JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

410.734.81 and 410.734.82
Practical Introduction to Metagenomics

Topic: Taxonomic analysis, phylotypes and binning

Instructor: Joshua Orvis

Introduction – Déjà vu?

In Week 06 we studied community profiling and defined it as “the determination of the abundance of each kind of microbe in a sample.” This included read alignment to reference genomes (with fragment recruitment plots) and ribosomal targeted sequencing.

Acknowledging that all of you will be busy toiling on your final projects, I want to extend that week's lesson to include other methods of phylogenetic study with your metagenomic data.

We'll start be talking about a few of these approaches generally and then talk about some specific tools which enable you to do it including Mothur, QIIME, FastUniFrac and the very new PhyloSift.

rRNA phylotyping

Typical steps (non-metagenomic approach):

- Collect DNA from environment
- PCR amplify rRNA genes using broad (so-called 'universal') primers
- Sequence
- Align to rRNAs from other organisms
- Infer evolutionary tree
- Unknowns “identified” by placement on tree

The main way that metagenomics has changed this is that rather than create and sequence targeted PCR products in step #2 the entire sample is just sequenced and the rRNA genes are searched within it.

Metagenomics has furthered phylotyping by allowing protein coding genes to be used rather than rRNA, such as RecA, HSP70, and EF-Tu. This can sometimes even be more accurate in relative abundance estimation due to variant copy number of the rRNAs compared with these marker protein coding genes.

We covered rRNA sequencing (especially 16S) in Week 06, so let's look at an example using proteins for phylogenetic analysis rather than rRNAs.

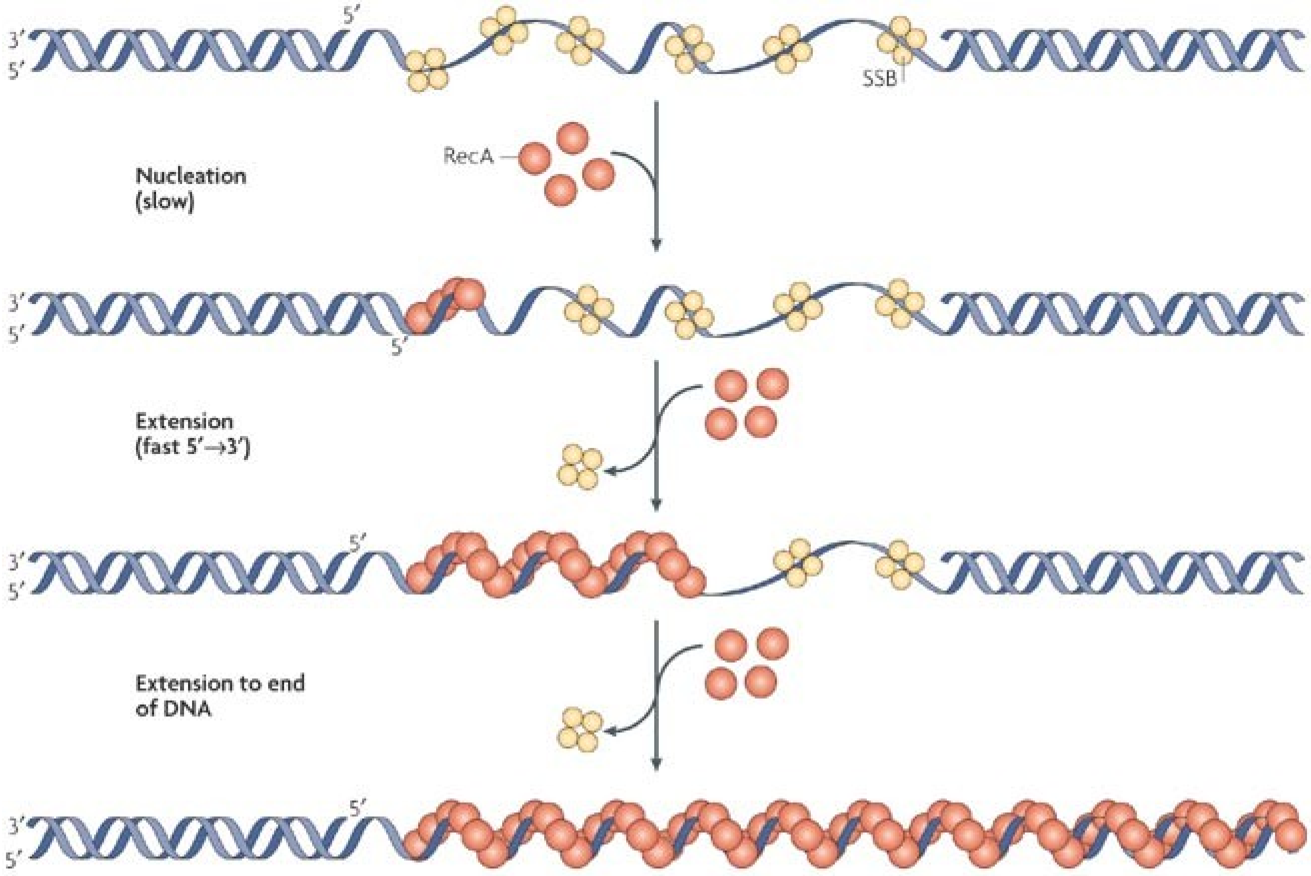
Jonathan Eisen talk 2012-06-23 "Phylogeny driven approaches to genomics & metagenomics"

RecA example

“The RecA Protein as a Model Molecule for Molecular Systematic Studies of Bacteria: Comparison of Trees of RecAs and 16S rRNAs from the Same Species”
Jonathan A. Eisen
J Mol Evol. 1995 December; 41(6): 1105–1123.

