Sequedex Documentation

Release 1.0-rc1

Joel Berendzen, Judith Cohn, Nicolas Hengartner, Mira Dimitrijevic, Benjamin McMahon

January 06, 2016

Contents

CHAPTER 1

Copyright notice

CHAPTER 2

Introduction to Sequedex

2.1 What is Sequedex?

Sequedex is a licensable software package for genomic analysis that was developed by an interdisciplinary team of scientists at Los Alamos National Laboratory with funding from the Laboratory-Directed Research and Development (LDRD) program. It is currently available for use with a fully functional, node-locked, six month demo license from the Technology Transfer Division of Los Alamos National Laboratory at http://sequedex.lanl.gov. The software is under active development, and information about future releases will be made available at this website, and by an email distribution to the list of users downloading the software. A publication describing the method behind Sequedex, its underlying scientific basis in evolutionary theory and protein structure, its validation on synthetic metagenomic data, and its application to ten soil metagenomics datasets has been published. [1] It was a recepient of a 2012 R&D 100 award. [2]. It is a very flexible piece of software, which can utilize multiple data modules and downstream analysis scripts. It is very fast, reading in signature lists of 5-500 million peptide signatures in 1-15 minutes, and subsequently processes genomic fragments at the rate of 6 Gbp/hr. It parallelizes without significant increase in memory requirements until I/O bound on multiple input files; parallelization works well on 64 processors.

Sequedex has a public repository, wiki, and mailing list on GitHub at http://github.com/joelb123/Sequedex. In the repository will be placed analysis scripts, documentation, and results of Sequedex runs on other data sets that you may compare your data to. Sequedex users should check the site from time to time. The html version of the documentation contains numerous web links and downloadable files.

2.2 What does Sequedex do?

Sequedex is a very flexible piece of software that can be used in conjunction with other software tools to solve a wide variety of problems. It requires a data module such as Life2550-40GB.0 or virus1252 and an input file of sequence data in either fasta, fastq format in plain text or compressed as fasta.gz or fastq.gz. Sequedex will then output four things:

- Genomic fragments in the correct reading frame, annotated by phylogeny (for example, the 2550 nodes defined by Sequedex's Life2550 tree) and function (for example, the 963 categories defined by the seed_0911 functional classification).
- The phylogenetic profile, with a file format described in *Output files - Who?* and reference organisms described in *Tree of Life, 2550 taxa* for Life2550 and *Viral tree, 1252 taxa* for virus1252 data modules.

[1] Berendzen, J., WJ Bruno, JD Cohn, NW Hengartner, CR Kuske, BH McMahon, MA Wolinsky, G Xie, "Rapid phylogenetic and functional classification of short genomic fragments with signature Peptides", BMC Research Notes, August, 2012 (http://www.biomedcentral.com/1756-0500/5/460/abstract)

[2] http://www.rdmag.com/Awards/Rd-100-Awards/2012/08/DNA-Defined-By-Ancestry/

- The functional profile, with a file format described in *Output files - What?* and reference functional categories defined in *Definition of functional classifications* for the seed_0911 categories.
- The 'Who does what?' matrix, with a file format described in *Output files - Who does what?*.
- Statistics about matching reads, with a file format described in *Output files - stats*.

From these outputs, and suitable reference data, such as that described in *Phylogenetic rollups* and *Reference data*, it is possible to answer a wide variety of questions, such as:

- Is an organism that is nearly identical to a sequenced reference genome present in the sample?
- Is a high abundance of a novel organism present in the sample?
- Can I perform host, geographic, or temporal attribution on a sample?
- Are phenotypic (drug resistance, virulence, metabolism) markers present?
- Are multiple samples clonally related?
- Can molecular mechanism be inferred from environmental or expression data?

2.3 How is Sequedex different from other sequence analysis packages?

- Sequedex identifies reads by matching to conserved amino acid signatures, rather than the nucleotide matching of most rapid read mappers. This results in better sensitivity for organisms not closely related to something in the reference database.
- Sequedex uses solid amino acid matches of length ten that imply gene homology, rather than the shorter patterns employed by binning tools. This results in the ability to profile sequence data sets both functionally and phylogenetically.
- Sequedex annotates individual reads with function and phylogeny for more detailed analysis or assembly, and places them in the correct reading frame for translation. Then enables a variety of downstream analysis such as targeted and de novo assembly, followed strain identification or analysis of particular genes.
- Sequedex combines a firm grounding in evolutionary theory and phylogenetic analysis with search-engine technologies.
- Sequedex replaces the question, 'How similar is this nucleotide sequence to those in a reference database?' with the question 'Where has this peptide signature been observed before?'.
- Sequedex tests phylogenetic specificity on a signature-by-signature basis, and is thus less sensitive to the confounding effects of domain swapping, paralogs, and horizontal gene transfer.
- Sequedex relys on an explicit RNA-polymerase based phylogeny to separate the effects of inheritence from functional pressure when interpreting genomic data.
- Sequedex can subsume functional classification schemes merely by searching example gene sets comprising a functional classification with phylogenetically-derived signature lists.
- Sequedex classifies gene fragments at a rate of 6.6 Gbp/hour, 250,000 times faster than BLASTX against NCBI's NR database on soil metagenomics data and runs on a laptop.

2.4 Who uses Sequedex?

Although the Sequedex software package was developed specifically to analyze large shotgun metagenomics data sets characterizing bacterial communities living in soils, just about anyone who generates or analyzes genomic data, or

studies systems where microbial communities are important is a potential user of Sequedex. Specifically:

Sequedex is written in Java and designed to run on modern laptop or desktop computers running Linux, Windows, or Macintosh operating systems

Sequedex is designed to run on computers with a moderate amount of memory that are readily available to data generators. Sequedex requires less than 4 GB of RAM for the 20 million signatures in the Bact403 data module, while the complete tree of life tree, with 160 million signatures will run in 29 GB of RAM. By enabling analysis on desktop (or laptop!) computers, we will enable analysis with rapid turn-around by the same people who best understand the metadata, sample preparation protocols, and expected relationship of the data to previously published studies.

In the process of developing Sequedex, the authors have collaborated with microbial ecologists, microbiologists, evolutionary biologists, medical doctors, epidemiologists, biochemists, and cellular biologists, exploring applications such as profiling microbial communities and microbiomes, identifying attributable changes in microbial communities and microbiomes with condition, enzyme mining, pathogen diagnosis, transcriptomic and proteomic analysis of different strains of algae in differential conditions, and forensic science. In each case, we have found Sequedex to compare favorable to existing methods in the hands of research scientists at the forefront of their field.

The developers have used Sequedex on all 48 cores of a modern server, to effectively run large datasets such as those derived from the human microbiome project or the diversity of environmental microbiomes deposited in NCBI's short read archive. Because of its speed and robust read-by-read classification on a phylogenetic tree, the ability of users to compare their samples to archival data is limited only by their willingness to download, curate, and understand reference datasets. Even the simple distance metrics described below show that meaningful comparisons across research groups, sequencing methodologies, and ecosystems can be made, in order to understand the phylogeny and function of metagenomics samples in comparison to existing datasets. We thus also expect Sequedex to be useful to genomic data centers as they struggle to organize the flood of genomic data into useful ecological, medical, and microbiological conclusions.

2.5 How is Sequedex used with other software?

Sequedex performs a role, phylogenetic and functional annotation of individual reads, that sits at the core of bioinformatics. While direct analysis of Sequedex output is often sufficient to answer the user's questions, it may also spur follow-up questions to obtain more detail or confidence in the results. It is helpful in understanding this to consider the process diagram below, for taking sequence data from the sequencer to various biological conclusions.

While the tasks in this diagram, such as aligning reads to a reference genome, de-novo assembly, or gene identification are often applied directly to the raw data, such directed application is often extremely expensive computationally and can lead to ambiguous results. By incorporating these applications in an analytical process with Sequedex, however, much more targeted questions can be asked, and on subsets of reads with known reading frames. Most exploration of even large next-gen sequencing runs can occur interactively, with most analysis questions requiring only a few seconds or a few minutes to run on a laptop or desktop computer. We guide the reader through numerous examples in this documentation.

2.6 How does Sequedex work?

The Sequedex software distribution contains a Java executable that rapidly searches genomic fragments against a classified signature list, and a data module consisting of a signature list, phylogenetic tree, and a functional classification. Sequedex first reads the signature list into a hash map, then sequentially matches each read of an input fasta file against this list, generating an output consists of three output files: a phylogenetic profile, a functional profile, and a matrix consisting of the number of reads tallied by both function and phylogeny.

The specific algorithm by which reads are assigned phylogeny and function is diagramed below. Each signature is placed on the tree at the most specific node covering all leaves at which the signature was observed. For example, if a particular 10-mer is observed in the genomes of the seven taxa indicated by blue dots in the figure below, that

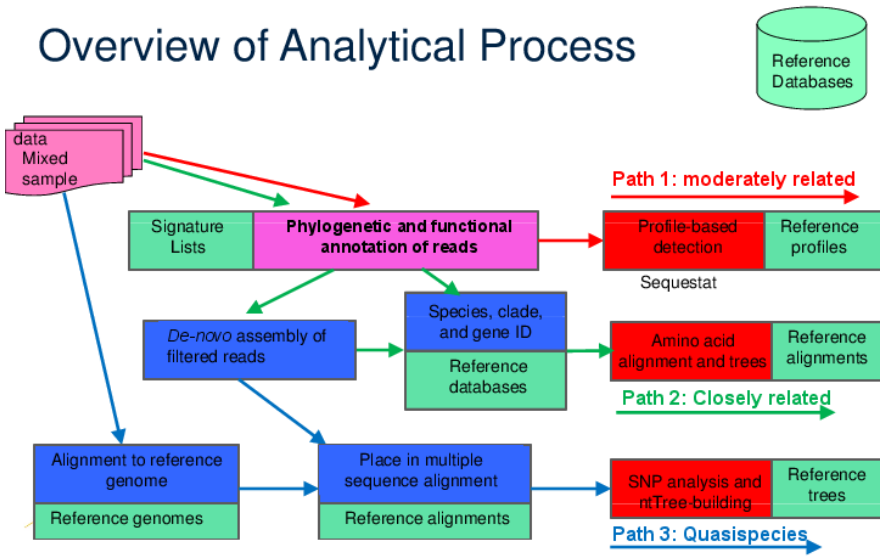


Fig. 2.1: A schematic of how Sequedex can fit into a larger analysis protocol by providing phylogenetic and functional annotation of reads. A complex sample can be analyzed for phylogenetic and functional content directly with Sequedex to give an overall profile of the sample's composition, even if it is composed of sequence from organisms only moderately related to anything in the reference database (Path 1 through protocol). Alternatively, Sequedex can be used as a filter to create phylogenetically or functionally (or both) filtered subsets of reads for *de novo* assembly and further analysis of the larger assemblies with, for example, blastn against NCBI's nt database or blastp against NCBI's nr database to identify organisms closely related to a reference organism (Path 2 through protocol). If these two analyses show high similarity and high abundance of reads to an organism for which a reference genome is available, a read mapper, such as BWA or Bowtie2 can be used to recruit and align reads for quasispecies analysis (Path 3 through protocol).

signature is assigned to the node indicated by the arrow. A similar algorithm, the least common ancestor, is employed for the frequent case of multiple signature peptides occurring in a single metagenomic read. Each read is placed on the tree at the most specific node consistent with the signatures contained in the read. For example, if a read contains signatures assigned to the seven nodes indicated in red on the figure below, the read is placed at the node indicated by the arrow. If conflicting specific signatures are contained in a read, it is assigned at the most specific node covering both conflicting specific assignments.

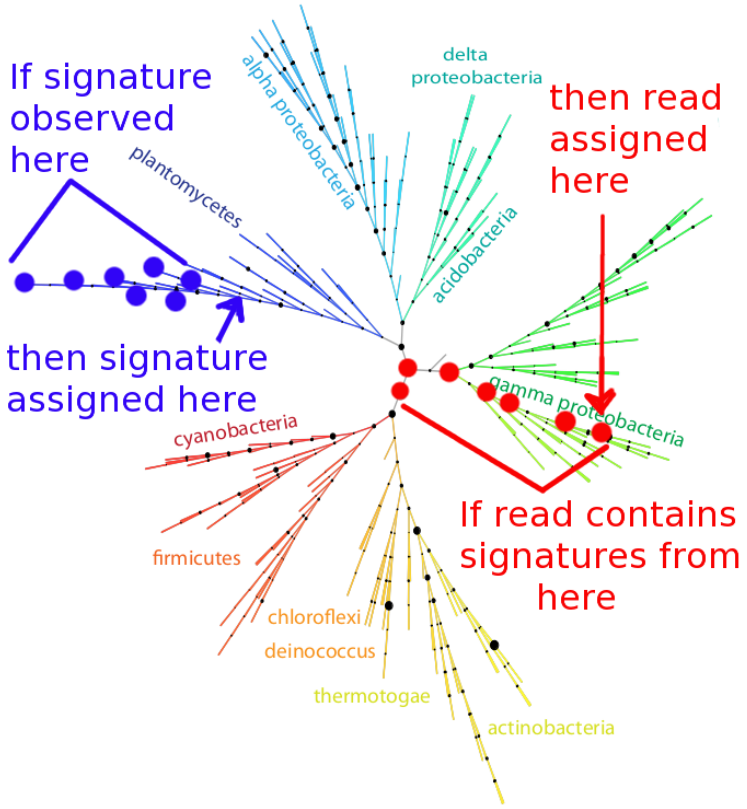


Fig. 2.2: Illustration of the 402-node bacterial reference tree derived from an alignment of the RNA polymerase beta and beta prime subunits, with the signature-assignment algorithm illustrated on the left in blue, and read-placement algorithm illustrated on the right in red.

Phylogenetic placement can be confounded by the ubiquitous processes of gene duplication, domain swapping, and horizontal gene transfer, as well as the differing gene inventories of even closely related bacteria. By choosing a signature-by-signature placement on the phylogenetic tree, we are eliminating the ortholog-identification step from the phylogenetic profiling process. In essence, our approach replaces the question of gene sequence similarity with the question, "Where have these signature peptides been seen before?". Since some signatures appear in dozens of organism, our decision to place the signature peptide close enough to the root of the tree to cover *every* observed instance of the signature can ve viewed as quite conservative. A specific phylogenetic assignment will only be made if **no** conflicting evidence is available, so observed phylogenetic signals represent self-consistency in our assumptions.

By selecting the 5% of 10-mer peptides in our 403 reference genomes, we have identified the 20 million peptides most-likely to appear in bacteria not in our reference set. In order to assign function to each signature peptide, we simply search example protein sequences for each of the signature peptides. The SEED classification [3] of protein function provides such lists for 1000 distinct classifications, as well as a hierarchical rollup to 28 different types of

[3] http://www.theseed.org

families. We were able to make functional assignments for 70% of our phylogenetically-derived signature peptides in this manner.

2.7 Sequedex's outputs

In this section, we show an example of each type of Sequedex's data being plotted.

Phylogenetic and functional profiles: Because each signature possesses both a phylogenetic and a functional assignment, it is possible to profile microbial communities on the basis of both phylogenetic and functional matches, where each read gets one count in the profile. By design, these profiles are not biased by the depth of sequencing, assembly, or the quality of annotation of reference genomes.

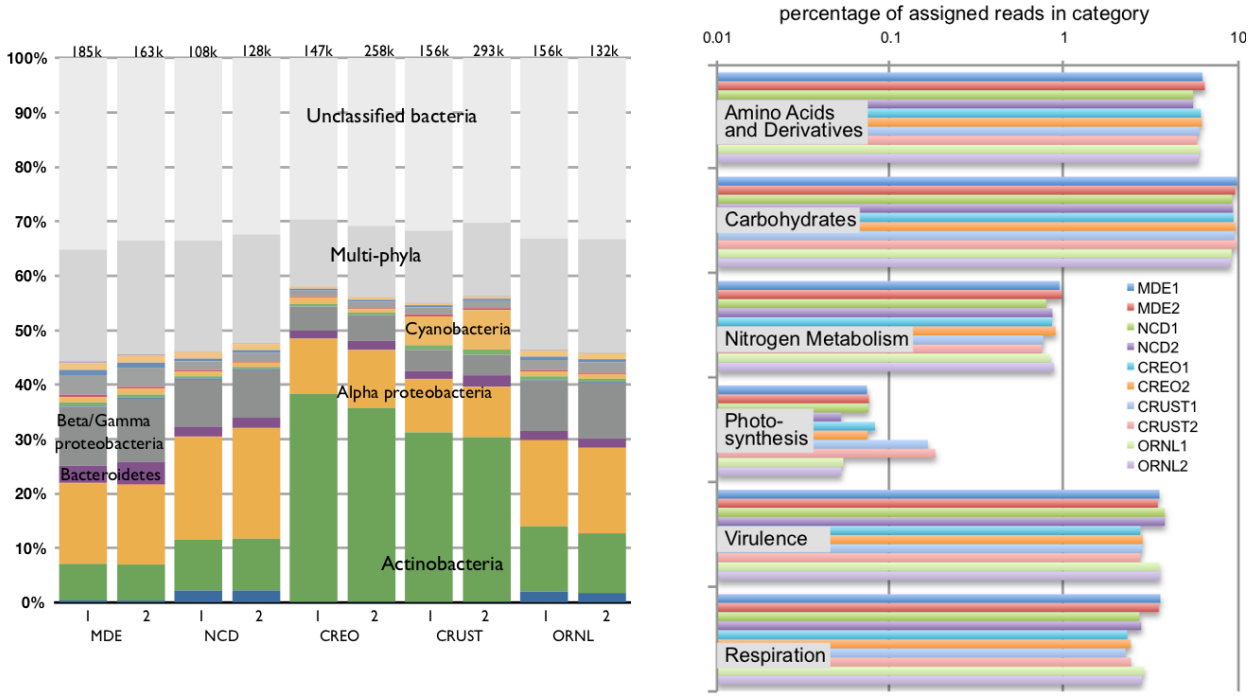


Fig. 2.3: Sample rollups of **(left)** phylogenetic and **(right)** functional profiles of the ten metagenomic communities. The resolution and structure of the phylogenetic and functional classifications depend on the classification of the signature peptides. This calculation used the 402-node bacterial tree and 962-category SEED functional classification. This data set is the one discussed at length in Reference 2.1.

The reference tree used for the Bact403 data module (with node numbers defined) can be downloaded in `phyloxml` or `pdf` format. The functional categories are described in `this csv file.`

Because possibility-space for 10-mer matches is so much larger than the number of signatures, the effect of sequencing errors will typically be to cause one or more signatures in a read to not be detected. Since this is a fairly benign mode of failure, quality scores not used by Sequedex at present.

'Who does what?' matrixes: By definition, every read recognized by Sequedex has a phylogenetic placement (although identification at the root node conveys no taxonomic information). Typically 70% of the identified reads are also given a functional assignment by Sequedex. As the above figure shows, it is frequently most convenient to visualize the phylogenetic profiles averaged over all functions and the functional assignments averaged across the phylogeny. In some cases, however, important information can be obtained by looking changes in the two-dimensional histogram

across both phylogeny and function. We refer to this as the 'Who does what?' matrix, and an example where this is informative is an analysis of a dental carries transcriptome data set figure. Since the bact403 tree contains 403 nodes and the seed_0911 classification contains 961 categories, this matrix has a large number (~400,000) of categories, and is most useful with data sets containing 10s or 100s of million reads.

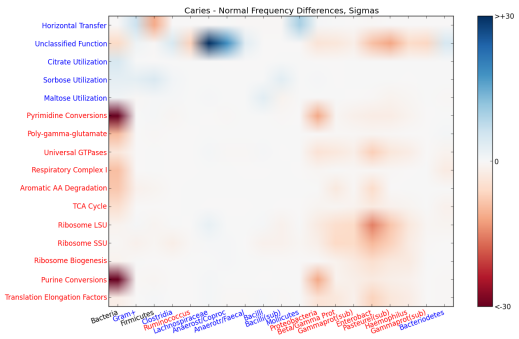


Fig. 2.4: Heat map of changes in expression level of the oral microbiome between 18 patients with dental carries and 18 patients without dental carries broken down by phylogeny (x-axis) and function (y-axis). Blue spots indicate populations that are elevated in patients with dental carries, while red spots indicate populations that are higher in patients without carries. The intensity scale, shown in the inset, is the difference in average between the two populations in units of the average standard deviation within each population.

Annotated reads: While a great deal can be learned from the histograms of counts, further analysis of reads can provide targeted assemblies, precise taxonomic assignments, characterization of quasi-species and mixed populations, and provide specific sequences of enzymes or regulatory proteins for structural modeling or more detailed assessments of gene function. Traditional analysis techniques have been used to identify such reads, but analyses are prohibitively slow, especially since researchers typically do not know ahead of time which phylogenetic taxa are present and whether the depth of sequencing is high enough and the richness of the sample in the phylogenetic region of interest is low enough to enable assemblies to be made. An example of reads identified by both phylogeny and function being compared to reference sequences is shown in the figure below.

FIGURE - Cyanobacterial rubisco reads identified by Sequedex from (i) Sanger sequencing of Global Ocean Survey metagenomic samples, (ii) 454 sequencing of FACE sites soil metagenomic samples, and (iii) reference genes from Genbank, compared together in a multiple sequence alignment.

2.7.1 What is the specificity and sensitivity of Sequedex when matching bacterial genes and gene fragments from novel bacteria?

The sensitivity and specificity with which Sequedex classifies short genomic fragments is determined by the signature list, phylogenetic reference tree, and functional classification contained in the data module. The Bact403 data module distributed with Sequedex was based on 403 reference genomes and their inferred evolutionary history, shown below and available here as a `pdf`.

Sequedex is comparable to BLASTX in sensitivity and specificity for matching bacterial genes from phylogenetically distinct organisms, and, and is intrinsically designed to provide conservative specificity when read-assignments are made to a phylogenetic tree. Because the signatures used for assignment by Sequedex are 30 base pairs in length, the method works nearly as well with short reads as long reads. We employ two validation strategies here: examination of matches to well-annotated but distantly-related organisms and analysis of draft genomes of novel soil bacteria.

The use of solid patterns of amino acids to annotate the function of proteins is startling in both its simplic-

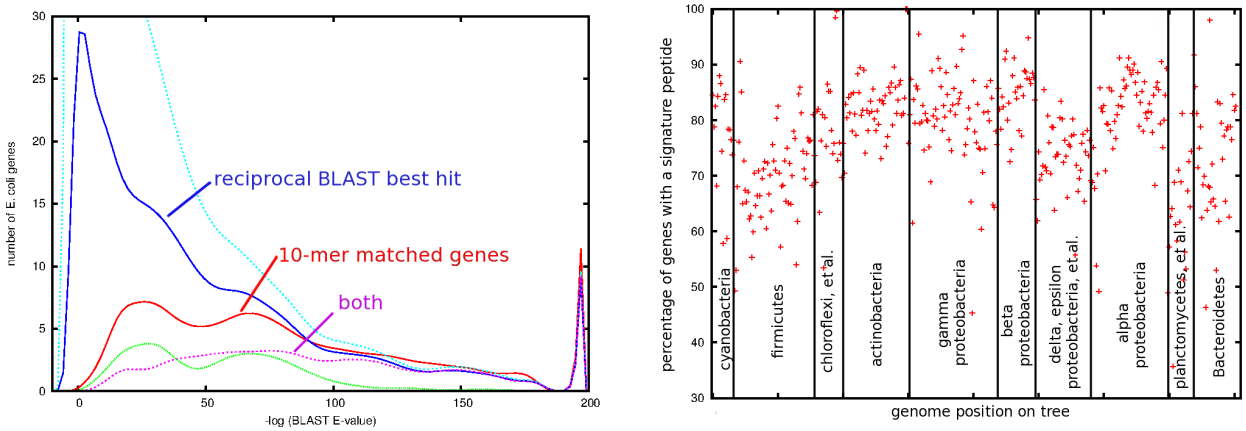


Fig. 2.5: **(left)** Distribution of Protein BLAST scores (-log(E-value)) for various sets of *E. coli* genes scored against genes in the *B. subtilis* genome. At the top, in cyan, is the distribution of the best-match BLAST scores for each of the 4145 genes in the *E. coli* genome. 1461 of these are also reciprocal best hits of the *B. subtilis* genome against *E. coli*; the distribution of these scores is shown in dark blue. 746 distinct pairs of *E. coli – B. subtilis* genes are connected by one or more *10-mer* matches; the distribution of BLAST scores for these matches is shown in red. In magenta is shown the distribution of BLAST scores for the 388 genes that are both reciprocal BLAST best hits and connected by one or more *10-mers*. At the bottom of the plot, in green, is the distribution for genes with matching *10-mers* and the word 'transporter' in either gene's annotation. The peak at the right of the plot indicates the 37 pairs of genes given an E-value of '0.0' by BLAST. **(right)** Fraction of genes containing at least one signature peptide in genomes across the 403 bacterial reference genomes. As described in the text, signature peptides are exact matches between genomes of different genera of length 10. The genomes are ordered along the x-axis according to their position in our bacterial phylogeny.

ity and effectiveness, and its use goes back more than 25 years (see [4] and [5] as well as the prosite website at http://prosite.expasy.org/prosuser.html). If the reader has access to a command line (standard in Linux and Mac operating systems and available as Cygwin for Windows), a few simple commands will illustrate the point.:

```
mkdir bacteria
cd bacteria
wget ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.faa.tar.gz
tar -zxf all.faa.tar.gz
for i in `ls -1 |tr '\n' ' '`;do cat $i/*.faa > $i/all.p.fas;
    tr -d '\r\n' < $i/all.p.fas |tr '>' '\n' > $i/all.1; done
grep GADDY */all.1
grep GG.R.GEME */all.1
```

If **wget** is not installed on your system, simply download the file by some other means and place it alone in a directory. In the case of GADDY, tens of thousands of response regulators will be identified. In the case of GG.R.GEME, you will identify the RNA polymerase in all of the bacterial genomes. Approximately 20 percent of the proteins containing the 5-letter GADDY motif are not response regulators, but the seven letters of GG.R.GEME has 100 percent specificity. Given these observation, it is perhaps not surprising that 10 letter motifs (much longer than anything in prosite) have a very high degree of specificity, as indicated when comparing the distribution of BLAST scores of *10-mer* - matched genes in *E. coli* and *B. subtilus*.

Given the high specificity of *10-mer* peptide exact matches, it is then natural to question the sensitivity of *10-mer* peptides in searching novel organisms. The right panel of the figure above shows that an average of 80% of the genes in our one-per-genera set of 403 representative bacteria contain a shared amino acid *10-mer* with another organism. Both logic and closer inspection of the data reveal that the genes not sharing an amino acid *10-mer* are not the well-understood proteins, and are not particularly robust choices for assigning phylogeny to a metagenomic read. Given that modern sequencers are capable of producing 100 million reads in a single sample, it seems prudent to discard the 20% of reads coding for such genes when trying to phylogenetically profile a microbial community. We are decreasing systematic bias at the cost of a slight increase in counting noise.

Most methods have a higher sensitivity when challenged with data closely related to that in the reference database used to construct the method. Hence, it is prudent to also check the sensitivity of Sequedex on novel organisms. In the next figure, we assess the sensitivity of Sequendex on synthetic reads derived from the 99% complete draft genomes of four soil bacteria cultured from a desert soil microbial community. On the left, we assess the read-length dependence of Sequedex analysis of novel genomic data and compare the results to three established methods. On the right, we look in more detail at how recruitment to nodes of a phylogenetic tree compares to the more typical method of recruiting to the leaves of a tree, with some post processing.

The relative insensitivity of the phylogenetic profile as the length of genomic fragment increases from 75 base pairs to 300 base pairs is visible in the top panel for the four organisms, and is one of the reasons Sequedex can be used to reliably compare data from different studies and sequencing platforms. Although the profile does not change with read length, panel (c) shows the sensitivity with which Sequedex identifies reads in comparison with three other methods for each of the four organisms and as a function of read length (75, 150, 300, and 600 base pairs). We do not attempt to characterize the phylogenetic specificity of assignment in this figure.

To better-understand how the different methods ascribe phylogenetic placement of the synthetic organisms, we present these results for *Herbaspirilium seropediacae* on a phylogenetic tree in panel (d), on the right side of the figure, averaging over the four read lengths. Although *Herbaspirlium* is on the tree (placed on the basis of its RNA polymerase gene), it was not in any of the reference databases. It is immediately evident that the BLAST-based methods, when run through MEGAN, all place the vast majority of reads overly-specifically, with a small fraction incorrectly assigned to leaves of the tree in the wrong family of bacteria. Sequedex, by contrast, assigns only about half of the reads to the appropriate node of the phylogenetic tree, where *Herbaspirilium* joins the phylogenetic tree. Rather than erroneously assigning the reads, Sequedex chooses to assign them simply as 'bacteria' or 'beta proteobacteria'. While increasing the counting noise in the phylogenetic profile somewhat, this behavior makes the comparison of phylogenetic profiles

[4] Doolittle R.F. (In) Of URFs and ORFs: a primer on how to analyze derived amino acid sequences., University Science Books, Mill Valley, California, (1986).

[5] Lesk A.M. (In) Computational Molecular Biology, Lesk A.M., Ed., pp17-26, Oxford University Press, Oxford (1988).

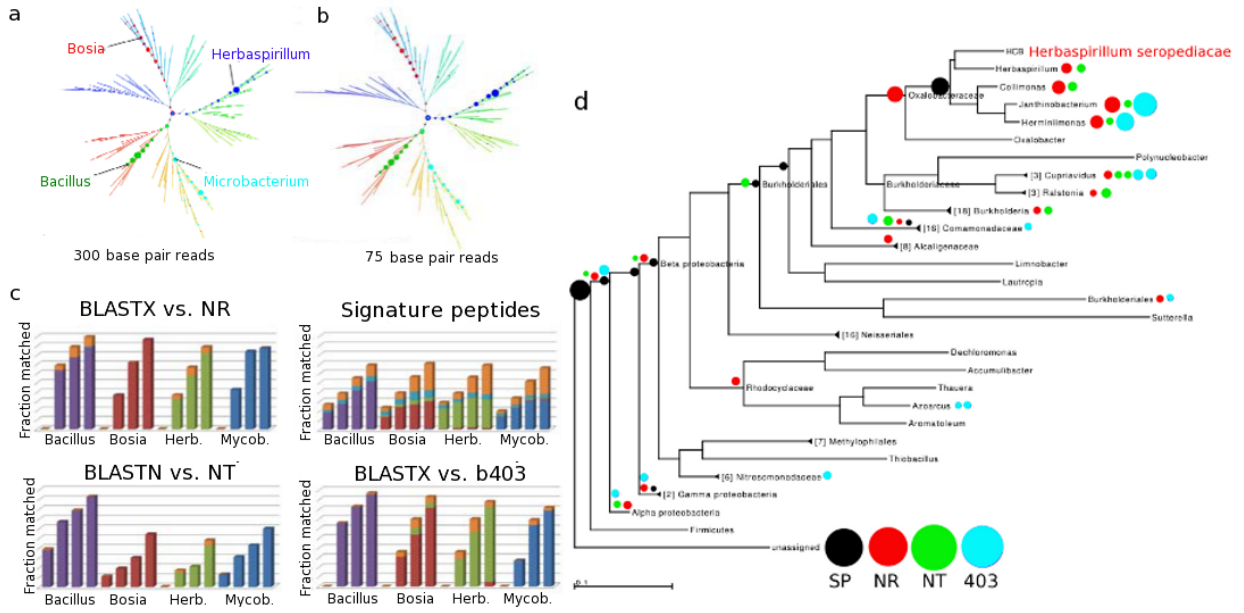


Fig. 2.6: Sensitivity of and specificity of simulated reads from draft soil genomes simulated reads were constructed using MetaSim from genomes of four soil bacteria. *Herbaspirillum seropedicae* and *Bacillus mojavensis* are species from genera represented in the BLAST databases NR and NT as well as our signature peptide database (SP). *Microbacterium trichotecenolytcum* represents a genus found in NR and NT but not in SP. Bosea thiooxidans is from a genus not found in any of the three. **(a)** Specificity of placement of simulated reads on the reference tree using our method. **(b)** Placement of 75-bp reads using our method. **(c)** Comparison of sensitivity of our method (top right panel) and MEGAN using three different BLAST databases: BLASTX and NR (top left) BLASTN and NT (bottom left), and BLASTX against the same genomes used in SP (bottom right). Simulated read lengths of 75, 150, 300, and 600 bp were used for each of the four genomes in each of the four panels. Colors indicate specificity of placement, with gold indicating non-specific placement near the root node in each case. **(d)** Details of specificity of placement of simulated 150-bp *Herbaspirillium seropediacae* for the 4 methods: our method (black), MEGAN4 with BLASTX against NR (red), MEGAN4 with BLASTN against NT (green), and MEGAN4 with BLASTX against the same genomes used in SP (cyan).

much less noisy, as the comparisons will be based only on signature peptides which have *never* been seen outside the specific location on the tree where the assignment is made.

2.7.2 How can Sequedex output be validated and understood?

Sequedex represents a novel method for classifying genomic data, serving a comparable roll to BLAST and defining protein families with hidden Markov Models. Although building off of the idea behind such well-established algorithms as PROSITE and BLOCKS, Sequedex leverages modern indexing and text-processing alogrithms, the relatively complete set of reference genomes, and an algorithmic incorporation of inferred phylogeny to produce a qualitative improvement in both the speed and robustness of genomic classification. Since it represents a novel algorithm, it will require a great deal of empirical application for users to gain a 'feel' for how to interpret the output of signature based analysis and how it handles the curve balls practical application throws its way.

