Post Lab Questions: Week 6

OneCard: 1955791

I. Check for Understanding

Toy Dataset

1. Describe the large-scale differences between the mapped reads from species 1 and species 2…

A large section of the reference genome is missing (possibly deleted) for species 2, but present in species 1. This region spans from ~10.5kb to ~40kb on the reference genome.

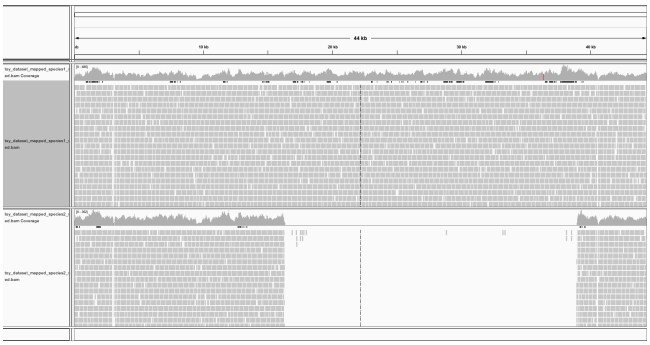

Figure 1. Toy dataset comparison of two species against the reference genome.

One possible biological mechanism that could explain this occurrence is horizontal gene transfer (HGT). A gene could have been horizontally transferred between the reference species and species 1, but not to species 2.

A dot plot comparing the two regions might look something like this. The sequences line up well until a large “in/del” appears.

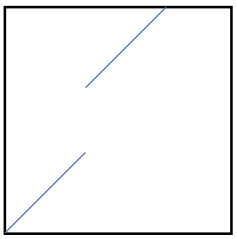

Figure 2. Dot plot mock-up, highlighting what a large in/del might look like.

2. Do you see evidence of misassemblies?

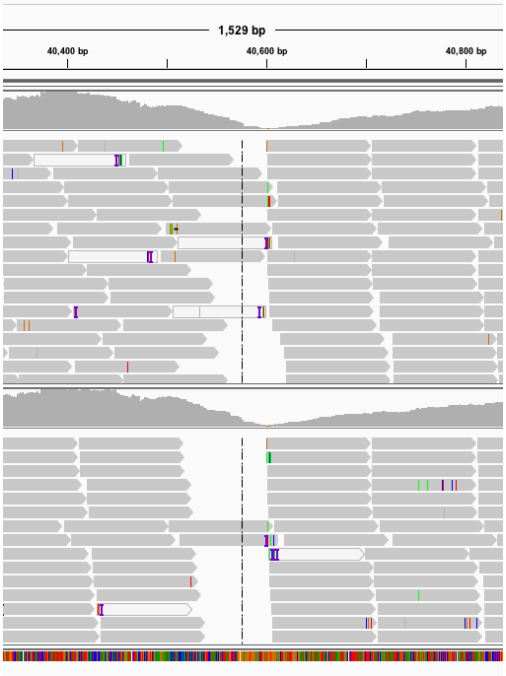

Figure 3. "Edges" where no reads overlap in toy dataset.

There are some “edges” where reads don’t overlap. These could be evidence of misassembly in the reference genome—no reads span that gap because it shouldn’t exist. It could alternatively indicate that species 1 and 2 have a gene that is missing in the reference lacks.

3. Do you see any evidence of single nucleotide polymorphisms?

There is fairly limited evidence of single nucleotide polymorphisms. IGV highlights the places where a base of read differs from the reference genome, but in these locations usually only one of the reads differs. This might indicate the variation is due to sequencing error rather than constituting a “SNP.”

Project Dataset

4. If you wanted to quantify the relative abundances…

When we assemble reads into contigs, repeated sequences are essentially ignored, since we expect many repeats to be generated during the sequencing process. Thus, there could have been 50 copies of a single gene or 500 copies—either way it will only appear once in the assembled contig.

