CSE 185 Quiz 1 Cheat Sheet - Spring 2019

Sequencing coverage

Coverage = #bases sequenced/genome size = (numreads*readlength)/(genome size)
Q: You sequence a genome of 3 billion bp with 10 billion reads of length 36bp. What is the average coverage?
A: 36*10B/3B = 120x

Q: How many 2x100bp paired end reads do you need to sequence to achieve average 50x coverage of an E. coli genome (5 million bp)?
A: 50*5000000/(2*100) = 1,250,000 read pairs

Variant calling from NGS + binomial distribution intuition

Q: You have 10 reads covering a position for which your sample is homozygous for the non-reference allele. In the absence of errors, on average how many reads will show evidence for the reference allele?
A: 0
Q: You have 10 reads covering a position for which your sample is heterozygous for the reference + a non-reference allele. In the absence of errors, on average how many reads will show evidence for the reference allele?
A: 50%
Q: You have 10 reads covering a position for which your sample is heterozygous for the reference + a non-reference allele. In the absence of errors, what is the probability to see 3 reads with the non-reference allele?
A: (10 choose 3)*0.5^3*0.5^7 = 11.7% # can also just write the math without giving the number

Kmer distributions

Kmer coverage = # kmers sequenced / # kmers in genome = (L-k+1)*N/(G-k+1), where L=read length, k=kmer size, N=num reads, G=genome size
Q: How many kmers of length 8 can you generate from a read of length 100?
A: L-k+1 = 100-8+1 = 93

Q: You sequence a genome of length 1 million using 3 million single end reads of length 36bp. Using kmers of length 31, how many times to you expect to see each kmer in your data? (i.e. what is the mean *kmer coverage*)?
A: sequenced kmers/#kmers in genome = 3000000*(36-31+1)/(1000000-36+1) = 18

Command line tools

Know basic usage of: cat, head, tail, cut, grep, ls, cd
Be comfortable using pipe “|” to combine commands, using “>” to redirect standard output to a file.

Example commands you might be expected to write:
head -n 20 file.txt # print first 20 lines of file.txt
head -n 20 file.txt | tail -n 10 # print lines 10-20 of file.txt
cat file.txt | cut -f 5,10 # print out columns 5 and 10 of file.txt
grep “chr5” file.fa # find lines in file.fa containing the string “chr5”
cd ../../ # navigate two directories above your current directory
ls *.bam | grep child > child_bam_files.txt # list all BAM files in the current directories and only print out those with the string “child” in the file name, and write the results to the file child_bam_files.txt

Other things we’ll assume you know:

File format basics: fasta, fastq, BAM/SAM, VCF
The meaning of Illumina quality scores
The concept of paired vs. single end sequencing and fragment/template/insert size
The difference between genome assembly and alignment