In later sections, we show the user how to validate with his or her own data. To this end, we see several ways Sequedex can be validated and understood:

- Compare results from analysis of files with nucleotide sequences of each gene from completed genomes of various organisms (ffn files)
- Compare results from analysis of individual genes from well-annotated organisms
- Generate and compare synthetic data
- Compare analysis of a variety of microbial communitites
- Output, assemble, align, and classify reads.

2.7.3 How will Sequedex be improved?

- Updated and improved trees (eukaryotes, mitochondria, plastids, archaea, viruses, phages, species-level resolution).
- Better functional classifications (eukaryotes, improved definitions of functional classifications, better annotation of gene function).
- Curation of signature lists to eliminate low complexity signatures, signatures that do not connect orthologs, and mis-called genes.
- Improve the output of peptide fragments on the basis of phylogeny and function, both unaligned and aligned to reference proteins. This will be useful both for assembly, enzyme mining, and more accurate phylogenetic assignments.
- Differentiate deeply branching signatures resulting from conservation and horizontal gene transfer.
- Compare metagenome richness and diversity on the basis of metagenomic data, with only weak dependency on the reference database.
- Identify and annotate signatures on the basis of metagenomic data, with only weak dependency on the reference database.

Sequedex is designed to operate with different signature databases. At present the only databases for public release are the one-per-genus bacterial tree: bact404, the one-per-species tree of life, Life2550, and the one-per-species viral tree, virus1252. This bact404 database was used to produce the data in the paper. Please contact the authors if you have unique access to data that could make use of the broader databases and if you are interested in collaborating on the initial papers.

CHAPTER 3

Installation instructions

Maintaining computational resources for analysis of sequencing data with competing methods has become cumbersome enough that an argument can be made to simply ship the data to a cloud computing platform. Sequedex shifts this balance considerably, enabling researchers to once again analyze their data on their laptop while traveling to conferences. Indeed, since it will typically be faster and easier to analyze data in place, perhaps before deciding whether further analysis is required or perhaps because new data suggests effects that should be present in archived data, users will find it convient to install Sequedex on platforms ranging from Macbook Airs to supercomputing clusters.

Because Sequedex is applicable to such a wide range of applications, we also chose to leverage three powerful and platform-independent programming languages: Java, Python, and JavaScript with html, together with some of their associated libraries. An additional complexity arises from the mixture of system wide and user-specific installations for Sequedex, Java, Python, and the web-browser. System-wide installations are often more convenient, but require that the user have appropriate permissions. We provide the utility, sequedex-bootstrap, which attempts to identify apropriate locations for the various dependencies, and another utility, sequedex-update, to check with the Sequedex website for newer versions of the various components. Upon initial use, sequedex-update will automatically download an appropriately sized data module.

For users behind a firewall, it should be possible to set the environment variable, http_proxy, to an appropriate value to enable internet access for, among other things, sequedex-update. Some users will find it appropriate to install Sequedex on computers without internet access, in which case several files, as well as the license files will need to be downloaded and copied by hand into the appropriate location. We provide specific instructions for such users below, in the section *Installation and updates without network access*.

While we have made every attempt to provide install scripts to work on all modern operating systems (Windows, Mac, and Unix), the popularity of Python, Java, and web browsers on the various platforms, configured by other applications makes it possible for computers to be in such a state that the Sequedex install scripts do not work appropriately. To account for this, we provide extensive configuration variables, accessed by sequedex-config, and debuging information, accessed through log files and sequedex-debug, to ensure the correct versions of software are accessible with the correct set of positions.

The computational expense of standard sequence analysis methods, such as Blast, HMMer, or assemblers have led most scientists to configure expansive compute clusters, and the temptation to install Sequedex in the same manner is significant. Sequedex is 100,000 times faster than BlastX running against the nr database from NCBI, and 64 files can be easily processed in parallel on a 64 core machine with 64 GB of RAM. Disk access to input files will be the limiting case in this situation, not memory or processor speed. Thus, rather then trying to force Sequedex into the expensive compute-environment needed by other bioinformatics tools, it almost certainly makes more sense to identify a modest computer node with rapid access to data, and run Sequedex jobs on it.

In addition to the installation instructions described here, users of Sequedex may find a wide variety of tools of use, both to visualize Sequedex output and to perform other analyses to confirm or further explore the Sequedex results. Although none of these tools are requred, we will refer to them as appropriate in the rest of this documentation. Introducstions and suggestions for using these tools are provided in the chapter on *Additional software tools and resources for use in conjunction with Sequedex*.

3.1 System requirements

- A computer with a 64-bit operating system such as Mac Snow Leopard or later or Linux 2.6 or later.
- A 64-bit Java 1.7 or 1.8 run-time installed. The Java 1.6 distributed with Mac OS X is not acceptable. Desktop users may download Java from the java.com website..
- A 64-bit Python distribution with working SciPy installed (either version 2 or 3 but tested under 2.7). The python distributed with Mac OS X is not acceptable, as it does not included needed components. Desktop users may download a free and complete python distribution (Anaconda) here.
- If your location uses a proxy, the environmental variable http_proxy must be set to use your proxy for all command-line sessions. For BASH users this can be done once by issuing the following command with appropriate substitutions:

  export http_proxy=proxyout.mycompany.com:8080
- 1 GB of disk space is required for download, installation, and testing.
- 4 GB of RAM. However, RAM is cheap and Sequedex will use bigger data modules that identify more reads if you have 8, 16, 32, or 48 GB installed.
- A windowing environment even on server systems to run the licensing GUI. The ability to forward X-window connections using ssh -Y is sufficient.

With the included data module, Life2550-4GB, Sequedex uses a maximum Java heap size of less than 4 GB. Larger 'Life2550' data modules, require more RAM, and provide a higher sensitivity, due to more extensive signature lists. Desktop computers can be purchases for under $1000 with 32 GB of RAM, and servers with 64 GB or more of RAM are widely available. Given the expense of acquiring sequence data, it seems likely that an investment of a 32 GB desktop computer for analysis would be warrented. We have found USB 3.0 drives for reference data (currently $120 per 3 TB drive) is also an effective investment.

Sequedex is designed to run multiple Java threads, with each thread processing one input file. Depending on whether the input file is compressed (.gz files are supported), whether classified reads are written to disk, and the I/O speed of your disk (USB 3.0 or Thunderbolt drives work well), we have observed the total performance of Sequedex to continue to improve, even with 48 threads (on a machine with 48 cores). Certainly, all available threads on dual- and quad-core processors would be effectively utilized.

3.2 Downloading and unpacking for Mac

The latest Mac-specific distribution is available at the Sequedex website. After downloading, unpack the file where you would like it installed, and add the directory sequedex/bin to your PATH environment variable. For user-installs, this might be in the user's home directory; for system-installs, this might be /usr/local:

```
tar -xzf sequedex-1.0.x-Mac.tgz
PATH=~/sequedex/bin/:$PATH
```

In order for sequedex to run, however, 64 bit versions both Java (version >= 1.7, and Python (either version 2.x or 3.x)must be installed and in the user's PATH environment variable. This can be tested in the command prompt as follows:

```
  $ python
Python 2.7.3 (default, Feb 14 2013, 10:33:11)
[GCC 4.6.3] on linux2
Type "help", "copyright", "credits" or "license" for more information.

  $ java -version
```

```
java version "1.7.0_17"
OpenJDK Runtime Environment (IcedTea7 2.3.8) (Gentoo build 1.7.0_17-b02)
OpenJDK 64-Bit Server VM (build 23.7-b01, mixed mode)
```

If your computer has more than 4 GB of memory and if you have a proper python installed as under "System Requirements", then you should type:

```
bin/sequedex-bootstrap
bin/sequedex-update
```

If the above was performed successfully, you can execute the:

```
sequescan
```

command and proceed with obtaining a demo license by using Utilities..Request License and following the instructions there. More details are available in the online documentation.

3.3 Downloading and unpacking for Linux

The latest Linux-specific distribution is available at the Sequedex website. After downloading, unpack the file where you would like it installed, and add the directory sequedex/bin to your PATH environment variable. For user-installs, this might be in the user's home directory; for system-installs, this might be /usr/local:

```
tar -xzf sequedex-1.0.x-linux.tgz
PATH=~/sequedex/bin/:$PATH
```

In order for sequedex to run, however, 64 bit versions both Java (version >= 1.7, and Python (either version 2.x or 3.x)must be installed and in the user's PATH environment variable. This can be tested in the command prompt as follows:

```
  $ python
Python 2.7.3 (default, Feb 14 2013, 10:33:11)
[GCC 4.6.3] on linux2
Type "help", "copyright", "credits" or "license" for more information.

  $ java -version
java version "1.7.0_17"
OpenJDK Runtime Environment (IcedTea7 2.3.8) (Gentoo build 1.7.0_17-b02)
OpenJDK 64-Bit Server VM (build 23.7-b01, mixed mode)
```

If your computer has more than 4 GB of memory and if you have a proper python installed as under "System Requirements", then you should type:

```
bin/sequedex-bootstrap
bin/sequedex-update
```

If the above was performed successfully, you can execute the:

```
sequescan
```

command and proceed with obtaining a demo license by using Utilities..Request License and following the instructions there. More details are available in the online documentation.

3.4 Downloading and unpacking for Windows 7 or 8

Sequedex has been tested under Windows 7, Java 7u60, and Python 2.7 (packaged in Anaconda 2.0.0.

1. Install the latest release of Java 7. Java 8 seems to work, but has not been tested as of yet. Java 6 is not supported because of memory management issues on some platforms and some kernel configuration settings.

To get a Java release for Windows, visit the Oracle Java Standard Edition Download site at http://www.oracle.com/technetwork/java/javase/downloads and scroll down to the "Java SE 7uXX" where "XX" is the latest release number. Click on the "Download" button under "JRE". You must click on "Accept license agreement" in the next page. Click on "jre-7u60-windows-x86.exe", then click on "Save File" in the popup that opens. Run the executable file in the usual way with defaults:

1. Click on "yes" when asked if you wish to allow the program to make changes to this computer.
2. Click "install" to install the JRE in the default location.
3. Click "close" to complete the installation.

Start a new Command Prompt window and issue the command:

```
java -version
```

and you should see some lines beginning with:

```
java version "1.7.0_XX"
```

where XX is the current release number.

2. Install the latest release of python 2.7, with the "setuptools", "scipy", and "bokeh" modules. Python 3.3 may also work, but has not been tested on platforms other than linux. We recommend the Anaconda python distribution from Continuum Analytics because it is free and includes the necessary modules out of the box. Visit the download site at http://continuum.io/downloads, then scroll down the page to "Windows installers" and click on "Windows 64-bit" to get the executable. We tested by installing for All Users; if you do not have administrative privileges you will need to ensure that the python executable ends up on the path for windows programs.

Start a new Command Prompt window and issue the command:

```
python --version
```

and you should see a line similar to:

```
Python 2.7.7 :: Anaconda 2.0.0 (64-bit)
```

3. If you are using Windows, you may need to install an unzip utility if you haven't already. Either WinZip or p7zip should work just fine. The latter may be downloaded from http://www.7-zip.org/download.html and installed using the usual procedures.

4. You may need to modify your system's environmental variables if any of the three things apply:

2. If you intend to use Sequedex from the Command Prompt window, which is required for updating and getting correctly-sized modules for your system (highly recommended)
3. If your location uses a proxy,and the variable HTTP_PROXY has not been set before. To check this, in a Command Prompt window issue the command:

```
set http_proxy
```

and you should see a value returned such as:

```
HTTP_PROXY=http://proxyout.mycompany.com:8080
```

If any of the 3 conditions above apply, follow the following steps:

1. From the Desktop, click the "Start" button, then click "Control Panel".
2. From the Control Panel click "System and Security", then click "System".

3. In the left-hand menu click “Advanced system settings” to open a new window.
4. Click the “Environmental Variables” button.
5. If you have administrator privileges, click as instructed below under “System variables”. If you do not have administrator privileges, click as instructed below under “User variables for ...”
6. Click on “Path” and click “Edit...”, or if it doesn’t yet exist as a User Variable click on “New...”. Go to the end of the PATH string definition and add the p7zip and Sequedex binary locations you will use, separated by semicolons. On most systems this looks like:

```
...;C:\Program Files\7-Zip;C:\Users\myusername\sequedex\bin
```

   Then click “OK” to close the popup.

7. If you need to define HTTPPROXY click on a new variable and in the popup for “New System Variable” for “Variable name” enter “http_proxy” without the quotes. Case does not matter. In the “Variable value” field enter the URL for your location such as:

```
"http://proxyout.mycompany.com:8080".
```

   Then click “OK”

8. Click “OK” to close the “Environmental Variables” window and “OK” to close the “System Properties” window.

Start a new Command Prompt window and verify that the path and proxy have been set as above. Failure to set the proxy when needed will result in obscure failures when Sequedex tries to update itself.

5. From the Sequedex website, download the latest Windows distribution and unpack it into the directory of your choice (typically your home directory on single-user systems). Issue the commmand:

```
sequedex-bootstrap
sequedex-update
```

and finally, launch the GUI with the command:

```
sequescan
```

3.5 Using Sequedex with Cygwin installed under Windows

Cygwin, described in *How to install Cygwin and what to include.*, provides a unix-like environment under a Windows operating system, and provides a much richer command-line experience than the DOS command prompt otherwise used to launch Sequedex on a Windows computer. Users should install Sequedex as above (although it is fine to put Sequedex in your cygwin home directory, the user-specific files will still be placed, by default, in the Windows home directory, typically /cygdrive/c/Users/username/.sqdx.

If the user places the appropriate versions of Java and Python in the Windows PATH environment variable, they will be available under Cygwin, and the command:

```
sequescan run -h
```

should work, largely creating a Unix-like environment for the downstream analyses described here.

3.6 Installation and updates without network access

3.7 Testing your installation

To test that Sequedex is installed properly, open the Sequedex GUI, click on the file chooser icon at the right edge of the "Input" text field. Mac users should find the testData directory described in the installation instructions above. Linux users will find the same testData directory at the top level of the Sequedex distribution directory (i.e. in the same place as the bin directory). Under the testData directory, users will find another directory called "synthetic" and should then select one of the files under this directory. Click "Run Sequescan" button...

Note: If you have not yet installed a license file, the following message will appear in the Progress window of the GUI:

```
You are currently unlicensed.   Limited analysis will be allowed for testing purposes.
```

Instructions for installing a license are provided below.

3.8 Running Sequedex on an example data file

To analyze a file of sequence data with Sequedex:

- Click on file chooser icon at the right edge of the "Input" line (Mac) or type 'sequescan' at the command prompt after correctly updating your PATH variable (Mac or Linux). You should get an interface like the one shown below.
- Select the testData directory you installed earlier
- Select one of the files under one of the sub-directories of "testData"
- Click "Run Sequescan"

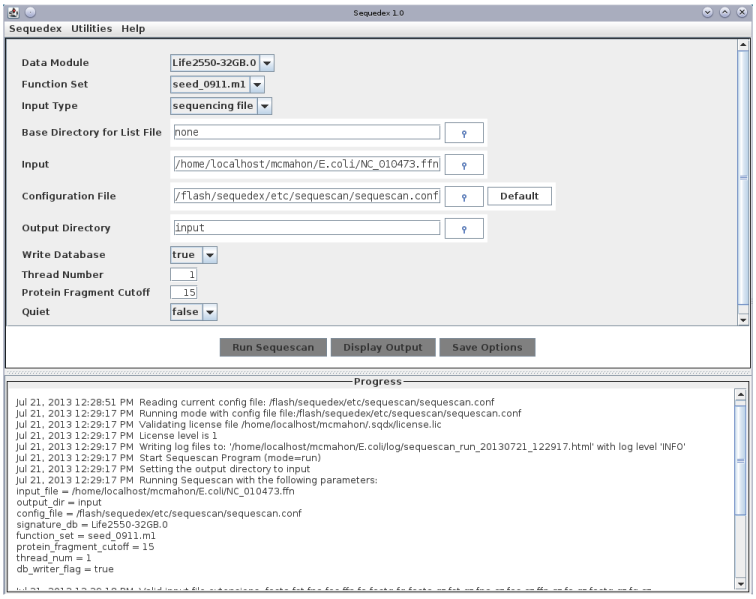


Fig. 3.1: The graphical user interface for sequescan.

After the "Run Sequescan" button is pressed, a few lines should appear in the "Progress" portion of the interface immediately. Sequedex will spend about one minute reading the bact403, and about 15 minutes reading the tol-all

signature list into memory. Once this is complete, processing of the data files will occur at a rate of a few billion base pairs per second. We will examine the run-time options and output files in the next chapter.

3.9 Obtaining a node-locked license file

Without a license file, Sequedex will only analyze a few tens of thousands of reads at a time, and batch file processing is disabled. This will enable the user to analyze bacterial genomes and portions of metagenomics data sets to verify that Sequedex is performing correctly. You may obtain a free 60-day node-locked demo license by submtting a license information file by e-mail. The terms of the demo license are available `here`.

To request a license, you must first create a license information file: start the Sequedex GUI and click on 'Utilities' at the top of the GUI, then on 'Request license'. Follow the instructions to generate a license information file. E-mail this license information file to to sequedex-demo-license@lanl.gov. A demo license will be returned to you immediately if your e-mail address has not requested a demo before. This file can be installed by choosing 'Install License' under the same 'Utilities' tab at the top of the sequescan GUI. If the license is correctly installed in the .sequedex subdirectory of your home directory, you will notice that mutiple threads and multi-file processing is enabled. If 'verbose' mode is selected, you will be notified that 'License level is 1' will be displayed as Sequescan starts its run. Since the license file is stored outside the Sequedex root directory, this file will not be overwritten when Sequedex is upgraded with a new distribution.

3.10 Installing new data modules and upgrading Sequedex - User-installs

New data modules can be installed into a Sequedex distribution by copying them into the sequedex/data/ directory.

A new Sequedex distribution can be installed directly with the install script if available (and if the existing version of Sequedex was set up in the default location) or by unpacking on top of the existing Sequedex distribution. Since the license file is located outside of the Sequedex distribution (in the .sequedex directory of the user's home directory) for User-install, it will be unaffected by a re-install of Sequedex.

3.11 Installing new data modules and upgrading Sequedex - System-installs

The license for Sequedex is node-locked, and not restricted to a single user. Consequently, the license file can either be located within the Sequedex distribution or copied into each user's home directory, into the ~/.sqdx directory. Note for Cygwin users, that the home directory is the Windows home directory (Typically c:Usershomeusername) and not the Cygwin home directory. For system installs, all users must have read and execute permissions to the Sequedex distribution, and new data modules can only be installed by someone with write permissions to the sequedex/data/ directory.

CHAPTER 4

Getting started

A note about downloadable files

In this and the following five chapters, we use a wide variety of real-world examples to illustrate how Sequedex can be visualized and understood as part of a broader work-flow. Not only are the analysis scripts useful templates for modification to meet the user's own specific needs, but reference trees and figures, as well as the coallated output from a broad array of publically available data sets provide a useful context for comparing his or her own data. While links are always provided in the midst of the discussion, it is also possible to download this entire documentation package, with web-pages, figures, and downloadable files at https://media.readthedocs.org/htmlzip/sequedex/latest/sequedex.zip. This file will unpack into a directory sequedex-latest, with a subdirectory _downloads. If the user places the contents of (or renames) sequedex-latest/_downloads into Sequedex-docs/dl/ of the user's home directory, most of the examples can simply be cut and pasted directly into the correct application in order to run. Otherwise, fairly straightforward changes will need to be made for many of the examples to account for the actual location of downloaded files.

4.1 Running Sequescan

Sequedex can be run in several modes with either a simple graphical user interface or directly from the command line. The platform-specific instructions for starting the graphical user interface and obtaining a license file were provided in the previous chapter, and a description of command line use of Sequedex is provided at the end of this chapter. If the license file is not obtained or installed incorrectly, Sequedex will process a limited number of reads (~10,000) and work only in single-thread mode. For the bulk of this chapter, we will assume the graphical user interface is being used, and that the user has some way of looking at the tab-delimited text output files, whether that be Excel, a text editor, or an analysis environment like R or Matlab.

Sequedex is distributed with several sets of test data in a tar archive or a zip file for demonstration and validation. In this chapter, we illustrate how to use Sequedex on these datasets. Two metagenomics data files are provided, one from a stool sample sequenced under the human microbiome program and two from a soil sample collected from a Maryland estuary. Also provided are synthetic data files constructed from the draft genomes of four novel organisms cultured from desert cyanobacterial mats, *Bacillus mojavensis, Bosea thiooxidans, Herbaspirillum seropedicae,* and *Microbacterium trichotecenolyticum.* The complete stool sample file can be obtained at http://www.ncbi.nlm.nih.gov/Traces/sra/?view=run_browser&run=SRR059346 and the Maryland estuary sample analysis can be compared to the analysis in Figure 7 of Berendzen, et al. [1] as well as additional files 5 and 6 of this work. The synthetic data sets were used in this same manuscript to explore the read-length dependence of Sequedex's sensitivity and specificity, illustrated in Figure 6.

[1] Berendzen, J., WJ Bruno, JD Cohn, NW Hengartner, CR Kuske, BH McMahon, MA Wolinsky, G Xie, "Rapid phylogenetic and functional classification of short genomic fragments with signature Peptides", BMC Research Notes, August, 2012 (http://www.biomedcentral.com/1756-0500/5/460/)

The test data set for the human microbiome sample is located in the unpacked archive directory testData/metagenomes/humanMicrobiome/gut/. If you are starting the Sequedex GUI from the command line, it may be easiest to navigate this directory before typing sequescan. Alternatively, you can start the sequescan GUI from the applications menu and navigate to the data directory with the dialog box. In either case, you should see a screen that looks like this when you have started analysis of the data set:

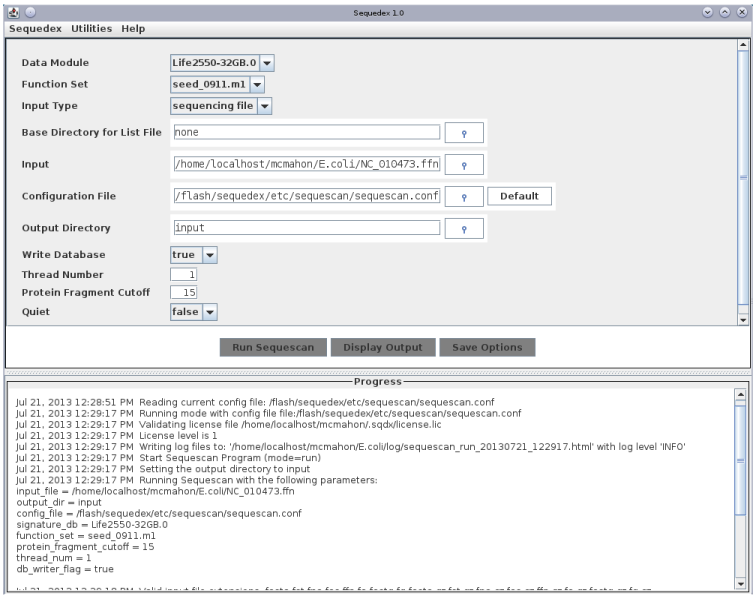


Fig. 4.1: **The graphical user interface for Sequescan.** At the top are two pull-down menus that enable the user to request and install a license and to quit the program. Three input selections allow the user to choose the signature data module, the functional data module and select whether the fasta input files will be chosen by name, by directory, or by an input file listing the complete path of files to be analyzed. To learn more about memory usage, please read *Alternative data modules and memory usage* before proceeding. If the user chooses to use a file with a list of input files (which can be in multiple directories), the next input specifies the base directory for the filenames (defaults to the filesystem root). Below this, the 'Input' box is used to select a data file, data directory, or a text file listing data files to be analyzed, depending on 'Input type' selected above. A configuration file is selected to specify additional program options. The top-level directory to which output files will be written can be selected. Finally, the number of threads used and whether verbose output messages are desired can be selected. Note that at present multiple threads can only be used when multiple files are simultaneously analyzed; each file is assigned a thread.

Download the nucleotide sequence of the called genes in *E. coli* from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__DH10B_uid58979/NC_010473.ffn). Click on the button to the right of the 'input' field and locate this file on your computer. After opening this file and clicking the 'Run Sequescan' button at the bottom of the dialog box, Sequescan will begin reading the bacterial signature list into memory, then classifying the reads in against this signature list. Output similar to that below will appear in the 'Progress' box at the bottom of the GUI:

```
Jul 21, 2013 12:28:51 PM  Reading current config file: /flash/sequedex/etc/sequescan/sequescan.conf
Jul 21, 2013 12:29:17 PM  Running mode with config file file:/flash/sequedex/etc/sequescan/sequescan.
Jul 21, 2013 12:29:17 PM  Validating license file /home/localhost/mcmahon/.sqdx/license.lic
Jul 21, 2013 12:29:17 PM  License level is 1
Jul 21, 2013 12:29:17 PM  Writing log files to: '/home/localhost/mcmahon/E.coli/log/sequescan_run_201
Jul 21, 2013 12:29:17 PM  Start Sequescan Program (mode=run)
Jul 21, 2013 12:29:17 PM  Setting the output directory to input
Jul 21, 2013 12:29:17 PM  Running Sequescan with the following parameters:
input_file = /home/localhost/mcmahon/E.coli/NC_010473.ffn
output_dir = input
```

```
config_file = /flash/sequedex/etc/sequescan/sequescan.conf
signature_db = Life2550-32GB.0
function_set = seed_0911.m1
protein_fragment_cutoff = 15
thread_num = 1
db_writer_flag = true

Jul 21, 2013 12:29:18 PM  Valid input file extensions: fasta fst fna fas ffn fa fastq fq fasta.gz fst
Jul 21, 2013 12:29:18 PM  It is assumed that input files contain DNA sequence; only A,C,T,G will be t
Jul 21, 2013 12:29:18 PM  Minimum protein fragment length is 15; if reads have less than 45 bp, you s
Jul 21, 2013 12:29:18 PM  Reading signature map from data module
Jul 21, 2013 12:35:38 PM  Reading functions from data module
Jul 21, 2013 12:35:38 PM  Adding /home/localhost/mcmahon/E.coli/NC_010473.ffn to the queue as fasta
Jul 21, 2013 12:35:38 PM  Adding analysis observer gov.lanl.sequtils.writer.SequencingFileWriter for
Jul 21, 2013 12:35:38 PM  Begin matching of reads in file /home/localhost/mcmahon/E.coli/NC_010473.fi
Jul 21, 2013 12:35:41 PM  NC_010473.ffn: 4441680 chars processed (100.0 percent completed)
Jul 21, 2013 12:36:10 PM  End Sequescan Program
```

The log directory will contain a subdirectory with an html file that can be opened with a web-browser, indicating progress, errors, and warnings associated with this Sequescan run. Output files will appear in the subdirectory 'NC_010473.ffn.sqdx' of the directory where the input file was, and contain four types of outputs in two seperate formats (tab-separated-variables and json / tsvj); we examine the four types of tsv files in the next section. These files can be imported into Excel, a text editor, or simply examined on the command line with the 'less' or 'cat' commands.

4.2 Output files - stats

The output file Life2550-32GB-stats.tsv contains basic statistics about the computation. The stats file resulting for the above run contains:

```
Sequescan Version       1.0
Log File        /home/localhost/mcmahon/E.coli/log/sequescan_run_20130721_122917.html
Data Input File /home/localhost/mcmahon/E.coli/NC_010473.ffn
Data Output Directory   /home/localhost/mcmahon/E.coli/NC_010473.ffn.sqdx
Local Time      07.21.2013_12:36:04
Percent of File Processed       100.0
Processing Complete     true
Thread Pool Size        1
Processing Rate (Gbp/hr)        4.241
Reads In        4128
Bases In        3904869
Reads With Fragments >= 15 AA   4128
Frames Processed        24768
Frames with Fragments >= 15 AA  24707
Fragments In    306270
Fragments >= 15AA       142158
Reads Assigned  4060
Single-Signature Reads  37
Single-Node Reads       102
Monophyletic Reads      3645
Non-Monophyletic Reads  276
Fragments Assigned      4577
Percent of Reads Assigned       98.4
Total Size of Matched Fragments (bp)    3976512
Fragments Assigned      3226
Total Size of Fragments Assigned (bp)   3180642
Input Time(ms)  33
```

```
Translate Time (ms)      1358
Match Time (ms) 1410
Assignment Time (ms)     154
Other Time (ms) 360
Total Time (ms) 3315
```

The input file is listed on the top line, followed by the fraction of the file processed, which in this case is 100%. For large files, intermediate results are written to output files, and this line indicates an estimated fraction processed. In the case of a missing or invalid license file, only a fraction is prcessed, and that fraction would be indicated here. The 'Processing Complete' line indicates 'true' if the processing completed normally, whether because the end of the input file was reached or because a valid license was not found.

The Thread Pool Size of one is indicated, together with a processing rate of 4.241 Gbp/hr. When processing multiple files simultaneously, it is possible to have a dedicated Java thread processing each file, all comparing to the same memory map. The processing rate provided is of the single thread dedicated to this particular file. Typical values for computers in use as of this writing are 3.5 to 7 Gbp/hr, with the rate per thread dropping roughly a factor of two when 40-60 processors are running together. Thus a 64 core machine capable of 3.7 Gbp/hr on a single thread was able to process fasta files at the cumulative rate of 100 Gbp/hr, and each output file had a processing rate of 1.7 Gbp/hr.

The 4128 genes and 3904869 base pairs processed are indicated next, from which the average gene length of 948 base pairs can be computed. Of the 4128 genes processed, all contained at least one region with 15 or more consecutive amino acids uninterrupted by an undetermined nucleotide or stop codon. This cutoff value of 15 amino acides necessary for further consideration by Sequedex can be changed in the configuration file. Sequedex at present does not recognize ambiguity codons.

From the 4128 genes, 24768 reading frames were processed, of which 24707 contained an amino acid region with 15 or more concecutive amino acids. Since many of the reads contained mulitple fragments, 142158 peptide fragments longer than 15 amino acids were identified, of which 4577 (less than half) were of length 15 amino acids or longer.

Over-all, 69,170 reads (98.4% of the processed reads) contained at least one signature peptide and thus could be assigned a phylogenetic placement.

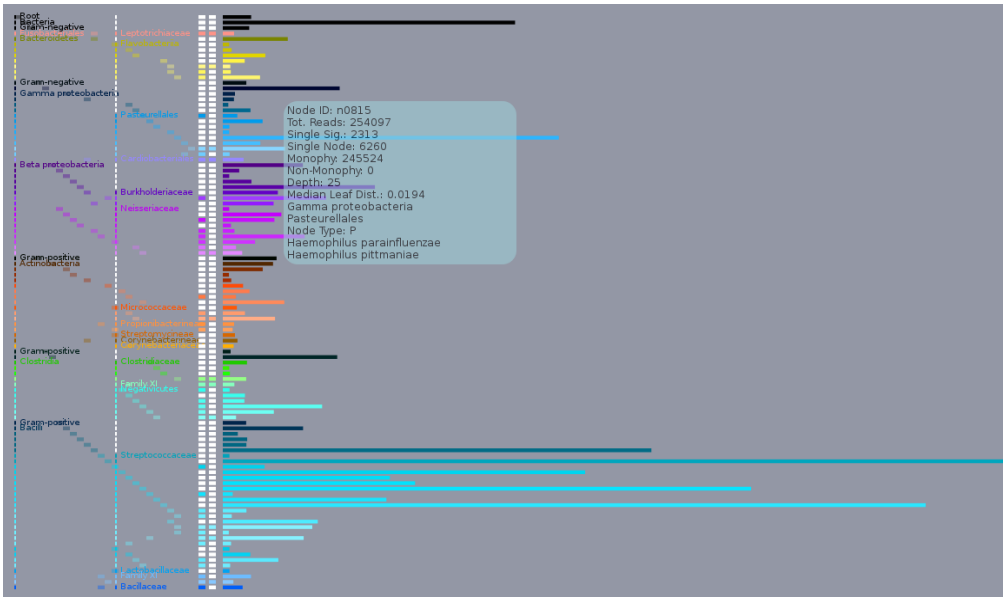
The next six lines break the total processing time (3.3 seconds) into input, translating, matching, assignment steps, as well as 'other'. From the output shown above, reading in the Life-32GB.0 data module took the majority of the clock time of 7 minutes 41 seconds, while the reads were processed at a rate of 4.7 Gbp / hour.

4.3 Graphical output: Sequinator

Sequedex produces an interactive html output file of the phylogenetic profile for each analyzed sample - a tool we call Sequinator. Simply open the html file produced in the appropriate sqdx output directory with a web browser. An output similar to the one below will appear. For an active version of this figure, click here.

Screenshot of the Sequinator demo, which graphically displays the node counts found in a Sequescan 'who' file from a dental caries study. A variety of known members of the oral microbiota are shown, one of which, *Haemophilus infleunzae* is highlighted on the screen. Since Sequedex assigns reads to the nodes of the tree, rather than the leaves, the output is slightly more complex than a simple table. The 100 horizontal bars on the right 3/4 of the plot have a length proportional to the top 100 most populated nodes on the tree. When you move the mouse over one of the bars, information about the node are displayed, as indicated on the screen shot for node 815. To the left of the bars is information on the phylum and family of the node, as well as the depth of the node from the root to the leaves of the tree. Detalied definitions of parameters in the box are explained in the **Who?** section, below. Approximately half of the nodes are immediately above one or two leaves, and these nodes are indicated by coloring in one or both of the columns immediately to the left of the population bars. These organisms are reasonable suggestions as to specific organisms that may be present in the sample.