5. Do you see evidence of single nucleotide variants?

Yes! Unlike with the toy dataset, there are places where many of the reads have a single nucleotide that differs from the reference. These SNVs could reveal variation among different populations of a given certain species—the reference genome could come from one population, and the reads where a SNV is present could come from another.

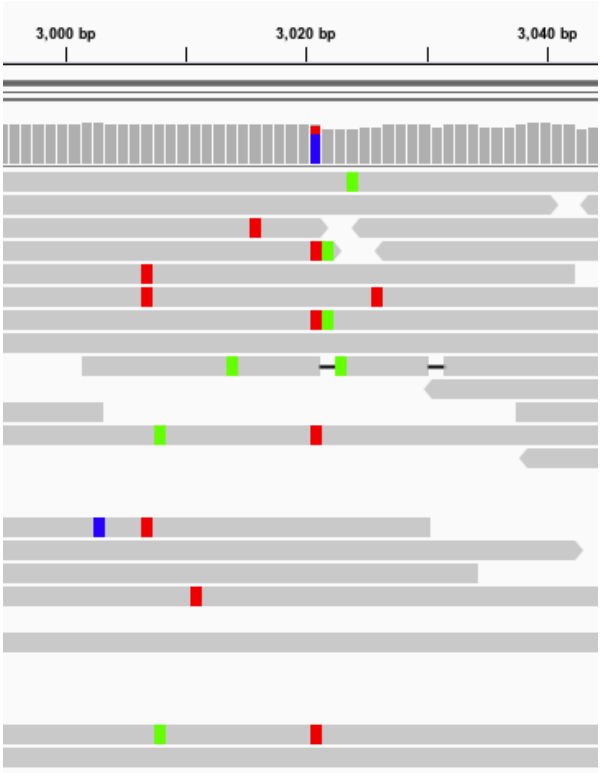

Figure 4. Evidence of SNPs in project dataset.

6. Go back to your Interproscan files and find two ORFs that you're interested in…

I don't have my Interproscan results back yet. Instead of using genes annotated by interproscan I did the following. To find an ORF that might be abundant in a sample, I chose a random ribosomal protein from the PFAM database and generated a "seed" fasta file. I made a blast database with my ORF file, and then blasted the seed file (as the query) against my ORF file (as the database), using the blastp command line tool. I set the evalue of this search to 1e-05. This blast returned a number of hits, and I chose the first one (c_000000056441_1). I predict that this ORF will have high coverage.

Since my metagenome comes from the mesopelagic zone, I expected that a photosynthetic gene would be less common in my sample. I selected the PsbK protein from the PFAM database, and blasted it against the same ORF database, in the same way as before. Unfortunately, this did not return any hits. I tried again with a "full" FASTA file from PFAM (rather than a "seed") but still got no hits. I tried a second photosynthetic protein family (TspO_MBR) and this time got a single hit—ORF c_000000054966_1. I would expect this ORF to be less abundant than the previous one.

7. Bar Graph.

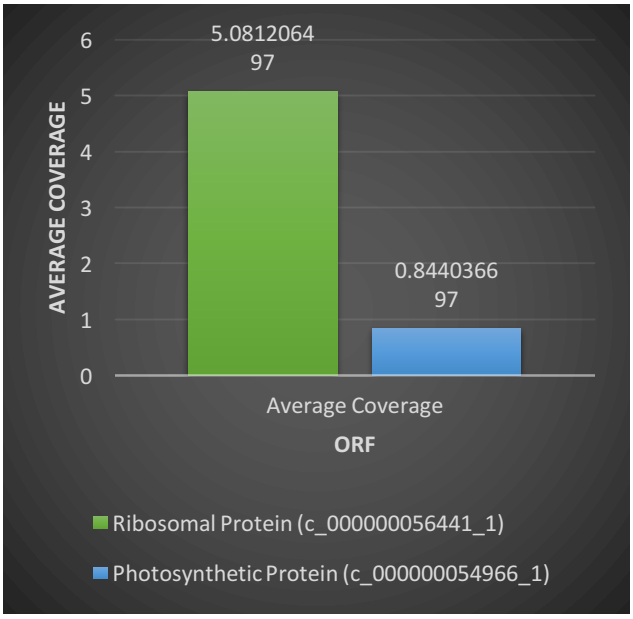

Figure 5 (officially named "Figure 1"). Average coverage for two ORFs associated with ribosomal and photosynthetic activity.

