CSE 185 Quiz 2 Cheat Sheet - Spring 2019

Population genetics

Allele frequency, Genotype frequency, Minor vs. major allele
Examples:
Q: You genotype a SNP in 1,000 people. 20 are homozygous for allele “A” , 240 are heterozygous for “A/T”, and the rest are homozygous for “T”. What is the minor allele frequency?
A: (20*2+240)/(2*1000) = 0.14 (note, A is the minor allele).

Q: For the same SNP as above, what is the genotype frequency of AT?
A: 240/1000 = 24%

Q: You genotype a SNP that is *tri*-allelic (there are 3 possible alleles, here A, C, or T. This is very rare). What are the possible genotypes a person can have?
A: AA, AC, AT, CC, CT, TT

Q: For the same SNP, you genotype 1,000 people. 10 have genotype AA, 100 have genotype AC, and 50 have genotype AT. What is the allele frequency of the “A” allele?
A: (10*2+100+50)/(2*1000) = 0.085

P-values

P-value definition, null hypothesis, multiple hypothesis correction
Examples:
Q: You are performing a GWAS for height with 5,000 people genotyped at 1.5 million SNPs. Before performing the GWAS, you accidentally randomly shuffle the sample phenotype (height) values, such that each sample is assigned a random height. How many SNPs do you expect to show a significant association with height (p<0.05)?
A: 0.05*1.5million = 50000

Q: For the same set of 1.5 million association tests as above, what do you expect the mean p-value to be?
A: 0.5. Since under the null hypothesis of no association, p-values follow a uniform distribution from 0 to 1.

RNA-seq

RPKM (reads per kilobase per million reads) , TPM (transcripts per million, can be interpreted as a fraction of all transcripts scaled by 1 million)
CIGAR Scores (M=match, I=insertion, D=deletion, N=gap)
Examples:
Q: You align an RNA-seq read to the reference genome. It matches for 30bp to exon 1, then spans a 1000bp intron, then matches for 15bp to exon 2. What is the CIGAR score describing this alignment?
A: 30M1000N15M

Q: For a read with CIGAR score 40M10I50M, what is the total length of the read?
A: 40+10+50 = 100 # 90bp total match the ref, plus 10bp insertion

Q: For a read with CIGAR score 40M10D50M, what is the total length of the read?
A: 40+50 = 90 # 90bp total match the ref. There is also a 10bp deletion from the reference

Q: You perform RNA-seq with 1 million reads. You find 300 reads map to gene A, which has length 5kb. What is the RPKM of gene A?
A: RPKM: 300/5/1 = 60

Q: You perform RNA-seq with 100 reads. Your sample has 3 genes total. Genes A, B, and C all have length 1kb. You align 10 reads to gene A, 50 to gene B, and 40 to gene C. What is the TPM of gene A?
A: RPK_A = 10/1 = 10, Scaling factor = (10+50+40)/1M = 100/1M, so TPM_A = 10/(100/1M) = 100,000.
(Note: TPM_B=500,000 and TPM_C=400,000. All TPMs should sum to 1 million).

Command line tools+File formats+troubleshooting

Know basics of UNIX commands: cat, head, tail, cut, grep, ls, cd, awk, sed, datamash

Know what the following bioinformatics tools are used for:

- Sequence quality control: fastqc
- Read trimming: sickle, nxtrim
- Alignment: bwa (DNA sequencing data), STAR (RNA sequencing data)
- Manipulating variant files: bcftools (for VCF files), plink (for VCF or plink files), tabix (to index VCF files), bgzip (to zip VCF files)

Know how to write the output of a command to a file:

- “My command…” > file.txt # writes output of a command to a file
- “My command…” >> file.txt # appends to an existing file

File formats: FASTA, FASTQ, SAM/BAM, VCF, BED

Also see link below for more in depth description of commands we have covered. But quiz will stick to this cheat sheet.
https://docs.google.com/document/d/15Y3UMdZguknp2U2hGBqrpZxtd9oYmcv9sDUR6rSNA2k/edit?usp=sharing

Examples:

```
cat file.txt | sed ‘s/chr//’ > newfile.txt # replaces the first instance in each line of file.txt with “chr” and writes to newfile.txt
cat file.txt | datamash mean 1 median 3 # prints out the mean of column 1 and median of column 3 of file.txt
```

Q: You would like to extract a subset of SNPs from a VCF file. What is an appropriate command line tool to use?
A: bcftools. Also: tabix could work if it is just one region. intersectBed (but we haven’t covered this). Could also string together some fancy grep commands.

Q: You are performing an analysis to extract SNPs from your VCF file mysnps.vcf.gz and get an error: “No index found for file mysnps.vcf.gz”. Suggest a command you could use to fix this error.
A: tabix -p vcf mysnps.vcf.gz.

Q: You are performing an analysis to extract SNPs from your VCF file mysnps.vcf.gz and get an error: “File not in gzip format”. Suggest a command (or commands) you could use to fix this error.
A: The file was likely named with “.gz” by mistake. You could do:

```
mv mysnps.vcf.gz mysnps.vcf # rename it correctly first
gzip mysnps.vcf
```

Q: You would like to perform a GWAS by testing for association between a set of SNPs and some phenotype. What is an appropriate command line tool to use?
A: plink