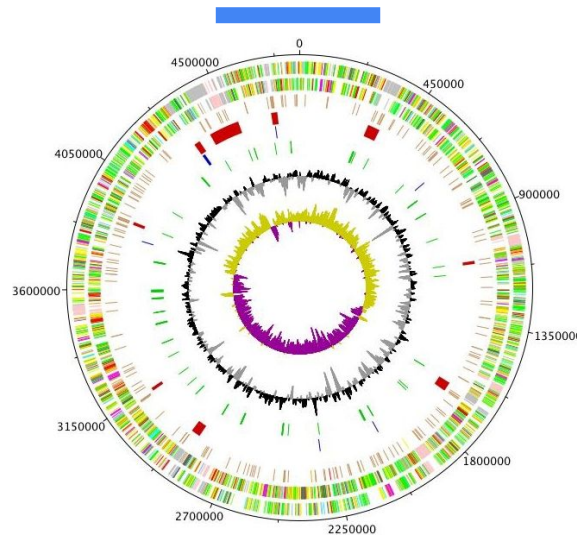


Phylogeny and genomic characterization of bee-associated *Lactobacillus* species and their closest relatives



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

Corbiculate apid
Flower
Other bee
Augochlora pura
Megalopta genalis
Halictus ligatus

L. kunkei clade

F4

F3

F5

- 

F3

5 *

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:

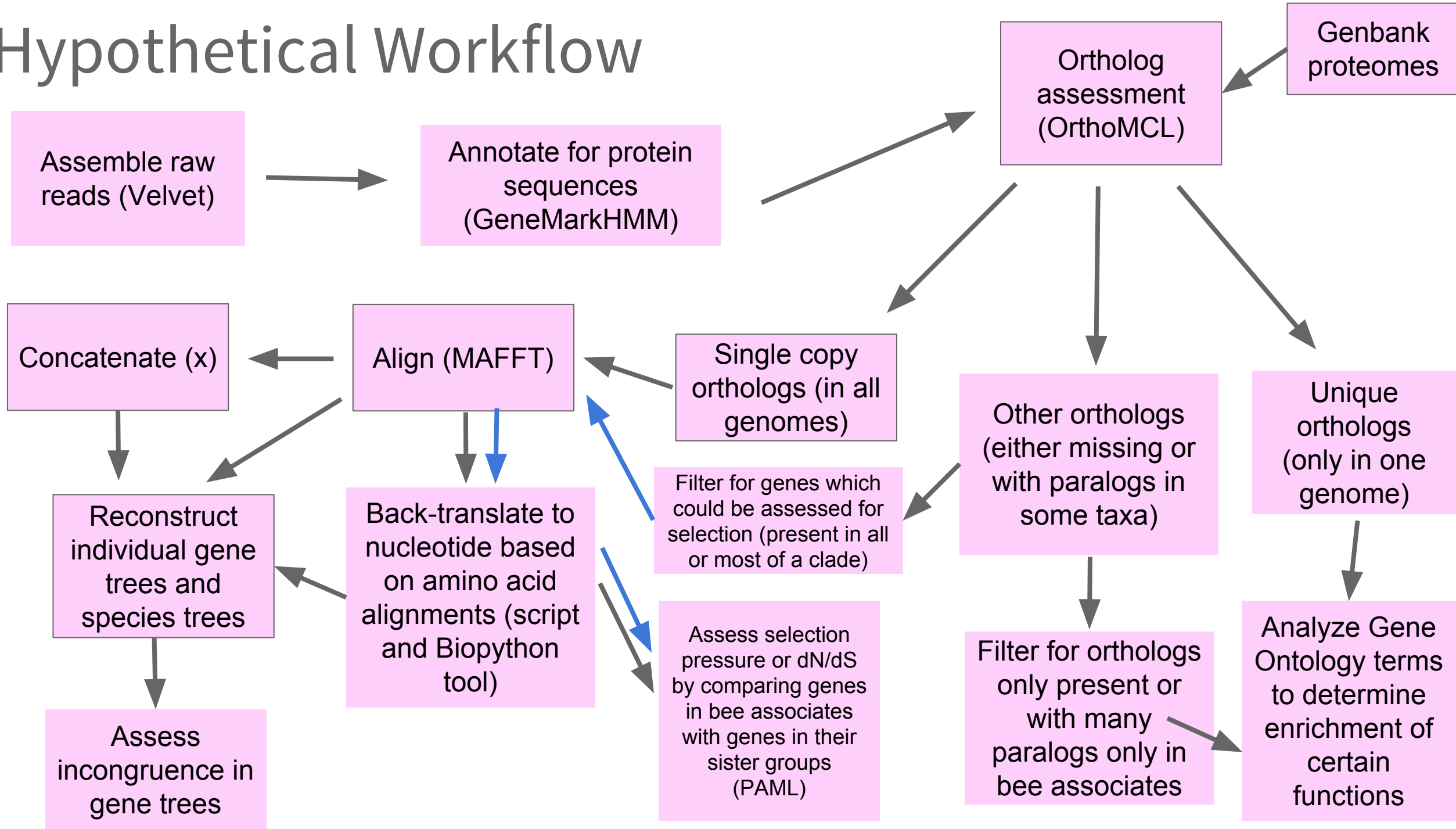
Create pipeline for analysis of bacterial genomes:

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.