8. As you scroll through the data file…

When I sort my file by coverage, I see that the first seven results have a coverage over 100, then there are a few in the 90s and 80s, followed about ORF's with coverage in the 70s. The average of the average coverage sizes is 5, and the median is 3. Once again, I don't have interproscan results yet, but I can search for the ORF with the highest coverage in my ORF file to get the protein sequence and then do a blastp. It appears to be a transposase of length 226.

9. Take a look at the mappings of each of the metagenomes…

My sample is from the 600m deep in the mesopelagic zone. I used it as the basis for a comparison against two other samples from similar depth (sample ERR598999 [600m], and sample ERR599008 [790m]). The alignment was fairly sparse (Figure 6).

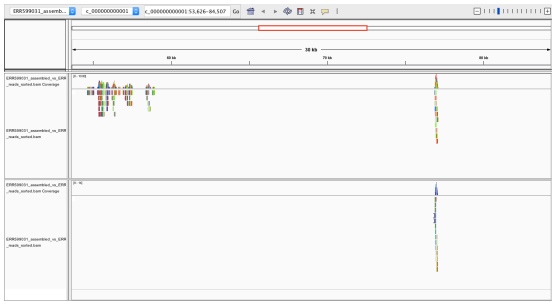

Figure 6. Screenshot of (sparse) mapping.

Both samples aligned with mine from ~9,000bp – 11,000 bp, though there was higher coverage for ERR598999. (Figure 7).

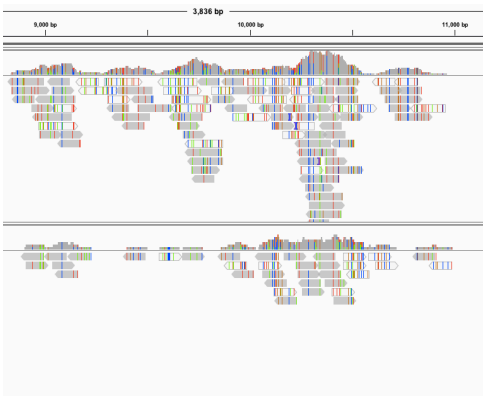

Figure 7. Other genomes mapped to mine. bp 9,000 - 11,000.

ERR598999 mapped to my genome again from ~55,000 – 59,000bp, but not ERR599008. Both line up again from ~76,000-77,100bp, both with coverage around 9-10. Finally, there is some alignment around 111,000-114,000bp (Figure 8).

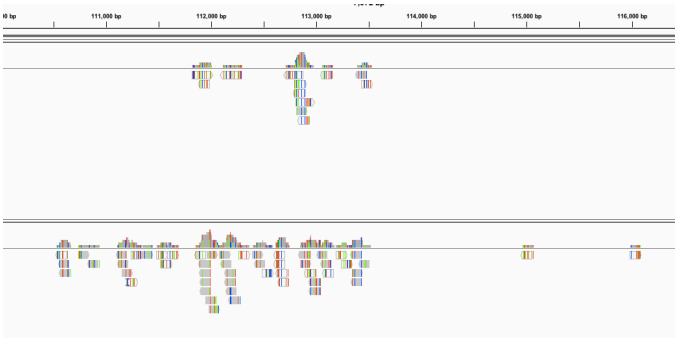

Figure 8. Mapping between 111,000-114,000bp.

There are a handful of other short regions where one or both of the metagenomes map back to mine, but these are the four most significant places.

9a. Describe the differences you observe in the *relative coverage*.

