Julie Garcia
April 4, 2018

Lab 8
Advanced Genomics and Genetics Analyses

In this lab, we will be working with an E.coli genome reference sequence. We will be mapping a set of 1 million paired end reads that I have simulated from this reference sequence (using wgsim from Samtools) in fastq format. Obviously, since these reads were not generated from an instrument, the quality values will all be the same, but we will still work with FASTQ files instead of FASTA files, so you have an idea of the code that you will have to use. Each read is 70mer and the insert size between the 2 mates is 500 bps. You will be using Bowtie to perform this mapping and Samtools to produce various summaries of the information.

Then we will work with human whole exome data truncated to only chromosome 22. We will utilize VarScan and a web annotation engine to call somatic variants between a liver tumor and normal healthy liver tissue specimen.

1.) We first need to index the reference sequence, so using the NC_008253.fa file, create the necessary index files for bowtie. Also get the 2 paired end FASTQ files called out1.fq and out2.fq.

```
bowtie-build –f NC_008253.fa NC_008253
```

2.) Now map the 2 mate files to the E.coli reference genome. For this function call, you need the following information:
   - seed length=24 (-l 24)
   - hits written in sam file format (--sam)
   - maximum insert size=600 (-X 600)
   - mates are in the forward/reverse orientation (--fr)
   - input files are fastq format (-q)
   - output the wall-clock time (-t)
   - seed sequence mismatches=3 (-n 3)
   - output the mapped and unmapped reads in separate files (--al out.align –un out.unmapped)

```
bowtie NC_008253 -1 out1.fq -2 out2.fq pair.sam -p 2 -l 24 --sam -X 600
–fr -q -t --al out.align --un out.mapped 2>&1 | tee out.log
```

3.) The seed length argument indicates how many bases are used in the primary alignment before extension. What are the tradeoffs between using a smaller seed length like 12 bases versus a longer one like 40 bases?

Smaller seed values cause bowtie to run slower, but you will be using more high-quality bases at the beginning of the reads. A 40 base seed would run much faster in bowtie, however, you will be getting into the lower quality bases at the end of the read.

4.) How many reads were aligned and what percentage is this?

```
#reads processed: 1000000
#reads with at least one reported alignment: 855814 (85.58%)
#reads that failed to align: 144186 (14.42%)
Reported 855814 paired-end alignments to 1 output stream
```

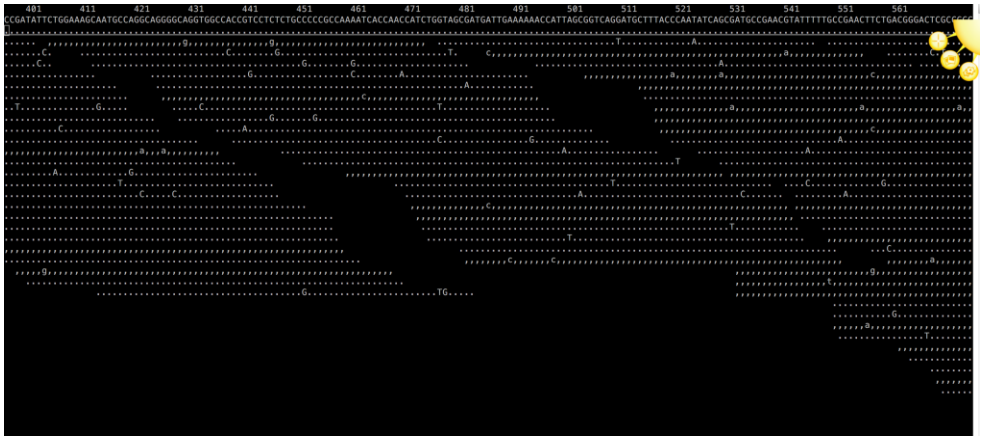
5.) Now using samtools, convert the sam to a bam file, sort the bam file, and index this sorted bam file.

```
samtools view -bS pair.sam > pair.bam
samtools sort -T /tmp/pair.sorted -o pair.sorted.bam pair.bam
samtools index pair.sorted.bam
```

6.) Use the tview function to visualize the mapped reads to the reference sequence.

```
samtools tview pair.sorted.bam NC_008253.fa
```

7.) Now scroll or jump to base position 400-650 and take a screenshot of the mapping.