The screenshot shows the GitHub interface for a repository named 'Gen220-Phylo-and-Evol-project-1' by user 'erg55'. At the top, there are buttons for 'Unwatch' (1), 'Star' (0), and 'Fork' (3). Below this is a navigation bar with links for 'Code', 'Issues' (0), 'Pull requests' (0), 'Wiki', 'Pulse', 'Graphs', and 'Settings'. A section for 'Team Members' lists 'Eric Gordon, Kaleigh Russell, Hoang Vuong, David Haisten' with an 'Edit' link. A summary bar shows '79 commits', '3 branches', '0 releases', and '4 contributors'. Below this is a bar with 'Branch: master', a 'New pull request' button, 'New file', 'Find file', 'HTTPS' dropdown, the repository URL 'https://github.com/erg55/Gen220-Phylo-and-Evol-project-1', a copy icon, a download icon, and a 'Download ZIP' button. The main content area shows a list of recent commits, including a merge pull request #21 from 'dhaisten/master'. The latest commit is '5171db9' from 'a day ago'. The commit list includes files like 'data', 'scripts', '.gitattributes', 'Ortho.phy.zip', 'README.md', 'ortholog_freq.txt', and 'ortholog_freq2.txt'. At the bottom, there is a 'README.md' section.

erg55 / Gen220-Phylo-and-Evol-project-1

Unwatch 1 Star 0 Fork 3

Code Issues 0 Pull requests 0 Wiki Pulse Graphs Settings

Team Members: Eric Gordon, Kaleigh Russell, Hoang Vuong, David Haisten — Edit

79 commits 3 branches 0 releases 4 contributors

Branch: master New pull request New file Find file HTTPS https://github.com/erg55/Gen220-Phylo-and-Evol-project-1 Download ZIP

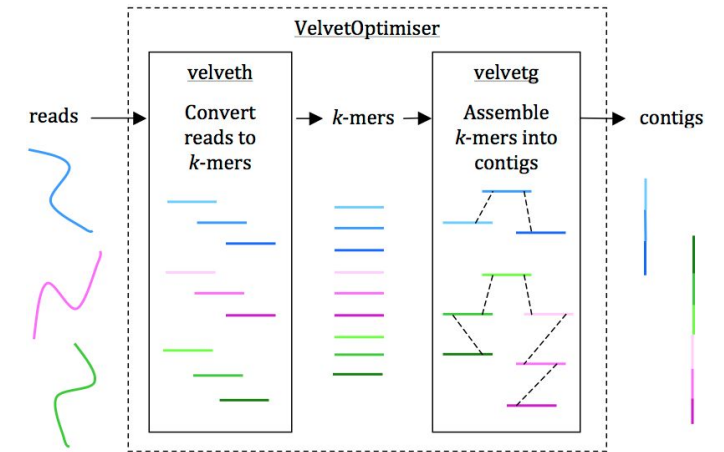
erg55 Merge pull request #21 from dhaisten/master Latest commit 5171db9 an hour ago

data	deleted individual	a day ago
scripts	convert_fa.sh	an hour ago
.gitattributes	djlfksdjfk	13 days ago
Ortho.phy.zip	david's stuff	a day ago
README.md	Update README.md	16 days ago
ortholog_freq.txt	added and fixed helveticus	4 days ago
ortholog_freq2.txt	added	3 days ago

README.md

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.