The colors of the bars reflect phylogenetic similarity. Deeply branching nodes (between the root and a phylum-level node) are black, while the rest of the nodes are colored in a manner that changes smoothly along the phylogeny (and



is similar from sample to sample). This helps identify the distinct clusters of populated nodes that often appear in metagenomic samples.

4.4 Output files - Who?

The output file who-Life2550-32GB.tsv contains the phylogenetic profile derived from the fasta input file. It contains eight columns of data, containing node-numbers defined with respect to a reference phylogeny, labels for these nodes, and histograms of reads assigned to these nodes in several different ways. Several sections of this output file are shown below:

indx	id	total	single_sig		single_node		monophyletic		non_monophyl		depth	min_l
0	n0000	151167	72994	59747	0	18426	0	4	0.0000	1.7234	Root	Root
1	n0001	874316	205831	377350	178906	112229	1	5	0.0823	1.6723	Bacteria	
2	n0002	154920	32755	18795	62076	41294	2	4	0.1198	1.5292	Gram-negative	
3	n0003	30464	6292	3310	17987	2875	3	3	0.2698	1.4586	Gram-negative	
4	n0004	1875	345	62	1320	148	4	3	0.3288	1.3869	Gram-negative	
5	n0005	90	29	35	26	0	5	2	0.3822	1.3777	Gram-negative	
6	n0006	16	2	14	0	0	6	1	0.9312	0.7069	Elusimicrobia	
7	n0007	7353	874	950	5013	516	6	3	0.5002	1.2885	Gram-negative	
8	n0008	42387	6604	13294	19396	3093	7	2	1.3383	0.4828	Fusobacterial	
9	n0009	3307	231	334	2742	0	8	1	1.5307	0.3248	Fusobacterial	
10	n0010	21530	1622	5090	14062	756	9	2	1.5909	0.3032	Fusobacterial	
11	n0011	471	59	55	357	0	10	1	1.6541	0.0753	Fusobacterial	
12	n0012	833	249	76	508	0	11	1	1.7189	0.0159	Fusobacterial	

The first seven columns of the phylogenetic profile contain the phylogenetic profile of the metagenomic sequence file processed:

- indx: Index starting at 0
- id: node id. consistent with `phyloxml` reference tree.
- rollup_category: general category name. May include several phylum for nodes near the root
- total: Total number of reads attributed to the node

- single_sig: Number of reads containing only one signatures
- single_node: Number of reads containing multiple signatures, but all of them belonging to the same node
- monophyletic: Number of reads containing multiple signatures that are consistent along the phylogeny
- non_monophyl: Number of reads containing multiple signatures that are inconsistent with the phylogeny

The last thirteen columns contain a variety of details about the nodes numbered in the first two columns, and are idential for all samples processed. The columns are:

- depth - the number of nodes from the root of the tree
- min_leaf_depth - the shortest distance to a leaf, in nodes
- root_dist - the phylogenetic distance to the root
- median_leaf_dist - the median phylogenetic distance to the leaves
- level_0_name - top level name in the Sequedex hierarchy: frequently corresponding to the phylum or order
- level_1_name - mid level name in the Sequedex hierarchy: frequently corresponding to the family
- taxon_id - the NCBI taxnomony ID of the node
- node_type - **I** for internal nodes, **S** for semi-preterminal nodes (above one leaf), and **P** for preterminal node (above two leaves).
- leaf_0_name - if the node is immediately above a leaf, the name of the leaf
- leaf_1_name - if the node is immediately above two leaves, the name of the second leaf
- color - a color code, which can be used to provide a consistent color-code in plots
- width - an integer characterizing the breadth of phylogeny covered by a node
- path - a list of all of the nodes on the path from the node to each node.

Visualization of the phylogenetic profiles can be aided by placing the counts onto the `phylogenetic tree`.

4.5 Output files - What?

Functional profiles are computed using a reference database of functional categories. The default functional classification is derived from that used by the SEED project (http://www.theseed.org) database as described in detail in Ref. [1] The functional profiles count the number of open reading frames associated to each functional category. Because many signatures occur in multiple SEED categories and each reading frame gets a total of one count, categories often are assigned a fractional count.

Approximately 70% of the phylogenetic signatures can be associated with one or more SEED subsystems.

The output file what-Life2550-32GBxseed_0911.m1.tsv contains the functional profile derived from the fasta input file. The carbohydrate section of this file is shown here::

```
indx    fragments       id      level1  level2  subsystem
0       19.1    si_0000 Amino Acids and Derivatives     Alanine, serine, and glycine    Alanine_biosy
1       15.1    si_0001 Amino Acids and Derivatives     Alanine, serine, and glycine    Glycine_Biosy
2       59.2    si_0002 Amino Acids and Derivatives     Alanine, serine, and glycine    Glycine_and_S
3       31.5    si_0003 Amino Acids and Derivatives     Alanine, serine, and glycine    Glycine_cleav
4       16.6    si_0004 Amino Acids and Derivatives     Alanine, serine, and glycine    Serine_Biosyn
...
63      65.7    si_0063 Carbohydrates   Central carbohydrate metabolism Dehydrogenase_complexes
64      10.9    si_0064 Carbohydrates   Central carbohydrate metabolism Dihydroxyacetone_kinases
65      72.2    si_0065 Carbohydrates   Central carbohydrate metabolism Entner-Doudoroff_Pathway
```

```
66      18.5    si_0066 Carbohydrates   Central carbohydrate metabolism Ethylmalonyl-CoA_pathway_of_C
67      31.6    si_0067 Carbohydrates   Central carbohydrate metabolism Glycolate,_glyoxylate_interco
68      85.4    si_0068 Carbohydrates   Central carbohydrate metabolism Glycolysis_and_Gluconeogenesi
69      28.9    si_0069 Carbohydrates   Central carbohydrate metabolism Glycolysis_and_Gluconeogenesi
70      34.7    si_0070 Carbohydrates   Central carbohydrate metabolism Glyoxylate_bypass
71      18.3    si_0071 Carbohydrates   Central carbohydrate metabolism Methylglyoxal_Metabolism
72      0       si_0072 Carbohydrates   Central carbohydrate metabolism Particulate_methane_monooxyge
73      77.6    si_0073 Carbohydrates   Central carbohydrate metabolism Pentose_phosphate_pathway
74      2.4     si_0074 Carbohydrates   Central carbohydrate metabolism Peripheral_Glucose_Catabolism
75      28      si_0075 Carbohydrates   Central carbohydrate metabolism Pyruvate:ferredoxin_oxidoredu
76      22.3    si_0076 Carbohydrates   Central carbohydrate metabolism Pyruvate_Alanine_Serine_Inter
77      80.9    si_0077 Carbohydrates   Central carbohydrate metabolism Pyruvate_metabolism_I:_anaple
78      109.1   si_0078 Carbohydrates   Central carbohydrate metabolism Pyruvate_metabolism_II:_acety
79      1       si_0079 Carbohydrates   Central carbohydrate metabolism Soluble_methane_monooxygenase
80      121.4   si_0080 Carbohydrates   Central carbohydrate metabolism TCA_Cycle
81      19.3    si_0081 Carbohydrates   Di- and oligosaccharides        Beta-Glucoside_Metabolism
82      23.2    si_0082 Carbohydrates   Di- and oligosaccharides        Fructooligosaccharides(FOS)_a
83      35      si_0083 Carbohydrates   Di- and oligosaccharides        Lactose_and_Galactose_Uptake_
84      10.2    si_0084 Carbohydrates   Di- and oligosaccharides        Lactose_utilization
85      88.9    si_0085 Carbohydrates   Di- and oligosaccharides        Maltose_and_Maltodextrin_Util
86      1.8     si_0086 Carbohydrates   Di- and oligosaccharides        Melibiose_Utilization
87      5.9     si_0087 Carbohydrates   Di- and oligosaccharides        Sucrose_utilization
88      1.6     si_0088 Carbohydrates   Di- and oligosaccharides        Sucrose_utilization_Shewanell
89      72.8    si_0089 Carbohydrates   Di- and oligosaccharides        Trehalose_Biosynthesis
90      13.7    si_0090 Carbohydrates   Di- and oligosaccharides        Trehalose_Uptake_and_Utilizat
91      0.3     si_0091 Carbohydrates   Di- and oligosaccharides        Unknown_oligosaccharide_utili
92      35.6    si_0092 Carbohydrates   Fermentation    Acetoin,_butanediol_metabolism
93      9.1     si_0093 Carbohydrates   Fermentation    Acetone_Butanol_Ethanol_Synthesis
94      62.6    si_0094 Carbohydrates   Fermentation    Acetyl-CoA_fermentation_to_Butyrate
95      26.4    si_0095 Carbohydrates   Fermentation    Butanol_Biosynthesis
96      21      si_0096 Carbohydrates   Fermentation    Fermentations:_Lactate
97      24.5    si_0097 Carbohydrates   Fermentation    Fermentations:_Mixed_acid
...
```

The columns of this file are::

```
*   indx :  Index, starting at zero
*   fragments: Number of fragment assigned to that SEED category
*   id : SEED identification number
*   level1 : Level 1 SEED category descriptor
*   level2 : Level 2 SEED category descriptor
*   subsystem : Level 3 or subsystem SEED category descriptor
```

More information about particular SEED subsystems can be obtained at http://www.theseed.org/SubsystemStories, but the user should note that the index numbers above (eg. si_0094) are only defined within the Sequedex distribution.

4.6 Output files - Who does what?

The output file tabulates the number of reads in each phylogeny **and** functional category. It is a matrix with 402 columns and 911 rows, corresponding to the phylogeny vector defined by the reference phylogeny and the functional vector defined by the SEED subsystems. A portion of this matrix is shown below, centered around the cyanobacteria and the carbohydrate metabolism subsystems::

```
id          n000     n001   n002    n003    n004    n005    n006     ...
...
si_0069     13      0.6     0.8     0       0.1     0       0
```

```
si_0070     9.7     0.7     0       0       0       0       0
si_0071     4.9     0       0       0       0       0       0
si_0072     0       0       0       0       0       0       0
si_0073     28.6    1.6     0       0       0.3     0       0
si_0074     0.4     0.4     0       0       0       0       0
si_0075     13.1    1.5     0       1       0       0       1
si_0076     5.3     0       0       0       0       0       0
si_0077     27      1.7     0.3     0       0       0       0
si_0078     26.7    2.6     0       0       0       0       0
si_0079     0       0       0       0       0       0       0
si_0080     26.3    2.8     0       0       0       0       1
si_0081     5.4     0.5     0       0       0       0       0
si_0082     5.6     0.9     0.3     0       0       0       0
si_0083     11.1    3.2     0       0       0       0.3     0
si_0084     1.4     0       0.3     0       0       0       0
si_0085     27.6    9.3     0.3     0.5     0.5     0       0
si_0086     1       0.3     0       0       0       0       0
si_0087     1.5     0.1     0.3     0       0       0       0
si_0088     0.3     0       0       0       0       0       0
si_0089     20.5    1.1     1.3     0       0       0       0
si_0090     3.5     0.7     0       0       0       0       0
si_0091     0       0       0       0       0       0       0
si_0092     7.9     3.3     0       0.5     0       0       0
si_0093     2.5     0.6     0       0       0       0       0
si_0094     14.8    1.8     0       0       0       0       0
si_0095     6.6     0.4     0       0       0       0       0
si_0096     6.8     0.5     0       0       0       0       0
si_0097     4.2     0.8     0       0       0       0       0
si_0098     0.3     0       0       0       0       0       0
...
```

Generally, this will be most useful on the larger data sets, as the counts are spread across approximately 400,000 bins. Given the even increasing number of reads generated by modern sequencers, we expect to have available sufficiently many reads to provide statistical resolving power. As discussed in the phylogeny and the functional profile output files, the results must be put into context with the used databases.

4.7 Obtaining annotated reads: Analysis of *E. coli* proteome

One way to validate Sequedex is to provide an input file derived from a single organism and observed the phylogenetic profile produced by Sequedex. Since Sequedex requires the input file to be nucleotide sequence, one cannot at present input amino acid sequences. Six-frame translations are automatically performed on input files, which must be sequeces of A, T, G, and C. NCBI does, however, distribute files with nucleotide sequences of each of the genes in their completed and draft genomes at their ftp site. We provide here instructions for analyzing *E. coli,* but they are readily extensible to any of the other genomes available from the same location.

To see the phylogenetic profile Sequedex produces from the *E. coli* genome, download the ffn file from NCBI, at ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__DH10B_uid58979/NC_010473.ffn and run Sequescan on it, as described above, except select 'true' for the 'Write Database' pull-down menu. This will cause Sequescan to write an output fasta file with annotated versions of all of the reads it identifies.

Examination of the Life2550-40GB-stats.tsv file reveals:

- 4,128 reads were processed, with 95 % of them phylogenetically identified, and 3,163 fragments assigned a function.
- After the signature list was read into memory, 4.6 seconds were spent translating and 2 seconds were spent matching to the signature list.

- More than 99 % of the reads contained multiple signature peptides.

Examination of the file reveals:

- A total of 12 nodes of the tree were assigned reads, with 90 % of them assigened to the most specific node above *E. coli*.
- Examination of `the Life2550 tree` shows that nearly all of the remainder are assigned to nodes on a path from the root to *E. coli*.

Examination of what-Life2550-40GBxseed_0911.m1.tsv reveals:

- 810 seed categories contained a match to at least one gene in *E. coli*.
- More than thirty genes were identified in four SEED categories: Flagellum, DNA Repair, LSU of the Ribosome, and DNA Replication.

Examination of db-Life2550-40GBxseed_0911.m1.fa reveals:

- Genes 2,3, and 4 of *E. coli* were correctly identified as in SEED category *si_0046* ; Serine - threonine metabolism.
- Seed category *si_0746*, identified RNA polymerase alpha, beta, beta prime, and omega subunits, as well as two hypothetical proteins and penicillin-binding protein 1b.

Once the user is familiar with this process, it is easy to substitute any of the 2300 other completed bacterial genomes available at this ftp site. For example, you can download a tar file with nucleotide sequences coding for genes in each of 2300 completed bacterial and archael genomes at ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.ffn.tar.gz. Running various individual proteomes through Sequedex will give the user some indication of the precision and sensitivity with which genes can be classified. Note that this is an excellent situation to use the 'list file' under 'input type' as the ffn files will be located in directories corresponding to the organism. Sequedex will separate the output files in a similar directory structure under the directory specified under 'Output (Top-Level)'

4.8 Analysis of multiple files in parallel with the GUI

Sequedex can process multiple fasta files in parallel, as a single batch. Besides eliminating the need to individually specify each fasta file with the GUI, batch processing of fasta files results in a significant speedup of computation time, because the signature list being read only once, and enables parallel processing of the fasta files with individual Java threads.

A set of sixteen synthetic metagenomic datasets were used in Berendzen et al. [1] to examine the read-length of the sensitivity and specificity of Sequedex on four separate clonal populations of bacteria. These files are provided with the Sequedex distribution in the testData directory under the 'synthetic' subdirectory. To process these files together, we can select the 'directory' option from the 'Input type' pull-down menu, locate the directory these fasta files reside in the 'Input' dialog box, and enter '4' or more in the box by the 'Thread Number' option. Upon clicking on 'Run Sequedex', the files will be processed in a batch, with the results appearing in separate subdirectories under 'data/output'.

Alternatively, the files to be analyzed can be listed, with their complete paths, in an input file which is selected in conjunction with the 'list file' option under 'input type' dialog box. On a Linux or Mac platform, it is possible to generate a complete path by including the full path in the 'ls' command as follows (which must conform to the path relevant to your own system):

```
ls -1 /home/username/sequedex/testData/synthetic/*.fna
```

4.9 The virus1252 data module

Sequedex also has a one-per-species viral tree of 1252 taxa and associated data module, available at http://sequedex.lanl.gov. To use this data module, simply download the data module and place it into the sequedex/data directory, then either select 'virus1252' from the pull-down menue (if using the Sequedex GUI) or specify it after the -d flag (if running Sequedex on the command line - see below).

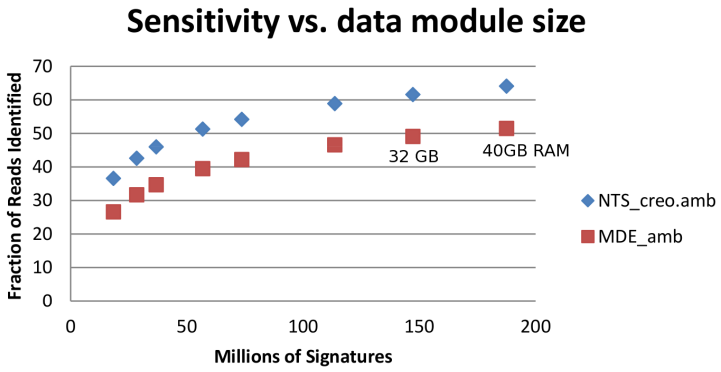
For example, to analyze this `virus.test.fasta` data file and run the command:

sequescan run -d virus1252.1 -s pfam27 -f 1 virus.test.fasta

to generate Sequedex output, which by default will appear in the subdirectory virus.test.fasta.sqdx. Note that the virus1252 data module uses functional classifications from pfam release 27, which contains families for many of the viral proteins. This provides greatly increased confidence in interpreting output in samples with a low abundance of reads.

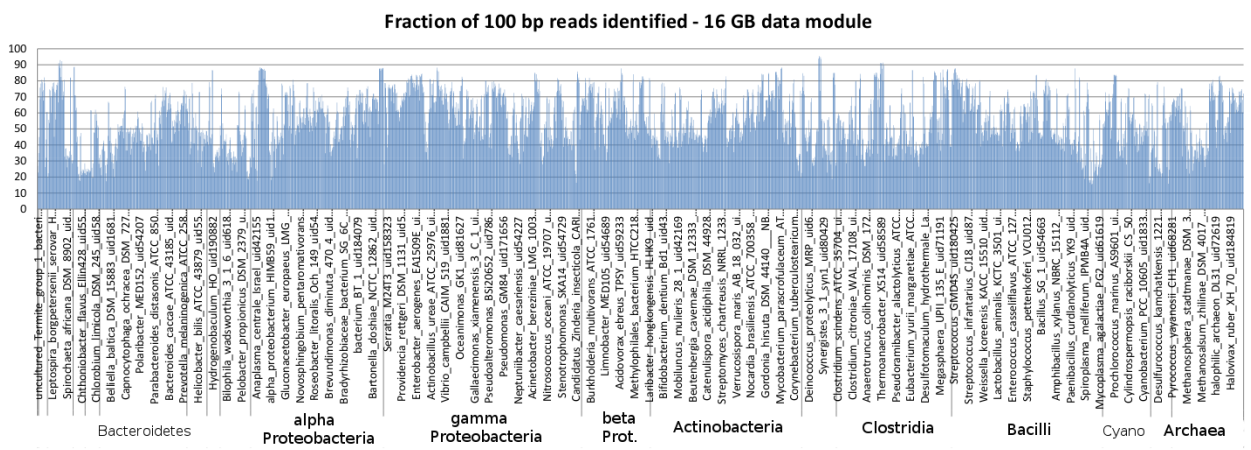
4.10 Alternative data modules and memory usage

Sequedex is designed to match nucleotide sequences against a list of signature peptides that each posses both a phylogenetic and functional annotation. The original manuscript on soil metagenomics used a comparatively small set of 20 million signatures derived from 403 bacterial reference genomes (one-per-genus) and functions defined by the SEED project, and this is the default data module described above. Sequedex version 1.0, released in the Summer of 2013, expanded this data set to one-per-species representation across all of the kingdoms of life, including 2550 taxa. The signature list was phylogentically thinned at the leaves to provide a data module of ~180 million signatures, fitting in under 40 GB of RAM. In addition to the 40 GB data module, additional data modules have been created and are distributed at the Sequedex website (http://sequedex.lanl.gov), designed to fit easily in 32,16, 8, and 4 GB of RAM. The sensitivity is shown for different data modules and across the phylogeny of reference genomes in teh next two figures.



The fraction of reads indentified in two soil metagenomics samples sequenced on a 454 sequencer for eight module sizes. Five of these modules are distributed at the Sequedex website. The modules all utilize the same trees and functional database, and are thinned by randomly discarding signatures from the 40 GB module.

The phylogenetic dependence of Sequedex's sensitivity across the bacterial and archael reference genomes, which were chopped into 100 bp reads and analyzed with Sequedex. The genome names are shown in the same order as they appear in the Sequedex reference tree. Note that most of the genomes are recognized with a sensitivity between 35% and 70%, providing a relatively small dependencey on whether or not the organism comes from a well-represented portion of the phylogenetic tree. Since signatures are required to be in two or more reference genomes, the sensitivity of Sequedex in detecting novel species and genera is not much less than its sensitivity against these reference genomes (in contrast to nucleotide-based detection methods and methods that do not use phylogeny as an importance filter).



We recommend that the user familiarize himself or herself with the software and output using a system with 4 or 8 GB of RAM and using the Life2550-4GB or Life2550-8GB data module distributed with Sequedex and set as default with the GUI. Additionally, smaller versions of tol-all are available. For many purposes, the lower sensitivity may not matter too much, but it is likely that the user will want to spend the $1000 necessary to acquire a desktop machine with 32 GB of RAM.

Sequedex enables the memory request of Java to be set either in configuration files or with an environment variable. See *Reference* for details.

By setting the memory size of 30 GB somewhat smaller than the system size (in this case, 32 GB), it is possible to use other applications while Sequedex is running, without significantly affecting the performance of Sequedex. The user may need to experiment to find optimal parameters for his or her individual needs.

4.11 Running Sequedex from the command line

Many users will find it advantageous to run Sequedex from the command line, which is availble to all users of Mac and Linux systems, and can be obtained for Windows by installing the Cygwin package http://www.cygwin.com.

However you acquire your command line, it is most convenient to change directories to where your data resides, the execute the particular Sequedex run. Help with the basic syntax can be obtained by typing:

```
sequescan run -h
```

which provides the following output:

```
Command line execution of sequescan run mode:
sequescan run [-h] [-q] [-c config_file] -d data_module [-o output_directory] [-s function_set] [-a r

Example:
sequescan run -d Life2550-4GB.0 -s seed_0911.m1 -f 1 /Users/jsmith/mgData

Option descriptions:
-h    mode help
-q    quiet option - less messages to console or progress window
-c    user-defined configuration file (overrides system configuation file)
-d    name of data module
-o    user-defined directory for data output (default is directory where input is located)
-s    name of function set
-a    minimum protein fragment length (overrides configuration file; default is 15)
-t    maximum number of threads in threadpool (default = 1)
```

```
-f    database writer flag (arguments:  0 = no, 1 = yes); analysis_writer_list in config determines t
-l    required if INFILE contains list of fasta/fastq files; if argument is none, list contains absol
     otherwise argument is base directory and paths in list are relative to base directory;
     when paths are relative to base directory and the -o option is set, output will inlcude relative

INFILE may be a fasta/fastq file, a directory with fasta/fastq files,
or a file containing a list of fasta/fastq files.
However, only fasta or fastq files or their gzipped (.gz) versions with an extension
in the config file parameter fa_ext_list will be processed.
```

Thus, running Sequescan from the directory with the fasta files with the command line:

```
sequescan run -v -d Life2550-8GB.0 -s seed_0911.m1 -f 1 -o out_dir infile.fas
```

will run Sequescan with the Life2550-8GB.0 data module and seed_0911.m1 functional classification of the signatures distributed with this release. Sequescan will match all of the reads in *infile.fas* and create the directory out_dir in which it will place the results (in a subdirectory), including the writing of a fasta file with each read containing a signature peptide, in the reading frame which results in that signature, annotated by phylogenetic and functional assignment.

Alternatively, the command line:

```
sequescan run -v -d Life2550-8GB.0 -s seed_0911 -f 0 -t 20 -o out_dir . >& err &
```

will run Sequescan as before, except processing all of the files with an appropriate file extension (by default, fasta, fst, fna, fas, and ffn, specified in sequedex/etc/sequescan/sequescan.conf) as a group, using 20 Java threads, without producing an annotated fasta file and again with the output appearing in subdirectories of out_dir. By placing '>& err &' at the end of the command, the output normally appearing on the screen will be sent to the file, 'err', and the process will run in the background, allowing the terminal window to be used for other commands while Sequescan is running. The output files can be written to a different directory tree if one or both directories are supplied with appropriate paths. This enables users to leave shared data directories untouched by output files produced by Sequedex without the need for recopying potentially large input files.

This remaining of this section describes the generated results from sequescan, which includes a phylogenetic profile, a functional profile, a matrix of function x phylogeny that informs on "who does what", and a fasta or fastq file of annotated reads. Visualization and interpretation of these profiles are discussed in *Initial analysis with Sequedex: Phylogenetic and functional profiles*, *Annotated reads*, and *Comparing Samples*.

4.12 The Sequedex configuration utility and sequescan.conf

Seuedex can be configured through environment variables accessible through a configuration file, described in *Environmental variables* and a configuration file, described in *Configuration options*.

The environmental variables, most easily accessed with the utility sequedex-config, enable alternate locations for files such as data modules, trees, and configuration files, and programs, such as python, java, or the web browser used by Sequenator. Sequedex detects reasonable defaults for these variables when sequedex-bootstrap is run upon initial installation, and sequedex-config provides a short description for how each variable was set.

- The amount of memory can be changed. If the memory is set to a value larger than

the available RAM on your computer, the computer may need to swap, described here, which, if it occurs to any significant amount, will severely degrade the rate at which Sequescan runs. If the memory is set to a small value, Sequedex, through Java, will attempt to manage the signature lists in a prohibitively small hash map with techniques such as those described here, again leading to severe degradation of performance.

The configuration file, sequedex/etc/sequescan/sequescan.conf, contains a variety of settings controlling default behavior of inputs and outputs of Sequescan. For example:

- Sequedex can read fasta, fastq, and gzipped fasta and fastq files. The file-type is identified by the file extensions, which can be changed, and are listed in the configuration file.
- The location and names of output files can be changed with variables in the configuration file. Some discussion of the logic associated with this option is available in *Directory structures for data, output files, and analysis*.
- The default minimum length of open reading frames to be processed can be defined both on the command line and in the configuration file. If the input sequence file has only 36 base pairs per read, the default setting of 15 amino acids will prevent any reads from being identified.

4.13 Directory structures for data, output files, and analysis

The choice of directory structures and file names for accessing, archiving, and analyzing large sequencing files can be challenging, and the user likely has a system in place. Because multi-threaded analysis of multiple files in one run can occur in parallel and with a single memory map is so efficient, Sequedex can be run by giving it the name of an input directory containing multiple sequence files and a top-level output directory (which can be the same as the input directory). Sequedex will then create separate output directories for each sequence file, naming it by simply appending '.sqdx' to the input filename. Analysis of the same data file with different data modules will go in the same output directory, but with the filenames distinguishing the data module used.

Because of the large file sizes, the user is encouraged to be aware of which disks are local and which are remote, where reading and writing to disk may be slow. Also, Sequedex is capable of uncompressing gzipped fasta or fastq files directly.

One logical way to organize further analysis is to keep files associated with each sample in that sample's directory, while cross-sample analysis can be one directory above that. Symbolic links can be used to create shorter and more meaningful names to sample directories, while preserving the original name, which is often necessary to backtrack to the original sample identity.

Many possibilities exist for organizing data, but the user is encouraged to put some deliberate thought into the problem before creating the large number of potential output files that can be created by the analysis enabled by Sequedex.

CHAPTER 5

Initial analysis with Sequedex: Phylogenetic and functional profiles

As described in *How is Sequedex used with other software?*, Sequedex can be applied to a wide variety of problems. While it is possible that a cursory examination of Sequedex output will be sufficient to address the user's question to the appropriate level of confidence, we expect most Sequedex users will need to combine their Sequedex analysis with other techniques and software packages. We provide examples of a variety of these analyses, including:

- the direct examination of phylogenetic or functional profiles with :program: *Excel*, described in this chapter.
- the visualization and characterization of these profiles, described in *Using Sequestat*
- the analysis of the individual reads retrieved, described in *Annotated reads*, and
- the multi-sample comparisons, described in *Comparing Samples*.

In these chapters, we will draw upon software tools and techniques that the user may not be familiar with or may not have installed on their system. To help these users, we provide *Additional software tools and resources for use in conjunction with Sequedex*.

It is assumed at this point that the user is comfortable running Sequescan and can obtain multiple output files for comparison. Please refer to the previous two chapters for installation, platform-dependent instructions on running Sequescan with both the GUI and the command line, and instructions on obtaining a license.

The ability of Sequedex to profile bacterial communities with metagenomic data leads naturally to the question of how a user's microbial community compares to previously sequenced microbial communities. In this chapter, we lead the user through the process of computing phylogenetic profiles and comparing them to publically available datasets. In additional to phylogenetic information, the sequences obtained in a shotgun metagenomics dataset can be used to obtain functional profiles of genes in a microbial community, identify the presence of particular genes of interest, or even determine which proteins are the most important determinants of classification within a set of microbial communities. In this chapter, we we lead the user through the process of computing functional profiles.

A note of caution is warranted here. In this and the next section we present phylogenetic and functional profiles, as well as analysis of these profiles across an extraordinary range of microbial communities with samples prepared by different researchers with a variety of purposes, sequenced by different methods with different read-lengths. We have primarily relied on publically available datasets with associated publications, although a few examples of unpublished work with experimental collaborators is also included. Our purpose in this user's manual is to provide illustrative examples of how phylogenetic and functional profiles can be visualized and analyzed for both self-consistency and biological insights, and provide analysis of a reference collection of data to which the user can compare his or her own data. The job of cross-checking these insights with further experimentation analysis is distinct and belongs in the peer-reviewed literature. Our intent here is to put this analysis tool into the hands of the people best-suited to perform these cross-checks - those who decide which samples will be prepared and know what biological phenomonena are expected.

5.1 Acquiring data files

Sequedex at present requires input sequence data files be in fasta, fastq, or gzipped fasta or fastq data format. Sequencing platforms typically output sequences in a format with quality scores, such as fastq, which can be used as input to assemblers. Because possibility space of 10-mers of amino acids is so much larger than the number of signatures, Sequedex is relatively insensitive to sequencing errors and the quality scores are ignored, and the fasta file format is suitable. If the user's data is in another format, the it must be changed into fasta or fastq format. While preprocessing of reads to remove duplicates is valuable, and minimal quality score filtering is probably useful, assembling the data into contigs before profiling the community is probably unhelpful, as it makes the profiles dependent on the depth of sequencing.

For users wishing to compare their data to others, metagenomics data can be acquired from several sources, including the Sequence Read Archive at NCBI, where the production phase of the human microbiome project can be found with the project number, SRP002163 and the study number SRP002163. Individual data sets can be downloaded from their ftp site with a web browser (or using wget) at ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR059/SRR059366/SRR059366.sra. Metadata about data sets, such as it is, can be obtained with the Run Browser at the Sequence Read Archive, http://www.ncbi.nlm.nih.gov/Traces/sra/?view=run_browser/. The human microbiome data is probably easier to obtain from the Human Microbiome Project website: http://www.hmpdacc.org.

Environmental metagenomics data sets can be found at NCBI with the taxonid 410658. A wide variety of individual projects can be found, including samples from Pru Toh Daeng Peat Swamp, in Southern Thailand: SRR023820, which can be downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP001/SRP001114/SRR023820/SRR023820.sra.

In order to process data sets from the Sequence Read Archive, it is necessary to extract fasta or fastq files from the .sra files obtained above. This can be done with the fastq-dump utility from the SRA Toolkit, available for Linux, Mac, and Windows platforms at http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software. Generation of a fastq or [fasta] file from the HMB dataset above requires the command:

```
fastq-dump [--fasta] SRR059366.sra
```

The Department of Energy's Joint Genome institute also maintains a repository of metagenomic sequence data at their integrated metagenomics (IM) site.

The CREO data from reference [1] can be downloaded in fasta format at http://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=MetaDetail&downloadTaxonReadsFnaFile=1&taxon_oid=2029527002&noHeader=1, while additional metagenomics datasets are listed at http://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonList&page=taxonListAlpha&domain=*Microbiome.

The Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis site (CAMERA) also maintains a repository of metagenomics sequence data at https://portal.camera.calit2.net/gridsphere/gridsphere, where we were able to download data from the Global Ocean Survey **[#f1]_**.

The human microbiome project has generated several terabytes of data from various locations on the bodies of a cohort of health people, which is described in detail at http://http://www.hmpdacc.org/ and shotgun metagenomics sets can be downloaded at their ftp site by, for example, typing:

```
wget ftp://public-ftp.hmpdacc.org/Illumina/anterior_nares/*.tar.bz2
```

Since these data files are compressed, tarred files, it is necessary to untar and unzip the files (:command: *tar -jxf SRS011105.tar.bz2*) and merge the paired-end files (:command: *cat SRS011105/*.fastq > SRS011105.fastq*) and recompress if desired (:command: *gzip SRS011105.fastq*) before analyzing with Sequedex.