As I'm sure everyone remembers from their molbio course, RecA is a crucial protein for maintenance and repair of DNA and has a homolog in every species currently known, making it a great candidate for molecular systematic studies.

Collections of unannotated proteins can be quickly searched for recA by using PFAM HMM ID PF00154 or even more standard BLAST searches.



Nature Reviews | Molecular Cell Biology

J Mol Evol. 1995 December; 41(6): 1105–1123.

Assembly of RecA filaments:
http://www.nature.com/nrm/journal/v8/n2/fig_tab/nrm2099_F2.html

RecA example

Phylogenetic trees of all currently (1995) available complete RecA proteins were inferred using multiple maximum parsimony and distance matrix methods.

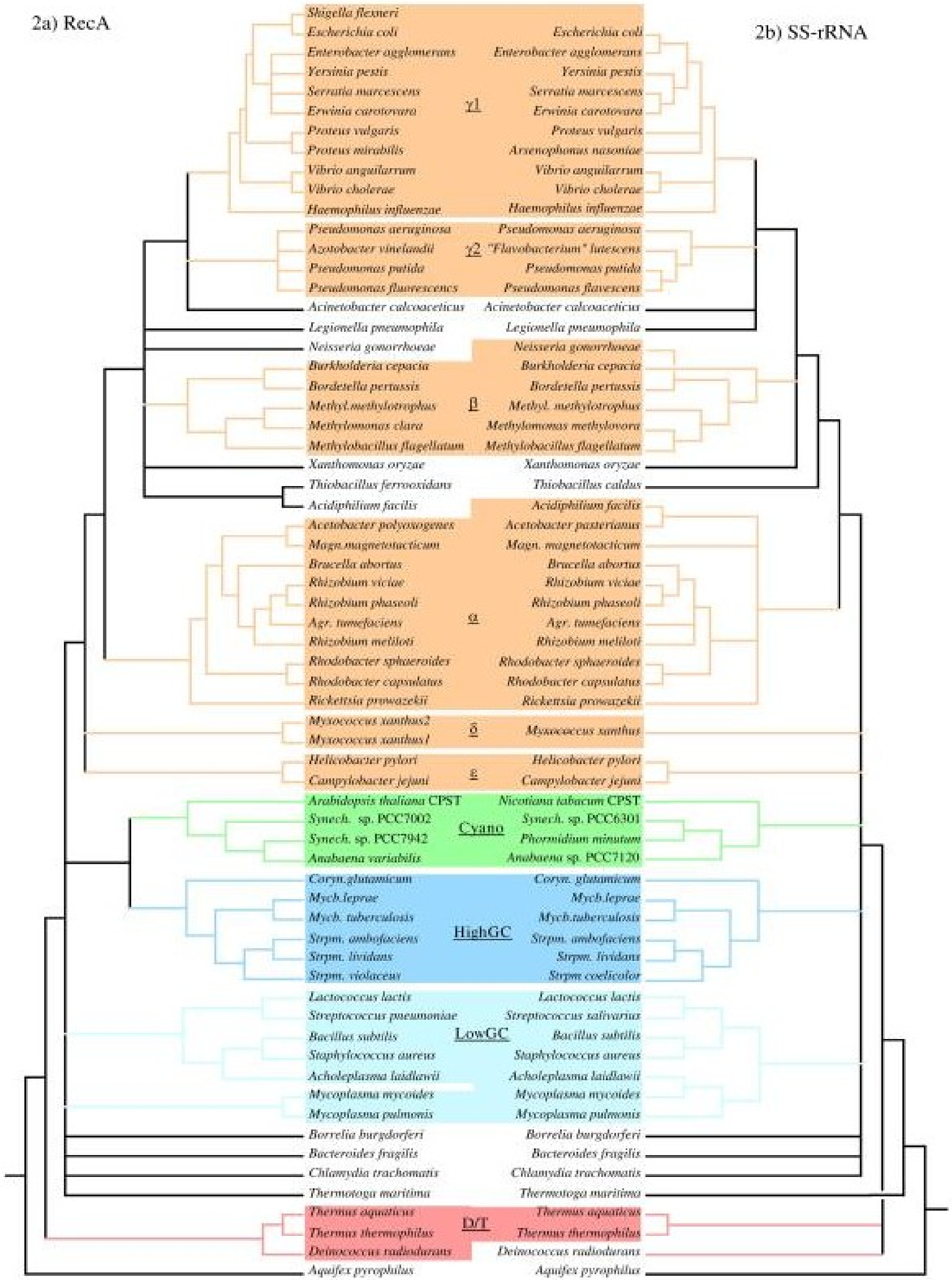
Statistical analysis and comparisons of trees generated by the different phylogenetic methods suggests that the RecA phylogeny is highly consistent and robust.

The RecA trees are compared to trees of SS-rRNA sequences from the same or very closely related species as represented in the RecA trees.

Overall, the trees of the two molecules are highly congruent.

This favorable result in using RecA as a marker has been bourne out in more recent studies as well, even providing better resolution at the subspecies level in some cases than rRNA markers. But because this isn't universally true, it might serve better as a supplemental method.

J Mol Evol. 1995 December; 41(6): 1105–1123.



RecA example

Phylogenetic trees of all currently (1995) available complete RecA proteins were inferred using multiple maximum parsimony and distance matrix methods.