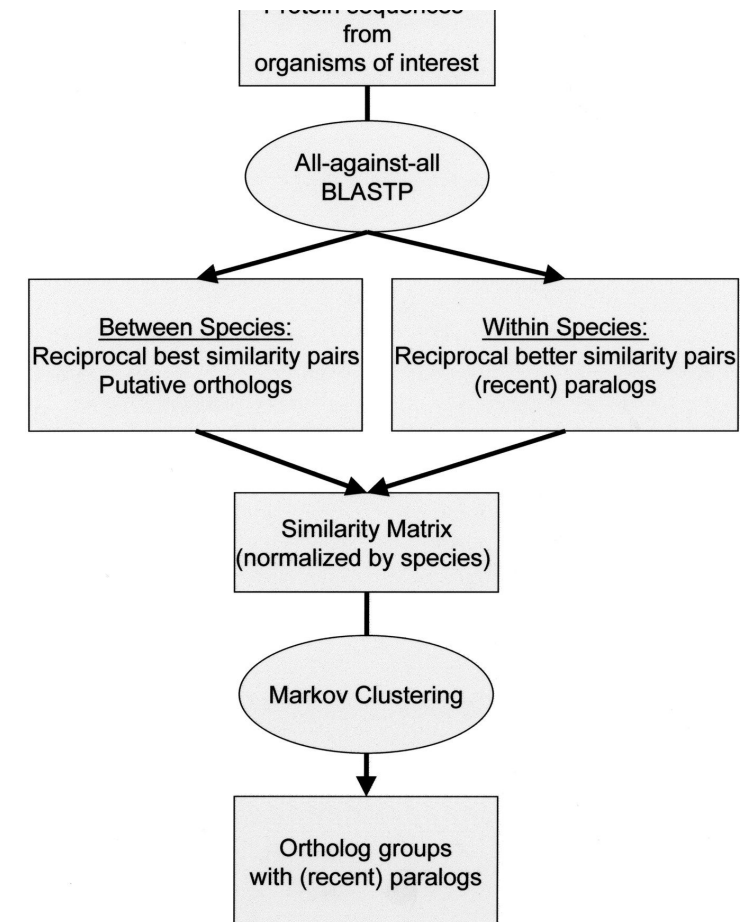
```
#version if can't get array job to work
```

```
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).

OG1.5_1049.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1060.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1061.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1062.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1064.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1066.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1067.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1070.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1072.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1073.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1075.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1078.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1081.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1083.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1084.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1085.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1088.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1089.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1091.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1093.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1094.fasta	Dec 1, 2015, 5:14 PM
OG1.5_1095.fasta	Dec 1, 2015, 5:14 PM