5.2 Phylogenetic rollups

For users wanting to look at output data as rapidly as possible, we supply three Excel spreadsheets with the collected results for `reference genomes chopped into 100 bp reads`, the `human microbiome data`, `environmental metagenomes`. The reference genomes were obtained from the completed and draft genomes at the NCBI ftp site, and are the same genomes used in the Life2550 reference tree.

When comparing phylogenetic profiles of metagenomes, it is helpful to produce a matrix of phylogenetic profiles, with the 2550 rows corresponding to the nodes on the bacterial phylogeny, and each sample assigned its own column, as well as a matrix of functional profiles, with the 963 rows corresponding to the SEED (http://www.theseed.org) functional categories. This can be done in many ways, including reading each output file into Excel and pasting the relevant columns into one sheet, but a simple shell-script will also do the job. If the user is comparing more than a handful of files, he or she will likely want to write a program to both combine the data files together and associate simplified labels with the longer, unique, labels assigned to ensure proper sample tracking. A simple script that does this is:

```
cd hmb_data/output
for i in SRR059330 SRR059331 SRR059338 SRR059339; do
awk '{print $2}' < $i.fq.sqdx/who-Life2550-16GB.tsv > $i.j;
awk '{print $2}' < $i.fq.sqdx/what-Life2550-16GBxseed_0911.m1.tsv > $i.k;
done
paste SRR059330.j SRR059331.j SRR059338.j SRR059339.j > ../hmb.stool.who
paste SRR059330.k SRR059331.k SRR059338.k SRR059339.k > ../hmb.stool.what
echo "stool1 stool2 stool3 stool4" > ../stool.lbl
```

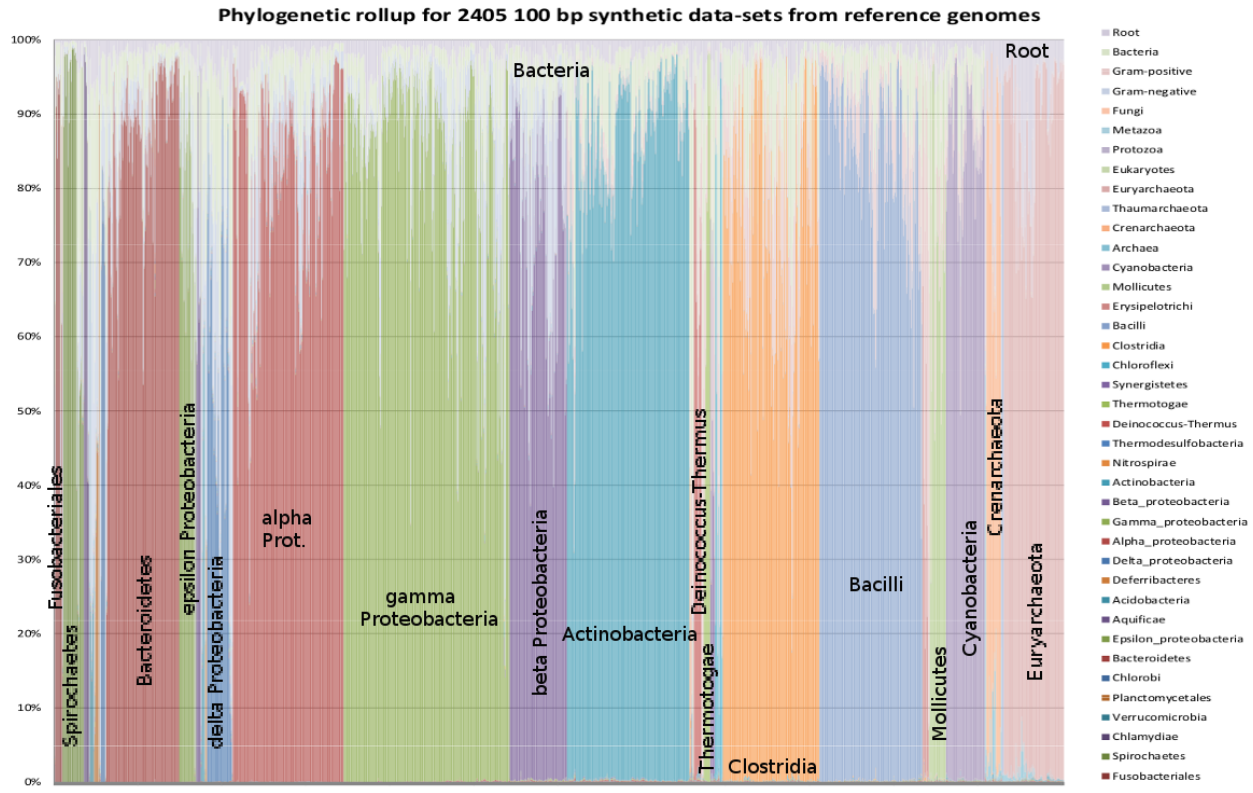
This will combine the output files from four stool metagenomics data sets into a matrix of counts for the who and what files. The user will, of course, need to supply directory and file names corresponding to his particular analysis. A spreadsheet may help to track the relationships between the sample names and labels. Note also that the paste command by default places tab characters between columns, while the echo command in the above script is placing spaces between labels. For the 100 bp synthetic data made from each of the 2405 bacterial and archeal genomes, with the results sorted phylogenetically and rolled up to the phylum level, one obtains:

When combined with the phylogenetic and data module size dependence of the fraction of reads recoginzed, shown in *Alternative data modules and memory usage*, the user can see that well-studied portions of the phylogeny, typically > 80 % of the reads are classified, with > 90 % having reasonably specific assignment. The tax in these figures are in the same order as in *Tree of Life, 2550 taxa*. The tools to explore phylogenetic assignments will be explored in much greater detail in the next chapters.

From these raw counts files, several types of profiles can easily be computed: phylogenetic rollups, normalized phylogenetic profiles, and normalized scalar products of the profiles. Once again, these quantities can be computed in many ways, but the gfortran (http:/gcc.gnu.org/wiki/GFortran/) `code` will suffice, and is described at the end of *The command line with standard tools (Linux, Mac, Cygwin for Windows)*. In particular, examination of the `tree` with archyoptrix (http://www.phylosoft.org) or an `output file` reveals forty roll-up categories, defined in the notation of Fortran 90 `here`. Only eight of the most highly represented categories are shown in the human microbiome plot, while the environmental metagenomes rollup, below, shows seventeen categories, with only 'multiphyla' and 'root' left off.

A description of the project aims and primary analysis results can be found in "A framework for human microbiome research" by the Human Microbiome Consortium, in http://www.nature.com/nature/journal/v486/n7402/full/nature11209.html.

The phylogenetic and functional profiles presented here can be compared to Figure 2 of "Structure, function, and diversity of the health human microbiome", also by the Human Microbiome Consortium, in http://www.nature.com/nature/journal/v486/n7402/full/nature11234.html. Although the Human Microbiome Consortium ran their datasets through BLASTX (see SRS016585 at http://www.hmpdacc.org/HMSCP/#data to verify that stool sample run SRR059346 is indeed primarily *E. coli*), the analysis in the HMB paper utilized a rapid-matching



scheme, described in Segata, et al., "Metagenomic microbial community profiling using unique clade-specific marker genes" Nature Methods 9:811-814 (2012) http://www.nature.com/nmeth/journal/v9/n8/full/nmeth.2066.html.

Although Segata, et al. used clade-specific genes to profile metagenomics data, while we used phylogenetic signatures of the majority of genes, numerous points of agreement between the analyses are evident. Note that we included a 'multiphyla' category in our figure, which was not included in their reported results. Specifically:

- For the stool samples, the ratio of bacteroidetes to clostridia varies from 60% clostridia to 95% bacteroidetes.
- For the tongue dosrum samples, firmicutes range from half to 10% of the identified species, with proteobacteria next most abundant, followed by bacteroidetes.
- For the buccal mucosa samples, firmicutes are even more prevalent, together with proteobacteria making up more than 90% of the assigned reads.
- For the supragingival plaque samples, roughly equal representation of firmicutes, actinobacteria, proteobacteria, and bacteroidetes occurs, with considerable variability in the relative abundance.
- For the anterior nares samples, actinobacteria makes up more than 90% of the assigned reads in some samples, with the firmicutes providing the bulk of the balance.
- For the posterior fornix samples, most of the samples consist of firmicutes (clostridia + bacilli), although three of the analyzed by Sequedex consisted of equal parts bacteroidetes and actinobacteria, as did one in Figure 2 of the HMB paper.
- For the retroauricular crease samples, most of the samples are almost entirely actinobacteria, while alpha proteobacteria make up a significant minority of several of the samples.

We will return to further analysis of the HMB dataset after examining a set of representative environmental microbiomes.

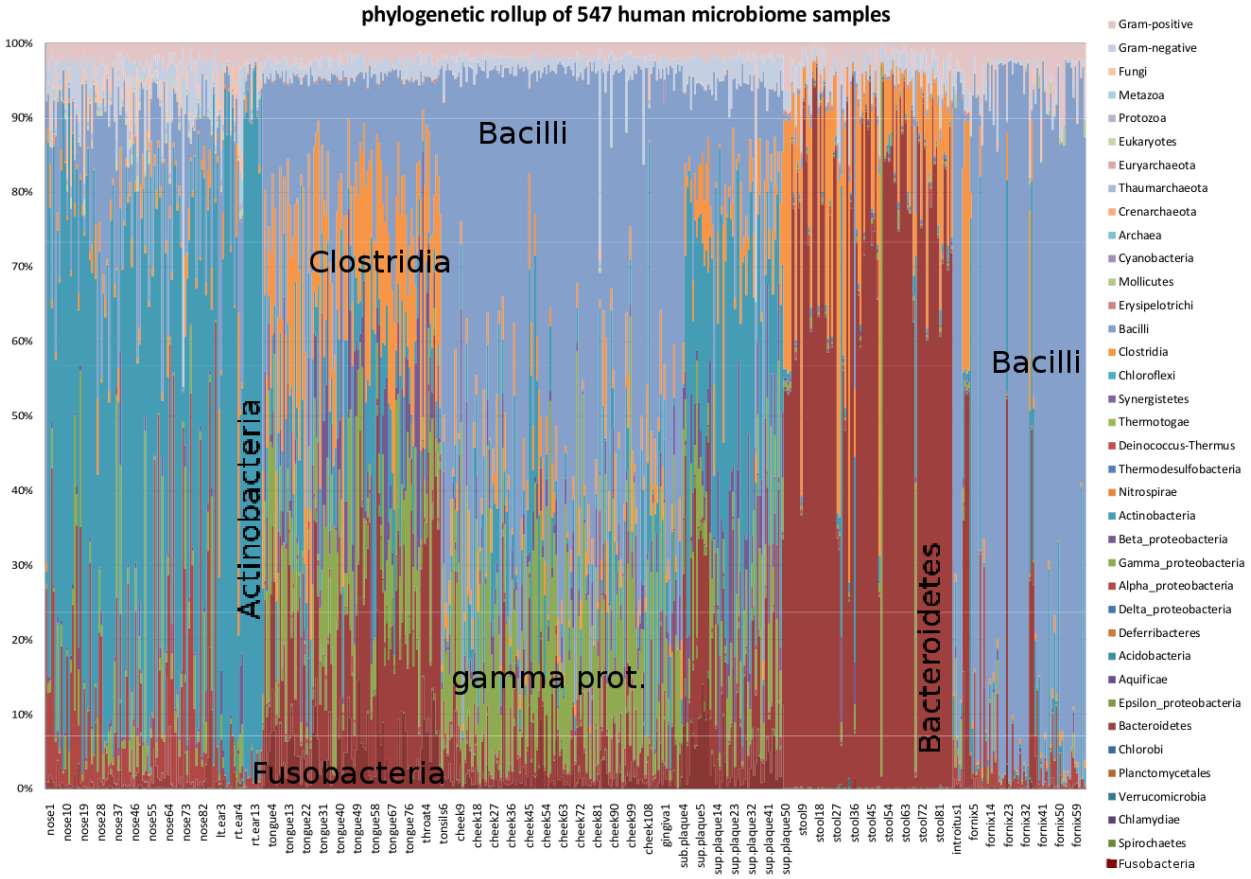


Fig. 5.1: Phylogenetic profile of 547 human microbiome samples, computed using the Life2550-40GB data module and Sequedex, rolled up with the above-defined classifications. The original plot and data can be found in the first panel of `this Excel spreadsheet`. For the most parts, the samples consist of pairs of replicates, and they are nearly indistinguishable from each other in this presentation.

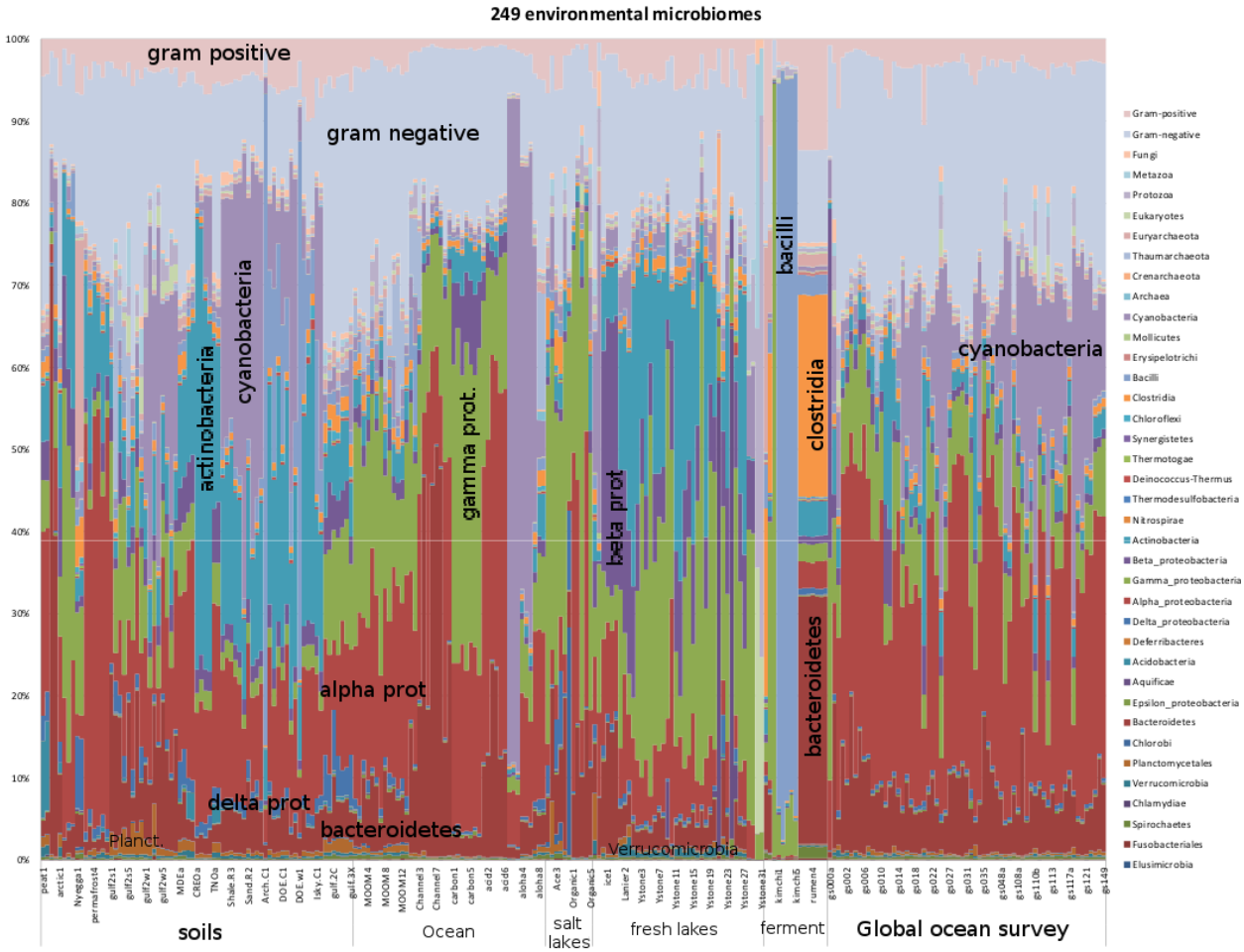


Fig. 5.2: Phylogenetic profile of 249 environmental microbiome samples, computed using the Life2550-40GB data module and Sequedex, rolled up with the above-defined classifications. The original plot and data can be found in the first panel of `this Excel spreadsheet`. The data are a compilation of 27 distinct studies, with citations provided below, in order from left to right on the figure. The data come primarily from the sequence read archive (indicated by the SRR number above the relevent columns in the spreadsheet. The Global Ocean Survey results can be obtained from CAMERA (https://portal.camera.calit2.net/) and the FACE data from JGI/IMG (http://img.jgi.doe.gov/), as well as several data sets kindly provided before publication and release by Cheryl Kuske at Los Alamos National Laboratory.

A cursory examination of the environmental microbiome profiles reveals distinctive and repeatable differences between studies. We present here references to the publications from each sample, and a brief summary of salient details.

peat Kanokratana, et al., "Insights into the phylogeny and metabolic potential of a primary tropical peat swamp forest microbial community by metagenomic analysis" *Microb. Ecol.* **61**:518-528 (2011). (http://www.ncbi.nlm.nih.gov/pubmed/21057783)

Permafrost1 Yergeau E, Hogues H, Whyte LG, Greer CW. "The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses." ISME J. 2010 Sep;4(9):1206-14. (http://www.ncbi.nlm.nih.gov/pubmed/20393573)

arctic Yergeau E, Sanschagrin S, Beaumier D, Greer CW, "Metagenomic Analysis of the Bioremediation of Diesel-Contaminated Canadian High Arctic Soils. PLoS ONE **7**:e30058 (2012) (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0030058)

Nyegga Stokke R, Roalkvam I, Lanzen A, Haflidason H, Steen IH., "Integrated metagenomic and metaproteomic analyses of an ANME-1-dominated community in marine cold seep sediments" Environ Microbiol. 2012 May;14(5):1333-1346 (2012) (http://www.ncbi.nlm.nih.gov/pubmed/22404914). Anaerobic methanotrophic archaea (ANME) and sulfur metabolising bacteria presumably include the delta proteobacteria visible in this sample.

Harvard FJ Stewart, AK Sharma, JA Bryant, JM Eppley, and EF DeLong "Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities" Genome Biology **12**:R26 (2011). (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129676/)

permafrost2 Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK. "Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw." Nature. 2011 Nov 6;480(7377):368-71. (http://www.ncbi.nlm.nih.gov/pubmed/22056985)

gulf2 Widger, et al., "Longitudinal Metagenomic Analysis of the Water and Soil from Gulf of Mexico Beaches Affected by the Deep Water Horizon Oil Spill" Preceedings of Nature hdl:10101/npre.2011.5733.1 (http://precedings.nature.com/documents/5733/version/1). The first eight samples are from sand, the second eight are from water.

face Berendzen, et al, "Rapid phylogenetic and functional classification of short genomic fragments with signature peptides" BMC Research Notes **5**:460 (2012). (http://www.biomedcentral.com/1756-0500/5/460/abstract). For a description of the FACE sites, see http://public.ornl.gov/face/results.shtml.

utah Kuske, CR, et al., Soil crusts in Utah field site. manuscript in preparation.

MOOM Canfield, DE, et al., "A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast" Science **330**:1375 (2010).

E. channel Gilbert, JA, et al., "Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel" Standards in Genomic Sciences **3**:183-193 (2010). Only functional analysis is presented, with SEED and MG-RAST.

carbon Mou, X., S. Sun, RA Edwards, RE Hodson, MA Moran "Bacterial carbon processing by generalist species in the coastal ocean" Nature **451**:708-711 (2008). (http://www.nature.com/nature/journal/v451/n7179/abs/nature06513.html)

acid Gilbert JA, Thomas S, Cooley NA, Kulakova A, Field D, Booth T, McGrath JW, Quinn JP, Joint I., "Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters." Environ Microbiol. **11**:111-125 (2009) (http://www.ncbi.nlm.nih.gov/pubmed/18783384)

aloha Martinez A, Tyson GW, Delong EF., "Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses." Environ Microbiol. **12**:222-238 (2010) (http://www.ncbi.nlm.nih.gov/pubmed/19788654)

Ace Lake Lauro, FM, et al., "An integrative study of a meromictic lake exosystem in Antarctica" ISME J. **5**:879-895 (2011) (http://www.nature.com/ismej/journal/v5/n5/abs/ismej2010185a.html)

Organic Lake Yau, S., et al. "Virophage control of antarctic algal host–virus dynamics" PNAS, USA **108**:6163-6168 (2011). (http://www.pnas.org/content/early/2011/03/24/1018221108.full.pdf+html) Primarily about viruses.

stromatolites Desnues C., et al., "Biodiversity and biogeography of phages in modern stromatolites and thrombolites" Nature **452**:340-345 (2008) (http://www.bio.sdsu.edu/faculty/kelley/31.pdf)

fanning Dinsdale, EA., et al., "Microbial ecology of four coral atolls in the Northern Line Islands" PLoS One. **3**:e1584 (2008). (http://www.plosone.org/article/info:doi/10.1371/journal.pone.0001584)

wastewater Sanapareddy, N., et al., "Molecular Diversity of a North Carolina Wastewater Treatment Plant as Revealed by Pyrosequencing" Appl Environ Microbiol. **75**:1688–1696. (2009). (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655459/)

ice Simon, C., A. Wiezer, AW Strittmatter, R. Daniel, "Phylogenetic Diversity and Metabolic Potential Revealed in a Glacier Ice Metagenome" App. & Env. Microb. **75**:7519-7526 (2009). (http://aem.asm.org/content/75/23/7519.full). In agreement that the major phyla are Betaproteobacteria, Bacteroidetes, and Actinobacteria.

Lake Lanier Oh, S., et al., "Metagenomic insights into the evolution, function and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem" App. & Env. Microbiol. doi: 10.1128 (2011). (http://aem.asm.org/content/early/2011/07/15/AEM.00107-11.short).

Yellowstone Inskeep, WP., "Metagenomes from High-Temperature Chemotrophic Systems Reveal Geochemical Controls on Microbial Community Structure and Function" PLoS ONE **5**:e9773 (2010). (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0009773). Not sure if this is the correct study.

biogas Jaenicke, S., et al. "Comparative and Joint Analysis of Two Metagenomic Datasets from a Biogas Fermenter Obtained by 454-Pyrosequencing" PLoS one **6**:e4519 (2011) (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0014519)

Leaf-cutter ants Aylward FO, et al. "Metagenomic and metaproteomic insights into bacterial communities in leaf-cutter ant fungus gardens" ISME J. **6**:1688-701 (2012). (http://www.ncbi.nlm.nih.gov/pubmed/22378535)

kimchi Jung JY, Lee SH, Kim JM, Park MS, Bae JW, Hahn Y, Madsen EL, Jeon CO, "Metagenomic analysis of kimchi, a traditional Korean fermented food" Appl Environ Microbiol. **77**:2264-74 (2011). (http://www.ncbi.nlm.nih.gov/pubmed/21317261)

rumen Hess, M., et al., "Metagenomic discovery of biomass-degrading genes and genomes from cow rumen" Science. **331**:463-7 (2011). (http://www.sciencemag.org/content/331/6016/463.short)

Global Ocean Survey Rusch, DB, et al. "The *Sorcerer II* global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific" PLoS Biology **5**:e77(2007) (http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0050077).

Another use for sequedex is to identify bacterial symbionts in eukaryotic samples, such as are readily visible in several of the transcripts from the Marine Microbial Eukaryote Transcriptome Sequencing Project.

5.3 Functional rollups

It is possible to produce functional profiles of the same samples used above for comparing phylogenetic profiles. Since we are using the SEED functional classification scheme, it is also possible to 'roll-up' the functional profiles into 28 categories of gene function. Inspection of the output file SRR059330.fasta.fun reveals the following roll-up categories, defined in the notation of Fortran 90 `here`.

The most striking observation about the functional profile across the 547 human microbiome samples is the consistent relative amplitude of the different functions from the different body sites. The two sample types with the greatest variation in functional rollup profiles are the posterior fornix and the anterior nares, consistent with Figure 2 of the Nature article referenced above (Nature **486**:207-214 (2012)). It is possible that this variability is a consequence of clonal populations of particular organisms that make up a significant population of some of the samples skewing the

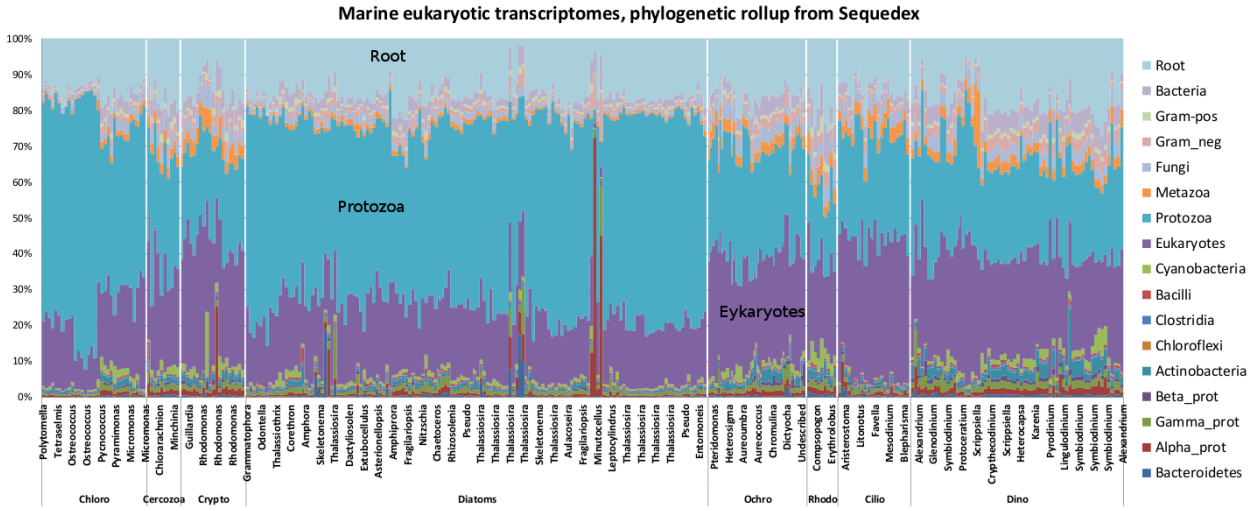


Fig. 5.3: Phylogenetic rollup of microbial marine transcriptomes. Since most of the transcriptomes are from lower eukaryotes only distantly related to completed genomes, most of the reads are identified simply as 'protozoal' or 'eukaryotic'. Nevertheless, the samples where reads are predominantly bacterial are easily visible in this figure.

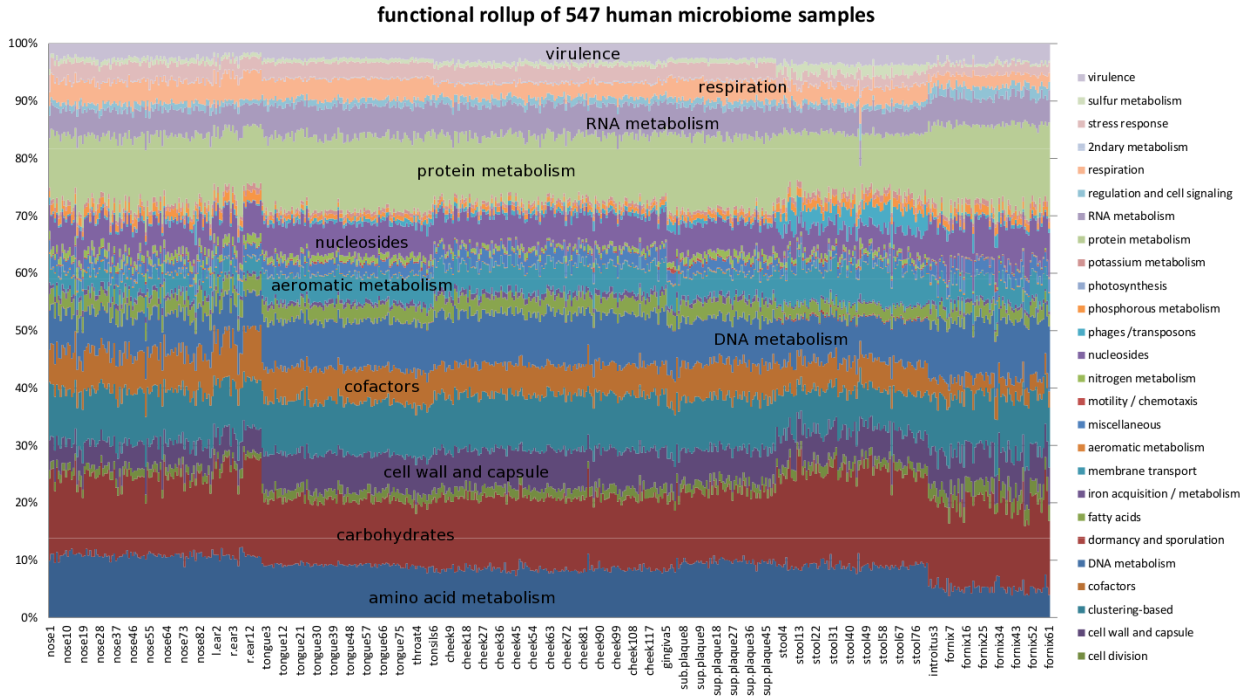


Fig. 5.4: Functional profile of 547 human microbiome samples, computed using the Life2550 tree, the SEED classification, and Sequedex, rolled up with the above-defined classifications. The functional profiles are provided above in an Excel spreadsheet.

results, but we have not explored this idea further. We observe in the environmental functional profiles, below, that these relative amplitudes are also broadly in agreement across diverse ecosystems. We will return to this discussion below, when we examine the functional profiles in greater resolution.

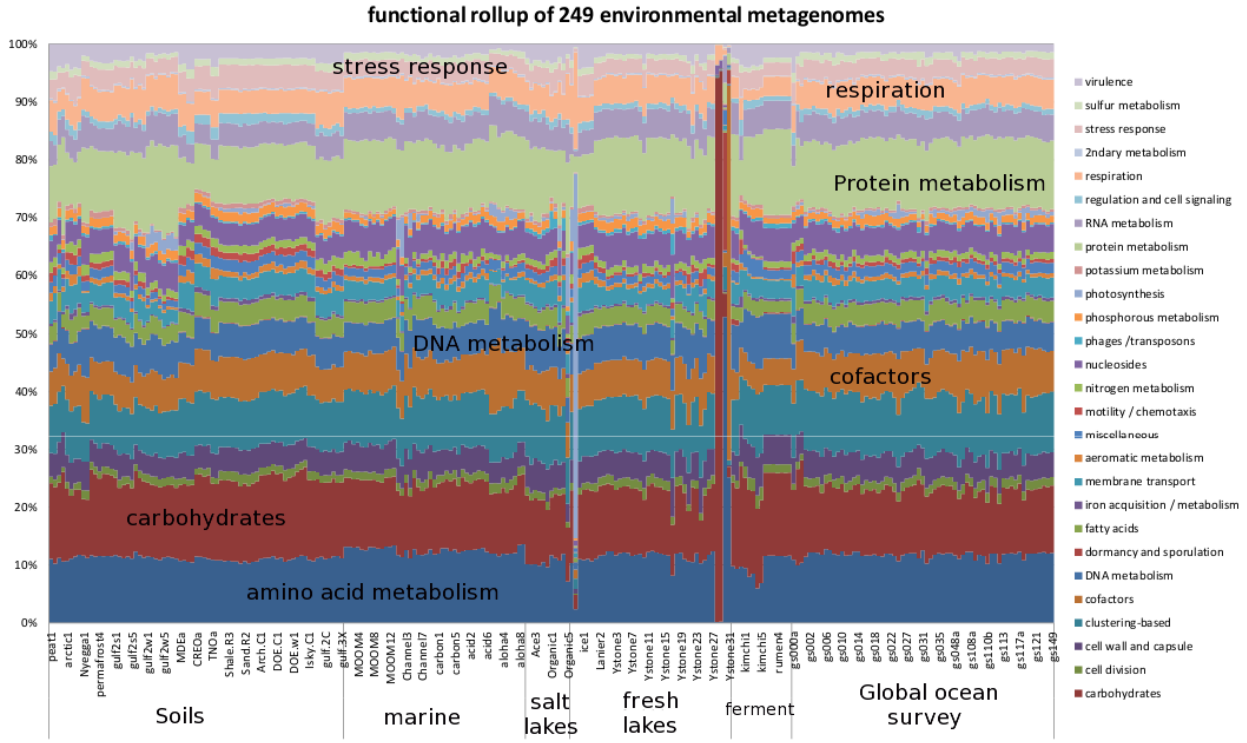


Fig. 5.5: Functional profile of 249 environmental microbiome samples with the same color-scheme as the HMB functional profile above, computed using the Life2550 tree, the SEED classification, and Sequedex, rolled up with the above-defined classifications. The functional profiles are provided above in an Excel spreadsheet. As with the human microbiome samples, the consistent relative fraction of each functional rollup is quite striking, as is the enormous dispersions that arise from this uniformity.