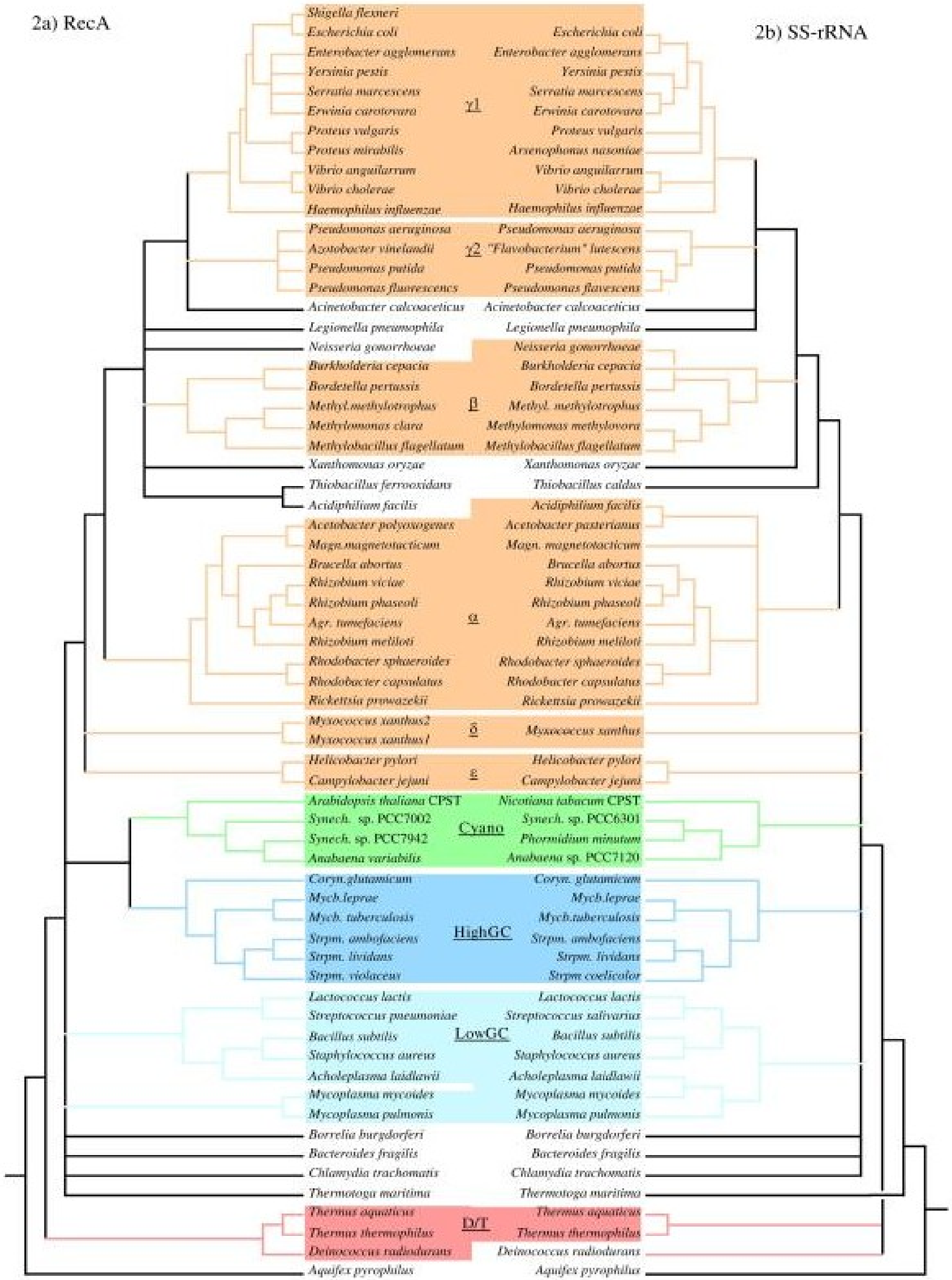
Statistical analysis and comparisons of trees generated by the different phylogenetic methods suggests that the RecA phylogeny is highly consistent and robust.

The RecA trees are compared to trees of SS-rRNA sequences from the same or very closely related species as represented in the RecA trees.

