Team Projects

*Please sign up with your name. Each team will be expected to create a GitHub repository for your team project. You can sign up for a github account here.*

*Report template:*
*https://docs.google.com/document/d/1zYikBlzpxgnjoNMtM82d6B1EJZaIx_HS1nxEAj9kPIw/edit*

# 1. Phylogenetics and Evolutionary Project 1

**Team Members:** Eric Gordon, Kaleigh Russell, Hoang Vuong, David Haisten
Emails: egord003@ucr.edu, david.haisten@email.ucr.edu, kruss002@ucr.edu, hvuon007@ucr.edu
**GitHub Repo.**

**Project Outline:**

For a set of species of bacteria, using the genomes.

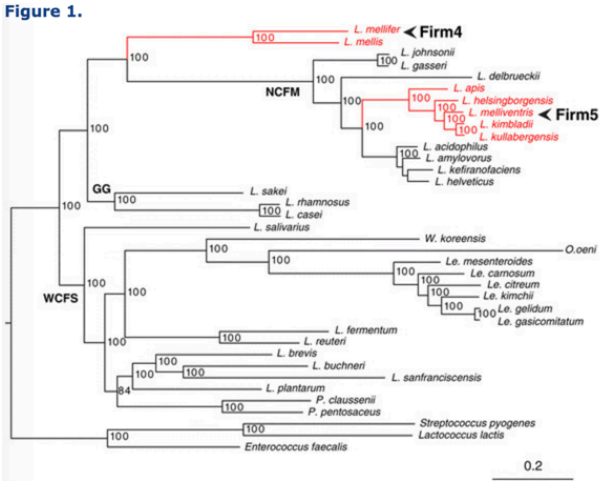
a. Identify the orthologs (or use existing mappings).
b. Align these orthologs at the Protein level
c. Map these alignments back to Coding Sequence
d. Compute the dN/dS score for genes (ranking of genes)
e. Make a tree of taxa

For a group of closely related microbial species which inhabit the same or different niches, what are the major genomic differences between species and can they be explained evolutionarily by any factor (perhaps by habitat). For instance microbes from a “species” designated by 16S rDNA phylotyping with >97% sequence identity may be extremely divergent when comparing ortholog divergence as well as comparing gene classes. We will create a pipeline for handling raw fastq data of a microbe’s genome to assemble, annotate and compare to other already annotated genomes. We will identify orthologs of this newly assembled genome for constructing alignments, making gene and species trees and computing dN/dS of genes. We will also compare unique genes found only in a single taxon to see if they provide insight into this species ecology. We expect to find selection pressure and/or specific presence of genes for particular species which may relate to that microbe’s biology.

Possible Ideas:
Pool together Lactobacillus Kunkeei + Lactobacillus from Firm4/5 clade then feed other lactobacillus (raw genomes) into more firm 5 genomes here
Lactobacillus ozensis as a potential raw genome to use.

Pipeline:
Assembly : Velvet (Eric)
Annotation Prokka (Eric)
Ortholog assessment: OrthoMCL (Kaleigh/Eric) 1 2 3 4 5
Alignment: ? (David)
Phylogenetics: RaxML (David)
dN/dS evaluation: PAML (Hoang)

**Figure 1.**



List of taxa for sampling: **G = genome available** In blue = data added to github

Firm 4:

L. mellifer (G)

L. mellis (G)

Environmental members of Firm 4 sister group:

L. johnsonii (G) = 9 sequences here

L. gasseri (G) = 14 assemblies

L. delbrueckii (G) = 27 assemblies

Also Firm 5 clade and sister group/

Firm 5:

L. apis (G)

L. helsingborgensis (G)

L. melliventris (G)

L. kimbladii (G)

L. kullabergensis (G)

wkB8 ( ) actually is annotated?

wkB10 ( ) actually is annotated?

Firm 5 environmental sister group:

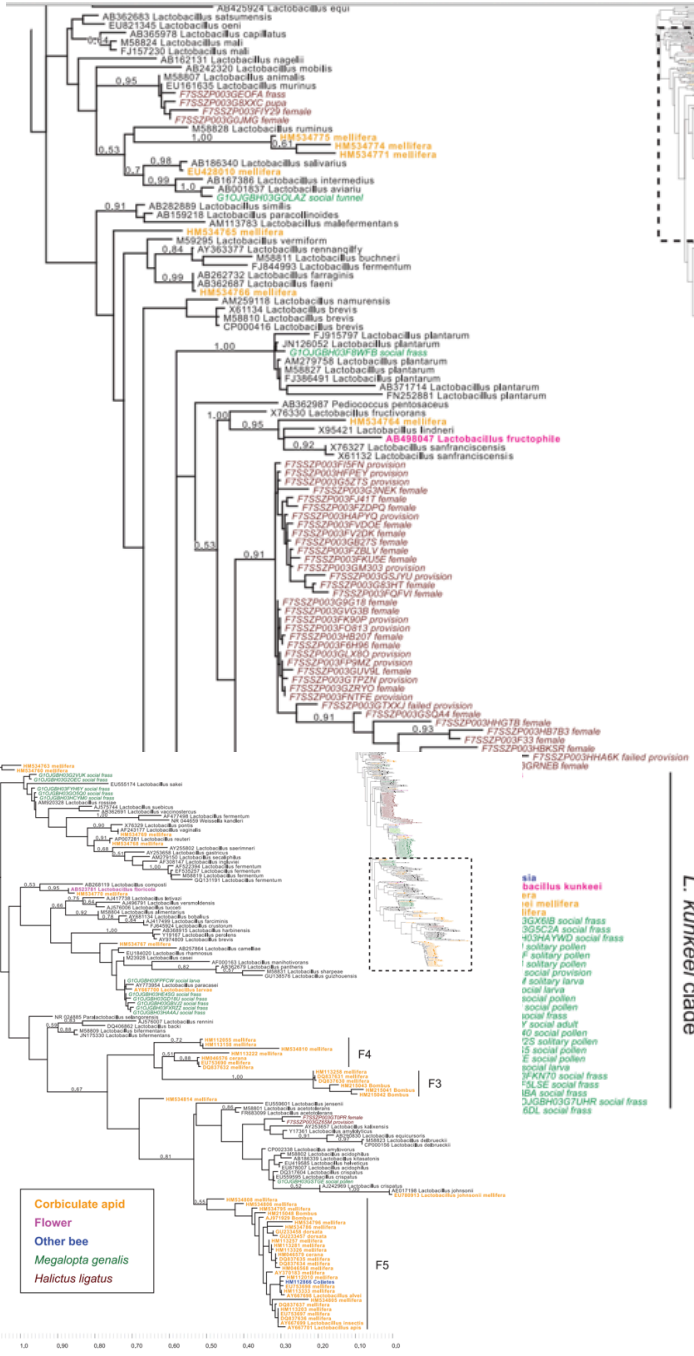
~~L. acidophilus (G)~~

~~L. amylovorus (G)~~

~~L. kefiranofaciens (G)~~

~~L. helveticus (G)~~

several **others**






L. kunkeii strains:

strains 1- 12 (**G's**)

**are there any flower sourced genomes? (?)**

Environmental sister strains to L. kunkeii:

L. ozensis (G)

L. sanfranciscensis (G)

Pediococcus pentosaceus (G)

L. fructophile / florum (G)

L. fructivorans (G)

L. lindneri (G)

Lactobacillus outgroups (maybe sister to everything else):

L. plantarum (G) 48 and some bioprojects

~~L. brevis~~

~~L. crispatus~~

170 total Lactobacillus species genomes

# 2. Phylogenetics and Evolutionary Project 2

**Team Members:** ✿*~~Austin Baker~~*✿, Alex Knyshov, **Ryan Perry**

**emails: abake005@ucr.edu**, aknys001@ucr.edu, rperr003@ucr.edu

**Project Outline:** Rank loci from a set of eukaryotic taxa for phylogenetic informativeness

Input data:

Step 1: Taxon selection

- Tab delimited taxon file
- Selects taxa from input phylip files

Step 2: Calculate % divergence within each locus

- Decide what statistic to use
- Assign values to file
- Sort new file by divergence value

Step 3: User selects min and max divergence value

- filter file by user selection
- select from files with matching ID from phylip files

Step 4: Tree filtering

- Gene tree filter?
- How well resolved are the trees? How do we measure that?

Step 5: Phylogenetics

- Concatenate data
- PartitionFinder
- RAxML analyzes

## 3. Transcriptomics - 2 teams :

**Team Members:** Tyler Dang (tdang004@ucr.edu) , Qihua Liang(qlian003), Mateo Espinoza (mespi010@ucr.edu), Andrea Rivera, Tianran JIA, Jui-Yu Liao

**Project: How does gene expression estimation vary based on different alignment methods and DEG stat methods**

[JS says: Notes: Please also look in the literature for similar attempts to compare expression estimate by different methods. This could be in worms, human, etc - usually when a new tool is produced there are attempts to compare accuracy. Also see Mortazavi et al 2008 for one of the first RNASeq papers and how they validated their results]

**Group 1: Tyler (tdang004@ucr.edu), Mateo (mespi010@ucr.edu), Andrea (arive019@ucr.edu)**
**Group 2: Qihua(qlian003@ucr.edu), Tianran (tjia003@ucr.edu), Jui-Yu (jliao010@ucr.edu)**

1. Download RNAseq dataset (2 conditions); 2 biological replicates (6-8 data sets)
    a. Model Organism: Yeast - see SGD for data downloads

    b. Conditions: Mutations? drug treatments? heat treatments?
2. Quality Control: trimming?
3. Alignment
    a. Alignment to genome:
        i. Bowtie2/Tophat(Group 1)
        ii. GSNAP
        iii. STAR
    b. Alignment to transcripts (choose some or all of these tools) - read more on these techniques [1], [2]. Note that eXpress is by same senior author as Kalisto, Kalisto is thought to be better / new so you may want to compare. Sailfish is a different approach, would like to know if it is comparable results?
        i. Kalisto
        ii. SailFish
        iii. eXpress
        iv. Salmon
4. Count reads: cufflinks, other tools
5. Normalization: cufflinks, egdeR, DESeq2
6. DEG statistics: edgeR
7. Functional Enrichment: (choose 1),
    a. Gene Ontology [relatively easy here] - can also use R code like this - see the GOStats R package - may want to use GO Slim instead of full GO.
    b. Pfam/Interpro Domain composition differences
    c. KEGG Pathway - do describe how you might compare?