The twenty functional categories with more than 1000 counts in the TNO sample are shown below with the subsystem number from the figure and the three levels of annotation. The top two levels are shortened to help make the lowest level fit on the page. Genes in each SEED subsystem are found by clicking on the subsytem title in section *Definition of functional classifications*. Many of the subsytems also have detailed explanations available at http://www.theseed.org.

```
si_0032 Amino Acids    Glutamine, GLN, ASP, ASN; ammonia    Glutamine,_Glutamate,_Aspartate_and_Aspar
si_0042 Amino Acids    Lysine, threonine, MET, Cys          Methionine_Biosynthesis
si_0078 Carbohydrates  Central carbohydrate metabolism      Pyruvate_metabolism_II:_acetyl-CoA,_aceto
si_0122 Carbohydrates  One-carbon Metabolism                Serine-glyoxylate_cycle
si_0205 Cell Wall      Unclassified cell wall and capsule   Peptidoglycan_Biosynthesis
si_0279 Clustering     Unclassified clustering-based        Bacterial_Cell_Division
si_0334 Clustering     Unclassified clustering-based        Conserved_gene_cluster_associated_with_Me
si_0357 Cofactors      Biotin                               Biotin_biosynthesis
si_0415 DNA Meta.      DNA replication                      DNA-replication
si_0448 Fatty Acids    Fatty acids                          Fatty_Acid_Biosynthesis_FASII
si_0526 Membrane tra.  Unclassified membrane transport      Ton_and_Tol_transport_systems
si_0589 Nitrogen       Unclassified nitrogen metabolism     Ammonia_assimilation
si_0602 Nucleosides    Purines                              Purine_conversions
si_0640 Phosphorus     Unclassified phosphorus metabolism   Phosphate_metabolism
si_0654 Potassium      Unclassified potassium metabolism    Potassium_homeostasis
```

```
si_0660 Protein         Protein biosynthesis                  Ribosome_LSU_bacterial
si_0790 Regulation      Unclassified regulation and cell sig cAMP_signaling_in_bacteria
si_0812 Respiration     Electron donating reactions           Respiratory_Complex_I
si_0929 Virulence       Resistance to antibiotics and toxic  Cobalt-zinc-cadmium_resistance
si_0939 Virulence       Resistance to antibiotics and toxic  Multidrug_Resistance_Efflux_Pumps
```

Functional rollups from genomic (DNA) transcriptomes have similar quantities of reads from each of the functional rollup categories. Examination of the functional rollup of the marine algae transcriptomes, however, show much more variable, and results that depend on growth conditions.

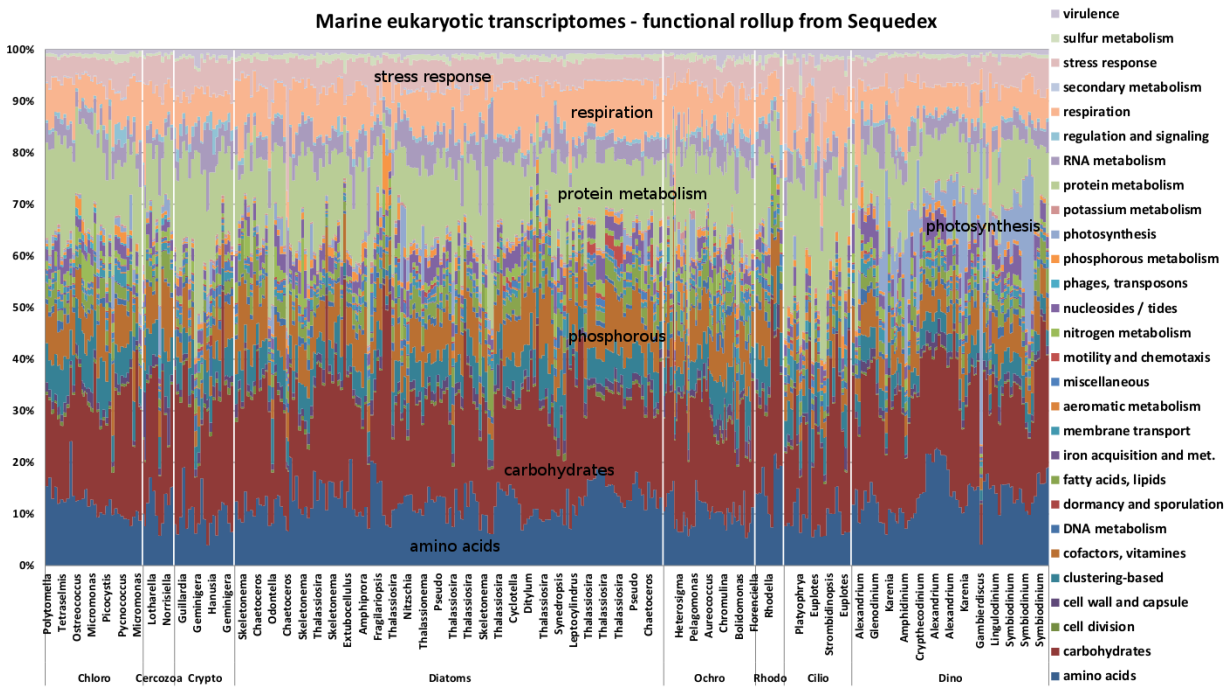


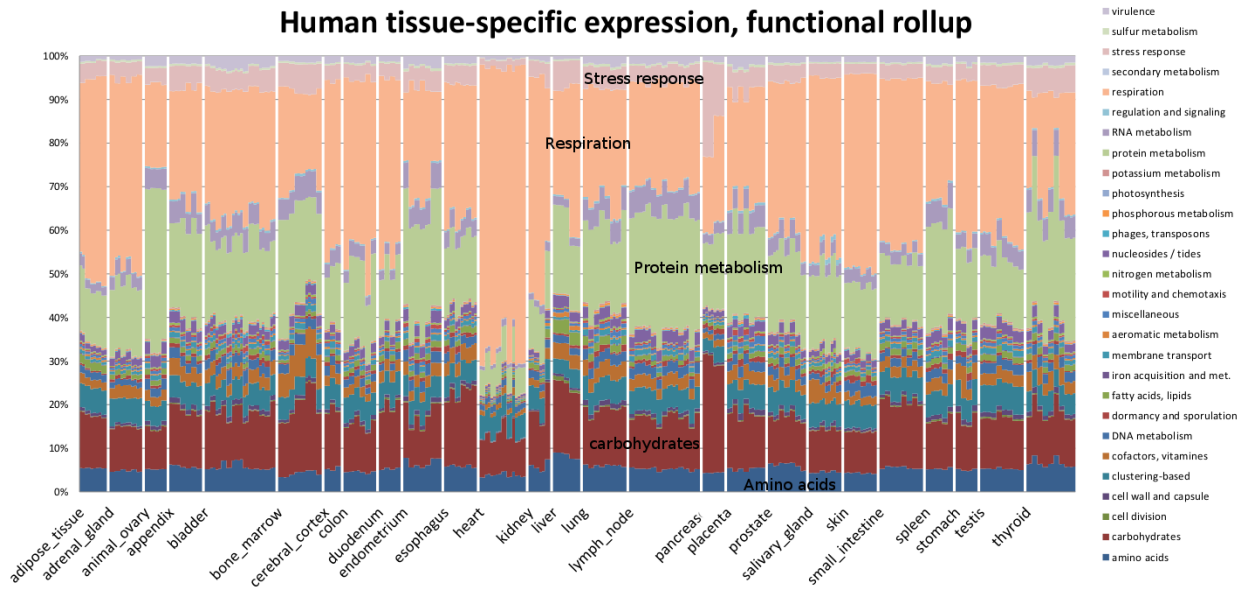
Fig. 5.6: Functional rollup of 330 marine eukaryotic algal transcriptomes. The transcriptomes are grouped phylogenetically, on the basis of the small subunit ribosomal RNA sequence obtained from the sample. It is evident that the condition-dependence of the functional rollups are significant, but the correlations of conditions with functional rollup will require closer examination.

When the transcriptomes are taken from a single organism and grouped by condition, such as with the Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics, by Fagerberg, *et al.*, the statistically significant differences in functional expression by grouping are immediately evident.

5.4 Normalized functional profiles

To investigate further, we examine the functional profiles of the larger data sets for the 400 human microbiome samples and 242 enviromental samples for individual SEED subsystems, answering the question, 'What type of nitrogen metabolism is present?' rather than 'Does the organism metabolise nitrogen?'.

Since most of the SEED subsystems cover metabolic and other functions necessary for all organisms, the consistency across samples is more reassuring than informative. To convince ourselves that the functional profiles are capable of



distinguishing samples from one-another, we normalize the functional counts (not including 'unassigned function' and compare two categories that are observed to vary across environmental and human microbiome samples.

Normalization occured with Fortran 90 code; see `F90 code`. One of the easiest functional categories to interpret is photosynthesis across the environmental metagenomes. By comparison, none of the human microbiome samples contained more than 0.25% photosynthesis, and this was primarily composed of hits to proteorhodopsin, rather than the photosystems.

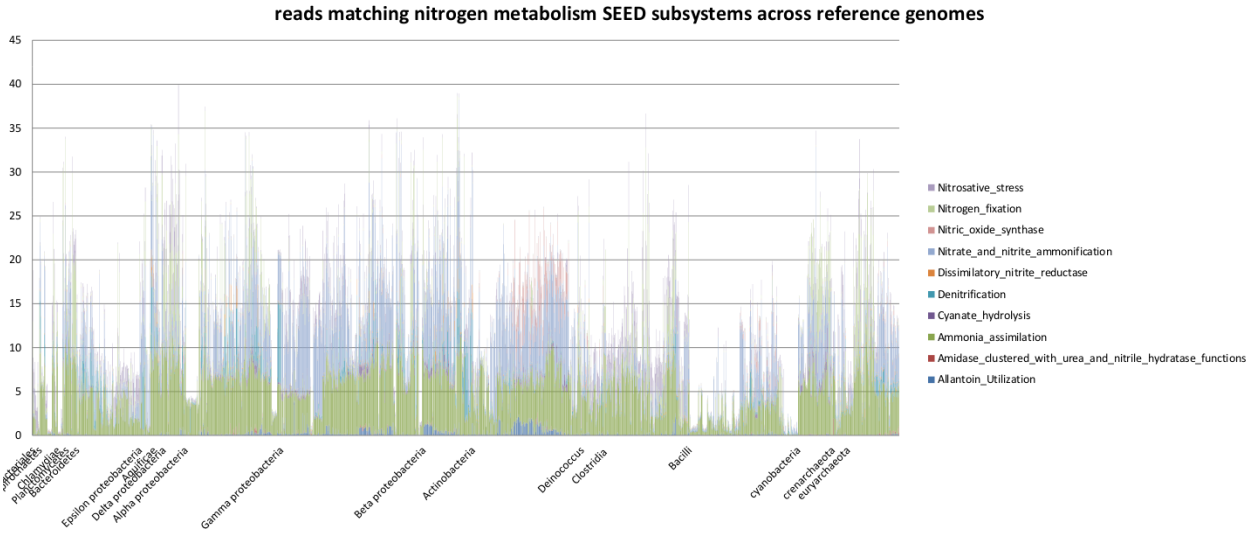


Fig. 5.7: Fraction of the functionally assigned reads encoding for nitrogen subsystems across the 2415 synthetic reference metagenomes.

When this figure is expanded and combined with insight from the microbiolgist, numerous useful insights can be gained, such as in these expanded views of three sets of subsystems across the cyanobacterial phylum, as in Proteomic profiles of five strains of oxygenic photosynthetic cyanobacteria of hte genus *Cyanothece*, by Aryal, *et al.*.

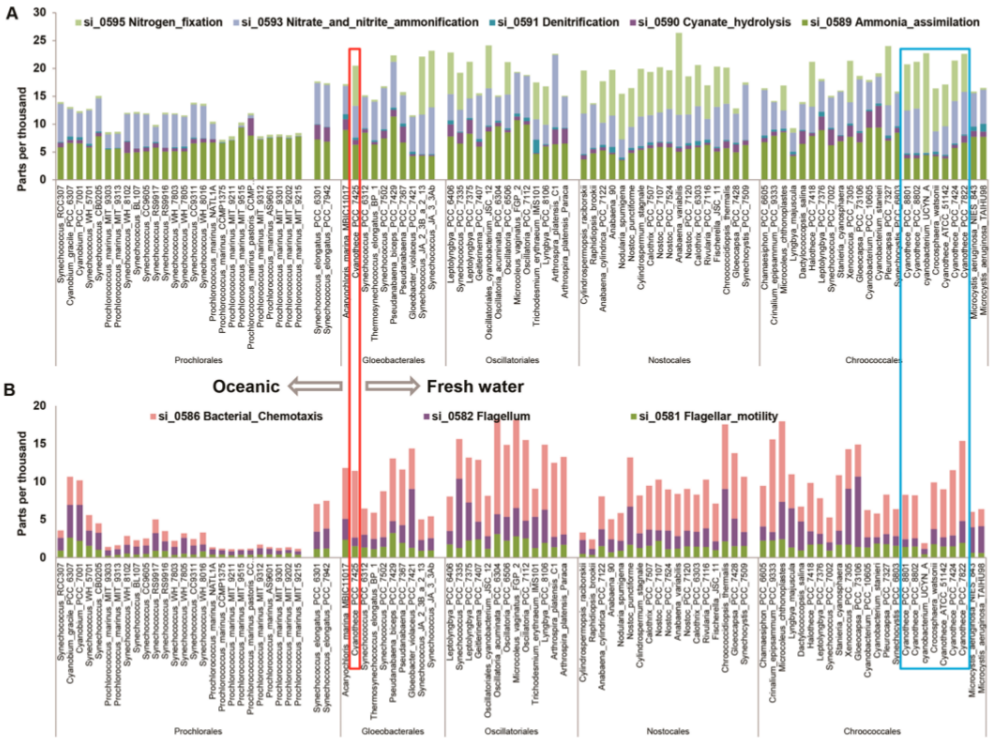


Fig. 5.8: Distribution of genes involved in various cellular functions across 94 cyanobacterial genomes. Shown are the genes involved in nitrogen metabolism (A) and motility and chemotaxis (B). The data were normalized within each genome so that the sum of all functional categories adds to 1000 counts, making the units on the amplitudes of each function parts per thousand. Genes associated with the *Cyanothece* are shown in boxes.

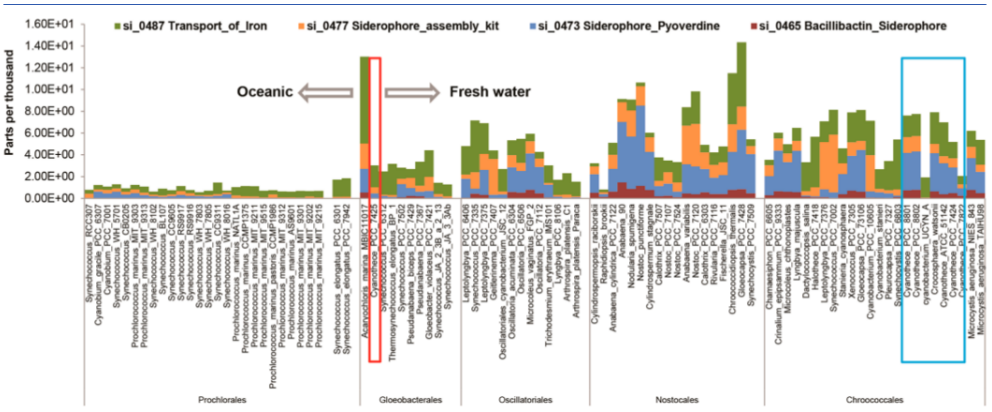


Fig. 5.9: Distribution of genes involved in iron acquisition and transport across 94 cyanobacterial genomes. Genes associated with different *Cyanothece* species are shown in boxes.

For the environmental metagenomes, we chose to examine the photosynthesis.

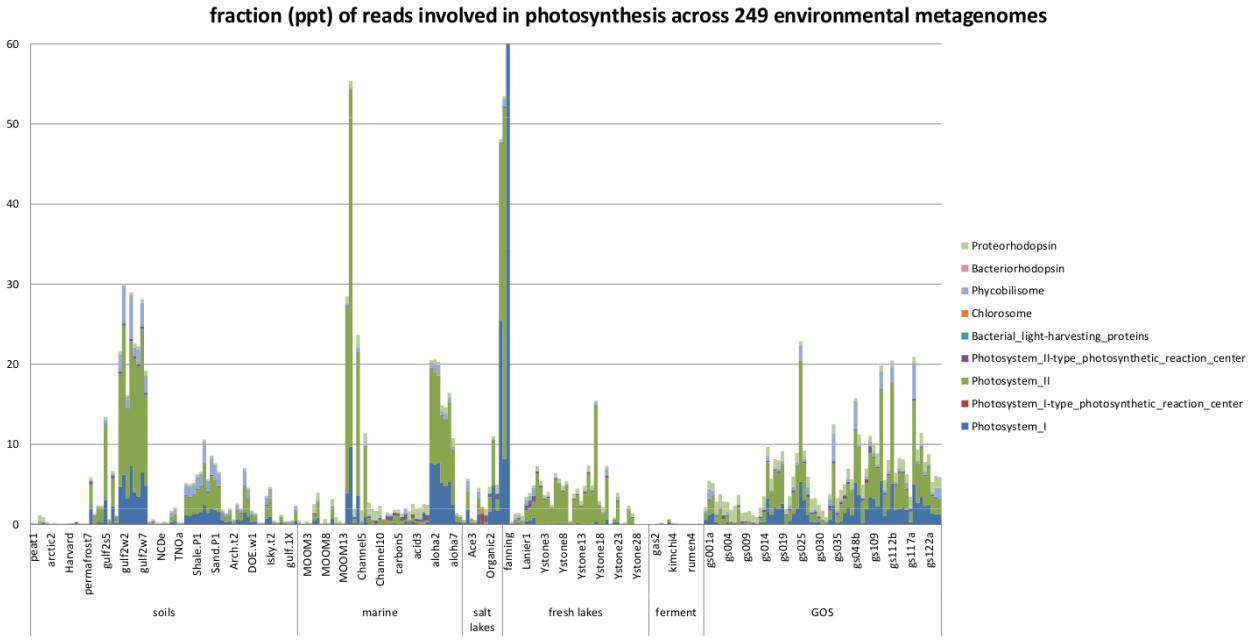


Fig. 5.10: Fraction of the functionally assigned reads encoding for phytosynthesis subsystems across the 249 environmental metagenomes.

For the human microbiome samples, we chose to examine the motility genes.

The set of output files can be viewed in Excel, and we supply two Excel spreadsheets with the collected results for the human microbiome data `hmb.xlsx` and environmental metagenomes `env.Life.xlsx`. The tab 'fnorm' in each spreadsheet contains the normalized functional counts. It is straightforward to scroll through this spreadsheet graphing various subsystems of interest, or sorting the functional categories based on counts in particular samples or difference in counts across various samples. With some thought, it is also possible to sort based on the signal-to-noise ratio with which the functional profiles distinguish ecosystems, with the noise defined by replicate samples.

Examination of the SEED subsystems from the human tissue transcriptomes shows the tissue-specific espression in more detail.

Thus, normalized functional profiles across reference data sets enable a 'top-down' approach to understanding functional classifications. The user may want to import results from their own metagenomics data sets into these spreadsheets to better understand the significance of their own results.

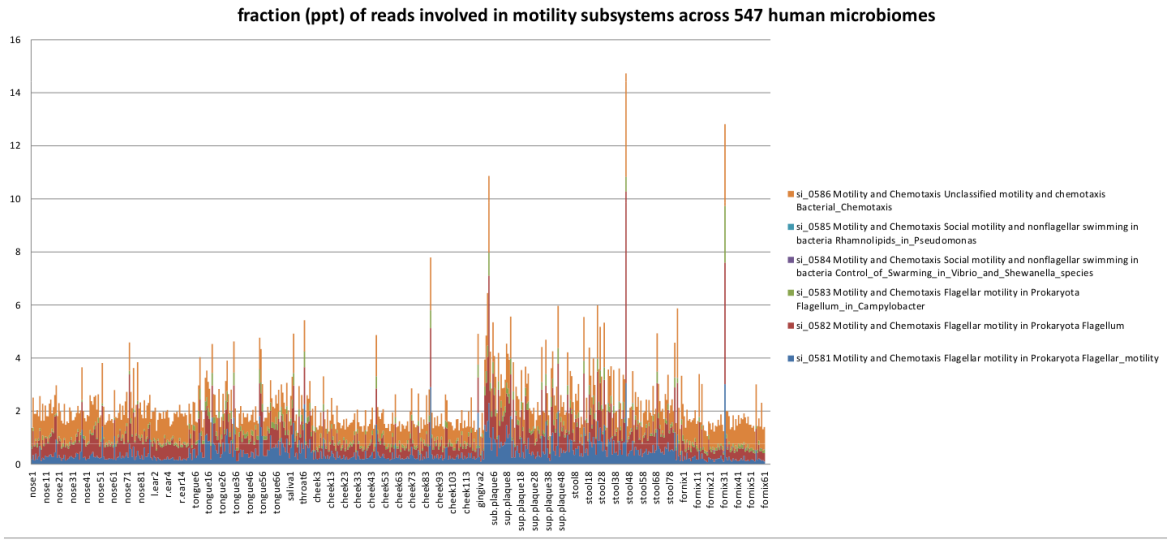


Fig. 5.11: Fraction of the functionally assigned reads encoding for phytosynthesis subsystems across the 527 human microbiome metagenomes.

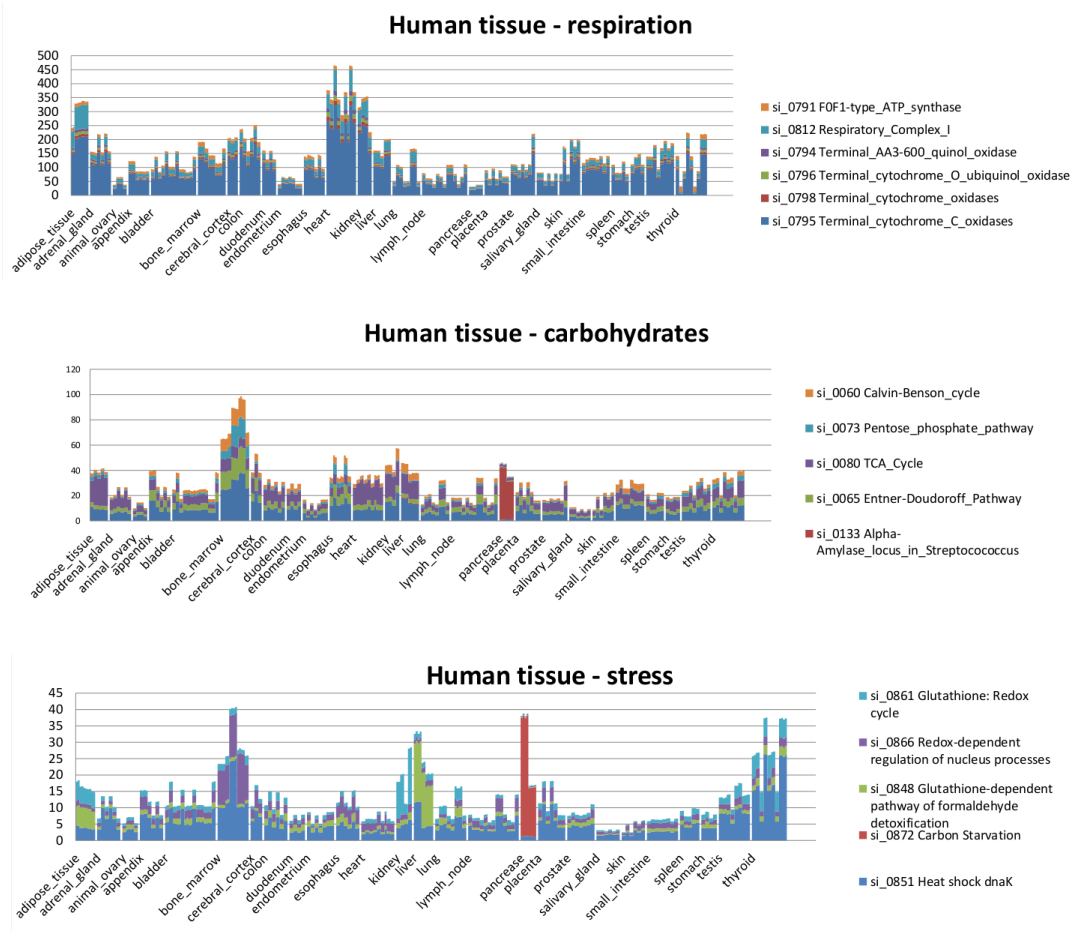


Fig. 5.12: Normalized counts for the tissue-specific expression.

CHAPTER 6

Using Sequestat

Sequestat is a suite of functions and visualizations written for analysis of Sequedex output in the statistical visualization package, R. Installation instructions and hints to getting started with R are provided in *Statistical analysis and graphing with R and R Studio*. RStudio provides a graphical environment with pull-down menus and more convenient ways to perform many tasks, but we provide instructions for the simpler, terminal window, interface to R. These instructions will work in both R and Rstudio.

The easiest way to get started with Sequestat in R is to first download `sequestat.r` and the nexus-format phylogenetic tree used with the Life2550 data modules `Life2550.nexus` and then examine sets of already-compiled output files from three large data sets:

- the set of `synthetic data from reference genomes` with `labels`,
- a set of `environmental microbiomes` with `labels`
- a set of `human microbiomes` with `labels`

Readers who have aready analyzed data sets, and have the Sequedex output files available on their desktop computers can skip ahead to *Reading Life2550 output files into Sequestat:* or *Reading virus1252 output files into Sequestat:*.

6.1 Visualizing phylogenetic data with Sequestat

Once R or Rstudio is started, the human microbiome data can be loaded into R or Rstudio and plotted with the commands:

```
source("~/Sequedex-docs/dl/sequestat.r")

filename="~/Sequedex-docs/dl/hmb.Life.who"
labelfile="~/Sequedex-docs/dl/hmb.Life.lbl"

dat=read.table(filename,header=F,sep="\t")
lab=read.table(labelfile,header=F,sep="\t")
lab=lab[!is.na(lab)]
nn=dim(dat)[1]-1
rname=c(paste("n000",0:9,sep=""),paste("n00",10:99,sep=""),paste("n0",100:999,sep=""),paste("n",1000
colnames(dat)=lab
rownames(dat)=rname
```

A set of PDF files that together tile the entire tree, Life2550, are available at _s.Life2550, above each of the sections, with node numbers, family, phylum, and taxon names. Sub-trees can then be viewed in R by typing the following:

```
Plot.profile(expts,1,"n0000",tip.frq=1,use.edge.length=T,type="phylo",show.tip.label=T)
```

where "n0000" is the node number defining the subtree to be plotted. The node number that is the second argument of the above function call is defined by the `phyloxml file`, that can be viewed with archaeopteryx, as described in *Exploring trees with Archaeoptrix, FigTree, or NJPlot*. Alternatively, section *Tree of Life, 2550 taxa* contains node numbers for a variety of families and phyla, and 23 pdf files detailing a spanning set of subtrees. To define subtrees by label in a way that covers the categories included in the high level rollup categories for the tree Life2550, copy the following lines into your R session:

```
root="n0000"; bacteria="n0001"; gramp="n1220"; gramn="n0002"; euks="n2401"; metazoa="n2447"; fungi="n
archaea="n2214"; crenarch="n2218"; euryarch="n2259"
elusi="n0005";fuso="n0008"; spiro="n00025"; chlamydia="n0075"; chlorobi="n0115"; bacteroidetes="n012
adeprot="n0301"; eprot="n0302"; aquificae="n0342"; acidobacteria="n0354"; dprot="n0369"; aprot="n042
rhiz="n0596"; rhodo="n0536"; sphingo="n0509"; rhodospirillaceae="n0470"; rickettsiales="n0430"
bgprot="n0691"; gprot="n0692"; enteric="n0705"; pasteurellales="n0793"; vibrio="n0818"; shewanella="n
altero="n0879"; gprot2="n0912"; moraxellaceae="n0980"; xanthomonads="n1011"; francisella="n1072";
bprot="n1083"; burk="n1086"; neisseria="n1197"; nitrosomonas="n1183"
actino="n1221"; coriobacteriales="n1222"; actinomyces="n1244"; streptomyces="n1334"; nocardia="n1387
chloroflexi="n1572";deino="n1512"; clostridia="n1591";negativicutes="n1784"; peptococcaceae="n1758";
clostridia="n1596";clostridia1="n1601";ruminococcus="n1650";thermoanaero="n1679"
lactobacillales="n1822";strep="n1826"; lactobacillaceae="n1875"
cyano="n2120";plasmas="n2065";staph="n1958";bacillus="n1988";paenibacillaceae="n2041";bacillales="n19
```

The abbreviations are somewhat arbitrary, with groups sometimes matching families or phyla, and sometimes more of a pneumonic; it is easy for the user to define his or her own families and labels. Once defined, however, the beta proteobacteria portion of the tree can be seen with the command:

```
Plot.profile(expts,1,bprot,tip.frq=1,use.edge.length=T,type="phylo",show.tip.label=T)
```

If you type the command:

```
expts$data[0,]
```

you will see the labels of the individual columns of data who's phylogenetic profiles are available in this session, and if you type ls(), you will see the list of defined functions and variables.

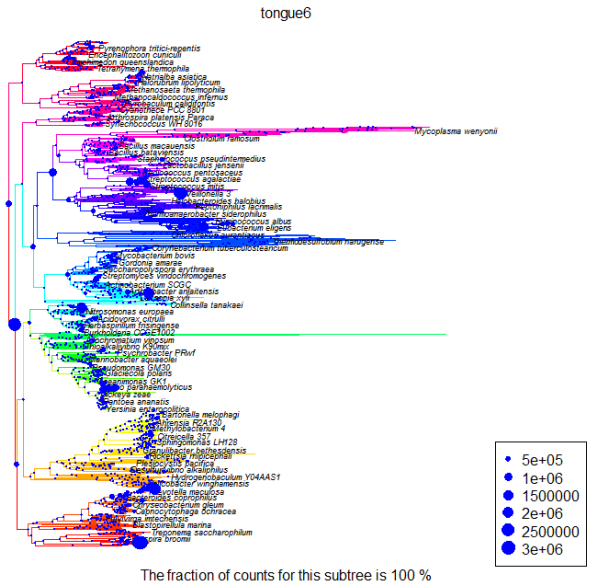
The user is invited to explore the full phylogenetic extent and plotting options available to this function by altering the various arguments to the Plot.profile function.