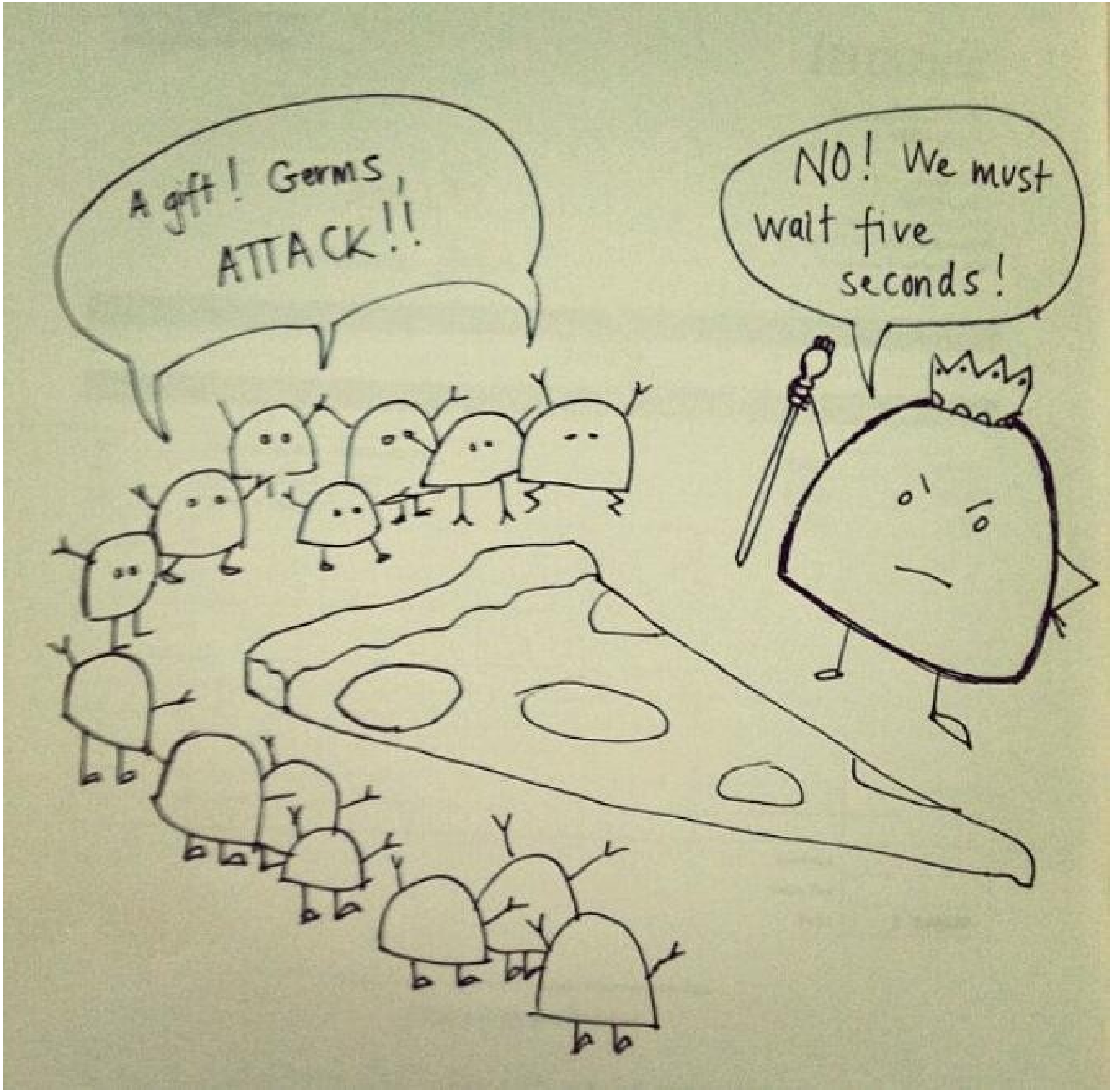
Overall, the trees of the two molecules are highly congruent.

This favorable result in using RecA as a marker has been bourne out in more recent studies as well, even providing better resolution at the subspecies level in some cases than rRNA markers. But because this isn't universally true, it might serve better as a supplemental method.

J Mol Evol. 1995 December; 41(6): 1105–1123.

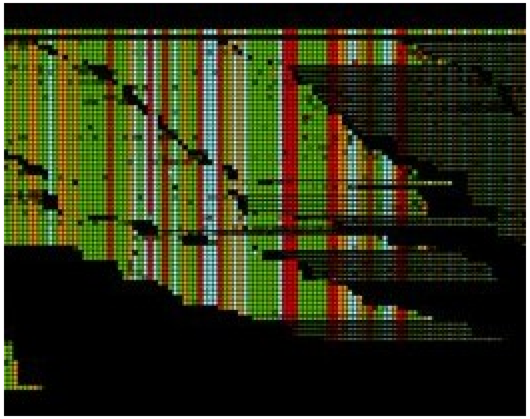


5-second rule



Binning

This is a generic term for an activity that sorts sequences (often reads) into 'bins' which represent species, genera or some taxonomic level of groups. Think of it like clustering, but not *necessarily* requiring alignment to form a cluster. The clusters are defined by properties which attempt to group them by taxonomic level, not sequence identity.



The best way to do this is by aligning the reads to a collection of reference genomes. This is great if you study many of the habitats which have a growing collection of reference genomes available in the sequence archives.

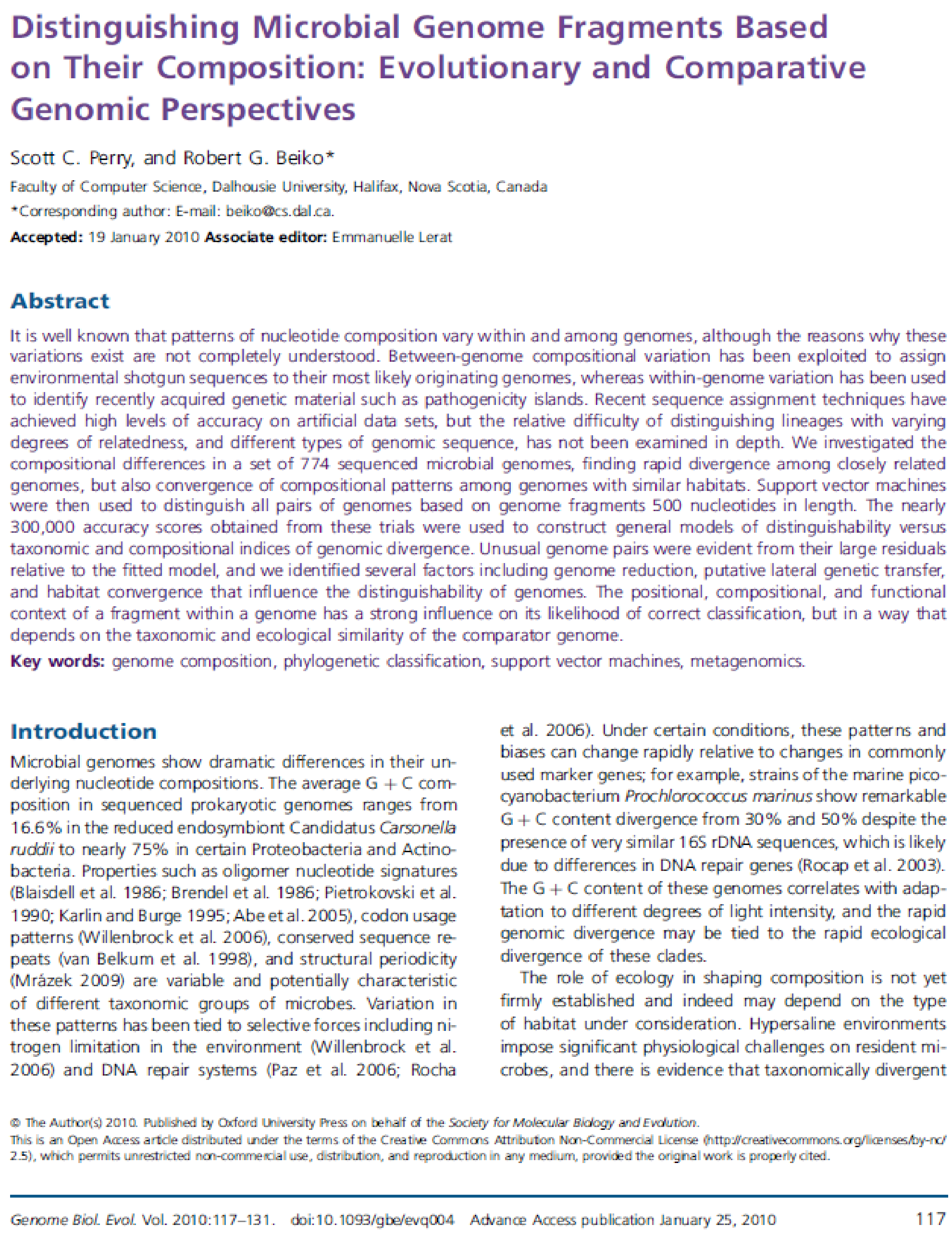
But what if you don't have reference genomes to use **or** you have a lot of reads that didn't align to your reference genome collection?

