Erika Maldonado-Rosado

1 March 2019

PanX

BINF 6203

**Introduction**

For this lab we are asked to use PanX, it is a program for pangenome analysis and production of core genome phylogenies. This program is available as a web server or a github package. The authors of PanX make a collection of precomputed comparisons available for student and researchers need. For this lab we are using the study of prochlorococcus by Kashtan to help us through this lab for comparisons of prochlorococcus genome.

**Method**

For this lab we are examine the collection of prochlorococcus available to us in PanX. When looking at the collection I used the web server due to it being easier to use. Using the web browser is very straight forward. Under species you can choose which species you want to look at and if it is not available you can type it in. There is core vs Accessory, which is displayed as a pie chart. There is gene count rank distribution, gene length distribution, gene cluster table where you can see different genes and their names. Lastly, there is sequence alignment and strain tree.

**Results**

**How many core genes and accessory genes are present in this collection?**

There are 1216 genes out of the 5410 genes that are in the core genes. There are 4194 genes out of the 5410 that are accessory genes.

**What is the average number of genes in each genome (you can probably approximate this by looking at a few examples)?**

There is about 1.7 mb of genes in each genome.   
**What is the most diverse gene in the core genome?**

The most diverse gene in the core genome is livF and its diversity is .56.

**What about the most diverse gene in the core genome that does not have any duplications?**

The most diverse gene in the core genome that does not have any duplications is a gene that does not have a name. It is known as a putative annexin. It has no duplication with a diversity of .447.

**What are the least diverse genes in the core genome?**

The least diverse gene in the core genome is non and it is known as 16S ribosomal RNA. It is duplicated with a diversity of .015.

**Look at the top five non-diverse genes. Are these genes that you would expect to be conserved in cyanobacteria, and why?**

The top 5 non-diverse genes are 16S ribosomal RNA, 23S ribosomal RNA, rbcS, psbA, and rbcL/\_cbbL. These are genes that I believe would be conserved, because they are important to photosynthesis.

**Compare the trees for the non-diverse ribosomal genes to the “species” tree produced by PanX. Do they have the same topology? If not, can you identify congruent regions between the trees?**

I got the first 10 ribosomal genes and I clicked on them to view them on strain tree and gene tree. They do have the same topology region.

**Take a look at the sequence alignments of these genes. Are the snp variants between the genomes scattered randomly, or are there obvious patterns of mutation?**

By looking at the sequence alignment you can tell that there are obvious patterns of mutation because some of the gene codes are very similar. However, there are a few genes that are slightly different and only similar in certain areas.

**Examine some genes in the accessory genome. It’ll be hard to decide exactly what to look at here, but for instance, take a look at psbM and psbF. What do these genes do? Are these genes in all the genomes? Do they have the same or different patterns of presence/absence in the core genome (i.e. are they present or absent in the same branches)? See if you can find a literature reference that explains what the “psb” system these genes are part of does, and what a complete set of psb genes would look like.**

psbM is a putative photosystem II reaction center of the M protein, while psbF is a cytochrome b 559, beta subunit. The psbM and psbF share the same amount of genomes that they are not in, which is 10. If they are both present, then they are both seen in all but 4. However, individually they are both not seen in the exact same ones. Psb are photosystem II promoters, they are encoding reaction proteins in photosystem II. A complete set of psb genes would show you where each reaction would occur in photosystem II, because each time a reaction would occur psb gene would be active.