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, and explain what this mapping tells us about the relative genome structure of the two genomes that we mapped. If we compared this genomic region in a dot plot, what would it look like? Describe at least one biological mechanism by which this may have occurred.**

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.

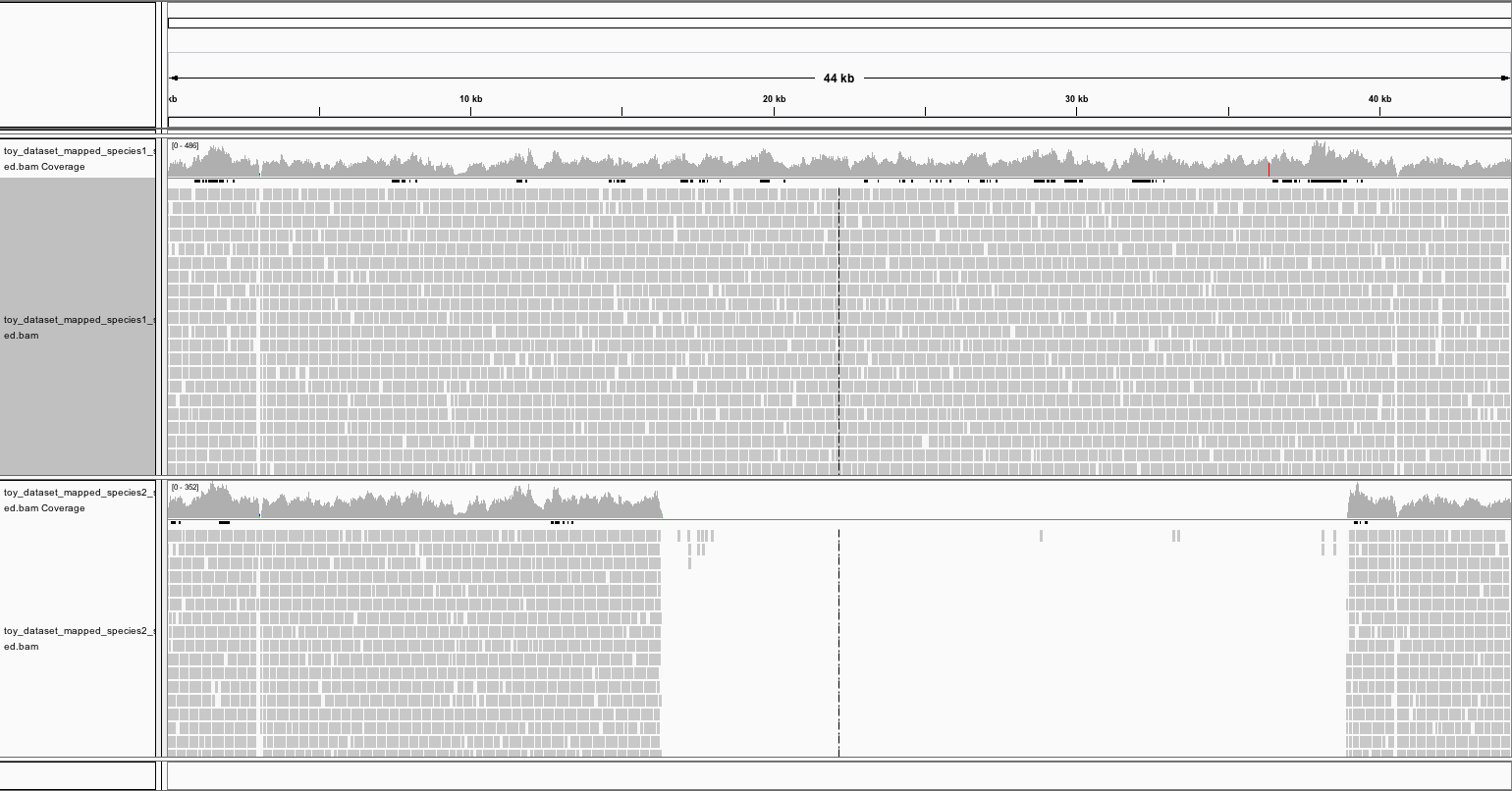


Figure . Toy dataset comparison of two species to reference genome.