Your first assigned reading (next slide) discusses the biological relevance of binning based on the composition of the sequences themselves.

Perry SC and Beiko RG.

An intensive analysis, this goes into great detail about the complexities and computational basis of binning based on sequence composition. While admittedly lengthly (and difficult), it is an illustrative example of the synthesis of computational/biological analysis with an eye on biological relevance.

“We investigated the compositional differences in a set of 774 sequenced microbial genomes, finding rapid divergence among closely related genomes, but also convergence of compositional patterns among genomes with similar habitats. Support vector machines were then used to distinguish all pairs of genomes based on genome fragments 500 nucleotides in length. The nearly 300,000 accuracy scores obtained from these trials were used to construct general models of distinguishability versus taxonomic and compositional indices of genomic divergence.”



Distinguishing Microbial Genome Fragments Based on Their Composition: Evolutionary and Comparative Genomic Perspectives

Scott C. Perry, and Robert G. Beiko*

Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

*Corresponding author: E-mail: beiko@cs.dal.ca.

Accepted: 19 January 2010 **Associate editor:** Emmanuelle Lerat

Abstract

It is well known that patterns of nucleotide composition vary within and among genomes, although the reasons why these variations exist are not completely understood. Between-genome compositional variation has been exploited to assign environmental shotgun sequences to their most likely originating genomes, whereas within-genome variation has been used to identify recently acquired genetic material such as pathogenicity islands. Recent sequence assignment techniques have achieved high levels of accuracy on artificial data sets, but the relative difficulty of distinguishing lineages with varying degrees of relatedness, and different types of genomic sequence, has not been examined in depth. We investigated the compositional differences in a set of 774 sequenced microbial genomes, finding rapid divergence among closely related genomes, but also convergence of compositional patterns among genomes with similar habitats. Support vector machines were then used to distinguish all pairs of genomes based on genome fragments 500 nucleotides in length. The nearly 300,000 accuracy scores obtained from these trials were used to construct general models of distinguishability versus taxonomic and compositional indices of genomic divergence. Unusual genome pairs were evident from their large residuals relative to the fitted model, and we identified several factors including genome reduction, putative lateral genetic transfer, and habitat convergence that influence the distinguishability of genomes. The positional, compositional, and functional context of a fragment within a genome has a strong influence on its likelihood of correct classification, but in a way that depends on the taxonomic and ecological similarity of the comparator genome.

Key words: genome composition, phylogenetic classification, support vector machines, metagenomics.

Introduction

Microbial genomes show dramatic differences in their underlying nucleotide compositions. The average G + C composition in sequenced prokaryotic genomes ranges from 16.6% in the reduced endosymbiont Candidatus *Carsonella ruddii* to nearly 75% in certain Proteobacteria and Actinobacteria. Properties such as oligomer nucleotide signatures (Blaisdell et al. 1986; Brendel et al. 1986; Pietrokovski et al. 1990; Karlin and Burge 1995; Abe et al. 2005), codon usage patterns (Willenbrock et al. 2006), conserved sequence repeats (van Belkum et al. 1998), and structural periodicity (Mrázek 2009) are variable and potentially characteristic of different taxonomic groups of microbes. Variation in these patterns has been tied to selective forces including nitrogen limitation in the environment (Willenbrock et al. 2006) and DNA repair systems (Paz et al. 2006; Rocha et al. 2006). Under certain conditions, these patterns and biases can change rapidly relative to changes in commonly used marker genes; for example, strains of the marine picocyanobacterium *Prochlorococcus marinus* show remarkable G + C content divergence from 30% and 50% despite the presence of very similar 16S rDNA sequences, which is likely due to differences in DNA repair genes (Rocap et al. 2003). The G + C content of these genomes correlates with adaptation to different degrees of light intensity, and the rapid genomic divergence may be tied to the rapid ecological divergence of these clades.

