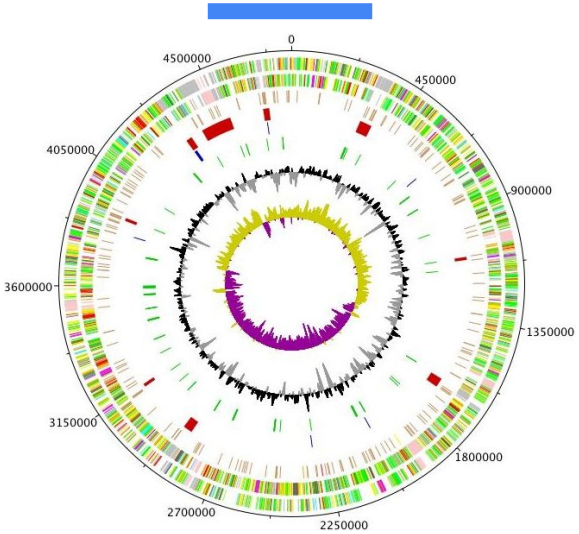
Phylogeny and genomic characterization of bee-associated *Lactobacillus* species and their closest relatives



Eric Gordon, Kaleigh Russell, Hoang Vuong, and David Haisten

Lactobacillus and bees

- Many *Lactobacillus* spp. are associated with bees and their environment (3 main clades * based on 16S data) but most in the genus are not associated with bees (Mcfrederick et al 2012).
- "Core gut microbiome": Firm4, Firm5, and *L. kunkeei* clade- important for nutrition.
- *Lactobacillus kunkeei* is a major symbiont of honey bees AND wild bees worldwide.
- It can be found in the honey bee gut, bee bread, honey, and honey bee larvae.

McFrederick, Q. S., Wcislo, W. T., Taylor, D. R., Ishak, H. D., Dowd, S. E., & Mueller, U. G. (2012). Environment or kin: whence do bees obtain acidophilic bacteria?. *Molecular ecology*, *21*(7), 1754.

Problems / Questions

- Looking at 16s data alone, may leave out relevant information on species relatedness and evolutionary history.
- Horizontal gene transfer?
- Are there particular genes important for living in bees?
- Do different species of bee associates have different niches within bees?
- Goals:

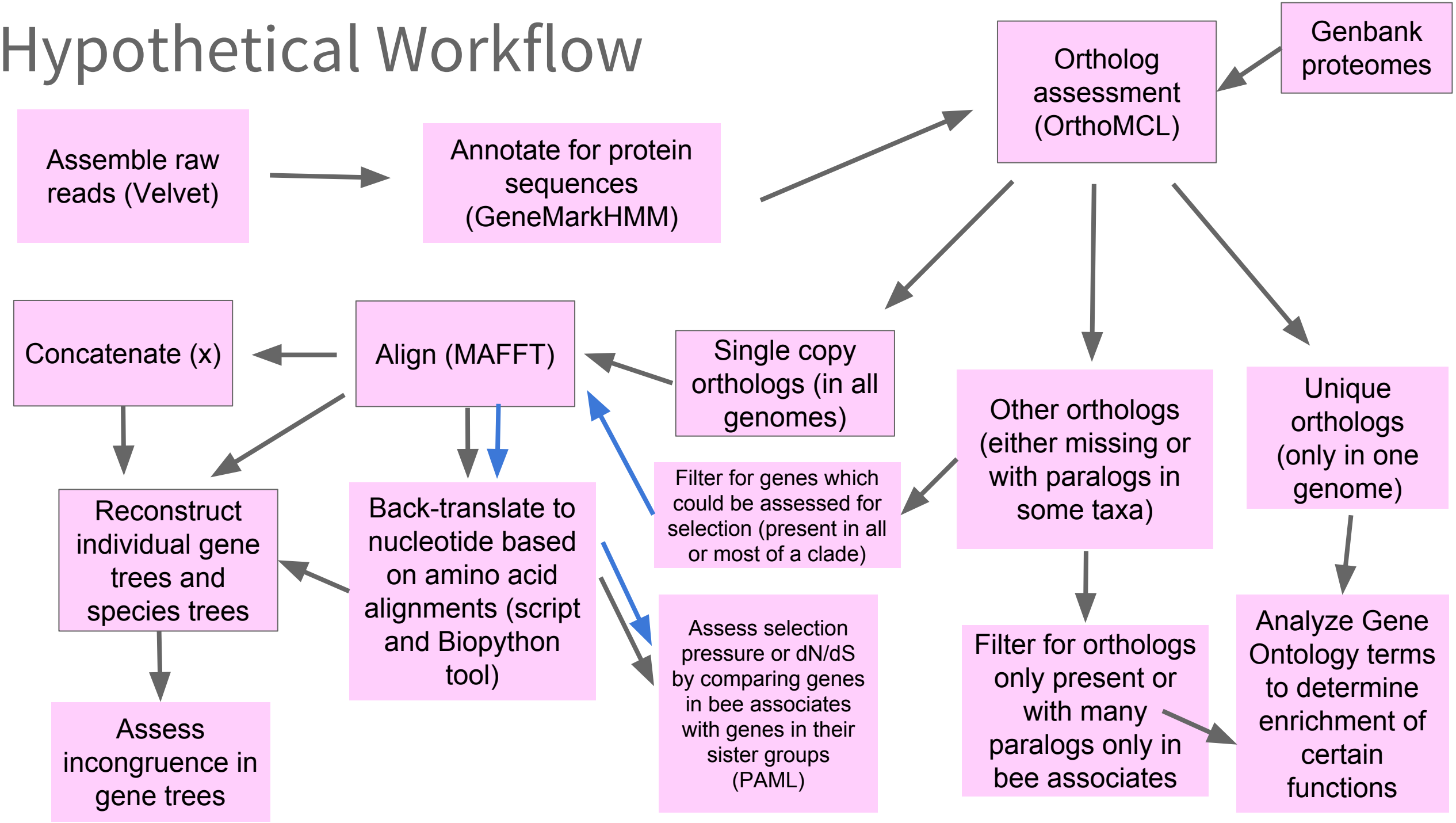
Create pipeline for analysis of bacterial genomes:

1. Assemble and annotate raw data
2. Identify orthologs
3. Construct alignments
4. Produce gene and species trees

Methods Overview:

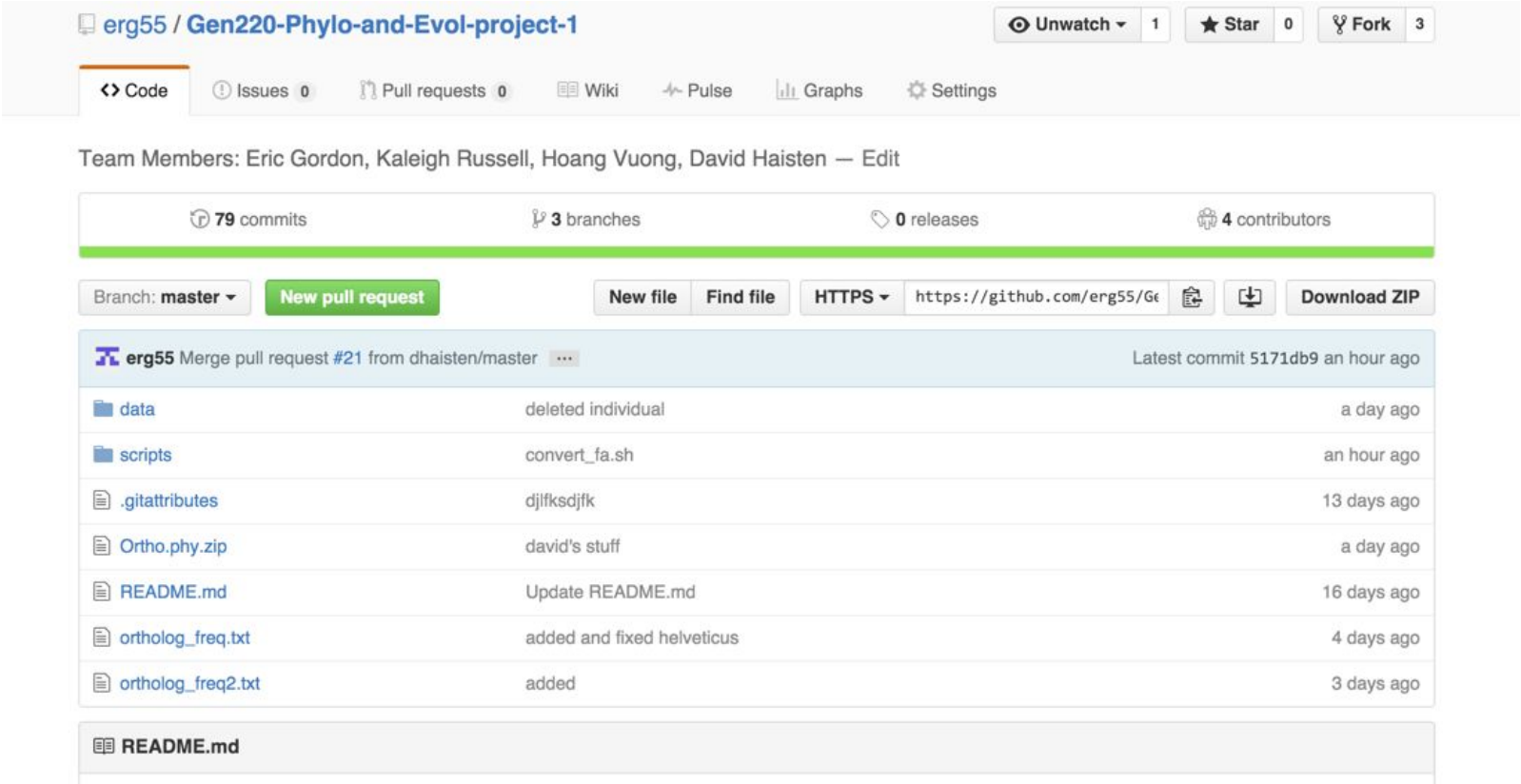
- Initially started with 36 annotated *Lactobacillus* spp. genomes from NCBI database -- including 12 strains of *Lactobacillus kunkeei* & 9 other species of bee associates along with 15 non-bee associated close relatives.
- *Lactobacillus brevis* as outgroup.
- Assembly and Annotation using Velvet and Prokka
- OrthoMCL assess orthologs of protein sequences
- Phylogenetic analysis using RaxML

Hypothetical Workflow



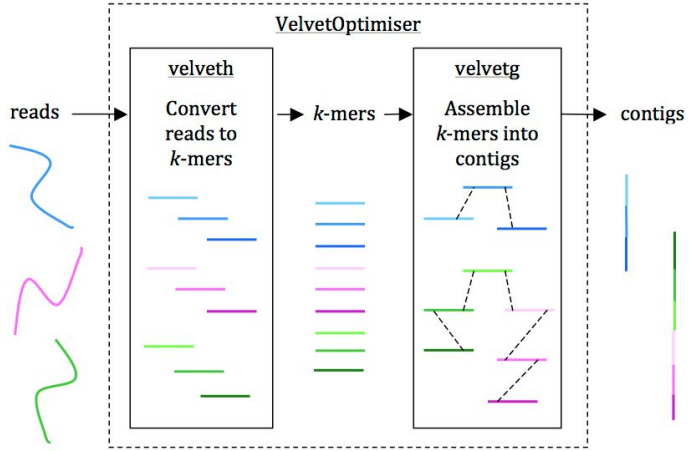
GitHub Repo

- Used GitHub repo to collaborate and share code and small data files
- Was not as straightforward as expected. Still slightly confused.