Conditions to consider testing [added by Jason Stajich]

1. Oxidative stress “The yeast Snt2 protein coordinates the transcriptional response to hydrogen peroxide-mediated oxidative stress.”
    - BioProject PRJNA184040 - 18 runs: 3 genotypes (WT, 2 mutants) x 2 conditions (0 time point and 0.5hr time point) x 3 replicates - downloadable here (but also see my post [blog post here] on how to download faster or wait for this in class) - single ended
2. Annotating Low Abundance and Transient RNAs in Yeast using Tiling Microarrays and Ultra High-throughput Sequencing - 16 seq datasets from 4 replicates x 4 genotypes (WT and 3 mutants) - single ended
3. Aging in yeast - here is a large study (61 experiments incl replicates) unpublished - paired end - (you could choose a subset of these to work from) - whole datasetset is here
    - Compare to results from this paper which was with Microarrays? “SIR2 and other genes are abundantly expressed in long-lived natural segregants for replicative aging of the budding yeast Saccharomyces cerevisiae.”

## ~~4. Transcriptomics Project 2: Intron Retention~~

~~**Team Members: Tianran JIA, Jui-Yu Liao**~~

~~**Project Outline: Intron Retention events - how many transcripts have unspliced introns, does this frequency change under stress conditions**~~

~~References:~~

- ~~Global analysis of the nuclear processing of transcripts with unspliced U12-type introns by the exosome - This paper focuses on U12 introns, but what about all introns?~~
- ~~Intron retention is a major phenomenon in alternative splicing in Arabidopsis - this used old EST technology can you do an updated version with Arabidopsis RNASeq?~~

---

~~**In simple terms:** Given a set of sampled RNAseq experiments for a species with an assembled and assembled genome. I suggest gathering data for *Schizosaccharomyces pombe* as they have small introns (note that once you have created the tool you can test many datasets). Animal systems are interesting but the assembly and searches may take a while.~~

~~a. Assemble into a consensus set of mRNA transcripts with Trinity~~
~~b. Extract the annotated introns from genome annotation file~~
~~c. Identify which transcripts likely have an unspliced intron~~
~~d. Calculate a rate/frequency/count for transcripts/genes and compare this overall rate/value between conditions.~~
~~e. Any statistics about the size, position, sequence characteristics of the retained introns.~~

## 5. Variant and/or Population Genomics

**Team Members:** Alex Rajewski (araje002), Alex Plong (aplon001), Chrissy Dodge (cdodg001), Joseph Carrillo jcarr022@ucr.edu

Project Outline:

- We will use a dataset of 18 naturally occurring Arabidopsis genomes that have been sequenced in 30-50 bp reads. http://mus.well.ox.ac.uk/19genomes/
- We will examine SNP patterns across the genomes to determine:
    - shared/unique SNPs among the genomes
    - how many SNPs produce synonymous vs nonsynonymous changes
    - plots summarizing the frequencies of particular mutations (INDEL lengths, SNP base transitions)
- With the rough geographic data in the dataset, we will use STRUCTURE and/or a tree building software (MrBayes, RAxML, PHYML) to see if there is significant clustering of genetic diversity by location.

- Compare mutagenized experiment to wild-type. Can you predict/describe the types of mutations most likely caused by the mutagen? (AlexR says: I dont conceptually understand how we would do this. If we wanted to do, like, mapping I'm with you, but otherwise I dunno how to explain this.)

# 6. Population Genetics of Tuberculosis

**Team Members:** Zach Piserchia, Elizabeth Deyett

**Project Outline:** Calculation of various useful Population Genetics measures from given data on alleles.

a. SNP frequency
b. percent polymorphism
c. dn/ds ratio (indication of positive selection if > 1)
d. average expected heterozygosity (H)
e. nucleotide diversity ($\pi$)
f. automate extraction of allele data from raw sequence data (if time allows)

The focus of our project is to create a program to serve as a useful tool for calculation of various population genetics measures from allele data. We aim to include SNP frequency, percent polymorphism, expected heterozygosity (since we are working on bacteria this is just to add functionality for future projects), nucleotide diversity, and dn/ds ratios. Once complete, we aim to apply this tool to Tuberculosis and examine different populations (particularly resistant strains) in different regions of the globe. Since dn/ds ratios > 1 indicate positive selection, we also aim to identify important genomic regions for Tuberculosis, likely involved in resistance or compensatory mutations.

Once this analysis is complete, we hope to also implement an automated means of extracting allele data from raw sequences if time allows. This would result in a very robust tool that can perform entire analyses (we may also add more measurements later) in population genetics from collected samples. In other words, if we gathered hairs from a mammal and extracted the DNA, everything from then on could be analyzed by the program, greatly facilitating experiments by researchers in the field.