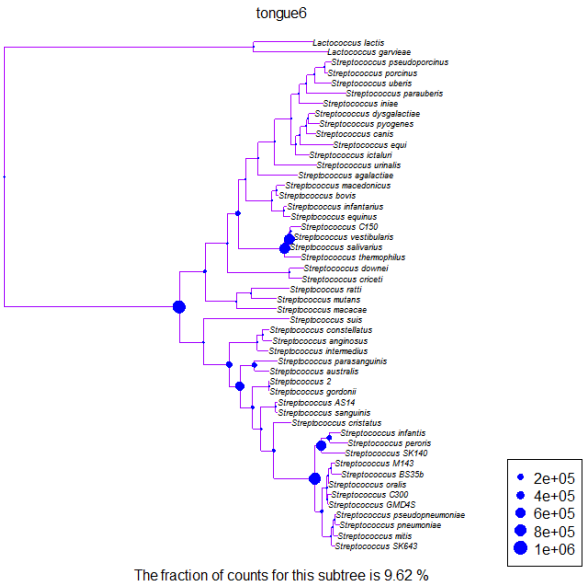
> hmb <- Read.tree("~/Sequedex-docs/dl/Life2550.nexus") hmb$data <- dat hmb$node.label <- rname
>
> hmb$data[0,] #list sample labels in order of column number
>
> Plot.profile(hmb,"tongue6","n0000",tip.frq=25,use.edge.length=T,type="phylo")
> Plot.profile(hmb,"tongue6",strep,tip.frq=25,use.edge.length=T,type="phylo")

results in figures like the following

This data set can be explored as above. Similarly, the environmental data can be loaded with the commands:

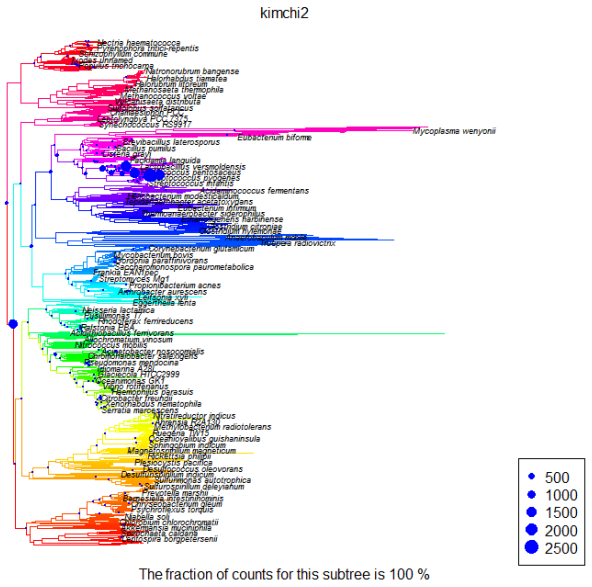
```
filename="~/Sequedex-docs/dl/env.Life.who"
labelfile="~/Sequedex-docs/dl/env.Life.lbl"

dat=read.table(filename,header=F,sep="\t")
lab=read.table(labelfile,header=F,sep="\t")
lab=lab[!is.na(lab)]
nn=dim(dat)[1]-1
rname=c(paste("n000",0:9,sep=""),paste("n00",10:99,sep=""),paste("n0",100:999,sep=""),paste("n",1000
colnames(dat)=lab
rownames(dat)=rname

env <- Read.tree("Life2550.nexus")
env$data <- dat
```

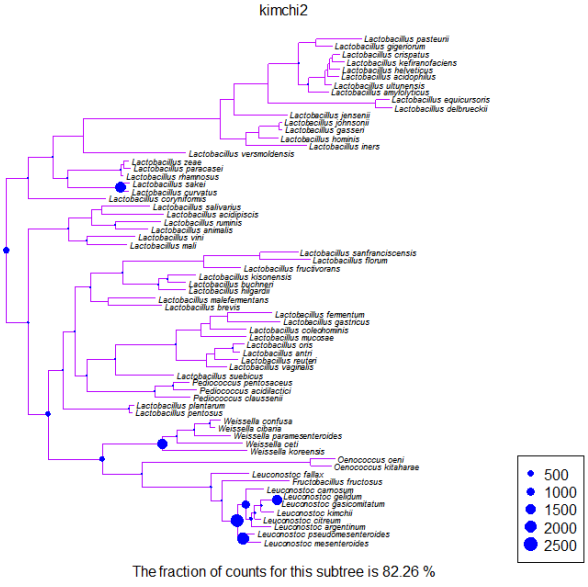





```
env$node.label <- rname

Plot.profile(env,"kimchi2",root,tip.frq=25,use.edge.length=T,type="phylo")
Plot.profile(env,"kimchi2","n1875",tip.frq=25,use.edge.length=T,type="phylo")
```

produces plots like



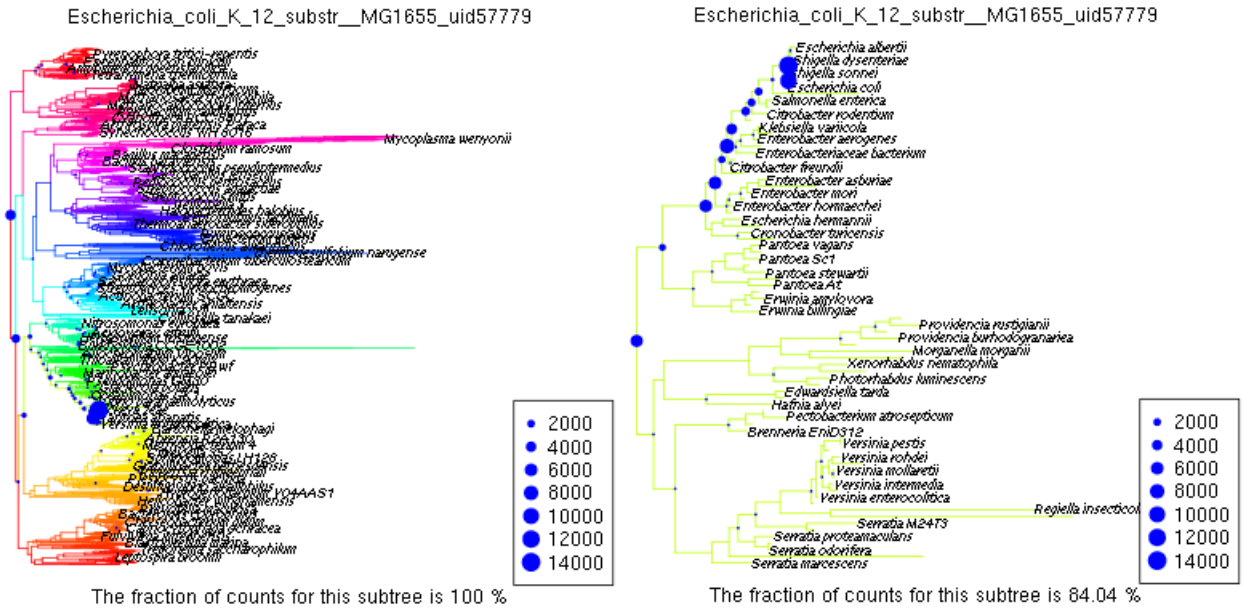


and the synthetic data evaluation with:

```
filename="~/Sequedex-docx/dl/syn.Life.who"
labelfile="~/Sequedex-docx/dl/syn.Life.lbl"

dat=read.table(filename,header=F,sep="\t")
lab=read.table(labelfile,header=F,sep="\t")
lab=lab[!is.na(lab)]
```

```
nn=dim(dat)[1]-1
rname=c(paste("n000",0:9,sep=""),paste("n00",10:99,sep=""),paste("n0",100:999,sep=""),paste("n",1000
colnames(dat)=lab
rownames(dat)=rname

syn <- Read.tree("~/Sequedex-docx/dl/Life2550.nexus")
syn$data <- dat
syn$node.label <- rname

Plot.profile(syn,"Escherichia_coli_K_12_substr__MG1655_uid57779","n0000",tip.frq=35,use.edge.length=
Plot.profile(syn,"Escherichia_coli_K_12_substr__MG1655_uid57779","n0707",tip.frq=2,use.edge.length=T,
```

produces plots like



These data sets can all be explored using the node numbers defined in *Tree of Life, 2550 taxa* or the phylogenetic groupings defined above, once the definitions are pasted into R. We assume the field separator in the read.table command needs to match the one in the input file.

6.2 Reading Life2550 output files into Sequestat:

To read in his or her own data from a set of directories of Sequestat output files, assumed to subdirectories of ~/expts/., with sequestat.r and the reference tree, Life2550.nexus in ~/Sequedex-docs/dl/, as described at the beginning of *Getting started*, a user should type:

```
source("~/Sequedex-docs/dl/sequestat.r")
expts=Read.phylo("~/expts","./Life2550.nexus")
Plot.profile(expts,1,"n0000",tip.frq=30,use.edge.length=T,type="phylo",show.tip.label=T)
```

at the R or Rstudio command prompt. At this point, a plot of the entire 2550 taxa tree of life should appear on the screen, with dots on the nodes indicating the number of reads in data set 3 assigned to each node. It is likely that the use will want to make trees of specific families. Plot.profile, which will display Sequedex output on either the entire reference tree, or of a subtree defined by a node number on the overall tree:

```
Plot.profile(expts,3,"n0000",use.edge.length=T,type="unrooted",show.tip.label=F)
```

6.3 Reading virus1252 output files into Sequestat:

When viewing output from the virus1252 data module wiht Sequestat in R, once again, source `sequestat.r` but `dl/virus1252.nexus` read in a set of directories of Sequestat output files, assumed to subdirectories of ~/expts/., with sequestat.r and the reference tree, virus1252.nexus in the working directory of R:

```
source("~/Sequedex-docs/dl/sequestat.r")
expts=Read.phylo("~/vexpts","~/Sequedex-docs/dl/virus1252.nexus")
Plot.profile(vexpts,1,"n0000",tip.frq=30,use.edge.length=T,type="phylo",show.tip.label=T)
```

At this point, a plot of the entire 1252 taxa viral tree should appear on the screen, with dots on the nodes indicating the number of reads in data set 3 assigned to each node. It is likely that the use will want to make trees of specific families. Plot.profile which will display Sequedex output on either the entire reference tree, or of a subtree defined by a node number on the overall tree:

```
Plot.profile(vexpts,3,"n0000",use.edge.length=T,type="unrooted",show.tip.label=F)
```

Sub-trees can then be viewed in R by typing the following:

```
Plot.profile(vexpts,1,"n0000",tip.frq=1,use.edge.length=T,type="phylo",show.tip.label=T)
```

where "n0000" is the node number defining the subtree to be plotted. The node number that is the second argument of the above function call is defined on the phyloxml tree `dl/virus1252.phyloxml`, that can be viewed with archaeopteryx, as described in *Exploring trees with Archaeoptrix, FigTree, or NJPlot*. Alternatively, section *Viral tree, 1252 taxa* contains node numbers for a variety of families and phyla, and 17 png image files detailing a spanning set of subtrees. To define subtrees by label in a way that covers the categories included in the high level rollup categories for the tree Life2550, copy the following lines into your R session:

```
root=0; dsDNA=1; dsDNA1=3; Baculo=5; Phycodna=51; Irido=63
dsDNA2=70; Papilloma=71; Polyoma=144; Pox=162
dsDNA3=180; Adeno=181; Herpes=215
RT=263; Caulimo=264; Retro=297; Hepadna=338
ssRNAp1=345; Calici=346; Nido=363
ssRNAp2=395; Flavi=395; Flavi1=396; Flavi2=432
ssRNAp2345=394; Tombus=446; Virga=470
ssRNAp4=503; Tymo=503; Tymo1=505; Tymo2=552; Tymo3=594
ssRNAp5=605; Picorna=606; Picorna1=608; Picorna2=687; Picorna3=714; Toga=721
ssRNAp6=735; Poty=735; Poty1=736; Poty2=805
ssRNAm=816; segRNAm=817; Arena=818; Bunya=843
Mononega=865; Mononega1=866; Mononega2=904;
Orthomyxo=931
ssDNA1=949; Parvo=949
ssDNA23=987; Gemini1=988; Gemini2=1230
dsDNA=1253; Reo=1253
```

The abbreviations are somewhat arbitrary, with groups sometimes matching families or phyla, and sometimes more of a pneumonic; it is easy for the user to define his or her own families and labels. Once defined, however, the beta proteobacteria portion of the tree can be seen with the command:

```
Plot.profile(expts,1,bprot,tip.frq=1,use.edge.length=T,type="phylo",show.tip.label=T)
```

If you type the command:

```
expts$data[0,]
```

you will see the labels of the individual columns of data who's phylogenetic profiles are available in this session, and if you type ls(), you will see the list of defined functions and variables.

The user is invited to explore the full phylogenetic extent and plotting options available to this function by altering the various arguments to the Plot.profile function.

6.4 Decomposing mixture into components with Sequestat

Even complex metagenomic samples often have dominant organisms, and when they are phylogenetically close to organisms in the reference database, the distiribution of reads across the nodes of the tree, above, can be used to identify what organisms are present. If this process identifies something closely related to an organism of particular interest to the user, follup analysis should be performed, as described in *Annotated reads*.

The strategy for decomposing the metagenomic phylogenetic profile into a mixture is to create a synthetic data set for the broadest possible array of curated organisms, thin this list, run Sequedex on this synthetic data with a particular tree (and ideally identical signature list), cluster this list, then fit the unknown sample to a linear combination of reference profiles.

- step 1: clean the raw profiles by removing profiles that either have less than 1000 hits (mainly divergent organisms) or profiles that have over 7.5% of reads that are not monophyletic (possibly contaminated organisms)
- step 2: cluster the remainin profiles into 800 groups
- step 3: use the lasso to model the normalized metagenomic profile asa linear combination of cluster profiles. Use only nodes that are at depth 4 or less from the leaves
- step 4: extract the top 100 clusters (use RSS to assess if this number is too small) that list is an upper bound of the possible profiles
- step 5: Use traditional model selection approach to select significant profiles (use k=1.25 be more inclusive than AIC)

Starting a new session in R or RStudio, and download `demo_protocol.r`, `syn.Life.who`, `syn.Life.lbl` `Life2550.nexus` into your working directory. Start R or RStudio in this directory and type:

```
source("~/Sequedex-docx/dl/demo_protocol.r")
```

Sequestat will chug along for a few minutes to completion. Typing sig.list will result in:

```
> sig.list
$`V202 0.31281`
[1] "Staphylococcus_caprae_C87_uid61125"
[2] "Staphylococcus_epidermidis_ATCC_12228_uid57861"
[3] "Staphylococcus_warneri_SG1_uid187059"
[4] "Staphylococcus_aureus_04_02981_uid161969"

$`V66 0.24039`
[1] "Sphingobium_SYK_6_uid73353"
[2] "Zymomonas_mobilis_ATCC_10988_uid55403"
[3] "Sphingomonas_wittichii_RW1_uid58691"
[4] "Sphingomonas_KC8_uid77733"
[5] "Sphingomonas_elodea_ATCC_31461_uid157063"
[6] "Sphingomonas_S17_uid66923"

$`V368 0.11183`
[1] "Novosphingobium_aromaticivorans_DSM_12444_uid57747"
```

```
[2] "Novosphingobium_nitrogenifigens_DSM_19370_uid64475"
[3] "Sphingomonas_LH128_uid174245"
[4] "Novosphingobium_AP12_uid171681"
[5] "Novosphingobium_Rr_2_17_uid170038"

$`V742 0.08731`
[1] "Staphylococcus_pettenkoferi_VCU012_uid179999"
[2] "Staphylococcus_arlettae_CVD059_uid175126"

$`V490 0.06027`
[1] "Staphylococcus_hominis_C80_uid61127"
[2] "Staphylococcus_haemolyticus_JCSC1435_uid62919"

$`V584 0.05848`
[1] "Staphylococcus_lugdunensis_HKU09_01_uid46233"

$`V420 0.04302`
[1] "Bacteroides_vulgatus_ATCC_8482_uid58253"
[2] "Bacteroides_plebeius_DSM_17135_uid54991"
[3] "Bacteroides_coprophilus_DSM_18228_uid55301"
[4] "Bacteroides_salanitronis_DSM_18170_uid63269"
[5] "Bacteroides_coprocola_DSM_17136_uid54879"

$`V821 0.03858`
[1] "Bacteroides_ovatus_3_8_47FAA_uid68195"
[2] "Bacteroides_xylanisolvens_CL03T12C04_uid181622"

$`V920 0.0318`
[1] "Staphylococcus_pseudintermedius_ED99_uid162109"

$`V790 0.02675`
[1] "Erythrobacter_NAP1_uid54197"
[2] "Erythrobacter_litoralis_HTCC2594_uid58299"
[3] "Erythrobacter_SD_21_uid54677"

$`V848 0.02402`
[1] "Bacteroides_finegoldii_CL09T03C10_uid181638"
[2] "Bacteroides_caccae_ATCC_43185_uid54521"
[3] "Bacteroides_faecis_MAJ27_uid86875"

$`V192 0.01953`
[1] "Macrococcus_caseolyticus_JCSC5402_uid59003"
[2] "Sporosarcina_newyorkensis_uid70561"
[3] "Solibacillus_silvestris_StLB046_uid168516"
[4] "Listeria_grayi_DSM_20601_uid55523"

$`V712 0.01661`
[1] "Staphylococcus_carnosus_TM300_uid59401"

$`V761 0.00867`
[1] "Novosphingobium_pentaromativorans_US6_1_uid78315"
[2] "Novosphingobium_PP1Y_uid67383"

$`V774 0.00743`
[1] "Sphingomonas_SKA58_uid54251"

$`V539 -0.02038`
[1] "Rhodospirillum_centenum_SW_uid58805"
```

```
[2] "Oceanibaculum_indicum_P24_uid176351"
[3] "Thalassobaculum_L2_uid182483"
[4] "SAR_116_cluster_alpha_proteobacterium_HIMB100_uid78325"
[5] "Candidatus_Puniceispirillum_marinum_IMCC1322_uid47081"

$`V985 -0.02118`
[1] "Phenylobacterium_zucineum_HLK1_uid58959"

$`V475 -0.03216`
[1] "Sphingopyxis_alaskensis_RB2256_uid58351"

$`V710 -0.05337`
[1] "Parvularcula_bermudensis_HTCC2503_uid51641"
[2] "Hirschia_baltica_ATCC_49814_uid59365"
[3] "Hyphomonas_neptunium_ATCC_15444_uid58433"
[4] "Parvibaculum_lavamentivorans_DS_1_uid58739"
```

Typing mGlbl will list the metagenomic samples loaded into this list, with the top portion of the output shown here:

```
> mGlbl
[1] "nose1"        "nose2"        "nose3"        "nose4"        "nose5"
[6] "nose6"        "nose7"        "nose8"        "nose9"        "nose10"
[11] "nose11"       "nose12"       "nose13"       "nose14"       "nose15"
[16] "nose16"       "nose17"       "nose18"       "nose19"       "nose20"
[21] "nose21"       "nose22"       "nose23"       "nose24"       "nose25"
[26] "nose26"       "nose27"       "nose28"       "nose29"       "nose30"
[31] "nose31"       "nose32"       "nose33"       "nose34"       "nose35"
[36] "nose36"       "nose37"       "nose38"       "nose39"       "nose40"
[41] "nose41"       "nose42"       "nose43"       "nose44"       "nose45"
[46] "nose46"       "nose47"       "nose48"       "nose49"       "nose50"
[51] "nose51"       "nose52"       "nose53"       "nose54"       "nose55"
[56] "nose56"       "nose57"       "nose58"       "nose59"       "nose60"
[61] "nose61"       "nose62"       "nose63"       "nose64"       "nose65"
[66] "nose66"       "nose67"       "nose68"       "nose69"       "nose70"
[71] "nose71"       "nose72"       "nose73"       "nose74"       "nose75"
[76] "nose76"       "nose77"       "nose78"       "nose79"       "nose80"
[81] "nose81"       "nose82"       "nose83"       "nose84"       "nose85"
[86] "nose86"       "nose87"       "nose88"       "nose89"       "l.ear1"
[91] "l.ear2"       "l.ear3"       "l.ear4"       "l.ear5"       "l.ear6"
[96] "l.ear7"       "l.ear8"       "r.ear1"       "r.ear2"       "r.ear3"
[101] "r.ear4"       "r.ear5"       "r.ear6"       "r.ear7"       "r.ear8"
[106] "r.ear9"       "r.ear10"      "r.ear11"      "r.ear12"      "r.ear13"
[111] "r.ear14"      "r.ear15"      "r.ear16"      "r.ear17"      "r.ear18"
[116] "tongue1"      "tongue2"      "tongue3"      "tongue4"      "tongue5"
```

To profile a different sample, simply re-run the last part of the demo_protocol.r code, with a new value for meta.idx. If you have labels defined, the command *meta.idx = match("stool1",mGlbl)* will select the correct column on the basis of the label. If you associate the tabular data with a tree, a described in s.read.sequedex or s.sets, you can compare the node plots to the lists of possible taxa. Here is the last part of the demo_protocol code, for cutting and pasting:

```
#=========================================================
#  apply algorithm to a selected metagenome

meta.idx <- 1
#meta.idx = match("stool1",mGlbl)

Y0 <- metaG[,meta.idx]/sum(metaG[,meta.idx])
Y0 <- Y0[ndepth < 251 ]
fit0 <- lars(PP,Y0,type="lasso",intercept=F,normalize=F,max.steps=300,use.Gram=F)
```

```
#-----------------------------------------------------------
#  step 4: extract the top 100 clusters

nvar <- 200
idx <- max((1:length(fit0$df))[fit0$df <= nvar])
rss <- fit0$RSS
bb <- fit0$beta[idx,]
cc <- bb[bb != 0]
big.list <- cc.name[bb != 0]
big.list <- big.list[sort.list(bb[bb != 0],decreasing=T)]

#-----------------------------------------------------------
#  step 5: Use traditional model selection approach to select significant profiles

vnames <- paste("V",1:dim(PP)[2],sep="")
colnames(PP) <- vnames
names(cc.name) <- vnames
mnames <- vnames[bb != 0]
mmodel0 <- formula(paste("Y0 ~ 0 + ",mnames[1]))
mmodel1 <- formula(paste("Y0 ~ ",paste(mnames,collapse=" + ")))
fit1 <- lm( mmodel0 ,data=as.data.frame(PP))
fit2 <- step(fit1, mmodel1, data=as.data.frame(PP),k=2)

sig.list <- cc.name[(names(fit2$coef))]
sig.list <- sig.list[sort.list(fit2$coef,decreasing=T)]
names(sig.list) <- paste(names(sig.list),sort(round(fit2$coef,5),decreasing=T))
```

CHAPTER 7

Functional analysis

Sequedex not only places each 10-mer on a phylogenetic tree, but it also searches an 962 example sets of genes for a functional assignment as well. For the seed_0911.m1 set of functional assignments, we used the SEED classification of functions, which have the added benefit of a well-defined hierarchical rollup. The names and sets of genes are ennumerated in section *Definition of functional classifications*, where clicking on each label provides the annotation for each gene included (across the phylogeny) in the subsystem. For the ribosome, category si_0962, both the large and small subunits from across the 1550 species in the tree of life were translated into all three forward reading frames to make 10-mer 'amino acid' signatures, while the bacteria and archaea also had the tRNAs translated in a similar fasion. In the event a gene with a 10-mer signature appears in multiple categories, a metagenomic read contains 10-mers from different genes in different categories, the read is apportioned equally among all categories.

In additional to the genomic (DNA) metagenomic data sets with the phylogenetic profiles examined in the previous chapter:

- the set of `synthetic data from reference genomes` with `labels`,
- a set of `environmental microbiomes` with `labels`
- a set of `human microbiomes` with `labels`

we provide transcriptomic (RNA) data sets from publically available data sets to illustrate the comparisons:

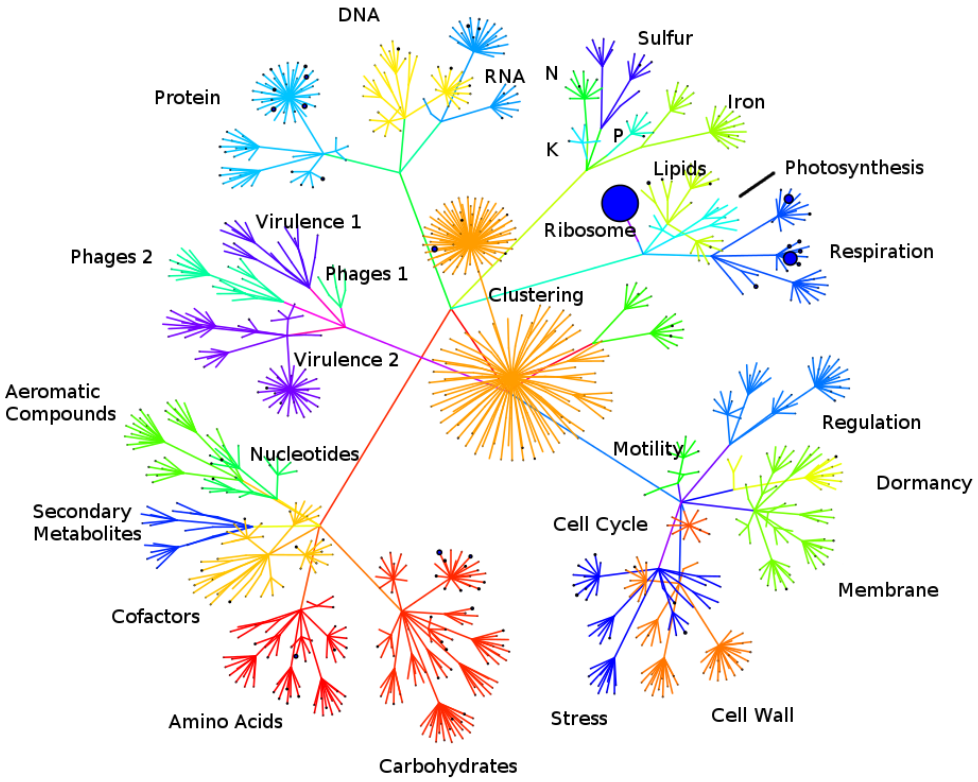
- a set of `human tisue-specific expression` with `labels`
- a set of `marine eukaryotic transcriptomes` with `labels`
- a set of `transcriptomes from plaque microbiomes` with `labels`

7.1 Visualizing SEED functional profiles with Sequestat

Graphical visualization can occur with Sequestat and the igraph package in R. You will first need to download `the functional definitions` (or use any what-Life2550-xxGB.0xseed_0911.m1.tsv file). Here, we assume both it and sequestat.r are available in the user's home directory (~/), and that the working directory of the R session is in a directory containing sequedex output files. Also, we assume the user is using the `.Rdata file with the graph layout in it`. Since we often do not want Ribosomal or Unclassified read counts to skew normalizations when plotting counts, we set them to zero. Analysis of all of the data sets will start with the following:

```
source("~/Sequedex-docs/dl/sequestat.r")
library(igraph)
load("~/Sequedex-docs/dl/seed.layout.Rdata")
```

This results in the graph layout below, to which we have added labels to aid in reading the graphs below, from which we leave the labels off.



Reading in particular Sequedex output files will begin with:

```
expt <- Read.functional("./", "~/Sequedex-docs/dl/what", type.ref="Life2550-40", data.type="what")
expt$layout=seed.layout
expt$data[963,]=0
expt$data[964,]=0
plot.graph(hatlas,1, simple.name= T,scol= "blue",sign= F, dimension= 2,cex=0.6)
```

To read in the tsv data sets listed above, download them and read them into the data structures as follows:

```
source("~/Sequedex-docs/dl/sequestat.r")
library(igraph)

syn=Read.graph("~/Sequedex-docs/dl/what", col=TRUE, sat=1)
data=read.table("~/Sequedex-docs/dl/syn.Life.what",sep="\t",header=F)
syn$data=data
load("~/seed.layout.Rdata")
syn$layout=seed.layout
syn$data[963,]=0
lbl=read.table("~/Sequedex-docs/dl/syn.Life.lbl",sep="\t",header=F)
lbl=lbl[!is.na(lbl)]
colnames(syn$data)=lbl
plot.graph(syn,1, simple.name= F,scol= "blue",sign= T, dimension= 2,cex=0.6)
Diff.graph(syn,1,2, dim=2, alpha= 0.000000001)

env=Read.graph("~/Sequedex-docs/dl/what", col=TRUE, sat=1)
data=read.table("~/Sequedex-docs/dl/env.Life.what",sep="\t",header=F)
env$data=data
load("~/seed.layout.Rdata")
env$layout=seed.layout
```

```
env$data[963,]=0
lbl=read.table("~/Sequedex-docs/dl/env.Life.lbl",sep="\t",header=F)
lbl=lbl[!is.na(lbl)]
colnames(env$data)=lbl
plot.graph(env,1, simple.name= F,scol= "blue",sign= T, dimension= 2,cex=0.6)
Diff.graph(env,1,2, dim=2, alpha= 0.000000001)

hmb=Read.graph("~/Sequedex-docs/dl/what", col=TRUE, sat=1)
data=read.table("~/Sequedex-docs/dl/hmb.Life.what",sep="\t",header=F)
hmb$data=data
load("~/seed.layout.Rdata")
hmb$layout=seed.layout
hmb$data[963,]=0
lbl=read.table("~/Sequedex-docs/dl/hmb.Life.lbl",sep="\t",header=F)
lbl=lbl[!is.na(lbl)]
colnames(hmb$data)=lbl
plot.graph(hmb,1, simple.name= F,scol= "blue",sign= T, dimension= 2,cex=0.6)
Diff.graph(hmb,1,2, dim=2, alpha= 0.000000001)

hatlas=Read.graph("~/Sequedex-docs/dl/what", col=TRUE, sat=1)
data=read.table("~/Sequedex-docs/dl/hatlas.Life.what",sep="\t",header=F)
hatlas$data=data
load("~/seed.layout.Rdata")
hatlas$layout=seed.layout
hatlas$data[963,]=0
lbl=read.table("~/Sequedex-docs/dl/hatlas.Life.lbl",sep="\t",header=F)
lbl=lbl[!is.na(lbl)]
colnames(hatlas$data)=lbl
plot.graph(hatlas,1, simple.name= F,scol= "blue",sign= T, dimension= 2,cex=0.6)
Diff.graph(hatlas,1,2, dim=2, alpha= 0.000000001)

ocean=Read.graph("~/Sequedex-docs/dl/what", col=TRUE, sat=1)
data=read.table("~/Sequedex-docs/dl/ocean.Life.what",sep="\t",header=F)
ocean$data=data
load("~/seed.layout.Rdata")
ocean$layout=seed.layout
ocean$data[963,]=0
lbl=read.table("~/Sequedex-docs/dl/ocean.Life.lbl",sep="\t",header=F)
lbl=lbl[!is.na(lbl)]
colnames(ocean$data)=lbl
plot.graph(ocean,1, simple.name= F,scol= "blue",sign= T, dimension= 2,cex=0.6)
Diff.graph(ocean,1,2, dim=2, alpha= 0.000000001)

caries=Read.graph("~/Sequedex-docs/dl/what", col=TRUE, sat=1)
data=read.table("~/Sequedex-docs/dl/caries.Life.what",sep="\t",header=F)
caries$data=data
load("~/seed.layout.Rdata")
caries$layout=seed.layout
caries$data[963,]=0
lbl=read.table("~/Sequedex-docs/dl/caries.Life.lbl",sep="\t",header=F)
lbl=lbl[!is.na(lbl)]
colnames(caries$data)=lbl
plot.graph(caries,1, simple.name= F,scol= "blue",sign= T, dimension= 2,cex=0.6)
Diff.graph(caries,1,2, dim=2, alpha= 0.000000001)
```

Alternatively, the graph can be looked at in three dimensions:

```
hatlas$layout <- layout.fruchterman.reingold(hatlas, dim=3)
plot.graph(hatlas,Val = 3, simple.name= F,scol= "blue",sign= T, dimension= 2,cex=0.6)
```

It is frequently the case that the ribosomal reads dominate the functional classifications. To eliminate the ribosomal reads from consideration, enabling a re-scaling of the other categories, type:

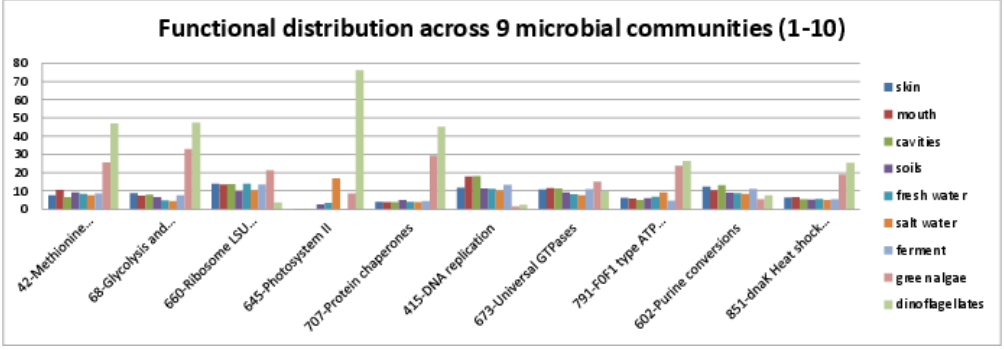
```
hatlas$data[963,]=0.
```

7.2 Top 100 functional categories

In this first section, we examine the functional categories with the most reads mapping to them across DNA sequenced across seven microbiomes and mRNA sequenced across two eukarotic marine samples. Functional categories will attract more reads for several reasons: first, because there may be many reads in them (eg. the citric acid cycle); second, because the genes may be highly conserved across numerous organisms (eg. the ribosome); and third, because the gene may be highly expressed in a particular microbial environment (eg. photosystem II). The seven groups of DNA sequenced across seven microbiomes include three groups from the human microbiome data shown early, grouped into the skin (ear and nose), the mouth, and 'cavities', which includes stool samples and vaginal samples. It also includes the environmental microbiomes, grouped into four categories; soils, fresh water, salt water, and fermented samples. The two transcriptome samples were the green algae and dinoflagellates from the eukaryotic marine algae transcriptome project, sponsored by the Moore foundation and sequenced at NCGR.

The goal of this section is to provide an understanding of how 10 percent of the more important functional categories are defined, represented in a variety of microbial environments, and relate to some of the other community resources available for functional annotation of proteins. We plot them in groups of ten, with the y-axis labeled in parts per thousand, and skipping the most prevalent functional category, si_0962, the large and small subunits of the ribosome.

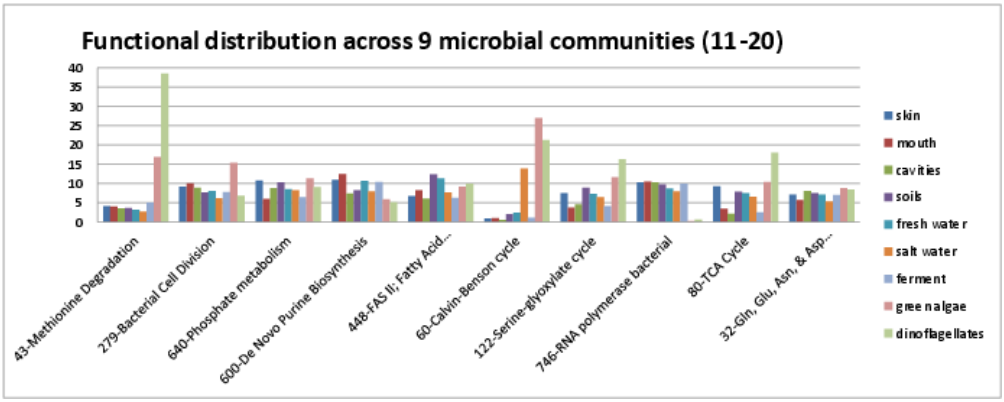
SEED description are available for many of these subsystems.



The ten most prevalent categories include:

- `si_0042; Methionine biosynthesis` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 270. The SEED description by Dmitry Rodionov describes a variet of pathways leading to methionine.
- `si_0068; Glycolysis and Gluconeogenesis` in the *Carbohydrates* rollup category. It can be found on Kegg map 10. The SEED description by Svetlana Gerdes and Ross Overbeek, describe glycolysis and gluconeogenesis.
- `si_0660; Ribosome LSU bacterial` in the *Protein Metabolism* rollup category. It can be found on Kegg map 3010.
- `si_0645; Photosystem II` in the *Photosynthesis* rollup category. It can be found on Kegg map 195.