The role of ecology in shaping composition is not yet firmly established and indeed may depend on the type of habitat under consideration. Hypersaline environments impose significant physiological challenges on resident microbes, and there is evidence that taxonomically divergent

Tool parade

I don't usually like doing a parade of tools but we're nearing the end of the semester and most of you are working on your final projects, so I'm not going to overload you with practical exercises on these.

But I do want you to be aware of the tools that are available to you both for your own use and because they are so often referenced in literature.

These tools include:

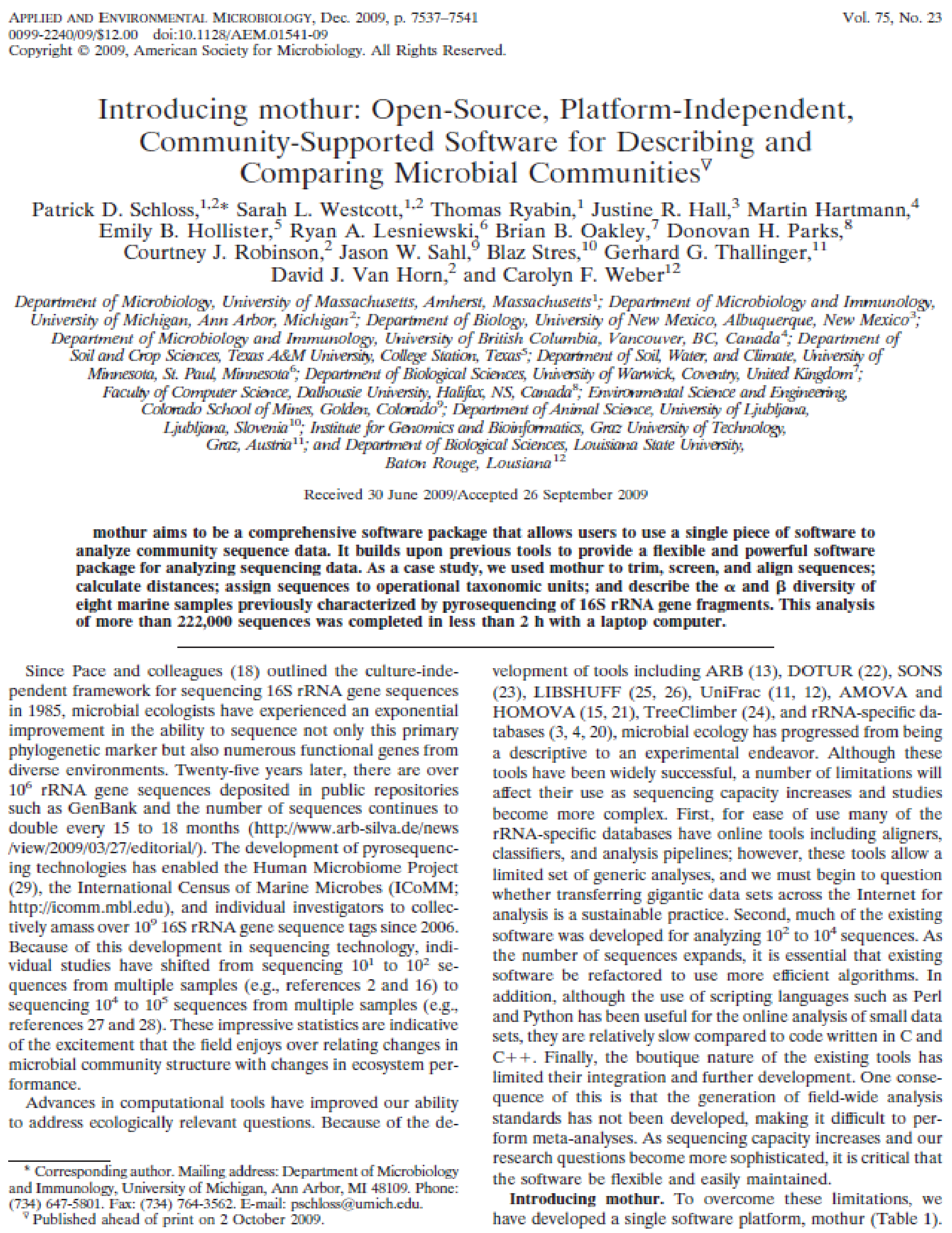
- Mothur
- Qiime
- FastUniFrac
- PhyloSift

Mothur

Mothur is a software suite for researchers in microbial ecology and offers the ability to go from raw sequences to the generation of visualization tools to describe α and β diversity.

Many of the tools incorporated within the suite were initially created elsewhere but have been optimized and integrated.

It contains too many tools to list out, but general categories include sequence/file manipulation, chimera identification, sequence filtering, alignment editing, noise reduction, etc.



APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2009, p. 7537–7541 Vol. 75, No. 23
0099-2240/09/$12.00 doi:10.1128/AEM.01541-09

Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities[∇]

Patrick D. Schloss,[1,2*] Sarah L. Westcott,[1,2] Thomas Ryabin,[1] Justine R. Hall,[3] Martin Hartmann,[4] Emily B. Hollister,[5] Ryan A. Lesniewski,[6] Brian B. Oakley,[7] Donovan H. Parks,[8] Courtney J. Robinson,[2] Jason W. Sahl,[9] Blaz Stres,[10] Gerhard G. Thallinger,[11] David J. Van Horn,[2] and Carolyn F. Weber[12]

