Joshua Schaaf

Computational Genomics - Week 5 Summary

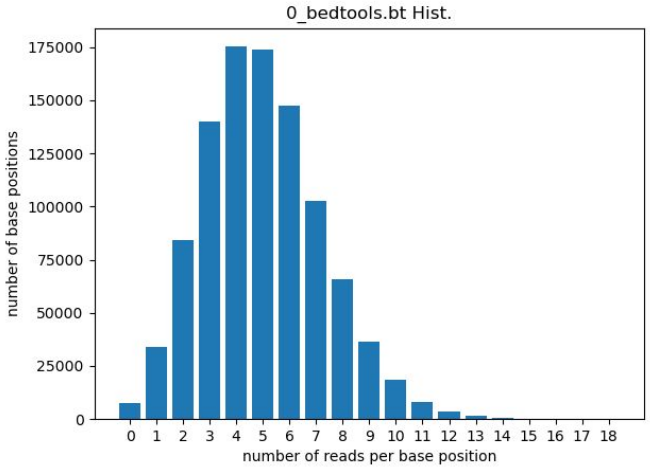
Bowtie was used to map art_illumina created reads with different percentages of SNPs to a megabase of GRCh38 chromosome 17 (bases 41628720-42628719). The bowtie results are in the following table. Bowtie processed all 50000 reads, with a sharp decrease in reads with one or more reported alignment as the SNP percentage went up (from 0 to 5%).

SNPs	**Reads Processed**	**Reads w/ >=1 Reported Alignment**	**Reads that Failed to Align**	**Reported Alignments**
-	50000	49996 (99.99%)	4 (0.01%)	49996
1%	50000	45155 (90.31%)	4845 (9.69%)	45155
2%	50000	32777 (65.55%)	17223 (34.45%)	32777
5%	50000	5701 (11.40%)	44299 (88.60%)	5701

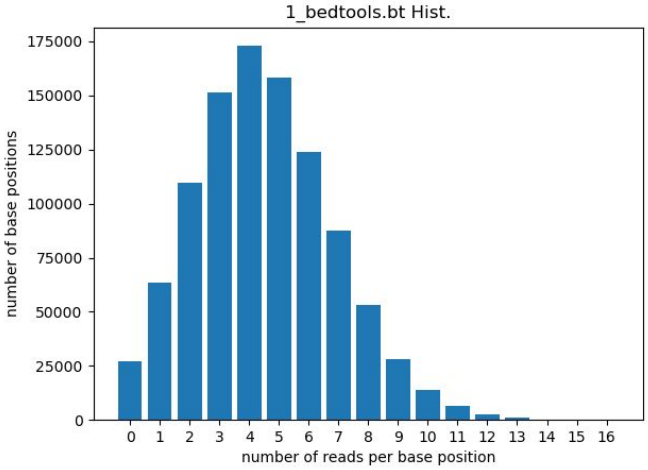
Data Table: bowtie results on art_illumina fastq files of GRCh38 chromosome 17 (bases 41628720-42628719) and increasing SNP percentage.

Samtools was used to convert the bowtie sam file into a bam file, and then samtools was used to sort the files, as well as create a samtools index file. With this sorted bam file, bedtools genomecov can create a text file describing the number of base positions that have a certain number of reads per base position. Because the art_illumina reads were made with 5x coverage, the histogram peak should be centered around 5, as there would be the most amount of bases in the genome that have 5 reads. A python file was made to quickly convert these text files to histograms with matplotlib.pyplot. The histograms follow this paragraph.

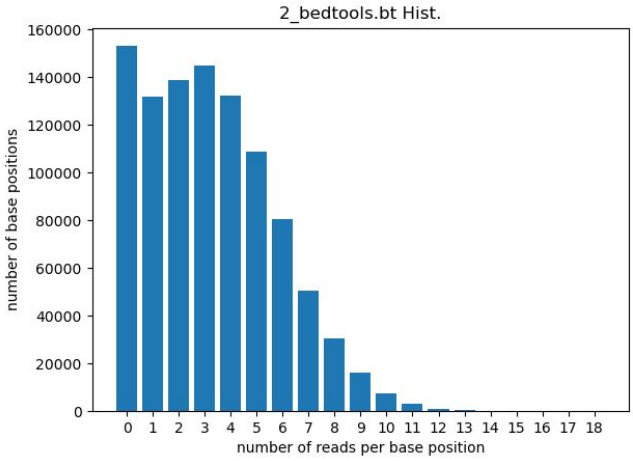
0% SNPs



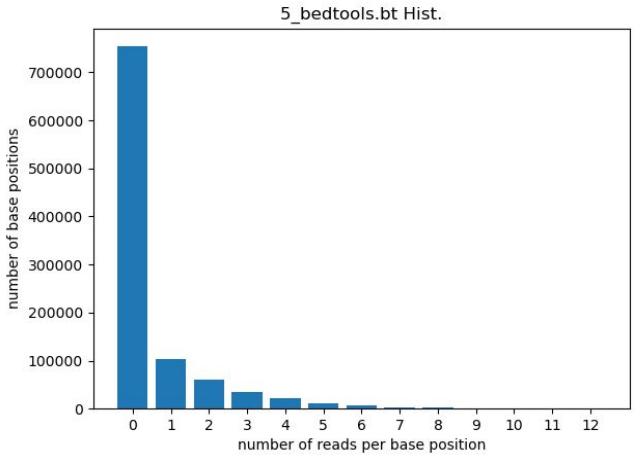
1% SNPs



2% SNPs



5% SNPs



In the above histograms, it is easy to see that the average number of reads per base position seems to shift to the left, towards zero, as the percentage of SNPs goes up. This is due to bowtie not mapping reads with multiple SNPs to the genome.