# Lab/Homework 5: Finding Genes

In this lab we will find genes in the genome of *Mycoplasma genitalium*. *M. genitalium* is one of the smallest bacteria, both in actual size and in genome size. It is parasitic, a human pathogen and, as it name suggests, it is sexually transmitted and can cause diseases in both men and women. The bacteria itself is of interest to genome researchers due to its small genome size, and is considered to be a "minimal organism".

Why would a parasitic organism require fewer genes than a free living one?

Which type of genes do you think are lost?

We will find out the number of genes in *M. genitalium* using the model we learned in class:
1. Estimate the expected number of open reading frames (ORFs) under a simple null model and use the result to infer whether the *M. genetalium* genome is unusually structured with respect to ORFs.
2. Estimate the ORF length distribution under the simple null model.
3. Use that distribution to choose a threshold for gene finding.

## Steps:

1. Get the genome file of *M. genitalium* from **`/opt/bcbio444/genefindinglab/NC_000908.gb`** (use the **`cp`** command). Use BioPython to convert from Genbank format to FASTA format. The BioPython library becomes available after you issue the command '**`module load python`**'. Attach the code you used and record the command you issued.

2. Use Python to determine the GC content of the *M. genitalium* genome. The human genome GC content is 41%. Is *M. genitalium* likely to have more or less ORFs than human? Longer or shorter

ORFs?

______________________________________________________________________

______________________________________________________________________

______________________________________________________________________

3. Since we want to perform inference (find the genes) in this genome, the population to which we want to extend our conclusions is the *M. genitalium* genome. We need a model that generates *M. genitalium*-like ORFs. We could use the iid Multinoulli model, but there is another model particularly useful for this case where we have no need to generalize to other genomes. The permutation model can generate whole genomes along with the ORFs in them by randomly shuffling the *M. genitalium* nucleotides. This model generates not quite iid nucleotides, because the probabilities for the next nucleotide depend on which nucleotides have already been placed. Write a function that randomly permutes the *M. genitalium* genome, produce 100 permuted genomes, and output them in FASTA format. Attach the code you used to perform this task, but do *not* attach the permuted data! Do you expect the ORFs to be longer or shorter in the permuted genomes? Why?

______________________________________________________________________

______________________________________________________________________

4. Now we need to find and translate the ORFs. Mycoplasma are weird: they use a slightly different codon table than most other organisms. UGA is commonly used as a stop codon. In mycoplasma, it codes for the amino-acid Tryptophan. This is a serious headache for anyone doing genetics in mycoplasma, but all we have to worry about is using the correct translation table. The translation table you should use is #4 here: https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi.

**Lab:** You almost have the code to write this yourself in Python, but others have already done the job. To speed up the lab, you will use the translation program **`getorf`**, part of the EMBOSS suite of software, which is available on `hpc-class` after you issue the command '**`module load emboss`**'. The command will be something like:

```
$ getorf -table 4 -find ??
```

What should you put after **`–find`**? Type '**`tfm getorf`**' to find out. Be sure to use the mandatory arguments **`-sequence`** and **`-outseq`** appropriately.

______________________________________________________________________

______________________________________________________________________

**Homework bonus:** Write python code to find and output all the ORFs in FASTA format. Please beware that an ORF may be in any of the six frames: three frames on each strand of the genome. Attach your code for credit.

5. The fundamental assumption of the permutation model is that the order of the nucleotides is random. Since ORFs are determined by the order of nucleotides, it seems reasonable to detect biologically functional ORFs as somehow unusual relative to ORFs generated by the random order

(permutation) model. How could you critique the assumptions of the permutation model?

We will use the number of ORFs as a test statistic. Generally speaking, what values of this test statistic reject the null model (hypothesis) of random order?

Count the number of ORFs in each of the 100 shuffled genomes. Report the mean number of ORFs found: ______________ Report the maximum number of ORFs found: ______________. Plot a histogram of the number of ORFs.

6. Count the number of ORFs in the **`NC_000908.fsa`** file. Can you reject the random order model for the *M. genitalium* genome?

Does the number of ORFs you found in *M. genitalium* genome make sense? Why or why not?

Are all of them true ORFs? Why?

7. We will now try to find the number of *true* ORFs in *M. genitalium* using a threshold derived from the null model. Our test statistic is the length of the ORF. Generally speaking, what values of the test statistic would indicate a true ORF?

Because our experimentally sampled unit is now the ORF (not the genome) and there are many ORFs in a single genome, we can estimate the ORF length distribution under the null distribution from a single genome permutation. Compute the ORF length null distribution from the first permuted genome. Estimate the 95% percentile (*i.e.*, the ORF length which is longer than 95% of the ORF lengths). Are the ORF lengths computed from the genome independent? Why or why not?

8. Percentiles are notoriously poorly estimated. You can get a better 95% percentile estimate (lower variance) by averaging the estimates from all 100 permuted genomes. What is the new estimate? (Note: Averaging the estimates is mathematically different from estimating the percentile from the combined sample of all ORF lengths from all 100 permutations. The difference would be especially notable when the ORF lengths are highly dependent within a single genome.)

9. Using the 95% percentile from question 8 as a minimum ORF length, how many ORFs are predicted to be in the *M. genitalium* genome? ____________________
How did you calculate that (share any code you wrote and command you issued)?

If the random order model were correct, how many of these ORFs would you expect to be false positives (not real ORFs)? How does the proportion change if there are true positives (real ORFs)? Why?

10. The annotators of *M. genitalium* have determined that there are 482 protein coding genes. Were you close to the mark? Why do you think that you deviated (if you did)?