```
>lcl|Lbre|741040218 unnamed protein product
MPRVKGGTYTRARRKVLKLAGYRGSKHRLFKVAKDQVMKGRQYAFDRKATKRNFRKLWIARINAAARMNGLSYSKLM
HGLKLANIDVNRKMLADLAVNDAAFAALAEQAKTALAA
>lcl|Lfru|497707398 unnamed protein product
MPRVKGGTTTRNRKRVLKLAGYRGAKHRLFKTAKDQVMKSQEYAFDRRANKGNFRKIWIARINAAATRNNGLSYSKFM
HGLKLANIDMNRKMLADLAVNDADAFSALAEKAKAALK
>lcl|Lfru|948626790 unnamed protein product
MPRVKGGTTTRNRKRVLKLAGYRGAKHRLFKTAKDQVMKSQEYAFDRRANKGNFRKIWIARINAAATRNNGLSYSKFM
HGLKLANIDMNRKMLADLAVNDADAFSALAEKAKAALK
>lcl|Fhon|1927065203 unnamed protein product
MPRVKGGTYTHARRKVLKLAGYRGKHSFLFKTAKDQVMKSREYAFDRRANKGNFRRLWIARINAAARMNGLSYSKLM
HGLKLSNIEMNRKMLADLAVNDEKAFASLAETAKKAIAK
>lcl|LAan|1927072621 unnamed protein product
MPRVKGGTYTHARRKVLKLAGYRGKHSFLFKTAKDQVMKSREYAFDRRANKGNFRRLWIARINAAARMNGLSYSKLM
HGLKLSNIEMNRKMLADLAVNDEKAFASLAETAKKAIAK
>lcl|LAce|1927069417 unnamed protein product
MPRVKGGTYTHARRKVLKLAGYRGKHSFLFKTAKDQVMKSREYAFDRRANKGNFRRLWIARINAAARMNGLSYSKLM
HGLKLSNIEMNRKMLADLAVNDEKAFASLAETAKKAIAK
>lcl|LAdo|937533902 unnamed protein product
MPRVKGGTYTHARRKVLKLAGYRGKHSFLFKTAKDQVMKSREYAFDRRANKGNFRRLWIARINAAARMNGLSYSKLM
HGLKLSNIEMNRKMLADLAVNDEKAFASLAETAKKAIAK
>lcl|LAFI|1927072363 unnamed protein product
MPRVKGGTYTHARRKVLKLAGYRGKHSFLFKTAKDQVMKSREYAFDRRANKGNFRRLWIARINAAARMNGLSYSKLM
HGLKLSNIEMNRKMLADLAVNDEKAFASLAETAKKAIAK
>lcl|LAko|1927074438 unnamed protein product
