Taxonomy_Matchup

Madison Hobbs

7/5/2017

Match up allCounts with the NCBI Taxonomy

Question 1: Are there duplicates?

Answer: yes.

There are no CDS/AS_CDS genes whose Geneid is duplicated in our data. However, there are 68 genes whose genename is duplicated. The CDS always have the same number as the AS_CDS. Most appear twice, but some appear 7 (insA) or 11 (insH1) times for CDS and AS_CDS each.

The vast majority (59) have the same bnumber across repetitions of a genename, but some do not, and these are: insA, insB1, insC1, insD1, insE1, insF1, insH1, insI1, and insL1. For the genes that have the same bnum across repetitions of a genename, they have a different number at the end of their Geneid : for example, for b4494 arpB, it has 1475:NC_000913.3 and 425:NC_000913.3.

Fortunately, the taxonomy repeats genes too! The only genes repeated in the taxonomy are CDS. There are no genes which are repeated in the taxonomy and which are not repeated in our data. 27 genes (matching genenames; 73 when matching bnums) are repeated in both the taxonomy and our data. Of the CDS/AS_CDS genes of our data, there are 41 genes which are repeated in our data but aren't repeated in the taxonomy data. All of these genes are to be found in the taxonomy, although they are not repeated.

For genes that were repeated in the taxonomy and which shared the same bnumber (these coincided exactly with our genes sharing numbers or not), the genes in the taxonomy would say a different "part" like for crl b0204, they have part=2%2F2 or part = 1%2F2.

Question 2: How well do our genes match up with the taxonomy?

There are no CDS or ncRNA genes which are in the taxonomy that are not in our data. There are also no CDS or ncRNA genes which are in our data that are not in the taxonomy! There are the same number of rRNA (22) and tRNA (89) in each data set, but I didn't expressly check them to be the same.

Question3: IGR's

NOTE: I removed rows in taxonomy of region "repeat region," "exon," "region," and "STS." They are easy for me to add back in if we'd like to see where those are being skipped over as well.

These are the IGR's which have genes between them:

- insL1 (b0016) and hokC (b4412) have mokC (b0018) between them.
- narI (b1227) and tpr (b1229) have rttR (b4425) and ncRNA rna40 between them.
- hcaT (b2536) and iroK (b4706) have hcaR (b2537) between them.
- hcaR (b2537) and hcaE (b2538) have iroK (b4706) between them.
- smpB (b2620) and intA (b2622) have ssrA (b2621) and tmRNA rna88 between them (of course it does; if you remember this is the gene with the highest read counts; in the millions!)
- ypjF (b2646) and ypjA (b2647) have psaA (b4645) and tmRNA rna89 between them.

- cspA (b3556) and hokA (b4455) have mokA (b4647) between them.
- mokA (b4647) and insJ (b3557) have hokA (b4455) between them.