**MIP 280A4: Microbial Sequence Analysis**

**Variant Calling, In-class exercise questions**

**Variant calling output questions**

1. How many variants were identified (how many variants are described in the lofreq vcf output file)? What bash command did you run to determine this? (2 pts)
2. What is the allele frequency of the first described variant? (1 pt)
3. Is linkage between variants described in the vcf file? (1 pt)
4. Are the variants SNPs, or InDels? Would you expect to see InDel variants here? (2 pts)

**Question about variants in Geneious**

1. Can you identify individual variants called in your VCF file in Geneious? (1 pt)
2. Are variant frequencies generally similar between Geneious and the VCF file? Report the variant frequency as reported by Geneious and in the VCF file for one variant (include its position in the genome too). (2 pts)
3. Can you visually identify linked variants by looking at reads in Geneious? How far apart do you think you can identify linked variants using this strategy? (2 pts)
4. Would long read sequencing be a good alternative approach to identify linked variants? Why or why not? (1 pt)
5. Are any of the variants non-synonymous (do they change the encoded amino acid)? Provide an example of one such variant (1 pt).
6. Are these intrahost or interhost variants? Explain your answer (1 pt)
7. Can you identify any InDel variants that \_should\_ have been called by lofreq that weren't called? (Hint: look near the beginning of the reference sequence.) (1 pt)
8. We ran bowtie2 in end-to-end mode, which doesn't permit soft-trimming of read ends. It forces the entire read to align. Can you identify any ends that should have been trimmed but weren't? (I.e. that contain low quality basecalls at their ends?) (1 pt)