8.) Why do you suppose that there were some reads that could not be mapped? Provide support for any possible reason you can think of.

My first thought as to why there would be unmapped reads is that some of the reads are low quality bases that have a lot of sequencing errors in them. Since the -n option on bowtie was set to 3, this means that any base that has more than 3 mismatches in the first 24 bases would be unmapped. The -e option is by default a score of 70, so if the total of all mismatch scores was less than this, it would be unmapped. However, these bases all have the same quality score, as stated in the intro.

Another reason that some reads may be unmapped, is that their sequence spans an exon boundary, and represents a splice junction, which would not be present in the genome due to splicing out of introns. If using TopHat, these reads would be set aside for later processing against the exome in order to map the splice junctions.

References:

1. Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. http://doi.org/10.1093/bioinformatics/btp120
2. http://bowtie-bio.sourceforge.net/manual.shtml#the--n-alignment-mode
3. RNA-Seq Lecture from this class

9.) Create a pileup format of this alignment and print the first 10 lines of your file. We won't be evaluating the SNP calls here because the quality values are identical for every base, but with reads that are not simulated, this file format provides valuable information at each base position that can be used to call indels and SNPs.

The version I installed said "pileup is no longer available" and to use mpileup instead.

```
samtools mpileup -f NC_008253.fa pair.sorted.bam > pair.pileup
```

```
gi|110640213|ref|NC_008253.1|   1       A       1       ^~.     2
gi|110640213|ref|NC_008253.1|   2       G       1       .       2
gi|110640213|ref|NC_008253.1|   3       C       1       .       2
gi|110640213|ref|NC_008253.1|   4       T       1       .       2
gi|110640213|ref|NC_008253.1|   5       T       1       .       2
gi|110640213|ref|NC_008253.1|   6       T       1       .       2
gi|110640213|ref|NC_008253.1|   7       T       1       .       2
gi|110640213|ref|NC_008253.1|   8       C       1       .       2
gi|110640213|ref|NC_008253.1|   9       A       1       .       2
gi|110640213|ref|NC_008253.1|   10      T       1       .       2
```

10.) Now let's work with the human whole exome sequence data for chromosome 22. Download the normal and tumor BAM files from the course website as well as the human hg19 reference sequence. Also install VarScan on your computer. This is a Java application, so all that is needed is a working instance of Java on your machine and you can launch the application jar file.

11.) Run Samtools and pipe the mpileup output into a VarScan call. To do this, we want to set the –B flag and both mapping and phred quality (base) scores to 30 in the Samtools mpileup call. Then for the VarScan somatic caller, output –output-snp format, set the min coverage to 50, min tumor coverage to 3, p-value to 0.01, and min variant frequency to 4%. How many somatic calls and germline calls were made?

```
samtools mpileup -Q 30 -q 30 -f hg19.fa normal_sort.bam >
normal_sort.pileup
samtools mpileup -Q 30 -q 30 -f hg19.fa tumor_sort.bam >
tumor_sort.pileup

java -jar VarScan.v2.3.9.jar somatic normal_sort.pileup
tumor_sort.pileup output.basename –output-snp –min-coverage 50 –min-
coverage-tumor 3 –p-value 0.01 –min-var-freq 0.04

4 somatic calls
753 germline calls
```

12.) From the output variant file (not the indel file generated), go to http://snp-nexus.org/ and format the appropriate input (Hint: set strand equal to 1 for all loci), then run the annotation engine to extract RefSeq, SIFT, PolyPhen, and COSMIC database information for your identified variants. Output a text file and identify how many variants are found in the COSMIC database? Using SIFT predictions, how many variants have high damaging effects? Using Polyphen, how many variants have 'probably damaging' effects? How many nonsynonymous variants were identified using RefSeq annotation (count all asterisk and non-asterisk instances)?

```
I formatted the data in excel and then copied it into snp-nexus.org.

42 variants were found in the COSMIC database.

21 variants were predicted by SIFT to have high damaging effects.

57 variants were predicted by PolyPhen to be probably damaging.

RefSeq found 232 nonsynonymous variants.
```