Assembly (Velvet) > Annotation (Prokka)

- Script for submitting array job to velvet per paired end Illumina data file for each genome.
- Velvet de bruijn graph assembler based on k-mers
- Prokka takes contigs from Velvet and rapidly annotates them using BLAST and HMMER3
- Feed newly assembled proteomes from Prokka into OrthoMCL.



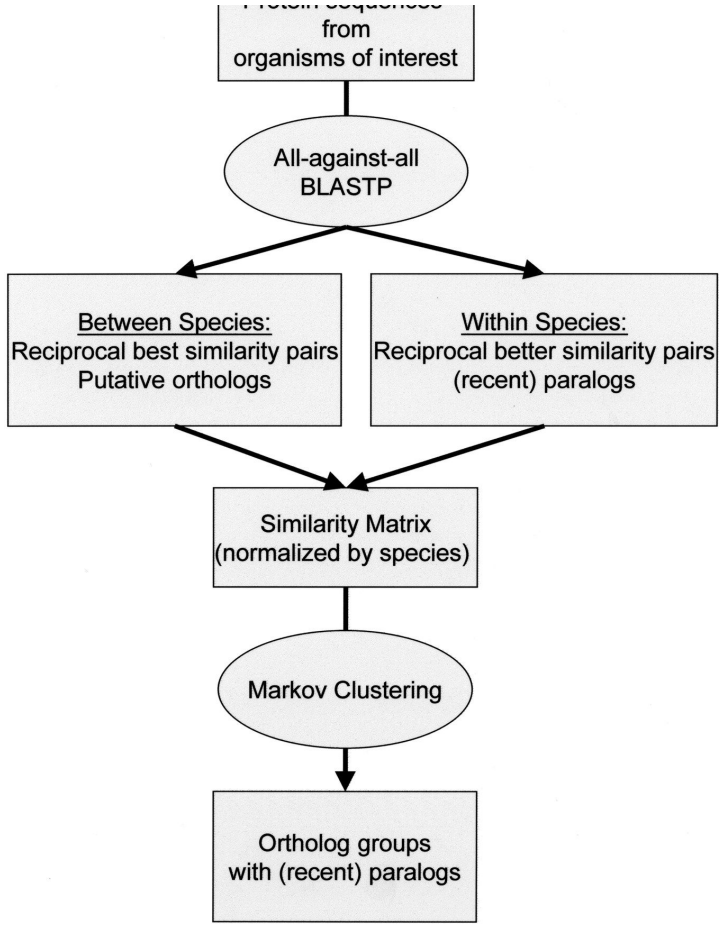
….In progress

Having trouble decoupling interspersed paired end data from SRA into two files (potentially because of new fastq format)

OrthoMCL

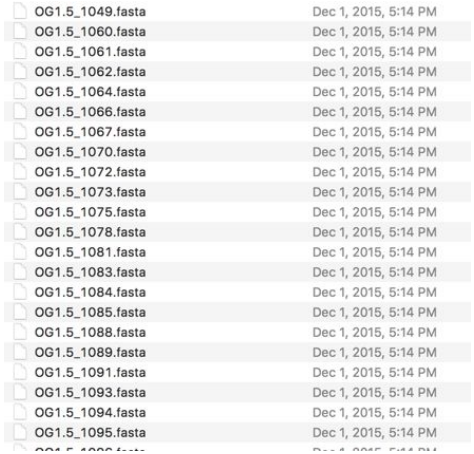
- Script to loop through fasta files organized in folders and rename by taxon and filtering out poor quality protein sequences (very short or with stop codons) with OrthoMCL and combine rest into one file
- Make blast db and Blast all-vs-all of proteomes. Keeping only hits with e-values less than e-5 and length >50%
- Load results into SQL database.
- Find orthologs, inparalogs, and co-orthologs through a series of 20 steps that create as many intermediate tables in MySQL
- MCL (Markov chain clustering) program
- Scripts for pulling out single copy orthologs and sequence fasta.

```
#version if can't get array job to work


module load orthomcl

# 38 files to be parallelized
unset folder

for i in {1..37}
do
folder=$(sed -n -e "$i p" arraylist.txt)

# folders must be a three or four letter unique abbreviation for the taxon, fasta file is in folder
for fastafile in ./"$folder"/*.fasta*
do
        orthomclAdjustFasta $folder " $fastafile" 2
done
done
```



Single copy orthologs

- 420 single copy orthologs (SCO) found.
- Accidental inclusion of proteomes of multiple strains per species for a couple of taxa when downloaded from genbank...removed from SCO assessment but then reincluded so these taxa may have artificial duplicate orthologs from each strain (will redo with newly annotated)
- First analysis containing all sequenced genomes of bee associates.
- More than previous studies which included distantly related outgroups (303; Ellegaard et al. 2015) but less than some others with either fewer taxa or more closely related taxa (530, 790; Tamarit et al. 2015).