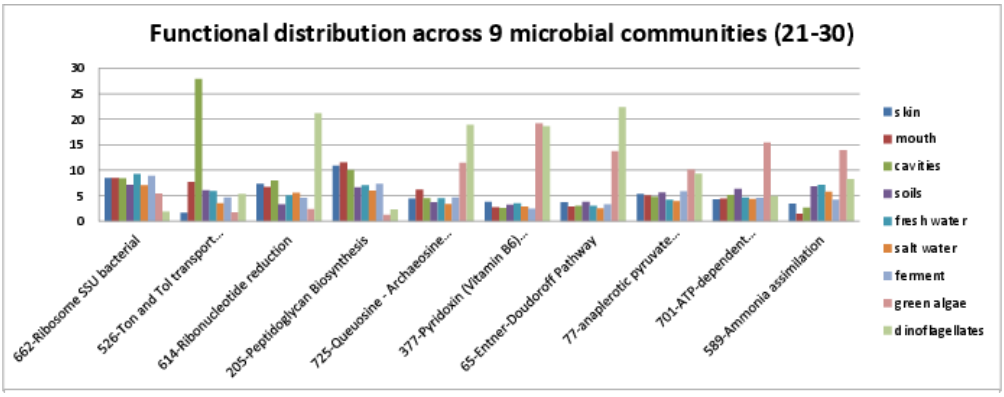
- `si_0707; Protein chaperones` in the *Protein Metabolism* rollup category. DnaJ is HSP-40 and dnaK is HSP-70. See the Wikipedia page and PMID 16952052.
- `si_0415; DNA replication` in the *DNA Metabolism* rollup category.
- `si_0673; Universal GTPases` in the *Protein Metabolism* rollup category. Caldon, et al. suggest that the 11 universal GTPases are either necessary for ribosome function or transmitting information from the ribosome to downstream targets for the purpose of generating specific cellular responses.
- `si_0791; F0F1 type ATP synthase` in the *Respiration* rollup category. It can be found on Kegg map 195.
- `si_0602; Purine conversions` in the *Nucleosides and Nucleotides* rollup category. It can be found on Kegg map 230.
- `si_0851; dnaK Heat shock cluster` in the *Stress Response* rollup category. DnaJ is HSP-40 and dnaK is HSP-70. See the Wikipedia page and PMID 16952052.



Categories 11-20 include:

- `si_0043; Methionine Degradation` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 270
- `si_0279; Bacterial Cell Division` in the *Clustering-based subsystems* rollup category. It can be found on Kegg map 4112.
- `si_0640; Phosphate metabolism` in the *Phosphorous Metabolism* rollup category. Some description can be found in Gebhard, et al..
- `si_0600; De Novo Purine Biosynthesis` in the *Nucleosides and Nucleotides* rollup category. It can be found on Kegg map 230.
- `si_0448; FAS II; Fatty Acid Biosynthesis` in the *Fatty Acids, Lipids, and Isoprenoids* rollup category. It can be found on Kegg map 61 for biosynthesis, Kegg map 62 for elongation. The SEED description by Andrei Osterman describes fatty acid biosynthesis through FAS II, which is largely homologous to FAS I.
- `si_0060; Calvin-Benson cycle` in the *Carbohydrates* rollup category. It can be found on Kegg map 710.
- `si_0122; Serine-glyoxylate cycle` in the *Carbohydrates* rollup category. It can be found on Kegg map 630.
- `si_0746; RNA polymerase bacterial` in the *RNA Metabolism* rollup category. It can be found on Kegg map 3020.
- `si_0080; TCA Cycle` in the *Carbohydrates* rollup category. It can be found on Kegg map 20.

- `si_0032; Gln, Glu, Asn, & Asp Biosynthesis` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 250.

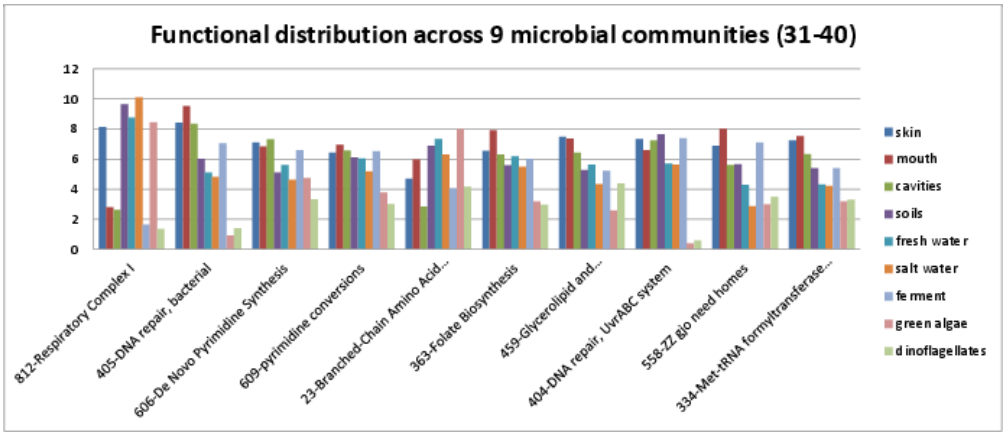


Categories 21-30 include:

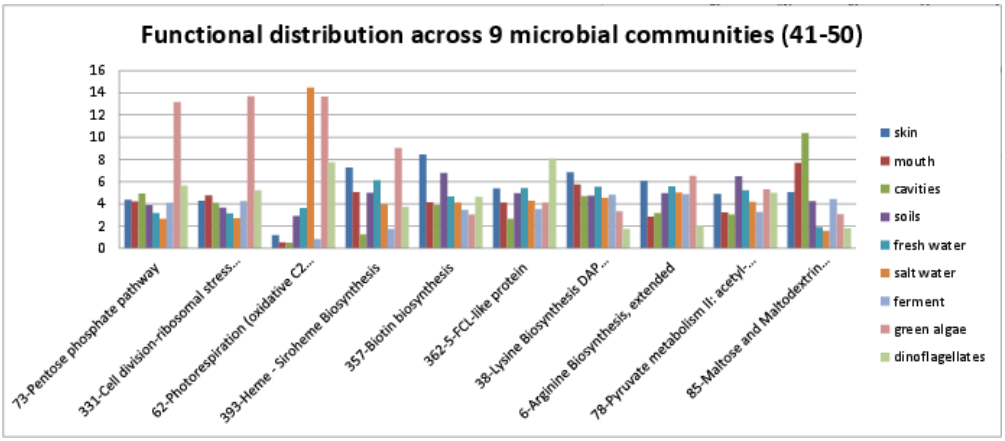
- `si_0662; Ribosome SSU bacterial` in the *Protein Metabolism* rollup category. It can be found on Kegg map 3010.
- `si_0526; Ton and Tol transport systems` in the *Membrane Transport* rollup category. Danese, et al. describe the Ton system to obtain iron in Brucella spp. Housden, et al. describe how Ton and Tol facilitate uptake of chelated iron or group B colicins through active transport.
- `si_0614; Ribonucleotide reduction` in the *Nucleosides and Nucleotides* rollup category. The Wikipedia article describes how ribonucleotide reductase converts RNA to DNA, maintaining an appropriate concentration of DNA throughout the cell cycle. The SEED description by Dmitry Rodinov desribes the three classes of ribonucleotide reductases.
- `si_0205; Peptidoglycan Biosynthesis` in the *Cell Wall and Capsule* rollup category. It can be found on Kegg map 550. The SEED description by Vassily Portnoy, Olga Vassieva, and Rick Stevens describes peptidoglycan biosynthesis.
- `si_0725; Queuosine - Archaeosine Biosynthesis` in the *RNA Metabolism* rollup category. The SEED description describes the synthesis and incorporation of the modified bases of tRNA, Queuosine and Archaeosine.
- `si_0377; Pyridoxin (Vitamin B6) Biosynthesis` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category. It can be found on Kegg map 750.
- `si_0065; Entner-Doudoroff Pathway` in the *Carbohydrates* rollup category. From Wikipedia, the Entner-Doudoroff pathway is a low-efficiency pathway to take glucose to pyruvate, found mostly in Gram-negative organisms, such as Pseudomonas, Rhizobium, Azotobacter, and Agrobacterium.
- `si_0077; anaplerotic pyruvate metabolism I: PEP` in the *Carbohydrates* rollup category. It can be found on Kegg map 20. See the Wikipedia article.
- `si_0701; ATP-dependent proteolysis in bacteria` in the *Protein Metabolism* rollup category.
- `si_0589; Ammonia assimilation` in the *Nitrogen Metabolism* rollup category. It can be found on Kegg map 910. The SEED description by Ed Frank, describes the glutamate dehydrogenase or GS-GOGAT pathways.

Categories 31-40 include:

- `si_0812; Respiratory Complex I` in the *Respiration* rollup category.
- `si_0405; DNA repair, bacterial` in the *DNA Metabolism* rollup category. The SEED description by Michael Kubal, describes a process of DNA base excision repair through detection, breakage, exonuclease activity, DNA polyerase, and DNA ligase.



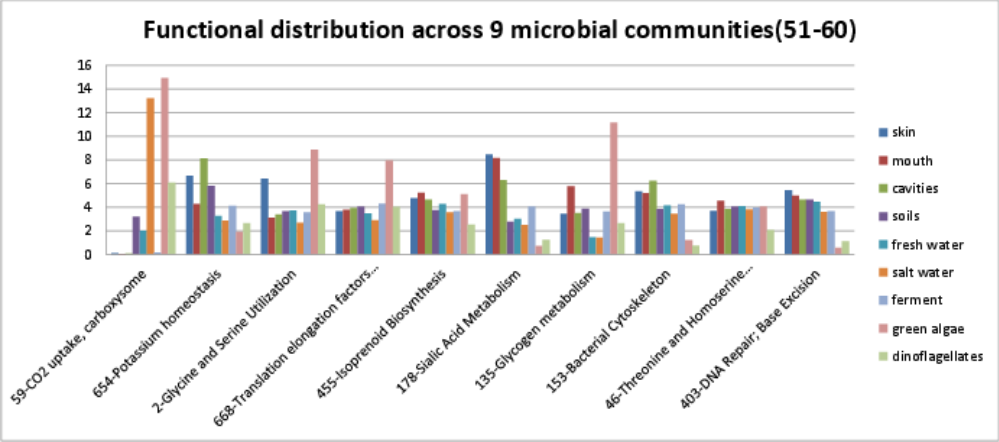
- `si_0606; De Novo Pyrimidine Synthesis` in the *Nucleosides and Nucleotides* rollup category. It can be found on Kegg map 240.
- `si_0609; pyrimidine conversions` in the *Nucleosides and Nucleotides* rollup category. It can be found on Kegg map 240.
- `si_0023; Branched-Chain Amino Acid Biosynthesis` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 290.
- `si_0363; Folate Biosynthesis` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category. It can be found on Kegg map 790. The SEED description by Valerie de Crecy-Lagard and Andrew Hanson describes folate biosynthesis.
- `si_0459; Glycerolipid and Glycerophospholipid Metabolism` in the *Fatty Acids, Lipids, and Isoprenoids* rollup category. It can be found on Kegg map 561 and Kegg map 564. The SEED description by Vasiliy Portnoy describes glycerolipid an diglycerophospholipid biosynthesis.
- `si_0404; DNA repair, UvrABC system` in the *DNA Metabolism* rollup category.
- `si_0558; ZZ gjo need homes` in the *Miscellaneous* rollup category.
- `si_0334; Met-tRNA formyltransferase gene cluster` in the *Clustering-based subsystems* rollup category.



Categories 41-50 include:

- `si_0073; Pentose phosphate pathway` in the *Carbohydrates* rollup category. It can be found on Kegg map 30.

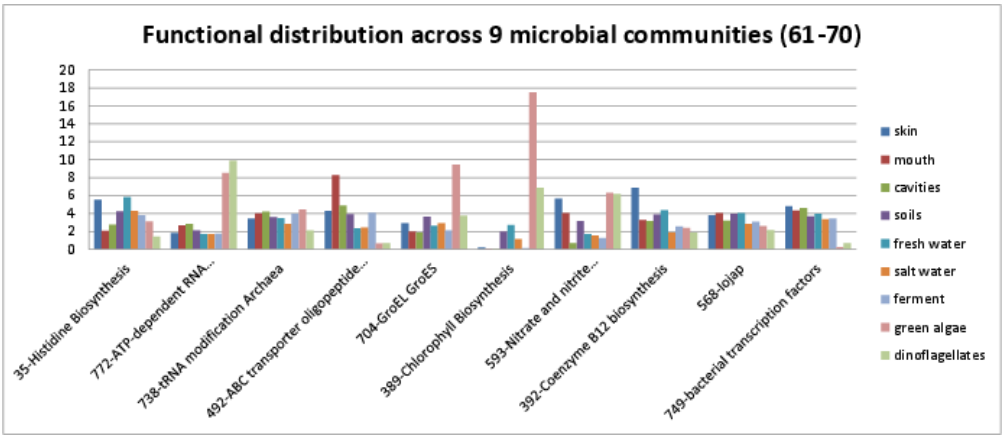
- `si_0331; Cell division-ribosomal stress proteins cluster` in the *Clustering-based subsystems* rollup category.
- `si_0062; Photorespiration (oxidative C2 cycle)` in the *Carbohydrates* rollup category.
- `si_0393; Heme - Siroheme Biosynthesis` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category. It can be found on Kegg map 860. The SEED description by Svetlana Gerdes describes tetrapyrrole biosynthesis.
- `si_0357; Biotin biosynthesis` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category. It can be found on Kegg map 780. The SEED description by Dmitry Rodionov, describes how biotin, vitamin H, which is an essential cofactor for a class of important metabolic enzymes.
- `si_0362; 5-FCL-like protein` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category.
- `si_0038; Lysine Biosynthesis DAP Pathway` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 300.
- `si_0006; Arginine Biosynthesis, extended` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 330.
- `si_0078; Pyruvate metabolism II: acetyl-CoA, acetogenesis from pyruvate` in the *Carbohydrates* rollup category. It can be found on Kegg map 620 and Kegg map 770.
- `si_0085; Maltose and Maltodextrin Utilization` in the *Carbohydrates* rollup category. It can be found on Kegg map 500.



Categories 51-60 include:

- `si_0059; CO2 uptake, carboxysome` in the *Carbohydrates* rollup category.
- `si_0654; Potassium homeostasis` in the *Potassium metabolism* rollup category.
- `si_0002; Glycine and Serine Utilization` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 260.
- `si_0668; Translation elongation factors bacterial` in the *Protein Metabolism* rollup category.
- `si_0455; Isoprenoid Biosynthesis` in the *Fatty Acids, Lipids, and Isoprenoids* rollup category. The SEED description by Olga Zagnitko, describes how the major terpenoid building blocks, isopentenyl diphosphate and dimethylallyl diphosphate, are produced by the mevalonate and non-mevalonate pathways.
- `si_0178; Sialic Acid Metabolism` in the *Cell Wall and Capsule* rollup category.

- `si_0135; Glycogen metabolism` in the *Carbohydrates* rollup category. It can be found on Kegg map 500.
- `si_0153; Bacterial Cytoskeleton` in the *Cell Division and Cell Cycle* rollup category.
- `si_0046; Threonine and Homoserine Biosynthesis` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 260.
- `si_0403; DNA Repair; Base Excision` in the *DNA Metabolism* rollup category.

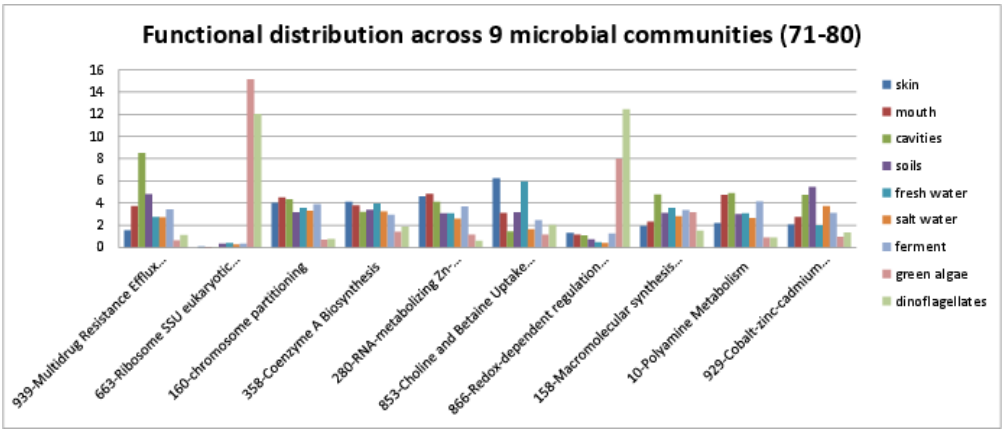


Categories 61-70 include:

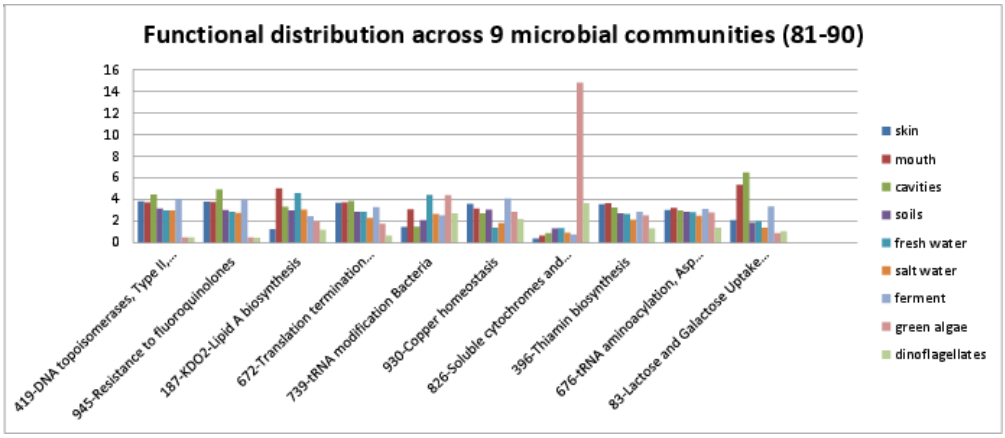
- `si_0035; Histidine Biosynthesis` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 340.
- `si_0772; ATP-dependent RNA helicases, bacterial` in the *RNA Metabolism* rollup category.
- `si_0738; tRNA modification Archaea` in the *RNA Metabolism* rollup category.
- `si_0492; ABC transporter oligopeptide (TC_3.A.1.5.1)` in the *Membrane Transport* rollup category.
- `si_0704; GroEL GroES` in the *Protein Metabolism* rollup category.
- `si_0389; Chlorophyll Biosynthesis` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category. It can be found on Kegg map 860. The SEED description by Svetlana Gerdes and Veronika Vonstein, describes the biosynthesis of chlorophyll, used in photosynthesis.
- `si_0593; Nitrate and nitrite ammonification` in the *Nitrogen Metabolism* rollup category. It can be found on Kegg map 910.
- `si_0392; Coenzyme B12 biosynthesis` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category.
- `si_0568; Iojap` in the *Miscellaneous* rollup category.
- `si_0749; bacterial transcription factors` in the *RNA Metabolism* rollup category.

Categories 71-80 include:

- `si_0939; Multidrug Resistance Efflux Pumps` in the *Virulence, Disease, and Defense* rollup category.
- `si_0663; Ribosome SSU eukaryotic and archaeal` in the *Protein Metabolism* rollup category.
- `si_0160; chromosome partitioning` in the *Cell Division and Cell Cycle* rollup category.
- `si_0358; Coenzyme A Biosynthesis` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category. It can be found on Kegg map 770.



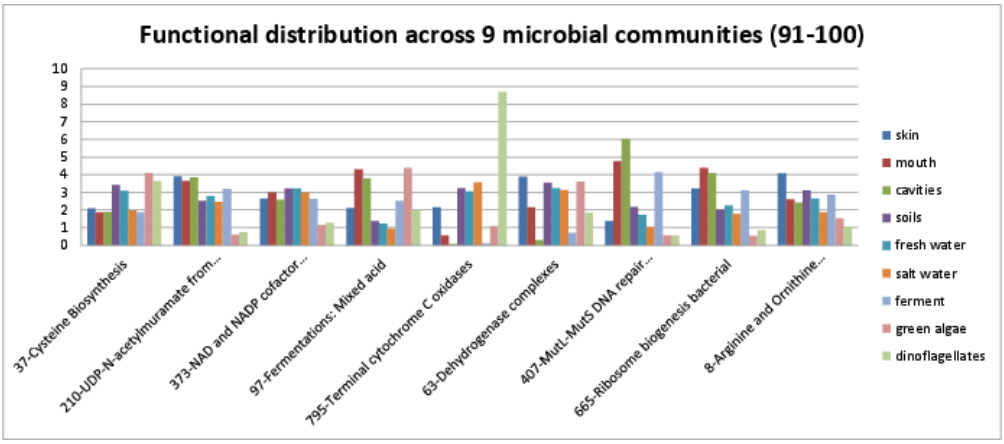
- `si_0280; RNA-metabolizing Zn-dependent hydrolases` in the *Clustering-based subsystems* rollup category.
- `si_0853; Choline and Betaine Uptake and Betaine Biosynthesis` in the *Stress Response* rollup category.
- `si_0866; Redox-dependent regulation of nucleus processes` in the *Stress Response* rollup category.
- `si_0158; Macromolecular synthesis operon` in the *Cell Division and Cell Cycle* rollup category.
- `si_0010; Polyamine Metabolism` in the *Amino Acids and Derivatives* rollup category.
- `si_0929; Cobalt-zinc-cadmium resistance` in the *Virulence, Disease, and Defense* rollup category.



Categories 81-90 include:

- `si_0419; DNA topoisomerases, Type II, ATP-dependent` in the *DNA Metabolism* rollup category. They are described in the Wikipedia article.
- `si_0945; Resistance to fluoroquinolones` in the *Virulence, Disease, and Defense* rollup category. Fluoroquinolones target DNA gyrase and topoisomerase IV. See Hooper, et al..
- `si_0187; KDO2-Lipid A biosynthesis` in the *Cell Wall and Capsule* rollup category.
- `si_0672; Translation termination factors bacterial` in the *Protein Metabolism* rollup category.
- `si_0739; tRNA modification Bacteria` in the *RNA Metabolism* rollup category.

- `si_0930; Copper homeostasis` in the *Virulence, Disease, and Defense* rollup category.
- `si_0826; Soluble cytochromes and functionally related electron carriers` in the *Respiration* rollup category.
- `si_0396; Thiamin biosynthesis` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category. The SEED description by Dmitry Rodionov describes how vitamin B1, thiamin, is synthesized.
- `si_0676; tRNA aminoacylation, Asp and Asn` in the *Protein Metabolism* rollup category.
- `si_0083; Lactose and Galactose Uptake and Utilization` in the *Carbohydrates* rollup category. It can be found on Kegg map 052.



Categories 91-100 include:

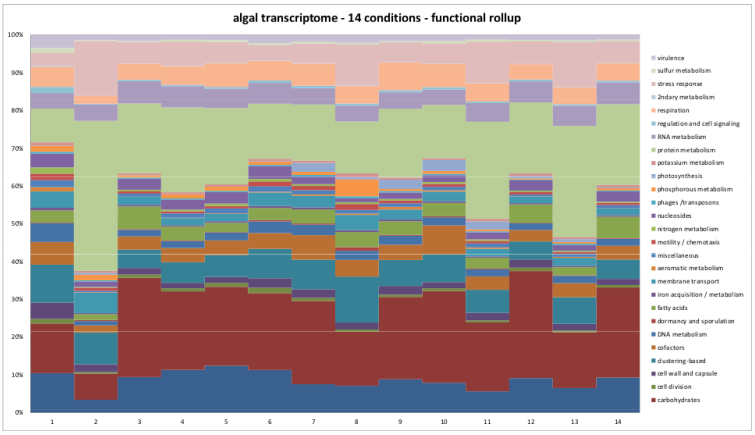
- `si_0037; Cysteine Biosynthesis` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 270.
- `si_0210; UDP-N-acetylmuramate from Fructose-6-phosphate Biosynthesis` in the *Cell Wall and Capsule* rollup category. The SEED description by Vasiliy Portnoy and Olga Zagnitko describes production of a major building block for biosynthesis of peptidoglycan.
- `si_0373; NAD and NADP cofactor biosynthesis ,global` in the *Cofactors, Vitamins, Prosthetic Groups, Pigments* rollup category. The SEED description by Andrei Osterman describes NAD and NADP biosynthesis.
- `si_0097; Fermentations: Mixed acid` in the *Carbohydrates* rollup category.
- `si_0795; Terminal cytochrome C oxidases` in the *Respiration* rollup category. It can be found on Kegg map 190.
- `si_0063; Dehydrogenase complexes` in the *Carbohydrates* rollup category.
- `si_0407; MutL-MutS DNA repair system, bacterial` in the *DNA Metabolism* rollup category. It can be found on Kegg map ko03430.
- `si_0665; Ribosome biogenesis bacterial` in the *Protein Metabolism* rollup category.
- `si_0008; Arginine and Ornithine Degradation` in the *Amino Acids and Derivatives* rollup category. It can be found on Kegg map 330.

7.3 Identifying enriched functions

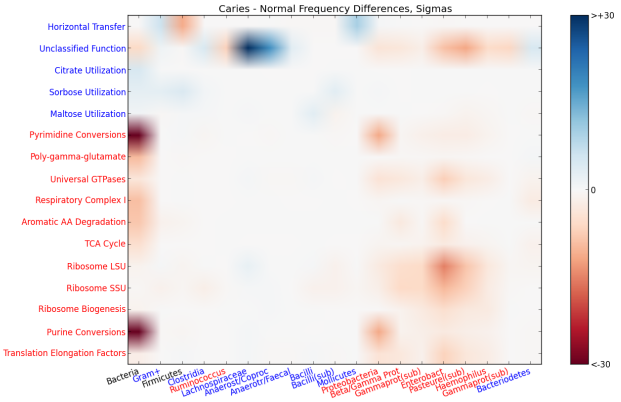
Using field replicates to identify important determinants of ecosystem function

face figure

Algal transcriptomes



Tooth decay study



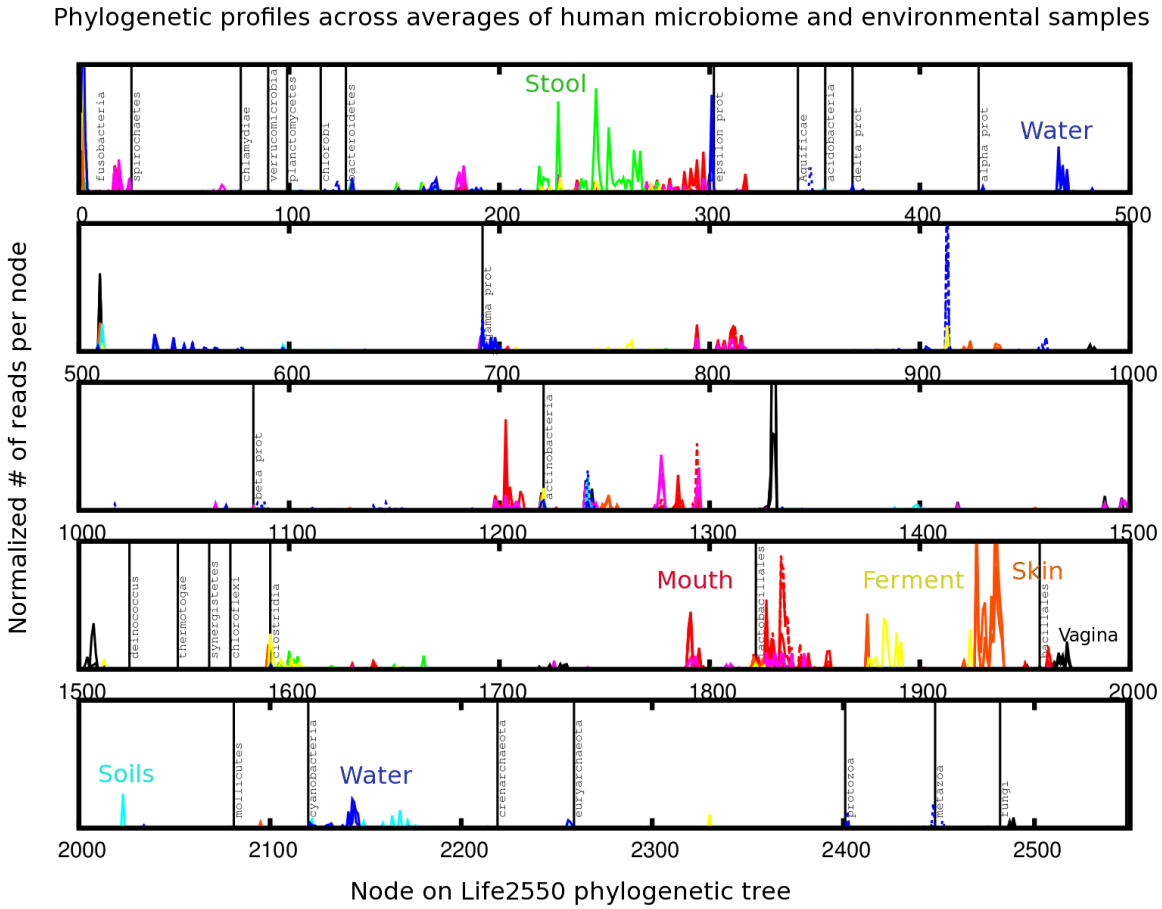
7.4 Visualizing distribution of functions

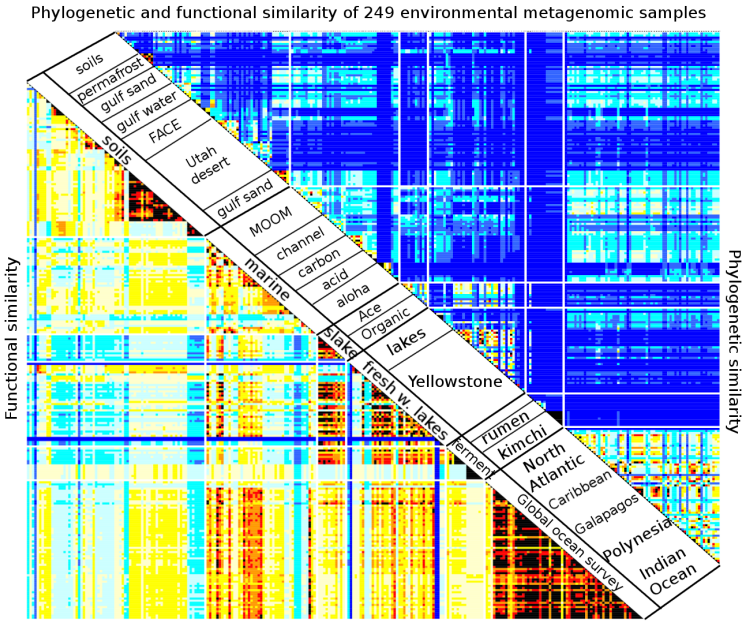
7.5 Profiling *Chlorella* kinases

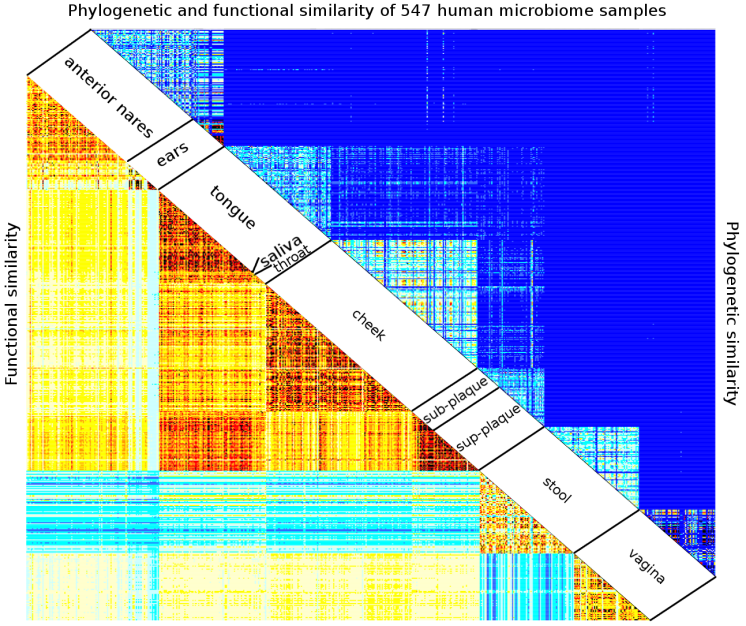
CHAPTER 8

Comparing Samples

In this chapter, we will show how to use R to compare two phylogenetic profiles using pairwise comparison, compare two or more types of environments using an anova analysis, compare multiple profiles using clustering and principal component analysis, and identify determinants functions and nodes that distinguish between two or more samples.