MPRVKGGTYTHARRKVLKLAGYRGKHSFLFKTAKDQVMKSREYAFDRRANKGNFRRLWIARINAAARMNGLSYSKLM
HGLKLSNIEMNRKMLADLAVNDEKAFASLAETAKKAIAK
>lcl|LAla|1927077663 unnamed protein product
MPRVKGGTYTHARRKVLKLAGYRGKHSFLFKTAKDQVMKSREYAFDRRANKGNFRRLWIARINAAARMNGLSYSKLM
HGLKLSNIEMNRKMLADLAVNDEKAFASLAETAKKAIAK
>lcl|LAmy|1489723002 unnamed protein product
MPRVKGGTYTRARRKVMKLAGYRGAKHMQFAASTQLFVSYYKAFDRRRKSEFRKLWIARINAAARMNGLSYSKLM
HGLKLAGVDMNRKMLADLAVNDKTFQAQLAETAKKALN
>lcl|LANi|1927076139 unnamed protein product
MPRVKGGTYTHARRKVLKLAGYRGKHSFLFKTAKDQVMKSREYAFDRRANKGNFRRLWIARINAAARMNGLSYSKLM
HGLKLSNIEMNRKMLADLAVNDEKAFASLAETAKKAIAK
```

Ellegaard, K. M., Tamarit, D., Javelind, E., Olofsson, T. C., Andersson, S. G., & Vásquez, A. (2015). Extensive intra-phylo-type 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).

OG_name	Pho	Laa	Lac	Lad	Laf	Lak	Lal	Lan	Lan	Lmi	Lmt	Lapi	Lapi	Lcpi	Lcpi	Lde	Lfo	Lga	Lhis	Ljoh	Lke	Lkin	Lkul	Llin	Lml	Lml	Lml	Lozi	Lsar	Ppe	YH1	wB3	wk1
OG1.5_1710	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1
OG1.5_1711	2	1	3	1	1	1	2	1	2	2	2	1	1	1	1	0	1	1	1	0	1	2	0	0	0	1	2	0	0	2	2	2	2
OG1.5_1712	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	2	1	1	0	1	1	
OG1.5_1713	0	0	0	0	0	0	0	0	0	0	0	2	0	1	1	1	2	1	1	2	2	1	1	2	0	1	1	1	0	2	1	1	
OG1.5_1714	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OG1.5_1715	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	1	1	1	1	0	0	1	1