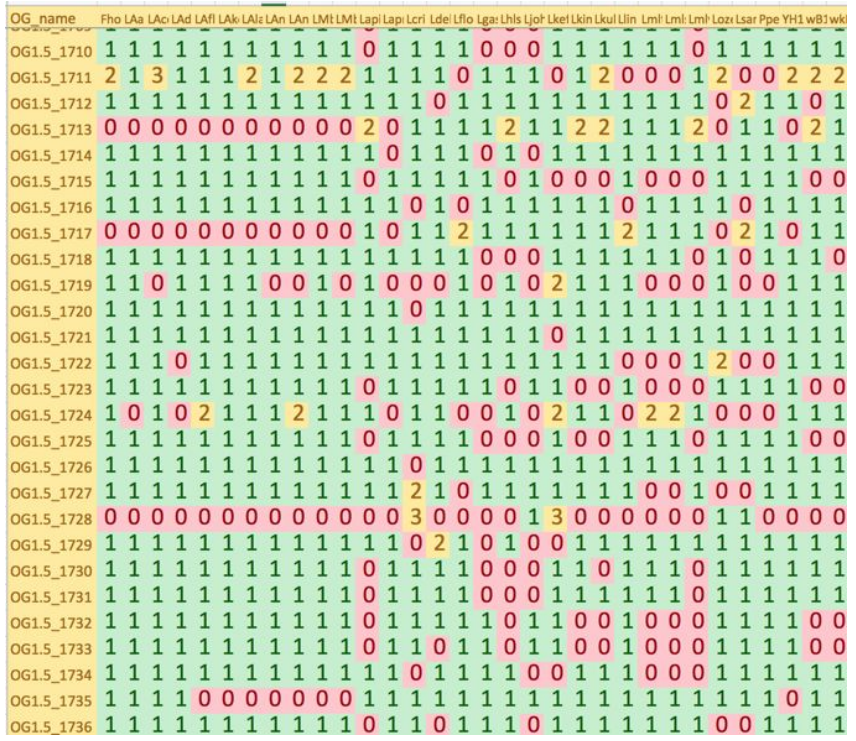


Ellegaard, K. M., Tamarit, D., Javelind, E., Olofsson, T. C., Andersson, S. G., & Vásquez, A. (2015). Extensive intra-phylotype diversity in lactobacilli and bifidobacteria from the honeybee gut. *BMC genomics*, *16*(1), 284.

Tamarit, D., Ellegaard, K. M., Wikander, J., Olofsson, T., Vásquez, A., & Andersson, S. G. (2015). Functionally Structured Genomes in Lactobacillus kunkeei Colonizing the Honey Crop and Food Products of Honeybees and Stingless Bees. *Genome biology and evolution*.

Additional OrthoMCL

- Cluster taxa based on similarity in number of shared ortholog for different ortholog groups?
- Define clades and look only within clades?
- Used custom script to pull single copy orthologs. Not simple to modify to pull out ortholog groups with more complex patterns.
- Scripts in development to parse out unique (might be easy) and other targeted genes by clade e.g., only present within individual bee associates (harder).



Gene Ontology (GO)

- Annotation information wiped out after OrthoMCL.
- Could use Biopython tools to retrieve again with accession number? Potentially, then couple gene identity with Uniprot to get GO terms to find enriched categories in unique bee associates orthologs
- Reach goal.

OrthoMCL output → AA Alignments

- 420 orthologs from raw OrthoMCL output needed to be converted into 420 individual alignment files and 1 Supermatrix

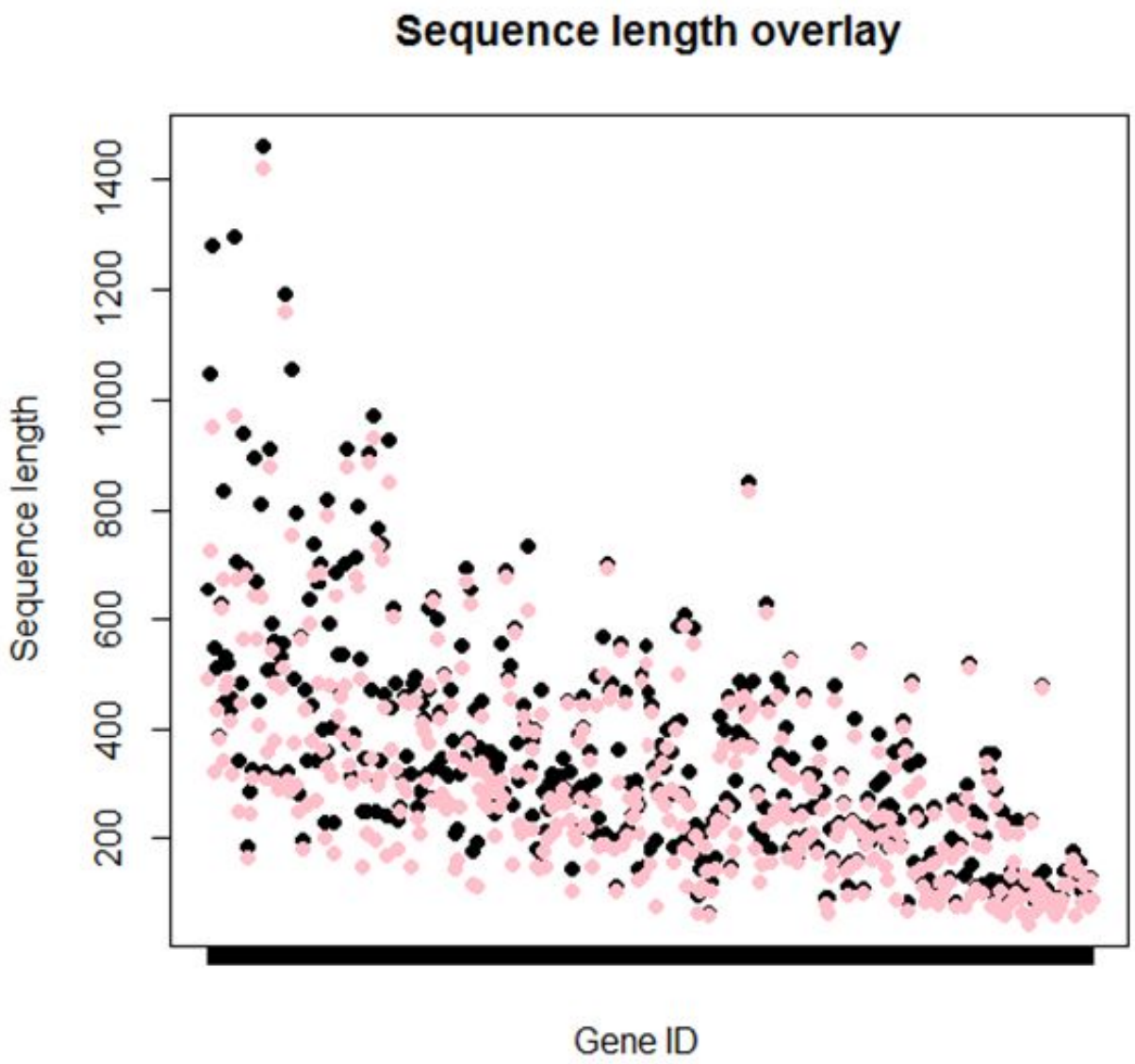
- Problem: Duplicate taxa
- Problem: OrthoMCL file headers

