Genome 373
Metagenomic alignment workshop
Day 2: June 2, 2017

First, finish running the analyses in Part 1 on at least one sample. Discuss with your group: Are there substantial differences in read mapping between samples? Can you conclude anything? Once you've finished, move on to the analyses below, which examine the gene content of your sample(s) in more depth.

Part 5: Generate gene counts and coverage

1) Run the following command to generate abundances of gene counts for each sample. The command "bedtools intersect" finds regions of overlap between two files of genomic coordinates - in this case, the read alignment information in the bam file and the gene locations in the bed file. The Perl script counts up reads whose alignment location overlaps a gene by at least 5 nucleotides. How would changing this parameter affect the resulting gene counts?

bedtools intersect -abam "$sampleID"_mapped_sorted.bam -b refs/pathogen_ref.bed -bed -wao | perl ../scripts/get_geneCov.pl 5 > "$sampleID".geneCov

If you see a message that says, "Died at ../scripts/get_geneCov.pl line 68, <> line xxxx", that means the code ran successfully. You should now have a new file ending in ".geneCov" that contains information on read counts for each gene. You can take a look at this file using **head**, **cat**, or **more**.

The 4 columns of the output file represent the following information for each gene:
geneID, geneLength, readCount, geneCoverage

The last column of numbers in the file represents gene coverage: the number of reads covering or mapping to each base of a gene, on average (i.e. geneCoverage = numReads*overlapLength/geneLength).

- Can you compare these gene coverage values between two different samples? Why or why not? What would you need to correct for?

Part 6: Compare abundances of each genome and sample based on marker genes

To calculate a more accurate estimate of the abundance of a species, microbiome researchers use read coverage across *marker genes*. Marker genes are genes that are always found in a particular species in a single copy, so the share of reads mapping to them is reflective of the concentration of the species in the original sample. The file "pathogen_ref.markers.geneInfo" contains lists of marker genes and coordinates for each of the 3 genomes analyzed here.

We can use the "join" command to merge this file of gene counts with the file of marker genes, to produce a new file that contains information on only the genes listed as marker genes, their counts in the sample, and their annotations (just copy and paste it the command below into your terminal):

join -j 1 -t $'\t' <(sort "$sampleID".geneCov) <(sort refs/pathogen_ref.markers.geneInfo) > "$sampleID".markers

- Open the resulting tab-delimited ".markers" file in Excel or Python and calculate the average and standard deviation gene coverage across all marker genes for each genome.
- A reasonable threshold for defining whether a taxon is "present" in a microbiome sample is an average marker gene coverage of 2 reads per base. Which species have evidence of being present in your sample(s)?
- How could you get this level of coverage if a species isn't actually present? (Hint: what would you want to know about other species in the sample?)

Part 7: Look for evidence of virulence genes

The presence of a known pathogen in your sample doesn't necessarily mean that it was a source of illness. *E. coli* is commonly found in the gut microbiome of healthy individuals (15% of healthy U.S. adults in the Human Microbiome Project). However, strains that possess virulence genes that encode a toxin or promote adherence can be very dangerous. The file "pathogen_ref.virGenes" contains a listing of virulence genes for 2 of the genomes analyzed here (unfortunately not for *Salmonella*), downloaded from the Virulence Factors of Pathogenic Bacteria Database.

As you did for marker genes, merge your gene counts file with the listing of virulence genes by running the code below (again, easiest to copy and paste):

join -j 1 -t $'\t' <(sort "$sampleID".geneCov) <(sort <(paste <(cut -f3 refs/pathogen_ref.virGenes) <(cat refs/pathogen_ref.virGenes))) > "$sampleID".virGenes

Using Excel or another program, examine the ".virGenes" file of virulence gene abundances, and make a bar plot of them if you have time. How much do their abundances vary? Which ones appear to be present or absent, and what are their functions? Are there differences between your two samples? You can also look at the read coverage distribution of any interesting genes in your IGV visualization.

Part 8: Wrap-up

What would you infer overall about your samples? Do you think you have sufficient information to make a diagnosis? Compare your results with the other groups.