In most places ERR598999 had higher coverage than ERR599008. This means that in the few places where genes *are* shared between the metagenomes, they are more abundant in the ERR598999 sample than in the ERR599008 sample. Interestingly, this higher-abundance sample was taken from a depth more similar to mine than the lower-abundance sample. It is also closer to me geographically.

b. Provide an example of different patterns of *single nucleotide variants* between the metagenomes…

There are a few places where one metagenome has SNVs that the other doesn't. Around 60bp for example, the ERR598999 sample has many copies of a 'T' in its reads, which differ from the reference genome, while the ERR599008 sample aligns more closely with the reference. This might suggest that this variant arose and propagated only on a particular population of a species carrying this

gene, which lives in the ERR598999 environment (South Pacific (near the Marquesas). In a few other places, the two samples are more unified in their disagreement with my reference genome, for instance at 10,343-10,345. Both samples seem to carry the same SNV that differentiates them from my assembly. This could indicate a sequencing error in my assembly, or a real difference between populations of a species in both regions vs. my own.

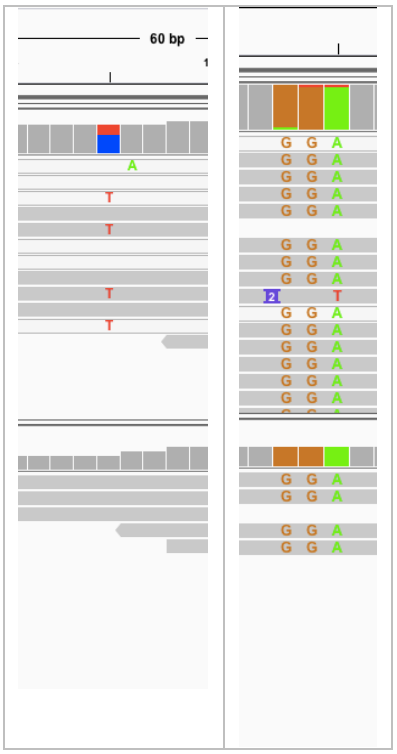

Figure 9. Two SNV examples.

II. Mini Research Question

10. Write either a question or generate a hypothesis about the *relative coverage* of this set of genes with respect to your project datasets. This question/hypothesis should include a comparison between your own project dataset and another dataset, and it should be couched within the larger ecological context.

I seem to have relatively high concentration of oxygen in my sample (1.63999 µmol/Kg) compared to the sample I was looking at from the Southern Ocean, near Antarctica (sample ERR599008). Thus, genes related to oxidative phosphorylation might be more abundant in that sample than in my own (though oxygen is not always required for oxidative phosphorylation.)

I still don't have interproscan results. A PFAM search for Oxidative phosphorylation yields 70 unique results, the first of which is MAM33 Mitochondrial glycoprotein. I'll assume the presence of this one gene relays some information about the ubiquity of oxidative phosphorylation genes generally. I downloaded the seed file for this protein and blasted it against my ORF file, to find matching ORFs. There were no matches.

I also planned to blast the same seed file against the ORF results for this other dataset, ERR599008. This would identify any matching ORFs in their dataset. However, this file does not appear to be in the shared class folder.

Since this assignment is late and I am progressing through it terribly slowly I won't continue searching for proteins in this pathway that I can find in my dataset and other. But I'll state that the strategy I *would* use would be to identify ORFs via blastp, using my ORFs and then another person's ORFs as the database, and a Pfam seed file as the query. Then I would search for those matching ORFs in my ORF_coverage.txt file, and in that of the other person, make note of the average coverage of each, and then calculate an average coverage for those averages. This would allow me to state which dataset (mine or that of another student) had higher coverage for the given gene, which is taken to be reflective of the oxidative phosphorylation pathway. This could be visualized as a single bar chart (as in an earlier question) or perhaps a series of stacked or overlapping bars, for each matching ORF. Something like this: