Practical Bioinformatics for Biologists (BIOS441/641) Course Project 2

Project title:
Comparative analysis of ecoli genomes [use Linux command-line tools]

Goals:
1) Practice Unix commands we have learned

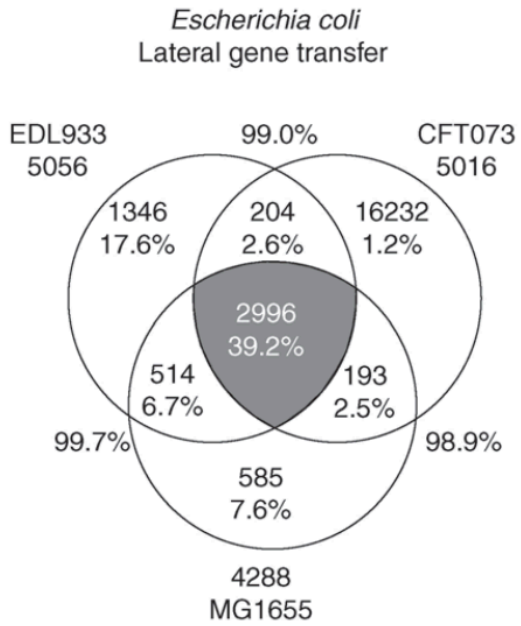
2) Learn to design Unix one-liner to pipe commands for processing biological plain text data

3) Learn how to perform comparative analysis of bacterial genomes

Background:
A bacterial genome is a pool of genetic elements with different evolutionary history (http://piel-l.org/blog/wp-content/uploads/2010/01/Bentley-2009-pan-genome.pdf). When you compare genomes of strains of the same species, some genes are present in all the genomes. They are called **core genome**. Some other genes are found in some but not all genomes. They are called **variable genome**. The remaining genes are only restricted to one genome. These genes are usually called orph**an** **O**pen **R**eading **F**rames (**ORFans**, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1559721/). All the three together is termed **pan-genome** (http://www.cbs.dtu.dk/CBS/courses/brazilworkshop/files/medini_COiGD_2005.pdf).

The origin of these ORFans has been a focus of comparative genomics research since 15 years ago. Some studies suggested that they are transferred from distant organisms and viruses (http://www.lcg.unam.mx/frontiers/files/frontiers/Daubin%20V.pdf, http://www.biomedcentral.com/content/pdf/1471-2148-6-63.pdf).

It is likely however that many ORFans found in the comparison of closely related strains might have homologs in more distantly related species.



*Here, the '**pan-genome**' is encompassed within the Venn diagram, and the '**core genome**' is represented by the shaded region that denotes the genes shared among all three genomes.*

Numbers inside the Venn diagrams indicate the number of genes (and percentage of total) found to be shared among the indicated genomes

Current Opinion in Microbiology 2005, 8:572–578
http://evolucion.fcien.edu.uy/Lecturas/Lawrence&Hendrickson_2005.pdf

My lab has developed a new algorithm implemented as a C computer program called *ORFanFinder,* http://cys.bios.niu.edu/orfanfinder/, http://cys.bios.niu.edu/yyin/teach/PBB/2016-ORFanFinder.pdf, which can take a BLAST result as input to classify the gene content of a bacterial genome into groups of ORFans restricted to genomes of the same species, genus, family, order, class and phylum. Here the BLAST result is from the comparison of a target genome against the NCBI nr database.

Figure 1: Define ORFans in the context of taxonomic ranks. "1" means the gene is found in but restricted to genomes of the same rank as the studied genome; "0" means the gene is not found in any genomes of that rank.

Restricted to genomes of the same →	Species	Genus	Family	Order	Class	Phylum	Super kingdom
$ORFan_A$	0	0	0	0	0	0	0
$ORFan_S$	1	0	0	0	0	0	0
$ORFan_G$	1	1	0	0	0	0	0
$ORFan_F$	1	1	1	0	0	0	0
$ORFan_O$	1	1	1	1	0	0	0
$ORFan_C$	1	1	1	1	1	0	0
$ORFan_P$	1	1	1	1	1	1	0

Questions:

This project is to study some sequenced *Escherichia coli* genomes to retrieve ORFan genes and study their associations with phages. Questions to be addressed include:

1. If just compare ecoli MG1655 with two other ecoli strains in terms of their gene contents, how many ORFans do they have?

2. If take ecoli MG1655 genome and compare it with all the sequenced bacterial genomes, how many ORFans will be found?

3. What makes the difference between the results from the two comparisons?

4. Do ORFans tend to be derived from phages and do younger ORFans have even a higher tendency for so?

Detailed steps:

****

Under your home, there is a hidden history file, I will also check this history file when I grade your report

****

Student	Strain	Student	Strain
Bracero	042, 536, MG1655	**LaCasse**	DH10B, MDS42, MG1655
Cooper	55989, ABU, MG1655	**Moreno**	E24377A, P12b, MG1655
Dean	APEC_O1, APEC_O78, MG1655	**Nissenbaum**	W3110, LF82, MG1655
Gautam	ATCC_8739, BW2952, MG1655	**Olawuni**	LY180, NA114, MG1655
Gavert	B_REL606, CFT073, MG1655	**Peterson**	O103_H2, O104_H4_2009EL_2050, MG1655
Gavlin	E24377A, ED1a, MG1655	**Pischl**	O104_H4_2009EL_2071, P12b, MG1655
Haws	ETEC_H10407, HS, MG1655	**Protity**	O104_H4_2011C_3493, PMV_1, MG1655
Holthuijzen	IAI1, IAI39, MG1655	**Ramirez**	O111_H__11128, S88, MG1655
Kelly	IHE3034, JJ1886, MG1655	**Schladweiler**	ABU, IAI39, MG1655
Serrano	CFT073, W3110, MG1655	**Xue**	MDS42, NA114, MG1655

1. Go to NCBI ftp site and download three ecoli folders (see above Table for your three genomes). Each folder corresponds to one particular ecoli strain and contains various data for the sequenced strain, like genomic DNA seq, predicted protein seq, annotation info, etc.

2. Make a folder in your home called project2 and move the three ecoli folders to there

3. Some of the ecoli folders have more than one faa files. That's because they have plasmids in addition to the main chromosome. For each of these ecoli folders, go in there and concatenate (cat command) the faa files as one single faa file, give it a name, like strain1.faa

4. For each ecoli strain, take its faa file (e.g. strain1.faa) as query to do BLASTP searches against the faa file of the other two strains (e.g. strain2.faa and strain3.faa). So you will have to do 6 BLASTP searches in total. Save the BLAST results with a name like strain1.vs.strain2.blst.out.

5. Create Unix one-liners to process each of the 6 BLAST output file (E-value < 1e-5) to do the following:
- How many proteins in the query strain have hits in the subject strain
- Save the protein IDs in a file
- How many proteins do not have hits in the subject
- Save the protein IDs in a file
- How many proteins do not have hits in any of the other two strains (these are ORFans)
- Save the protein IDs

At the end, make **Table 1** as follows:

Strain (# of proteins)	Hits in 042	Hits in 536	Hits in **MG1655**	ORFans
042	4000			
536		4500		
MG1655			5000	

Draw a venn diagram as **Figure 1** (include the ORFan numbers from Table 1) in PowerPoint as shown in slide 5 of http://cys.bios.niu.edu/yyin/teach/PBB/project2-venndiagram.pdf

6. For MG1655, I have already done the BLASTP search against nr database (can't let you do this as it took a really LONG TIME to finish and my server will likely be dead if all of you are running BLAST on it simultaneously for a few days). The file is at /disk1/ecoli/ecoliNRblast.bl. Take this file as input to run ORFan_Finder program.

The program is at http://cys.bios.niu.edu/orfanfinder/download/. The document is the INSTALL and README files when you uncompress the package. Download the program and install it in your project2 folder and run it to predict ORFans of different ages (you need to read the document to learn how to install and run it). You will need MG1655's tax ID, which you can search for it at NCBI's website choosing Taxonomy database.

7. Create Unix one-liners to extract the IDs for ORFans of different age groups (phylum and below). Save them as different files. Count how many are in each group and make **Table 2**. Compare the number of strict ORFans with the number of MG1655 ORFan in Table 1. Explain why they are different? How many ORFans in Table 1 now becomes older ORFans?

ORFan groups	# of proteins	Uncharacterized	Phage/prophage	Transposase
Strict ORFan				
Species ORFan				
Genus ORFan				
Family ORFan				
Class ORFan				

Phylum ORFan				
Total ORFan				

8. For each age group, take the IDs of ORFans in step 7 and extract them from the ptt file. Save all the files as .txt file. Transfer them to your Windows using SSH Secure File Transfer client (http://cys.bios.niu.edu/yyin/teach/SSHSecureShellClient-3.2.9.exe) or MAC (scp command). Open them in excel and make an excel sheet for each group (**Data S1 to S6**). Plot their length distribution (**Figure 2**).

9. Based on the info given in the .txt files that you saved, use Unix one-liners to count how many of them are annotated as uncharacterized protein, phage or prophage related, transposase and add the numbers in **Table 2**?

Check list for your report:
Table 1, Table 2, Data S1 to S6, Figure 1, Figure 2.
All questions are answered

Report format:
Introduction: Use one sentence to explain pan-genome, core genome, variable genome, ORFans, respectively. See references above. Explain what you want to study in this project: see questions to be answered above.

Methods: Feel free to use the above detailed steps with your own modification/addition. **Explain what you are doing for that step and why you are using that particular tool(s) in that step**.

Results: Explain in details each Figure, Table and Data that you have made. What is the data in there and what the data tells you?

Conclusion: Use one sentence to answer each of the questions above.

Create a zip file with the report document (no need to include the fig/tab in the text, but do cite each of them in the right place of the text), all the figures and tables, as well as Data files. Name the zip file "YourLastName-project2.zip" and sent it to me. Screen shots are encouraged but not required.

Report should be in word doc file.

Check out one example from previous student Bill Wysocki:
http://cys.bios.niu.edu/yyin/teach/PBB/WysockiProject1Writeup.pdf

Report is due through email by the midnight of 12/13. Total 20 points.

Use office hours or email me for questions.

DO NOT COPY OTHERS WORK!!!