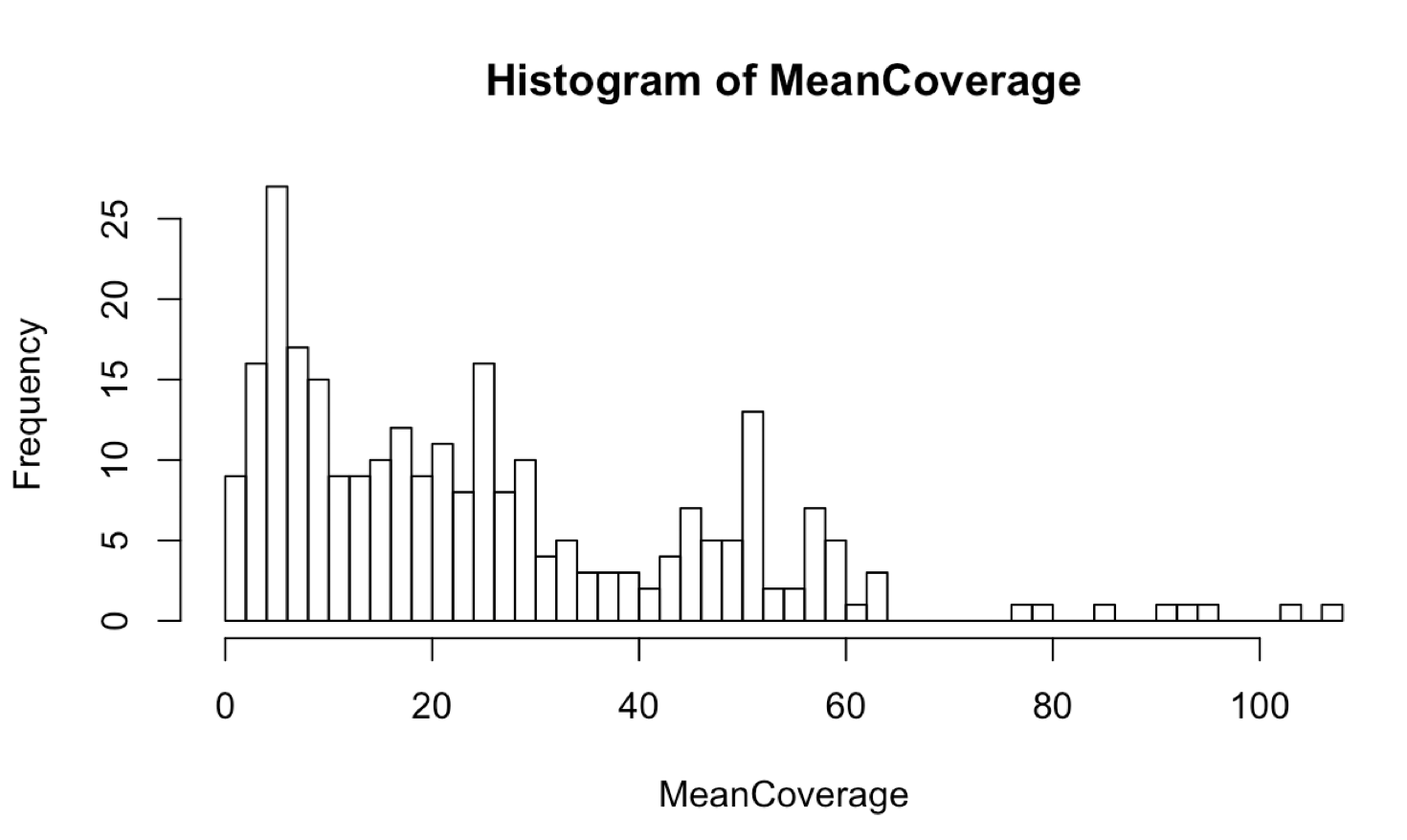


Filter Progeny

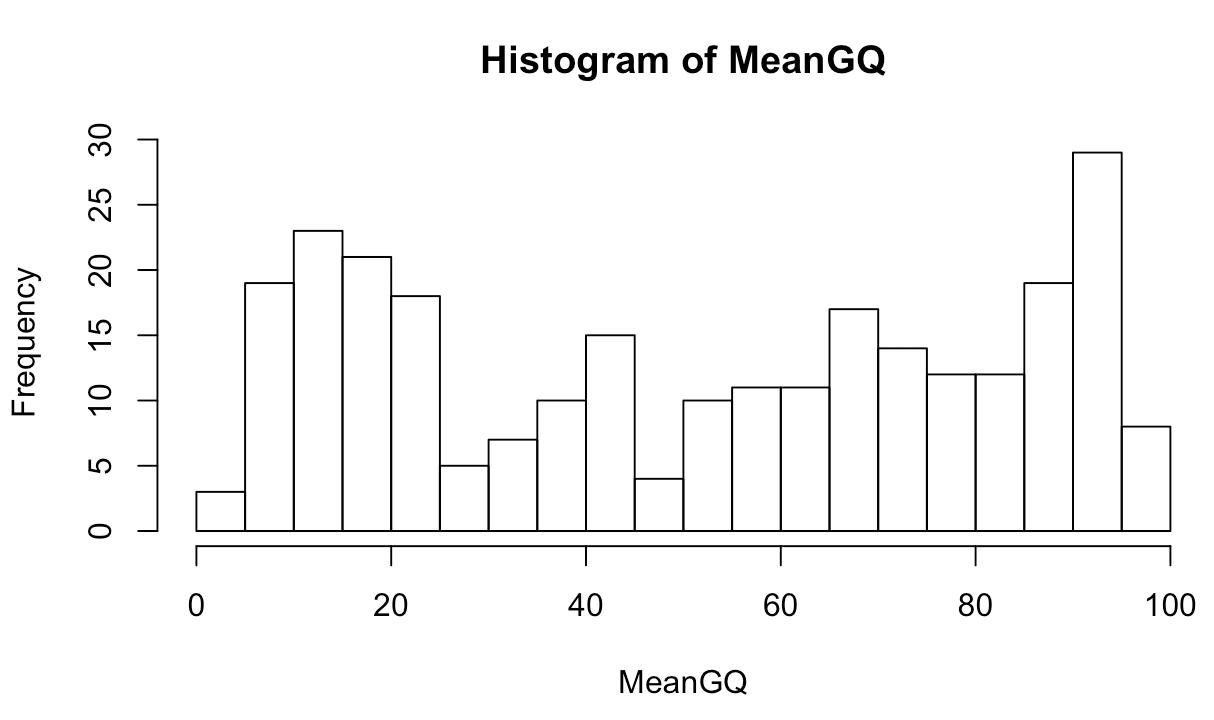
1. Filter progeny based on % missing data, coverage and GQ score (this ends of filtering out the same progeny, ie. those with high missing data have low coverage and low GQ scores)
   1. Keep progeny with Missing Data % < 80%



* 1. Mean coverage > 3



* 1. Mean GQ > 10



1. Filter non-clonal progeny defined as having any windows with above background heterozygosity

Coding SNP calls for final cleaned data

1. SNPs in progeny matching MKK2835 parent are coded as 0
2. SNPs in progeny matching NHP1337 parent are coded as 1
3. SNPs with “0/1”,”1/0”, “./1”, “./0”, “1/.”, “0/.” and “./.” are coded as NA

Identify unique recombinant progeny

1. Set cut-off for relatedness amoung progeny
   1. Choose single progeny and generate dataset repeating this progenies sequence with same total number of potential progeny
   2. For each simulated progeny, choose a percent heterozygous sequencing error rate distribution for clonal progeny and recode a random sampling of SNPs as other genotype
   3. Calculate % sequence similarity across all simulated progeny
2. Cluster progeny and include 1 from each cluster in final call set (this is done on coded data, not raw SNP calls)