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts[1]; Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan[2]; Department of Biology, University of New Mexico, Albuquerque, New Mexico[3]; Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada[4]; Department of Soil and Crop Sciences, Texas A&M University, College Station, Texas[5]; Department of Soil, Water, and Climate, University of Minnesota, St. Paul, Minnesota[6]; Department of Biological Sciences, University of Warwick, Coventry, United Kingdom[7]; Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada[8]; Environmental Science and Engineering, Colorado School of Mines, Golden, Colorado[9]; Department of Animal Science, University of Ljubljana, Ljubljana, Slovenia[10]; Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria[11]; and Department of Biological Sciences, Louisiana State University, Baton Rouge, Lousiana[12]

Received 30 June 2009/Accepted 26 September 2009

mothur aims to be a comprehensive software package that allows users to use a single piece of software to analyze community sequence data. It builds upon previous tools to provide a flexible and powerful software package for analyzing sequencing data. As a case study, we used mothur to trim, screen, and align sequences; calculate distances; assign sequences to operational taxonomic units; and describe the α and β diversity of eight marine samples previously characterized by pyrosequencing of 16S rRNA gene fragments. This analysis of more than 222,000 sequences was completed in less than 2 h with a laptop computer.

Since Pace and colleagues (18) outlined the culture-independent framework for sequencing 16S rRNA gene sequences in 1985, microbial ecologists have experienced an exponential improvement in the ability to sequence not only this primary phylogenetic marker but also numerous functional genes from diverse environments. Twenty-five years later, there are over 10^6 rRNA gene sequences deposited in public repositories such as GenBank and the number of sequences continues to double every 15 to 18 months (http://www.arb-silva.de/news/view/2009/03/27/editorial/). The development of pyrosequencing technologies has enabled the Human Microbiome Project (29), the International Census of Marine Microbes (ICoMM; http://icomm.mbl.edu), and individual investigators to collectively amass over 10^9 16S rRNA gene sequence tags since 2006. Because of this development in sequencing technology, individual studies have shifted from sequencing 10^1 to 10^2 sequences from multiple samples (e.g., references 2 and 16) to sequencing 10^4 to 10^5 sequences from multiple samples (e.g., references 27 and 28). These impressive statistics are indicative of the excitement that the field enjoys over relating changes in microbial community structure with changes in ecosystem performance.

Advances in computational tools have improved our ability to address ecologically relevant questions. Because of the development of tools including ARB (13), DOTUR (22), SONS (23), LIBSHUFF (25, 26), UniFrac (11, 12), AMOVA and HOMOVA (15, 21), TreeClimber (24), and rRNA-specific databases (3, 4, 20), microbial ecology has progressed from being a descriptive to an experimental endeavor. Although these tools have been widely successful, a number of limitations will affect their use as sequencing capacity increases and studies become more complex. First, for ease of use many of the rRNA-specific databases have online tools including aligners, classifiers, and analysis pipelines; however, these tools allow a limited set of generic analyses, and we must begin to question whether transferring gigantic data sets across the Internet for analysis is a sustainable practice. Second, much of the existing software was developed for analyzing 10^2 to 10^4 sequences. As the number of sequences expands, it is essential that existing software be refactored to use more efficient algorithms. In addition, although the use of scripting languages such as Perl and Python has been useful for the online analysis of small data sets, they are relatively slow compared to code written in C and C++. Finally, the boutique nature of the existing tools has limited their integration and further development. One consequence of this is that the generation of field-wide analysis standards has not been developed, making it difficult to perform meta-analyses. As sequencing capacity increases and our research questions become more sophisticated, it is critical that the software be flexible and easily maintained.

Introducing mothur. To overcome these limitations, we have developed a single software platform, mothur (Table 1).

[*] Corresponding author. Mailing address: Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109. Phone: (734) 647-5801. Fax: (734) 764-3562. E-mail: pschloss@umich.edu.

[∇] Published ahead of print on 2 October 2009.

QIIME was written to primarly support processing of high-throughput amplicon sequencing data such as targeted rRNA sequencing, but has been expanded to support shotgun metagenomic samples and other types of data.

It is available as a web interface or command-line tool, and takes users through the following steps using their raw sequencing read data:

- Operational taxonomic unit (OTU) picking
- Taxonomic assignment
- Phylogenetic tree construction
- Statistical analysis and validation
- Generation of summary graphics

It has been developed with scalability of large dataset needs in mind and has been made available as a virtual image (VM) you can run on Amazon's EC2 (for the cloud lovers) or you can get the VirtualBox image and run it yourself.

http://www.qiime.org

FastUniFrac

The older UniFrac has been updated to handle metagenomic datasets where hundreds of microbial communities are possible. It was developed in Rob Knight's lab who, not to be confused with the Fruit of the Loom founder, is an almost intimidatingly productive colleague at the University of Colorado. He is very active in the Human Microbiome project.

From the project website, FastUniFrac allows you to:

- Determine if the samples in the input phylogenetic tree have significantly different microbial communities.
- Cluster samples to determine whether there are environmental factors (such as temperature, pH, or salinity) that group communities together.
- Determine whether system under study was sampled sufficiently to support cluster nodes.
- Easily visualize the differences between samples graphically, with support for three dimensional exploration of datasets and with multiple subcategory coloring.

Registration is free for this web-based resource, which allows you to analyze up to 100000 unique sequences.