OG1.5_1716	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1
OG1.5_1717	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	2	1	1	1	1	1	1	2	1	1	1	0	2	1	0	1	1	1
OG1.5_1718	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	0	1	0	1	1	1	0	1	1
OG1.5_1719	1	1	0	1	1	1	1	1	0	0	1	0	1	0	0	0	1	0	1	0	2	1	1	1	0	0	0	1	0	0	1	1	1
OG1.5_1720	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OG1.5_1721	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
OG1.5_1722	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	2	0	0	1	1	1
OG1.5_1723	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	0	1	0	0	0	1	1	1	1	0	0	1
OG1.5_1724	1	0	1	0	2	1	1	1	2	1	1	1	0	1	1	0	0	1	0	2	1	1	0	2	2	1	0	0	0	1	1	1	1
OG1.5_1725	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1	0	0	1	0	1	1	0	1	1	1	1	0	0	1
OG1.5_1726	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OG1.5_1727	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	0	1	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	1
OG1.5_1728	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	3	0	0	0	0	0	1	1	0	0	0	0	0	0
OG1.5_1729	1	1	1	1	1	1	1	1	1	1	1	1	0	2	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OG1.5_1730	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1
OG1.5_1731	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1
OG1.5_1732	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	0	1	0	0	1	1	1	1	0	0
OG1.5_1733	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	0	0	1	0	0	0	1	1	1	1	0	0	1
OG1.5_1734	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1
OG1.5_1735	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
OG1.5_1736	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1



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.

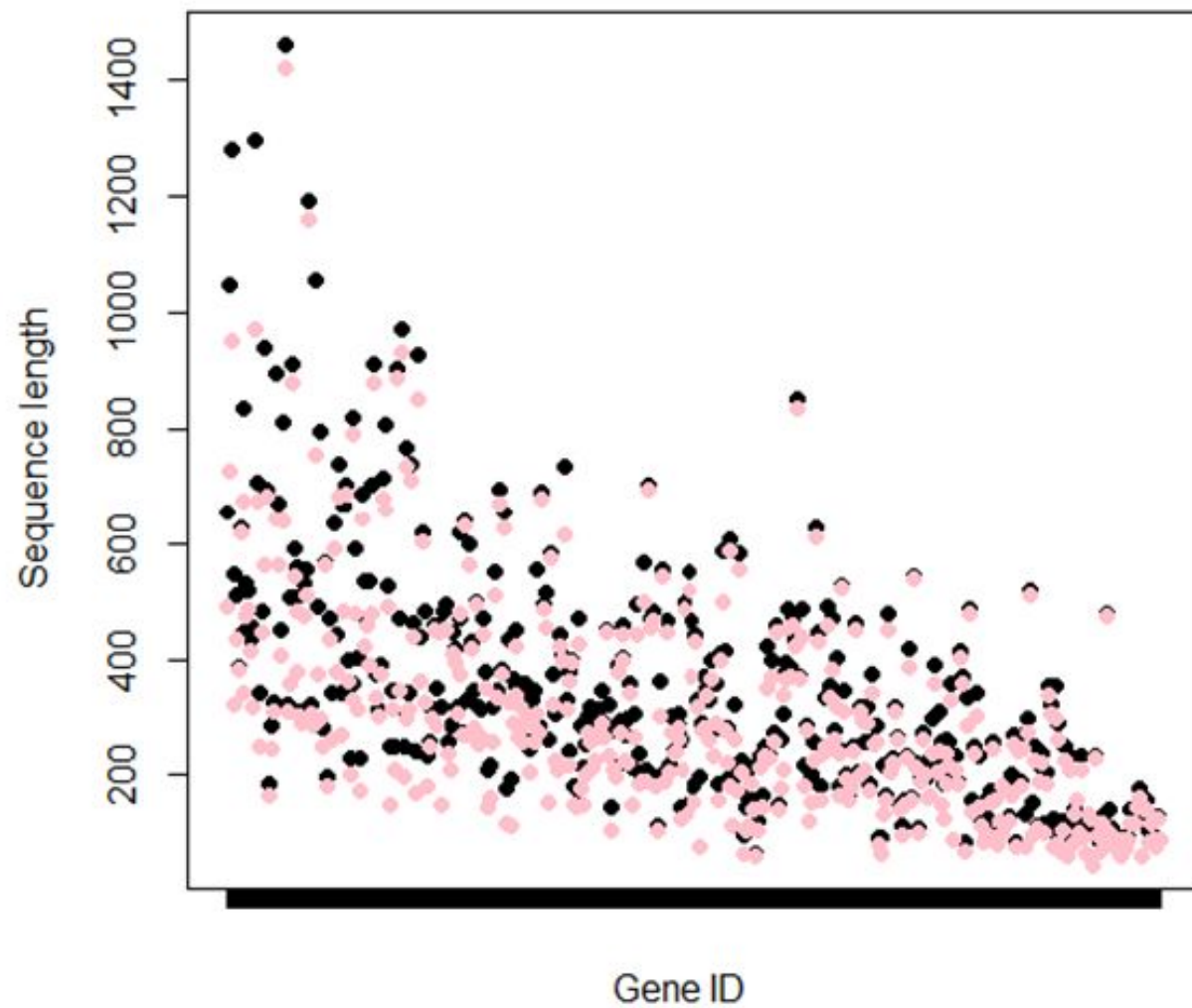
OrthoMCL output → AA Alignments

- 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

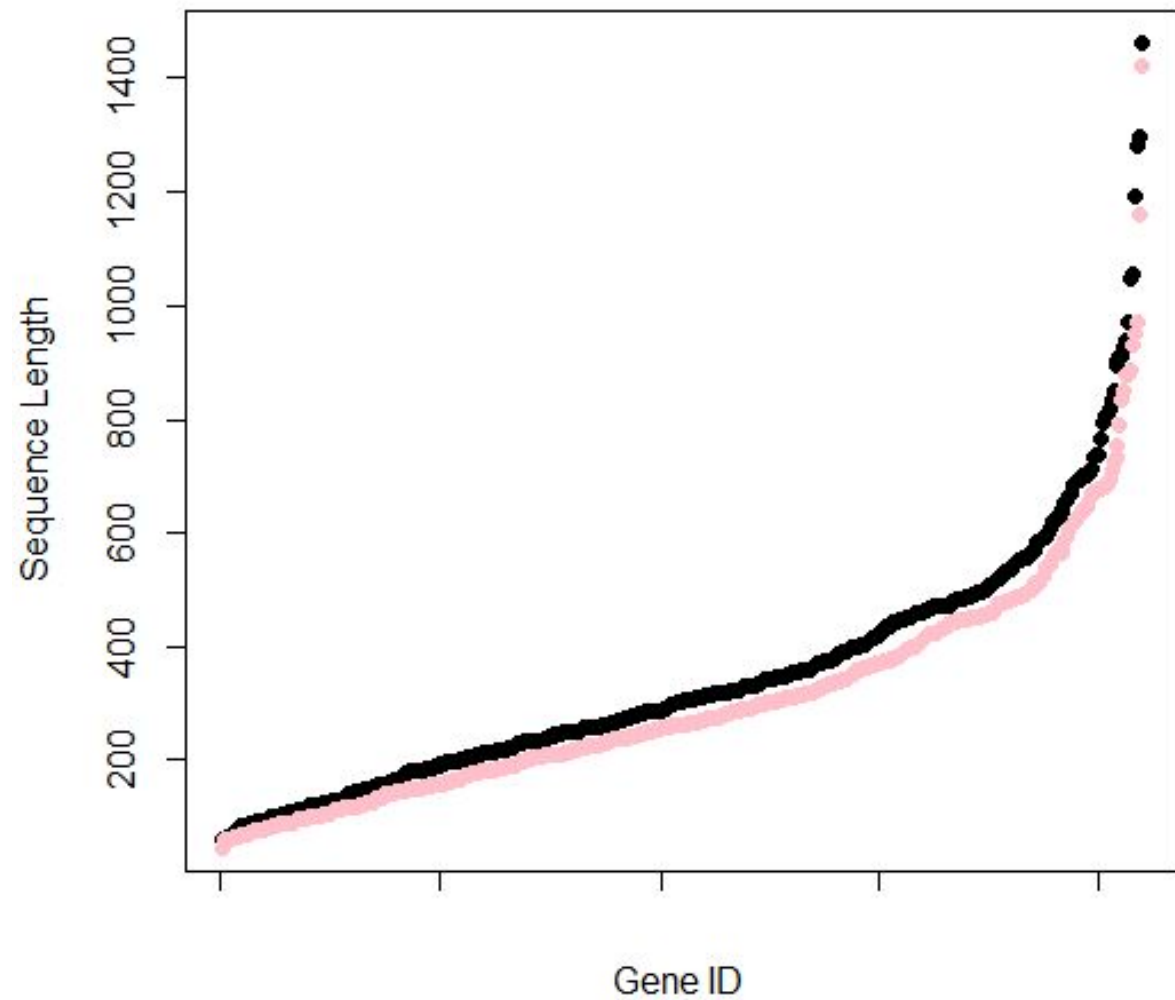
OrthoMCL output ➡ AA Alignments

- 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 Overlay



OrthoMCL output ➡ AA Alignments



- Duplicate taxa? No problem, **seqmagick**
- Need individual gene alignments converted from fasta to phylip? seqmagick
- **FASconCAT-G**, indispensible versatile tool: concatenation, 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

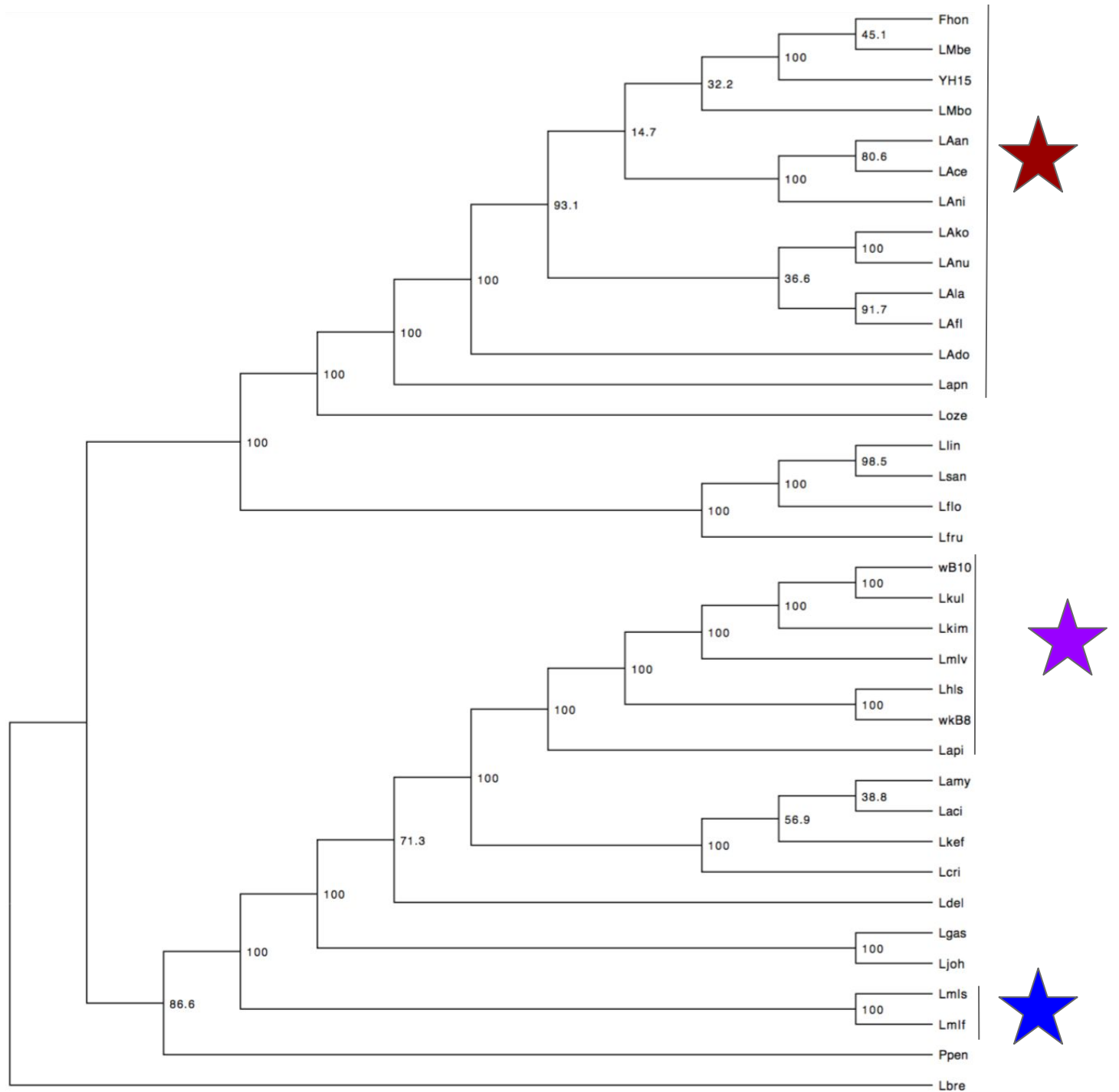
Quartet score: 1789445

Normalized quartets: 0.92

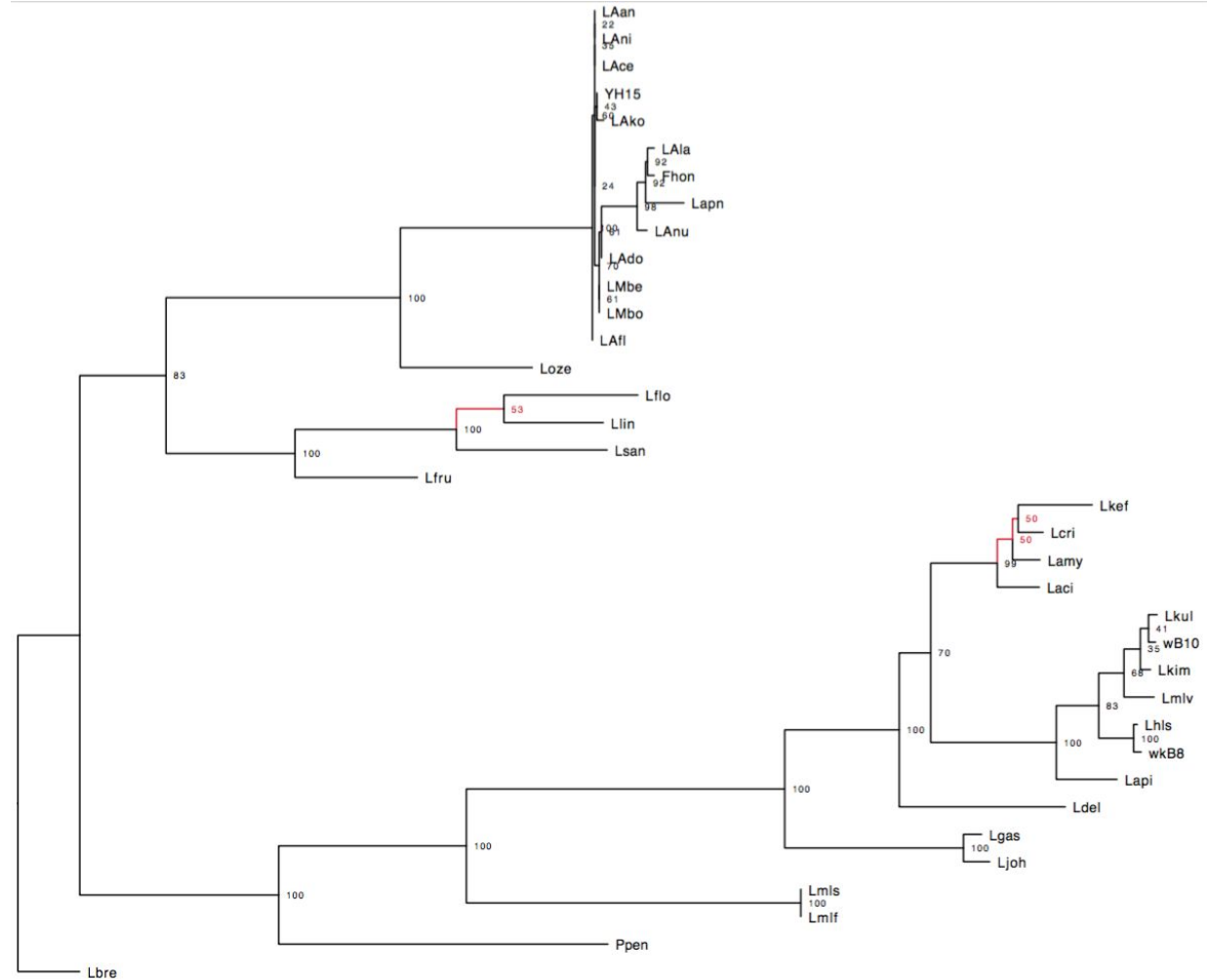
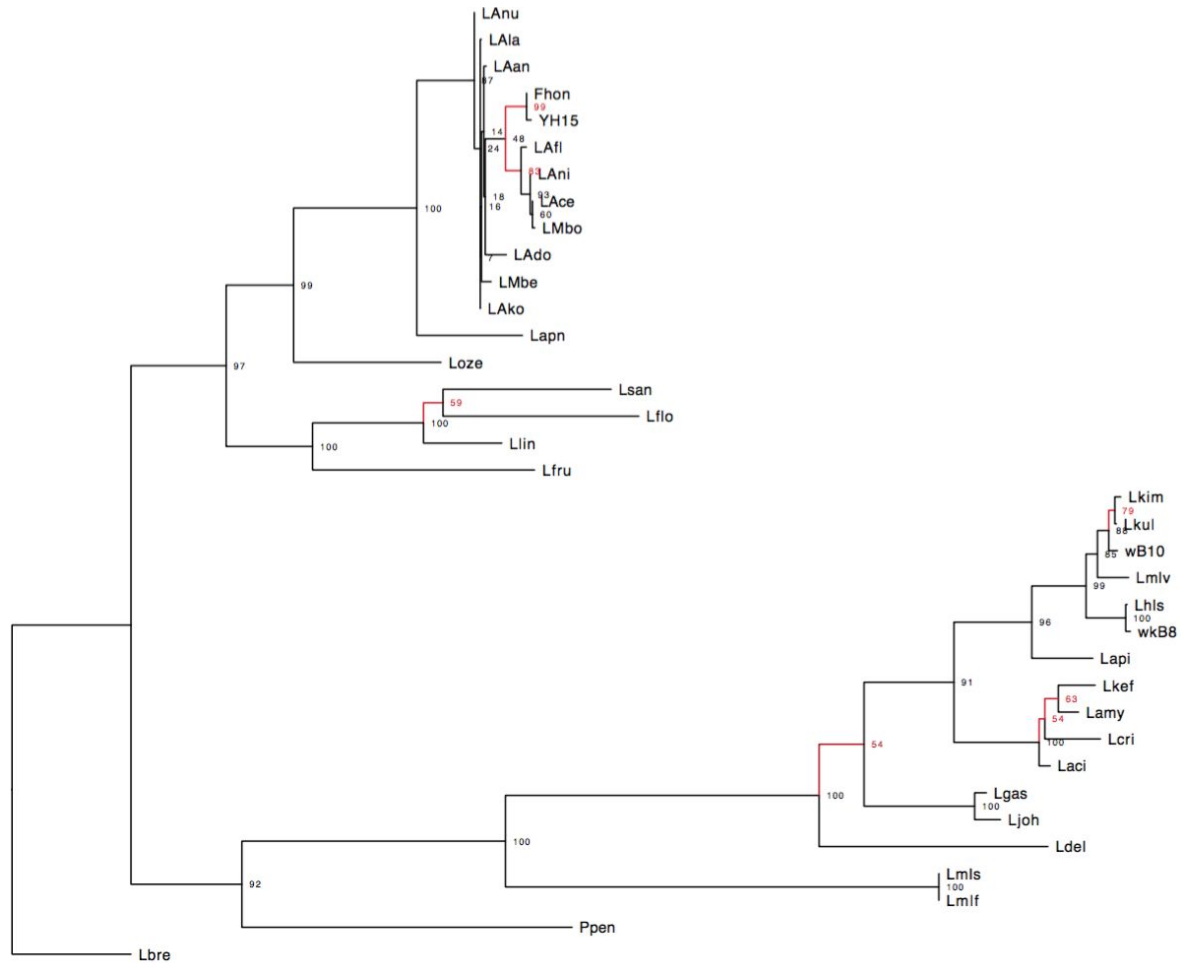
L. kunkeei clade

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  
ALNCTNLQDGEPCNECANCTAANECSMDTMEIDAASNMGVDTEPDEPDEKVEYAPTEGKY
```



```
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,
```