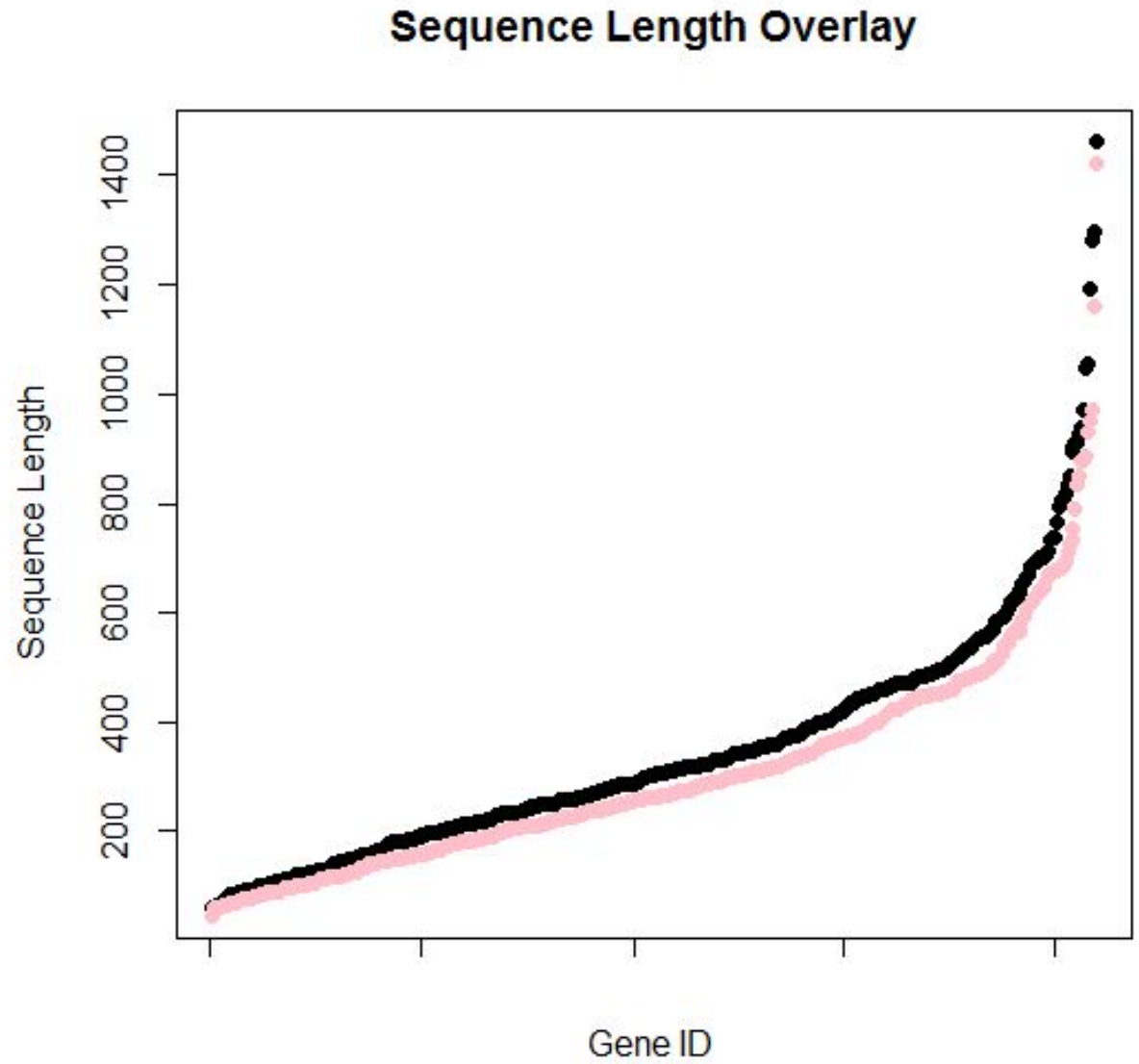
- Solution: trimAL 2, MAFFT v7.266, FASconCAT-G, seqmagick

OrthoMCL output → AA Alignments

- Single “Pipeline” written for conversion of 420 orthologs from raw OrthoMCL output into 420 individual alignment files and 1 Supermatrix

- Solution: trimAL 2, MAFFT v7.266, FASconCAT-G, seqmagick
- Installed on Mac OS X: requires biopython, GNU version of SED, Homebrew makes life easy
- Simple Unix loops


Sequence length overlay
Sequence length
200
400
600
800
1000
1200
1400
Gene ID


Sequence Length Overlay
Sequence Length
200
400
600
800
1000
1200
1400
Gene ID

OrthoMCL output → AA Alignments

- Duplicate taxa? No problem, **seqmagick**
- Need individual gene alignments converted from fasta to phylip? seqmagick

- **FASconCAT-G**, indispensible versatile tool: concatentation, translation, back translation, exclusion of 3rd codon positions, number of parsimony informative sites, file conversion, partition files...ideal for analysis pipelines...