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.

**2. Do you see evidence of misassemblies? If so, describe where you see evidence for this and what this evidence looks like.**

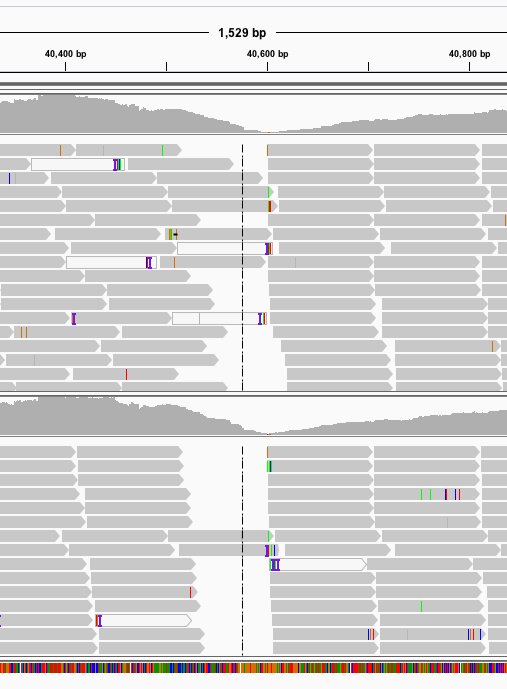


Figure . "Edges" where no reads overlap in toy dataset.

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 also indicate that species 1 and 2 simply contain a gene that the reference lacks.

**3. Do you see any evidence of single nucleotide polymorphisms? If so, describe where you see evidence for this and what this evidence looks like.**

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 of specific genes in your sample, why couldn't you simply count the number of times your gene appears in your assembly?**

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? Biologically speaking, what does this indicate? (Keep in mind this is a metagenome from a population of individual organisms vs an assembly, not an individual vs. an individual.)**

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.

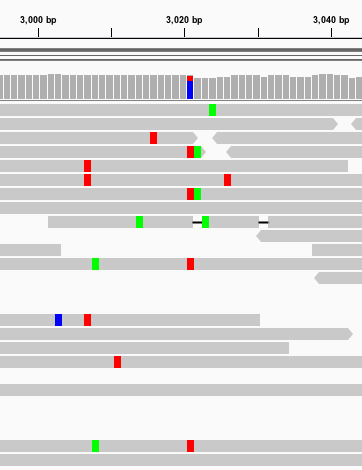


Figure . Evidence of SNPs in project dataset.

**6. Go back to your Interproscan files and find two ORFs that you're interested in. Choose one that you think might be really abundant in a sample (a housekeeping gene, for example, that might be really common) and choose one that you think might be more specialized and only found in specific types of microbes. Describe the ORFs you chose and which one you predict to have higher coverage.**

**7. Make a bar graph showing your results and submit it as 'Figure 1' for this week's lab writeup.**

**8. As you scroll through the data file reporting the average coverage of all of your ORFs, which ORF had the highest coverage? What did it encode, according to your Interproscan file? Speculate on why that gene may have had the highest coverage of all the genes in your dataset.**

**9. Take a look at the mappings of each of the metagenomes to your first contig. (Please include a screenshot of this mapping so I can see what you're looking at.)**

**a. Describe the differences you observe in the *relative coverage* between the different metagenomes, and explain what this means biologically.**

**b. Provide an example of different patterns of *single nucleotide variants* between the metagenomes, and explain what this means biologically. What might this tell us about microbes of the same species living in these different habitats?**

# 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.**

**Describe your results and create at least one graph to visualize those results. This should represent a mini 'Results' section in a lab report or paper. Interpret your results within the context of the ecosystem you are investigating. This should represent a mini 'Discussion' section in a lab report or paper.**

Text here.