**MCB 525, Fall 2016**

**2.3: Genotyping**

In Section 2.2, we created a filtered sam file of reads that mapped well to the BcgI reference. In this section, we will count the bases at each position, identify polymorphic sites, and decide whether each individual is homozygous or heterozygous at each polymorphic site.

**Objectives:**

* Gain experience running perl scripts on the command line
* Evaluate two variant calling pipelines
* Understand vcf and bam file formats

**Protocol:**

1. First, we count the A’s, C’s, T’s, and G’s (and N’s) at each position:

$ ../scripts/SAMBasecaller.pl CC7\_flt.sam BcgI\_ref.fa 3 <output\_file\_for\_basecalls.tab>

#this script reads in an input sam file and the reference, discards loci covered by <3 reads, and writes counts of A, C, T, G, and N frequencies to an output file

1. Based on these nucleotide frequencies, call each sample as heterozygous or homozygous at each position in the reference:

$ ../scripts/NFGenotyper.pl CC7\_basecalls.tab 0.05 0.25 15 > CC7\_genotypes.tab

#If an alternate allele is covered at <5% of the total coverage, this script calls the genotype as homozygous. If alternate allele has >25% frequency, it calls a heterozygous genotype. If between 5% and 25%, it considers the genotype ambiguous and does not call the genotype at that postion. Loci with <15x coverage are ignored.

1. Repeat Step 2.3a-b with the other sample.
2. The output file of NFGenotyper contains every position in the reference, but we are only interested in polymorphic sites. To filter for only polymorphic sites:

$ ./CombineGenotypes.pl sample1\_genotypes.tab sample2.genotypes.tab >both\_genotypes.tab

#combines genotypes from samples 1 and 2.

$ ./PolyFilter.pl both\_genotypes.tab 2 y >snps.tab

#selects loci for which more than 2 genotypes were observed, and writes them to an output file

*How many single nucleotide polymorphisms were identified between CC7 and your unknown sample?*

1. Sometimes we may see multiple SNPs on a single tag. These close-together SNPs are unlikely to be separated by recombination and therefore cannot be considered independent in downstream analyses. Let’s filter to choose only one SNP from each tag:

$ ../scripts/OneSNPPerTag.pl snps.tab y >selected\_snps.tab

# For sites with multiple snps, this script selects the one with the least missing data

*How many unique tags had at least one polymorphism?*

1. The scripts used above in Steps 2.3a-e have been designed specifically for analysis of 2bRAD data, but many other freely available variant callers exist (e.g. samtools, GATK) that enable quite sophisticated variant filtering based on variant quality, forward-reverse read balance, minimum and maximum depth, etc. In Steps 2.3f-h, we use a different variant calling process and compare the numbers of SNPs called from both pipelines. First, we need to convert the filtered sam file from yesterday to a sorted bam file. Bam files are compressed, non-human-readable versions of sam files.

$ samtools view -b CC7\_flt.sam | samtools sort - >CC7.bam

$ du –sh \*

#the ‘|’ is the pipe character, and passes the output from the left as input for the program on the right.

#du –sh \* will tell you the size of every file in your current directory. How do the sizes of the sam and bam files you have created compare?

1. Repeat Step 2.3f with the other sample.
2. Now, we will compare the bam files using the mpileup tool from the samtools package and pass the output to the call function from bcftools to create a file of raw variants.

$ SGE\_Batch -c 'samtools mpileup -uDf ../reference/BcgI\_ref.fa CC7.bam sample2.bam | bcftools call -vmO v -o var.raw' -r b2

#mpileup does need the reference to be indexed (see Step 2.2f)

#for more information about the available options, type samtools mpileup or bcftools call without any arguments.

1. Use vcftools to filter genotypes with a given minimum coverage (we’ll choose 15x for now). The output is a vcf or ‘variant call format’ file that contains many header lines starting with ‘##’ or ‘#’ and then a single row for each variant. More on vcf formats can be found here: <http://www.1000genomes.org/wiki/Analysis/vcf4.0/>

$ vcftools --vcf var.raw --minDP 15 --recode --out var.15

$ ../scripts/VCFPolyFilter.pl var.15.recode.vcf var.15.flt.vcf 20

# VCFPolyFilter.pl filters output from vcftools to contain only SNPs that are called in both samples, are polymorphic across the two samples (i.e. not sites heterozygous in both), and excludes sites for which the quality score of the SNP is <20.

$ cat var.15.flt.vcf | grep -v '#' | awk '{print $1}' | sort | uniq | wc –l

# count only 1 tag when there are multiple SNPs close together. This number is more comparable to the final number of SNPs from 2.3e

$ grep –v ‘##’ var.15.flt.vcf | head

#a quick look at the interesting part of the vcf file