Phylogenetic Analyses

- **PartitionFinder**: Used for model test
- **Partitioned Supermatrix**: 420 partitions with individual AA substitution models
- **Unpartitioned Supermatrix**
- 420 Amino Acid based gene trees reconstructed via **RAxML**
- Gene tree based coalescent analyses via **ASTRAL II**

Coalescent Hypothesis

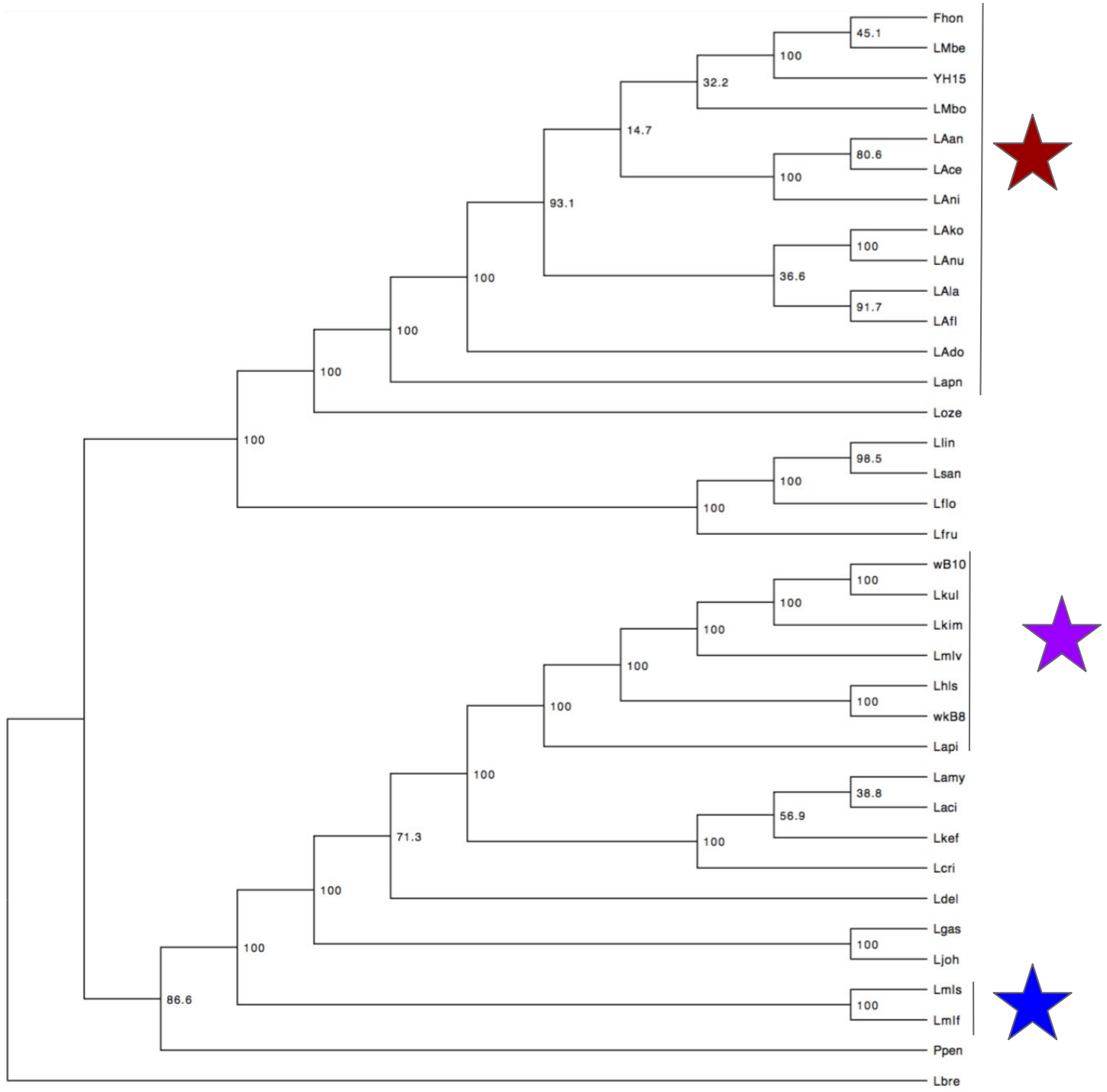
Quartet score: 1789445

Normalized quartets: 0.92

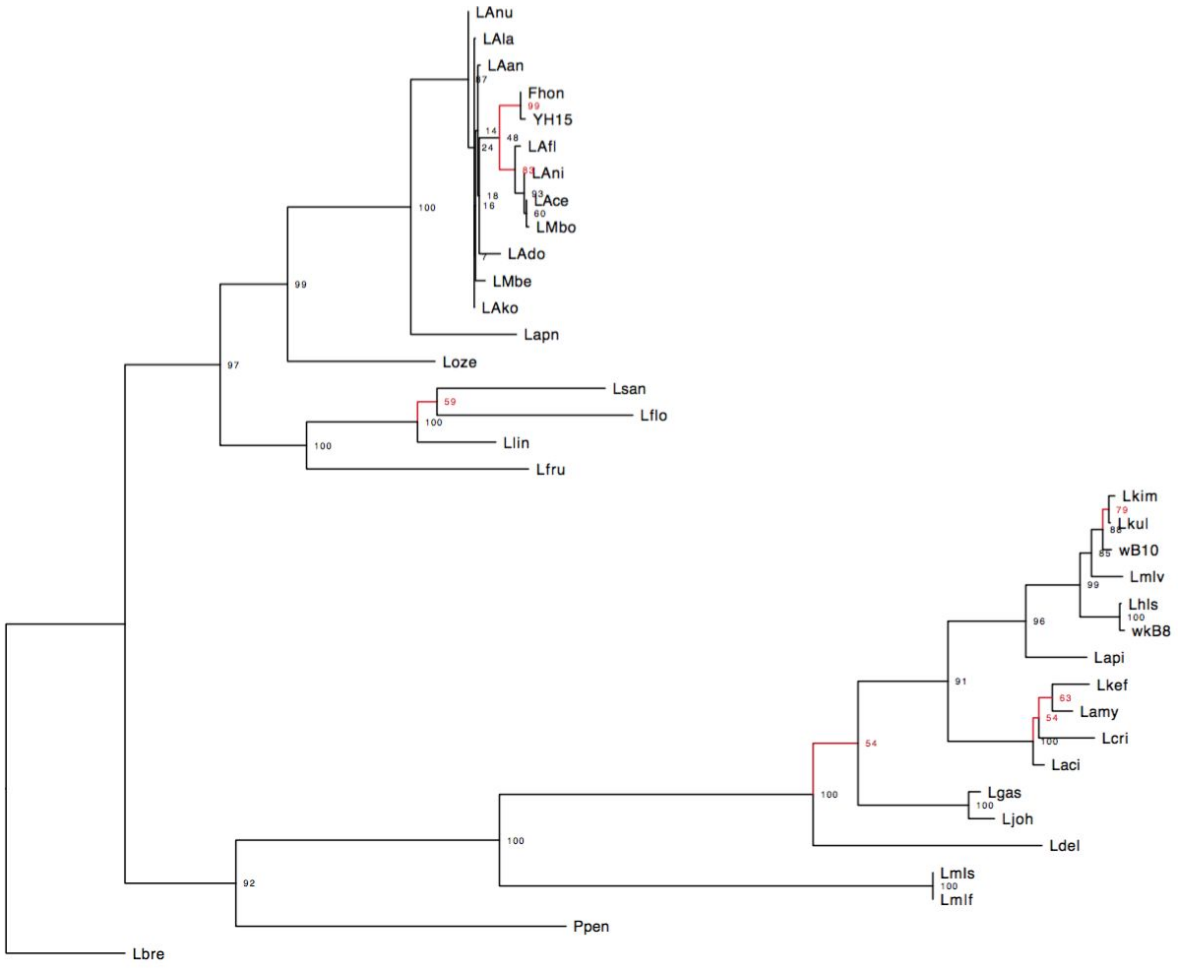
L. kunkeei clade

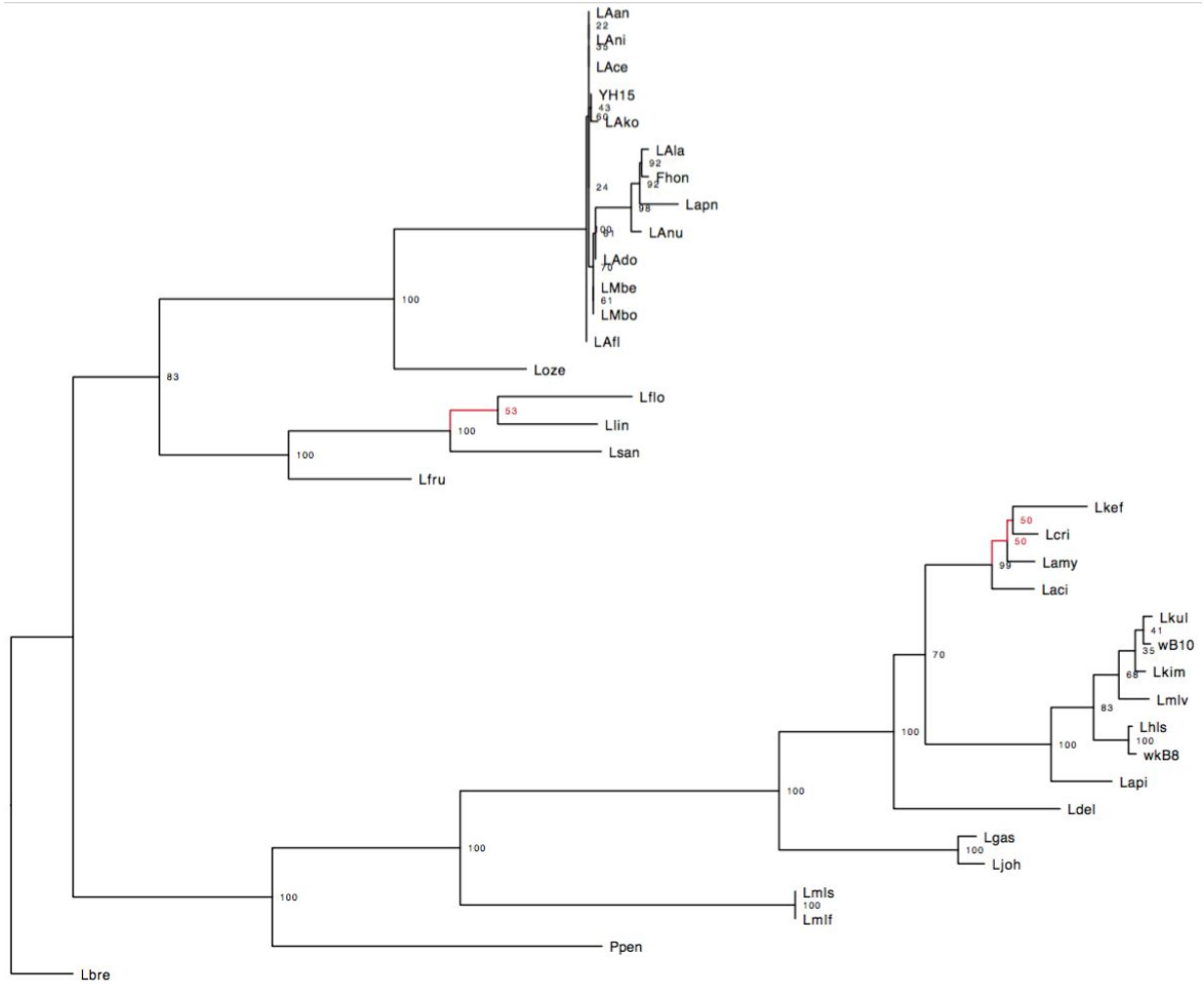
Firm 5 clade

Firm 4 clade



Individual Gene Trees: Low Incongruence?