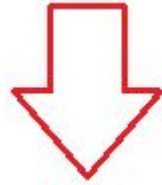
Orthologs organized by group; one entry from each taxa ('awk' and 'sed')

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
ATGGTTTTTTAGTAAAAAATAAACGGACATTAATTAGTCTTGTTGCTTTAGTTTCTTTA
GTTTCTTGTTGGTGCAGTATTTACAACACCGGTTAGTGCAGATACATCAAGTAGTTATCGC
AATAATGAAGTGAATTTAGATGTTAAATCTGCAATTGCAATTGATAGTAATTCGGGGCAA
ATTTTGTATGCTAAAAATGCTGATAAGACTTTACCAATTGCTTCAATGACAAAGTTAATT
ACAGTTTATTTAACTTTAAATGCAATTAAAAATAAAAAATTATCTTGGAATCAAAAGGTG
AAGCCAACTGCTTCAATTGTAAAAGTAGCTAATAATGCGGAATATTCAAATGTACCGCTT
AAGATGGGGCATTCTTATACTATTTCGTCAGCTTTATCAAGCAACTTTAATTGAATCAGCT
AATGGGGCCGCAATGCTTTTGGGGCCAACTATTGCTGGTTCACAAAAGAAATTTATTGAT
CAATGCGTGCCCAAGTTAAAAAATGGGGGATTGAAGATGCCGAGATTTATACGGCATGT
GGTTTACCTAATGGTAATGTAGGTAAAGATGCCTATCCTGGTGTAATAAGAATGCTGAA
```

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