Detailed and extensive tutorials are available and include input sample data sets.

http://bmf2.colorado.edu/fastunifrac/

PhyloSift

mining the global metagenome

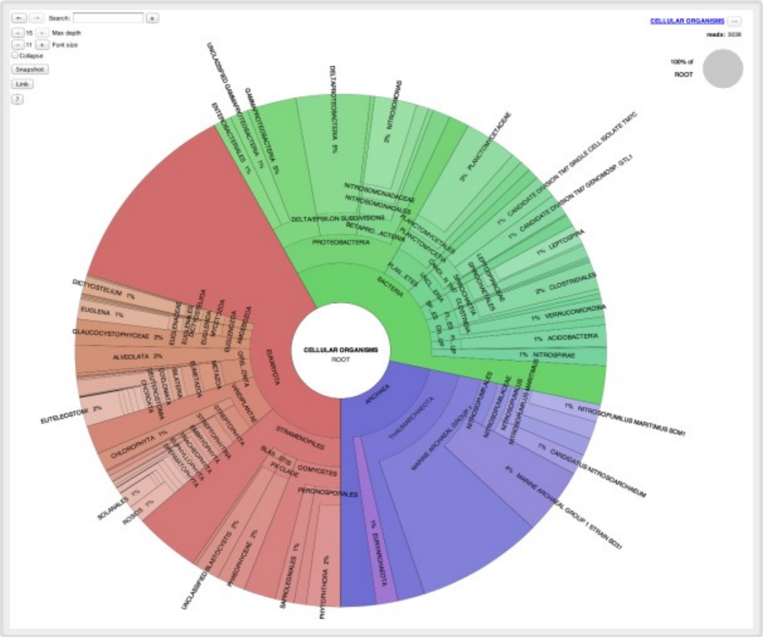
This is the newest of these tools and comes from Jonathan Eisen's lab, a lecture of whom is part of this week's lesson in Blackboard. On the PhyloSift site his developers dub him the “master of microbial evolution of chief oversight on all our fiendish plans for data analysis.”

PhyloSift is a suite of software tools to conduct phylogenetic analysis of genomes and metagenomes.

Using a reference database of protein sequences, PhyloSift can scan new sequences against that database for homologs and identify the phylogenetic relationship of the new sequence to the database sequences. During this procedure, high quality alignments of codon and amino acid sequence are generated.

You can run PhyloSift on your sequences and choose to use a Core Marker Set (~40 markers) or the much larger Extended Marker Set (several GB in size) or even a Custom Marker Set.

The output includes parsable text files and visual output in the form of an interactive Krona viewer.

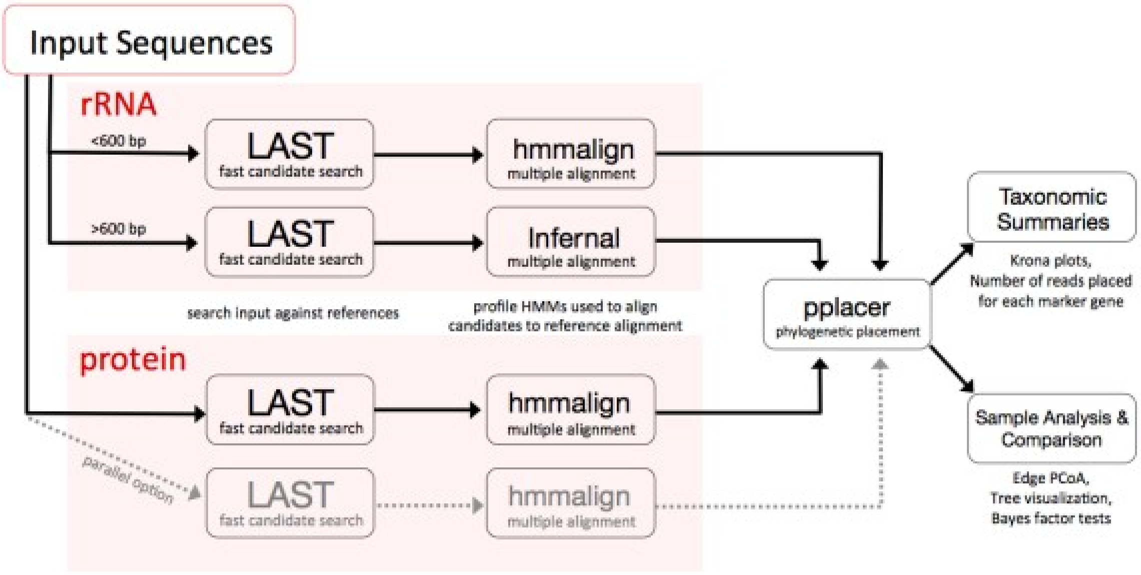


http://phylosift.wordpress.com/

PhyloSift

mining the global metagenome

The PhyloSift client workflow has minor deviations depending on the nature of the input data.



The website is well documented (and is the first bioinformatics tool site I've seen driven by WordPress!)

There's even a tutorial using Illumina data from the Human Microbiome Project.

http://phylosift.wordpress.com/

Further reading

Taxonomic binning of metagenome samples generated by next-generation sequencing technologies
http://m.bib.oxfordjournals.org/content/early/2012/07/31/bib.bbs031.abstract

A novel abundance-based algorithm for binning metagenomic sequences using l-tuples.
http://www.ncbi.nlm.nih.gov/pubmed/21385052

SPHINX--an algorithm for taxonomic binning of metagenomic sequences.
http://www.ncbi.nlm.nih.gov/pubmed/21030462

Unsupervised statistical clustering of environmental shotgun sequences
http://www.biomedcentral.com/1471-2105/10/316?fmt_view=classic

What's in the mix: phylogenetic classification of metagenome sequence samples
http://www.mendeley.com/c/4510348135/g/1628103/mchardy-2007-whats-in-the-mix-phylogenetic-classification-of-metagenome-sequence-samples/

A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio
http://www.mendeley.com/research/robust-accurate-binning-algorithm-metagenomic-sequences-arbitrary-species-abundance-ratio/

MetaABC--an integrated metagenomics platform for data adjustment, binning and clustering
http://www.mendeley.com/research/metaabc-integrated-metagenomics-platform-data-adjustment-binning-clustering/