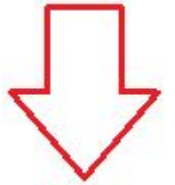
Nucleotide Analysis: The problem and starting point

- Starting point: Protein Sequences from Bioproject page of the chosen Lactobacillus taxon for analysis
- Fed into OrthoMCL
- Output: Protein Sequences sorted by ortholog group with GI #

- Problem: GI# record of a protein sequence entry provides ascension number of origin sequence, this origin sequence is only the contig or scaffold that contains the sequence coding the protein, and not the actual sequence

From Protein Sequences to Nucleotide Sequences (Retaining Ortholog Group Organization)

```
>Lactobacillus aci 489644671 unnamed protein product
MAYQALYRKWRPRTFDSVIGQEAITDTLKNAIKRGKVSHAFLFAGPRGTGKTSCAKIFAK
```



```
hvuong@pigeon:~/GEN220_2015/Git/data/Orth_aligned2/Fixed$ less noline.fasta.new |
 head -n 5
499573729
58337864
503203834
503407681
499988013
```

```
499573729,58337864,503203834,503407681,499988013,56
4,497708702,948625644,927065584,927073764,927070455
8088,927076921,937537498,797157825,927064882,575000
7160491,41582892,503619612,797154153,797151167,9488
797156644,797152894,948850649,345504528,116103593,
29,58337864,503203834,503407681,499988013,505287233
8702,948625644,927065584,927073764,927070456,937533
7076921,937537498,797157825,927064882,575008684,311
,41582892,503619612,797154153,797151167,948967592,
```

Orthologs organized by group; one entry from each taxa (‘awk’ and ‘sed’)

All entries parsed down to just GI, then converted to comma separated values using:

```
with open('noline.fasta.new') as f:
    pepid =[]
    for line in f:
        line = line.strip()
        pepid.append(line)
```

From Protein Sequences to Nucleotide Sequences: NCBI Entrez in Biopython, *Entrez Direct*

```
CDS             1..432
                /gene="vanY"
                /locus_tag="LBA1603"
                /coded_by="complement(NC_006814.3:1596356..1597654)"
                /note="serine-type D-Ala-D-Ala carboxypeptidase"
                /transl_table=11
                /db_xref="GeneID:3251454"
```

```
>Lactobacillus acidophilus NCFM
ATGGTTTTTAGTAAAAAAATAAAACGGACATTAATTAGTCTTGTTGCTTTAGTTTCTTTA
GTTTCTTGTGGTGCAGTATTTACAACACCGGTTAGTGCAGATACATCAAGTAGTTATCGC
AATAATGAAGTGAATTTAGATGTTAAATCTGCAATTGCAATTGATAGTAATTCGGGGCAA
ATTTTGTATGCTAAAAATGCTGATAAGACTTTACCAATTGCTTCAATGACAAAGTTAATT
ACAGTTTATTTAACTTTAAATGCAATTAAAAATAAAAAATTATCTTGGAATCAAAAGGTG
AAGCCAACTGCTTCAATTGTAAAAGTAGCTAATAATGCGGAATATTCAAATGTACCGCTT
AAGATGGGGCATTCTTATACTATTCGTCAGCTTTATCAAGCAACTTTAATTGAATCAGCT
AATGGGGCCGCAATGCTTTTGGGCCAAACTATTGCTGGTTCACAAAAGAAATTTATTGAT
CAAATGCGTGCCCAAGTTAAAAAATGGGGGATTGAAGATGCCGAGATTTATACGGCATGT
GGTTTACCTAATGGTAATGTAGGTAAAGATGCCTATCCTGGTGTAAATAAGAATGCTGAA
```

PAML analysis to come later!

Using GI as input; feed into Biopython and Efetch code using Entrez package in Biopython to pull out Genpept records

Utilize genpept parser program in Biopython(Pipermail emailing list):

Pulls out Nucleotide ascension number from “CDS” feature in record
Pulls out Sequence ranges and parses down the entry
Labels the header of sequence based off elements of the Genpept records

Short discussion

- Current OrthoMCL Used: “All or nothing approach”
  - Of the orthologs found, most if not all are primarily housekeeping genes
- High levels of conservation due to analyses so far limited to only housekeeping genes

Future Directions (To-Do)

- Inclusion of raw genomes into analysis
  - Comparison of annotation quality
- Ortholog Nucleotide Sequence alignment using aligned protein sequence (back-translation alignment)
- Feed nucleotide alignment for tree constructions, by individual ortholog group or concatenation
- Nucleotide/Protein alignment comparison
- dN/dS analysis of ortholog group(s) of interest

Future Directions (Out of reach)

- Analysis of Unique Ortholog groups by taxa
  - Or by clade of taxa or by group of taxa's environmental niche
  - Includes: dN/dS analysis of ortholog group(s) of interest

- Gene Ontology analysis by taxa/clade

- Horizontal Gene Transfer Detection tests via topology

Acknowledgements

Jordan Hayes - for assistance with setting up MySQL database, and advice on using the cluster and running OrthoMCL

Jason for advice, help with scripts and qsub jobs.

“Peter” and Animesh Agrawal - Biopython code for acquiring sequence from a given Genpept record, accessed from Biopython “pipermail” mailing list archive