Phylogenetic and functional similarity of 249 environmental metagenomic samples
Functional similarity
Phylogenetic similarity
soils
permafrost
gulf sand
gulf water
FACE
Utah desert
gulf sand
MOOM
channel
carbon
acid
aloha
Ace
Organic
lakes
Yellowstone
fresh w lakes
ferment
rumen
kimchi
North Atlantic
Caribbean
Galapagos
Polynesia
Indian Ocean
Global ocean survey
marine
lakes


Phylogenetic and functional similarity of 547 human microbiome samples
Functional similarity
Phylogenetic similarity
anterior nares
ears
tongue
saliva
throat
cheek
sub-plaque
sup-plaque
stool
vagina

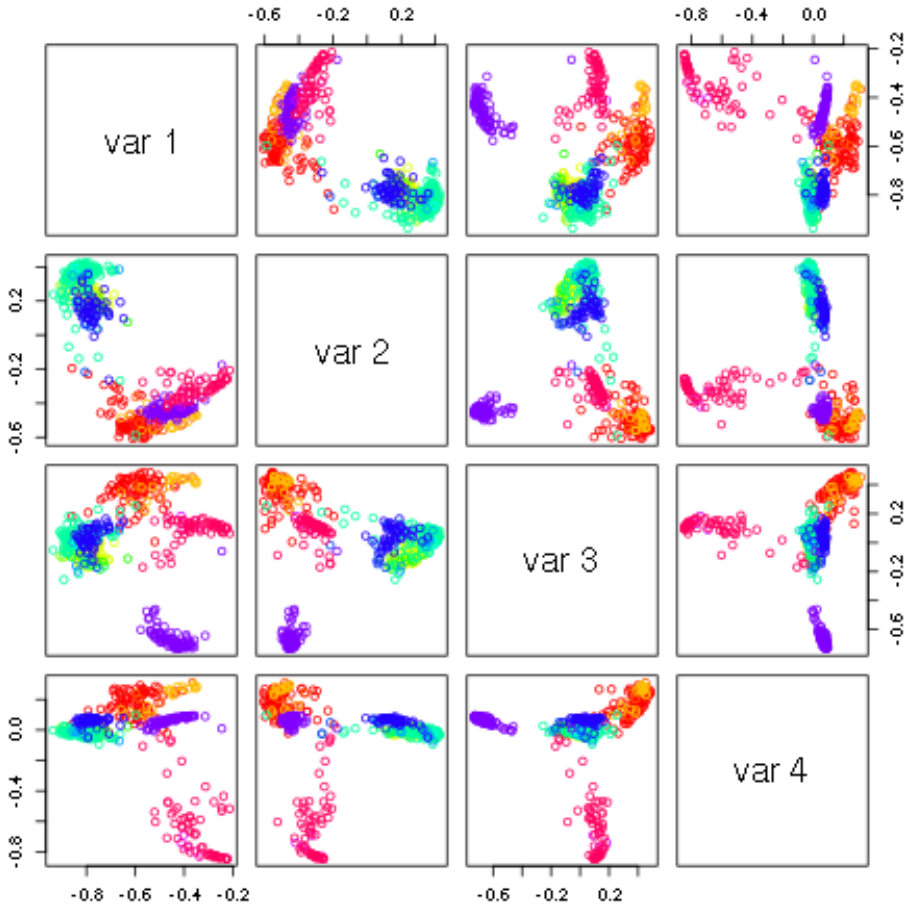
8.1 Scalar product of profiles

8.2 Principal component analysis

Principal component analysis is a common method to gain a broad understanding of how samples relate to one another. Starting with the human microbiome data from s.sets, we normalize the data with the commands:

```
hs <- apply(hmb$data,2,sum) # Hdat <- HmB raw data
Hnorm <- hmb$data/matrix(hs,dim(hmb$data)[1],dim(hmb$data)[2],byrow=T) #normalizing data -> all the c
HD <- sqrt(t(Hnorm))
HE <- eigen(t(HD) %*% HD)
H <- HD %*% HE$vector

nnn <- substr(rownames(H),1,4)
n1 <- unique(nnn)
v1 <- rainbow(length(n1))
names(v1) <- n1

pairs(H[,1:3],col=v1[nnn],xlim=c(-.9,.8),ylim=c(-.9,.8))
```

produces the following image.



to identify representative samples from the first two principal components, type:

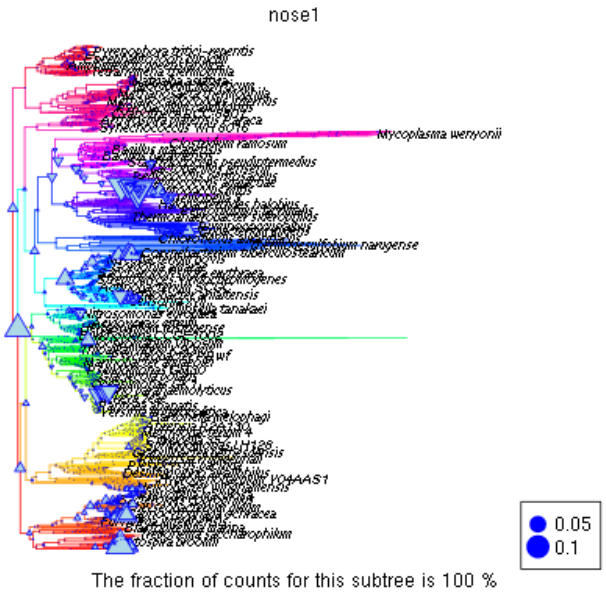
```
plot(H[,2],H[,1],col=v1[nnn],pch=19)
legend(0.35,-0.2,n1,col=v1,pch=20)
identify(H[,2],H[,1])
```

8.3 Ploting significant differences between two profiles

If we select samples 351 (subgingival plaque 8) and 235 (cheek 27) as reflective of the difference between cheek and plaque, the difference in phylogenetic profiles can be viewed by:

```
jnk=hmb
jnk$data[,1]=Hnorm[,351]-Hnorm[,235]
Plot.profile(jnk,1,"n0000",tip.frq=25,use.edge.length=T,type="phylo")
```

producing



where upward triangles reflect phylogenetic categories more prevalent in the plaque than on the cheek.

8.4 ANalysis Of VAriance

In this example, we analyze the 547 human microbiome. Of interest is to determine what distinguishs the sample types from one another, and what constitutes 'normal' variation within a sample type. To answer this question, we perform an analysis of variance for each node:

```
Hstab=asin(sqrt(Hnorm))    #stabilize the variances
cavity=factor(nnn)         #label the sample types
```

```
anova.cavity=lapply( apply( Hstab, 1, FUN=function(y,v) lm(y~v), v=cavity ), FUN=anova )    # perform
Pval      <- sapply( anova.cavity, FUN=function(x) x[1,5] )                                 # assess t
```

We note that the anova analysis is readily extended to comparing more than two types of metagenomic samples.

So far, we have shown what type of signal we are able to extract from the phylogenetic profiles. A question of equal interest is to quantify the noise in the profiles over repeated sampling. We have acquired ten samples soil bacterial samples in the Utah desert near Moab, five from sandy soils and five from shale. The samples have been taken at various distances apart, ranging from a few centimeters to tens of meters, to hundreds of meters.

From 16S surveys, we know that phylogenetic profiles of soil metagenomes show variations above and beyond sampling variability. Here, we will show how to quantify sample-to-sample variability. Such quantification are important when seeking to make inferences about the profiles.

We can use a chi-square goodness of fit test, to test for the uniformity of the proportion of reads recruited at a given node.::

```
# get data
data3 <- data.phylo[,48:52]
data4 <- data.phylo[,53:57]

ChiSq <- function(x){
   n.x <- length(x)
   zz  <- sum((x - mean(x))^2/mean(x))
   return(zz)
}

chi.sand  <- apply(data3,1,FUN=ChiSq)
chi.shale <- apply( data4,1,FUN=ChiSq)

boxplot( list(sand=chi.sand[!is.na(chi.sand)], shale=chi.shale[!is.na(chi.shale)]),log="y")
abline(h=qchisq(0.99,4),lwd=3,col=2)
```

The boxplot shows that at most nodes, the assumption of homogeneity across the replication does not hold. That is, the variations in the profiles at nearby locations exceeds simple sampling variability. To investigate the dependence of that extra variation on the nodes, we can plot the chi square statistics in sands and shale.::

```
plot(chi.sand,chi.shale,xlab="sand",ylab="shale",pch=20,log="xy")
abline(c(0,1))
```

We see that nodes that have larger sample-to-sample variations in sands also tend to have larger sample-to-sample variations in shale. We can visualize which nodes have the largest sample to sample variation (say in sandy soils).::

```
siz <- 2*sqrt(chi.sand/max(chi.sand,na.rm=T))
library("ape")
source("plotOnPhylo.r")
tree <- PlotOnTree(siz,1,20)
```

This display indicates that the root node is the most variable. Why? Furthermore, we can zoom into the branch that shows a string of relatively large variability to see if it is a poorly sampled region of the tree of life.

8.5 Clustering multiple profiles with R

Multiple phylogenetic profiles can be grouped for similarities using k-means clustering. We demonstrate this type of analysis on all the phylogenetic profiles simultaneously. This provides an opportunity to investigate the hypothesis

that metagenomes from similar environments have similar phylogenetic profiles. Our clustering analysis is done in several steps.

We start by transforming each profile into a probability distribution, taking the square root to stabilize the variance, and approximatively center each node::

```
#  normalize each column and take a square root transform
#  and remove columns with small counts

D <- dim(data.phylo)
n <- apply(data.phylo,2,sum)
keep <- n > 40000
N <- matrix(n[ keep ],D[1],sum( keep ),byrow=T)
profile.phylo <- sqrt(data.phylo[, keep ]/N)
mprofile.phylo <- matrix(apply(profile.phylo^2,1,mean),D[1],D[2])
rprofile.phylo <- profile.phylo - sqrt(mprofile.phylo)
```

The k-means clustering uses a random starting value for the cluster centers. As a result, each time k-means is used, a different solution is found. To ensure reproducibility, we set the random seed and sort the clusters in decreasing size.::

```
set.seed(9111)
#  number of cluster
n.cluster <- 18
kmr.phylo <- kmeans(t(rprofile.phylo),n.cluster)

# sort clusters in decerasing size and sort the
# the profiles to put together profiles in the same cluster
idx0 <- kmr.phylo$cluster
tabl <- sort(table(idx0),decreasing=T)
lab.new <- 1:max(idx0)
names(lab.new) <- names(tabl)
idx1 <- lab.new[as.character(idx0)]
idx.sort <- sort.list(idx1)
rclust.phylo <- rprofile.phylo[,idx.sort]

# transform the data to improve the dynamic range
rclust.phylo <- sign(rclust.phylo)*log(abs(rclust.phylo))

# plot the grouped profiles
image(1:sum(keep),1:D[1],t(as.matrix(rclust.phylo)),
       xlab="environment",ylab="profile",col=topo.colors(20))
clust.boundary <- cumsum(tabl)
abline(v=clust.boundary+0.5,col=1,lty=1)
```

One can cross-walk the cluster labels with the labels in the dataset.::

```
kmr.label2 <- split(label2[keep],idx1)
kmr.label1 <- split(label1[keep],idx1)
```

The correlation between the profiles (normalized dot-product) allows for the visual display of how close profiles within a cluster and profiles in other clusters, are.::

```
#-----------------------------------------------------
#  inner product matrix

rprofil2.phylo <- as.matrix(rprofile.phylo[,idx.sort])
dD <- t(rprofil2.phylo) %*% rprofil2.phylo
nd <- apply(rprofil2.phylo,2,FUN=function(x) sqrt(sum(x^2)))
cC <- dD/outer(nd,nd,FUN="*")
```

```
brks <- seq(-1,1,length=20)

image(1:sum(keep),1:sum(keep),cC,xlab="",ylab="",
   breaks=brks,
   col=heat.colors(19))
abline(v=clust.boundary+0.5)
abline(h=clust.boundary+0.5)
abline(c(0,1))
```

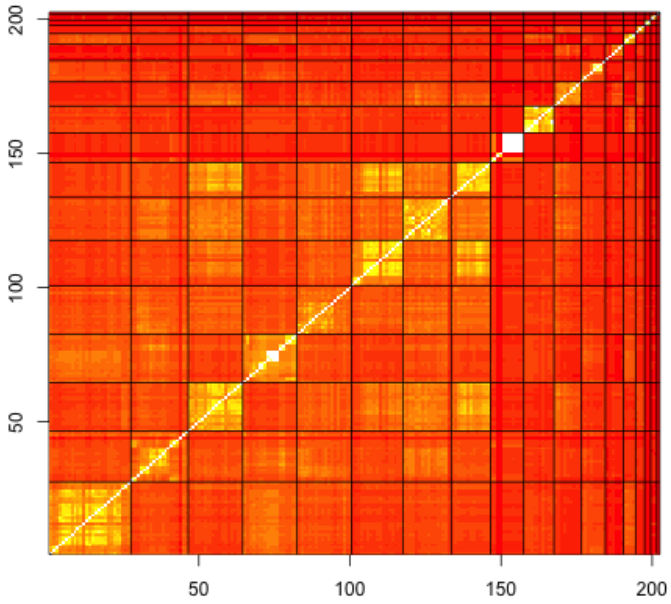


Fig. 8.1: Distance between profiles, grouped in clusters.

8.6 Visualizing the clusters with R

Principal Component Analysis is a useful statistical tool for dimensionality reduction. The two first eigenvectors are the two orthogonal directions of largest variation, and provide a basis to produce 2d plots of our 402 dimensional profiles. It is instructive to color the each profile by its cluster id. In addition, each point can be given a colored halo corresponding to its environmental type. Such a plot can help display the clusters and may help with the interpretation.::

```
#--------------------------------------------------
#  exploratory data analysis with principal component

PC <- prcomp(t(rprofile.phylo),center=F)
rC <- kmr.phylo$center %*% PC$rotation

#  look at what the clustering returns
#  and compare it with the labels
```

```
set.seed(321)
ccol <-  sample(rainbow(n.cluster))
ccol2 <- sample(rainbow(length(table(label1))))
names(ccol2) <- names(table(label1))

par(mfrow=c(1,1))
plot(PC$x[,1],PC$x[,2],xlab="pc1",ylab="pc2",
    col=ccol[idx0],pch=20,cex=0.75)
points(rC[,1],rC[,2],col=ccol,pch=18,cex=1.5)
points(PC$x[,1],PC$x[,2],col=ccol2[label1],cex=1.25,lwd=2)
legend(-0.35,-0.2,names(ccol2),col=ccol2,pch=1)
```

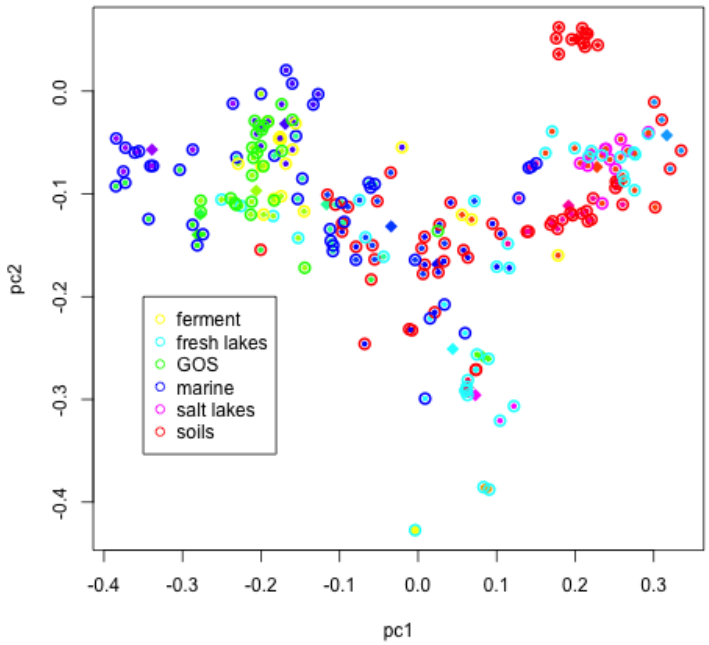


Fig. 8.2: Environmental samples dispersed by first two principal components, colored by cluster.

It can be useful to look at similar plots for a few more eigenvectors. The following R `script` script does just that.

8.7 Niche determinants, gulf oil

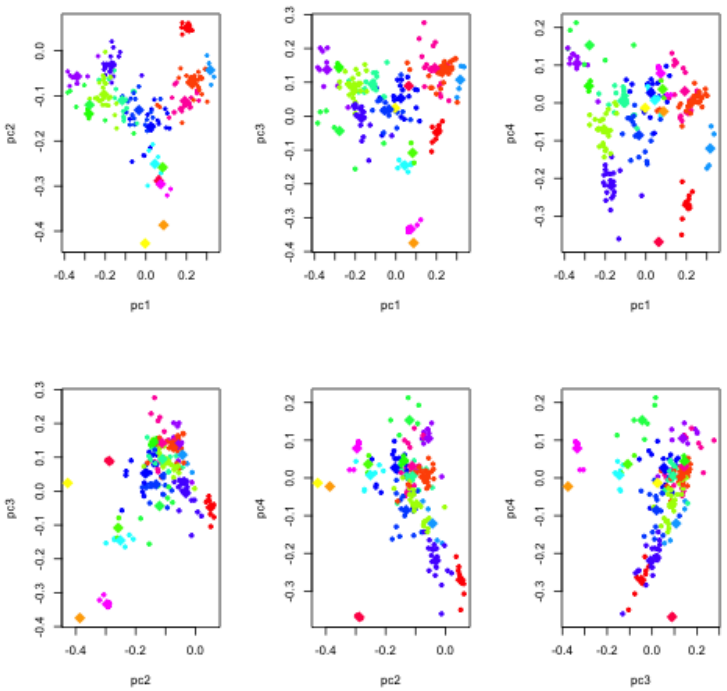
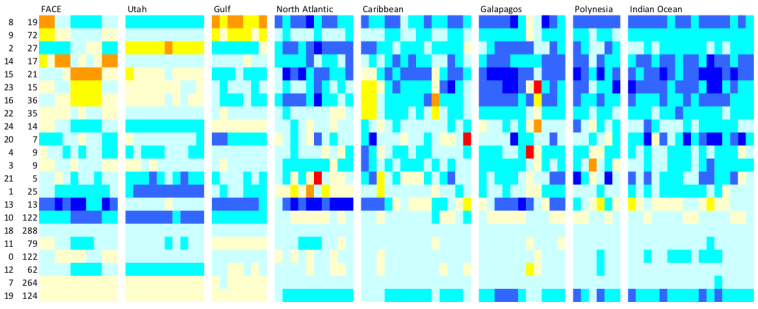


Fig. 8.3: Environmental samples dispersed by next several components, colored by cluster.



CHAPTER 9

Annotated reads

9.1 Obtaining annotated reads

Running Sequescan with the database writer flag enabled (-f 1 on the command line, selection box on the GUI version):

```
sequescan run -d Life2550-40GB.0 -s seed_0911.m1 -f 1 -t 8 -o out . >& err &
```

will produce an output file of annotated reads, written in the correct reading frame to be translated into a peptide sequencing including a signature peptide. Reads matching signature peptides in multiple reading frames will be output once in each frame. If the input sequence data are fasta or fasta.gz, the output will be fasta (.fa). If the input sequence data are fastq or fastq.gz, the output will be fastq (.fq). For example, here are the first three reads output from the Mock even illumina community using the Life2550-40GB.0 data module:

```
@SRR172902.38_3 node=n0990      assign=3        unq_k=5 func=si_0083
NACCTCAAGTTTGAAAAAGAGCACAACTGGACCAATTATCCAAAAGGTGTCCTTCATTTCTTGCAANAAGCTGGGC
+SRR172902.38 USI-EAS376:1:1:2:817 length=75
::BCCB@@CBCCCB?BAACBCB;BA@BBCBBBCA?CBBBCC@=@/@B;B?CCBC?;@BBBBBCB67%-;BBBBAA>
@SRR172902.39_4 node=n0586      assign=3        unq_k=6 func=si_0533
TCTATGATTNCAGCGGCATCAATGGCAGGTGTTCCTCTCACTAATGGCTTTATTTCTAAGGAAATGTTCTTTACC
+SRR172902.39 USI-EAS376:1:1:2:208 length=75
9>9:;A:<5%=ABBBB@;9B?BAABB?BBBA@BBA@B@B=BAAABBBBAB>BA0@BBBBCCBBBBBBBBBBB>BB
@SRR172902.40_5 node=n0900      assign=2        unq_k=3 func=si_0566
NNCGGCGCGCGNTGGGGGGACGGGGAGTGCGTGACCGACAAGGGCGTGGTGACAAGCCGCAAGCCCGACGACCTGCC
+SRR172902.40 USI-EAS376:1:1:2:1659 length=75
###############B@7>:*,:(>@6@9<+:A<+6=>19:(>B@4@6@B?B<<AAABBBABBB>ABBABB?BBABAB
```

since every Sequedex-annotated read will contain a nodeID, this can be changed to fasta format with:

```
grep -A1 node= db-Life2550-40GB.0xseed_0911.fq |tr '@' '>'| grep -v ^-- > reads.fas
```

producing:

```
>SRR172902.38_3 node=n0990      assign=3        unq_k=5 func=si_0083
NACCTCAAGTTTGAAAAAGAGCACAACTGGACCAATTATCCAAAAGGTGTCCTTCATTTCTTGCAANAAGCTGGGC
>SRR172902.39_4 node=n0586      assign=3        unq_k=6 func=si_0533
TCTATGATTNCAGCGGCATCAATGGCAGGTGTTCCTCTCACTAATGGCTTTATTTCTAAGGAAATGTTCTTTACC
>SRR172902.40_5 node=n0900      assign=2        unq_k=3 func=si_0566
NNCGGCGCGCGNTGGGGGGACGGGGAGTGCGTGACCGACAAGGGCGTGGTGACAAGCCGCAAGCCCGACGACCTGCC
```

note the 'Ns' at the start of two of the three reads, shifting into the correct reading frame.

If you have the EMBOSS utilities installed, it is possible to obtain a fasta file of the peptides identified in each read:

transeq =frame 1 reads.fas pep.fas

at which time the coverage can be estimated by searching for RNA polymerase reads (functional category si_0746) or a universally conserved amino acid motif in the RNA polymerases, GG.R.GEME:

grep si_0746 pep.fas|wc grep GG.R.GEME pep.fas|wc

9.2 Phylogenetic filtering and assemblies of genomes

9.3 Functional filtering

Reads from a particular functional category can be obtained by specifying also the functional category, such as the one for the ribosomal (and tRNA) reads:

```
grep -A1 node= db-Life2550-40GB.0xseed_0911.fq |tr '@' '>'| grep -A1 si_0962| grep -v ^-- > out.fas
```

These reads can be used with the ribosomal database (see *The Ribosomal Database Project*) or placed into alignments to gain a better understanding of the community composition of samples and how they relate to database of reference genomes and one another.

9.4 Translated reads

9.5 Gene-specific BLAST databases

9.6 Placing reads into a multiple sequence alignment: conserved part of RNAP

9.7 Obtaining specific genes

9.8 Targeted assemblies of genes

see *Velvet and de-novo assembly*

9.9 Nucleotide alignments

9.10 Obtaining ribosomal (and tRNA) reads

9.11 Aligning to a reference genome

CHAPTER 10

Reference

10.1 Command line options for sequescan

The arguments for sequescan can be listed by typing:

```
~/sequedex/bin/sequescan run -h
```

at the comand line, resulting in the following output:

```
Command line execution of sequescan run mode:
sequescan run [-h] [-q] [-c config_file] -d data_module [-o output_directory] [-s function_set]
          [-a min_prot_frag_length] [-t thread_num] [-f database_writer_flag] [-l] INFILE

Example:
sequescan run -d Life2550-4GB.0 -s seed_0911.m1 -f 1 /Users/jsmith/mgData

Option descriptions:
-h    mode help
-q    quiet option - less messages to console or progress window
-c    user-defined configuration file (overrides system configuation file)
-d    name of data module
-o    user-defined directory for data output (default is directory where input is located)
-s    name of function set
-a    minimum protein fragment length (overrides configuration file; default is 15)
-t    maximum number of threads in threadpool (default = 1)
-f    database writer flag (arguments:  0 = no, 1 = yes); analysis_writer_list in config determines t
      (currently fasta/fastq file)
-l    required if INFILE contains list of fasta/fastq files; if argument is "none", the list contains
      otherwise argument is base directory and paths in list are relative to base directory;
      when paths are relative to base directory and the -o option is set, output will inlcude relativ

INFILE may be a fasta/fastq file, a directory with fasta/fastq files, or a file containing a list of
However, only fasta or fastq files or their gzipped (.gz) versions with an extension in the config fi
fa_ext_list will be processed.

Parameters:
  INFILE is the input file in FASTA format.  Complete paths, relatives paths, and symlinks
  may be used here.
```

10.2 Environmental variables

The environment variables can be seen by typing:

```
sequedex-config
```

producing:

```
Global Configuration Variables:
                SEQUEDEX_USERDIR (user directory location) set to "/home/localhost/username/.sqdx/" f
              SEQUEDEX_LOGLEVEL (debug|info|warning|error) set to "info" from default.
           SEQUEDEX_LOGFILE_DIR (location of launcher log file) set to "/home/localhost/username/.sq
           SEQUEDEX_SEES_STDOUT (if False, pop up windows) set to "True" from default.
                  SEQUEDEX_HOME (top-level installation directory) set to "/home/localhost/username/
                   SEQUEDEX_ETC (location of system config files) set to "/home/localhost/username/s
                   SEQUEDEX_LIB (location of library files) set to "/home/localhost/username/sequede
                   SEQUEDEX_BIN (location of executable files) set to "/home/localhost/username/sequ
                  SEQUEDEX_DATA (location of data modules) set to "/home/localhost/username/sequedex
                   SEQUEDEX_DOC (location of documentation and help files) set to "/home/localhost/u
               SEQUEDEX_CONTRIB (location of contributed files and data) set to "/home/localhost/use
                  SEQUEDEX_JAVA (absolute path to java executable) set to "/usr/bin/java" from outpu
    SEQUEDEX_CHECK_JAVA_VERSION (if True, check java version) set to "True" from default.
  SEQUEDEX_REQUIRE_JAVA_VERSION (java version must be above this) set to "7" from default.
      SEQUEDEX_PLATFORM_MEMSIZE (amount of system RAM in GB) set to "31" from reported system memory
                SEQUEDEX_PYTHON (absolute path to python executable) set to "/usr/bin/python" from f
             SEQUINATOR_COMMAND (command run upon "Display Output") set to "/home/localhost/username
             SEQUINATOR_BROWSER (path to browser, with arguments) set to "/usr/bin/firefox" from pla
              SEQUINATOR_SERVER (write and serve files) set to "True" from default.
              SEQUINATOR_CLIENT (start browser) set to "True" from default.
          SEQUEDEX_HAS_INTERNET (has web access for update) set to "True" from default.
                SEQUINATOR_DATA (location of data library) set to "/home/localhost/username/sequedex
          SEQUINATOR_SEARCH_PATH (path searched for output) set to "/home/localhost/username/:/home/l
                SEQUINATOR_HOST (IP address for sequinator to use) set to "127.0.0.1" from default.
                SEQUINATOR_PORT (IP port for sequinator to use) set to "52707" from default.
    SEQUINATOR_MAX_BROWSER_TABS (max tabs for sequinator to open) set to "10" from default.
               SEQUINATOR_TITLE (format string for sequinator titles) set to "%(filename)s" from def
      SEQUEDEX_LAUNCHER_VERSION (sequedex-launcher version number) set to "1.0.10" from internal ver
SEQUEDEX_LAUNCHER_DEFAULT_COMMAND (default command) set to "sequescan" from default.
                   SEQUEDEX_WWW (URL for updates and sequinator) set to "http://sequedex.lanl.gov" f
                     http_proxy (system http proxy setting) set to "None" from default.
                      ARCHY_ETC (location of Archaeoptryx config file) set to "/home/localhost/usern
            ARCHY_DEFAULT_TREE (default tree for archy) set to "/home/localhost/username/sequedex/c
                 SEQUESCAN_ETC (location of sequescan conf files) set to "/home/localhost/username/
          SEQUESCAN_HEAPSIZE_MB (java heapsize in MB for sequescan) set to "30000" from user config
            SEQUESCAN_JAVA_ARGS (full java arguments to sequescan) set to "-Xms1000m -Xmx30000m" fro
Data Module Configuration Variables:
  Module virus1252:
                  minimumMemSize: 4
                      moduleName: virus1252
                     installDate: Sat Jun 21 08:30:56 2014
                        filename: virus1252.1.jar
                         version: 1
                     nextMemSize: 1000
  Module Life2550:
                  minimumMemSize: 16
                      moduleName: Life2550
                     installDate: Sat Jun 21 08:30:55 2014
                        filename: Life2550-16GB.0.jar
```

```
                        version: 0
                    nextMemSize: 32

Use the sequedex-config command for setting these variables.
```

To check a specific environmental variable, add the variable as an argument to sequedex-config:

```
sequedex-config SEQUESCAN_HEAPSIZE
```

and to change a particular variable, provide both the variable name and a new value:

```
sequedex-config SEQUESCAN_HEAPSIZE 3000
```

10.3 Configuration options

The configuration options for sequescan can be seen by typing:

```
cat sequedex/etc/sequescan/sequescan.conf
```

producing:

```
; Windows style configuration file (INI file) where
; Properties:  name=value
; Sections:    [section]
; created by M. Bussod
; last modified 03/18/2014 by J. Cohn
; should always be current version


; Parameters in this configuration file always evaluate to string values unless the parameter
; name ends in one of the following suffiexes:
;    *_int:     evaluates to an integer type.
;    *_bool:     evaluates to a boolean type.
;    *_float:     evaluates to floating type.
;    *_intList:     evaluates to a comma-separated list of integer types.
;    *_boolList:     evaluates to a comma-separated list of boolean types.
;    *_floatList:     evaluates to comma-separated list of floating types.

; extensions of possible output files (files generated during a run depend upon options selected):
;    .tsv - tab-separated files as documented in the Sequescan design document.
;    .fa   - nucleic acid fasta file of matching reading frames
;    .fq   - nucleic acid fastq file of matching reading frames
;    .log - logging file.
;    .json - JSON file
;    .json and .tsv files for the most part have the same content in different formats


; runtime values used by this config file
;    DBNAME         ; signature data module name for this analysis from the -d option
;    SCHEME        ;  function scheme for this analysis from the -s option

; environmental variables used by this config file (if not set, launcher script will set default valu
;    SEQUEDEX_HOME      ; path to the Sequedex distribution directory
;    SEQUINATOR_COMMAND ; full path to the command which launches the sequinator program (currently a
;    SEQUEDEX_USERDIR   ; default is currently ~/.sqdx on Linux and Mac
;    SEQUESCAN_ETC ; path to the Sequescan etc directory (which is where, for example, the default se
```

```
; Strings enclosed in matching pairs of percent signs will be passed for environmental variable expan
; Where paths are relative in config file, on Linux these will be relative to working directory
; when using Mac app, they will be relative to /? (since working directory for app is /)


[global]
nCPUS_int=1       ; allows usage of n processing threads (if valid license only);  default is 1;
                  ; overridden by command line option
min_prot_frag_len_int=15  ; minimum length of protein fragments in amino acids between stop codons
                          ; overridden by command line option -a
config_file_version=1.0  ; config file version number


[licensed_features]
system_license_file=%SEQUESCAN_ETC%/license.lic
       ; if user cannot write to this location,
       ; license should be installed in $SEQUEDEX_USERDIR/license.lic
           ; gui will install license in $SEQUEDEX_USERDIR/license.lic
       ; program will always look for license in $SEQUEDEX_USERDIR first, then in system_license_file
write_db_bool=F   ; analysis writer(s) are only added if set to T (true)
                  ; write_db_bool=T is ignored if no valid license
                  ; overridden by command line option

[input]
fasta_ext_list=fasta,fst,fna,fas,ffn,fa,fastq,fq    ; allowed file extensions for input sequencing fi
                                                                                          ; fas

[output]
log_dir=log       ; output directory for sequescan log file relative to the top-level output director
out_dir_ext=sqdx   ; extension for lowest level output directory
who=who-%DBNAME%         ; count of reads assigned to each interior nodes.
what=what-%DBNAME%x%SCHEME%  ; fractional count of reads assigned to functional scheme classificatior
whoDoesWhat=wdw-%DBNAME%x%SCHEME%   ; matrix of fractional counts of reads assigned by functional sch
     ; Rows are scheme classifications columns are phylogeny.
     ; Sum across columns is the what vector.
stats=%DBNAME%-stats            ; file of general and phylogeny statistics needed for normalization
whatstats=whatstats-%DBNAME%x%SCHEME%
db=db-%DBNAME%x%SCHEME%    ; core name for database output file (currently only options are for fasta
             ; if run with no functions, %SCHEME% will be substituted with "none"

progress_writer_list=gov.lanl.sequtils.writer.ProgressFileWriterJ,gov.lanl.sequtils.writer.ProgressFi
   ;list of summary/stats file writers to be used for each sequence file analyzed
analysis_writer_list=gov.lanl.sequtils.writer.SequencingFileWriter
   ; list of analysis writers to be used for each sequence file analyzed
analysis_output_type=same_as_input  ;  analysis writer output type - same_as_input, fasta (.fa), fast
      ; currently fastq will only work with fastq input but can force fasta output from fastq input
analysis_top_node_int=0     ; analysis writer output will include reads assigned to this node
      ; and all children nodes under it - 0 means all nodes - not yet working but will be overridable
progress_interval_long=1000000      ; if set to 0, will print summary statistics at end of processing
                           ; if n > 0, summary statistics will be written after every n reads have be
result_display_path=%SEQUINATOR_COMMAND%
```

CHAPTER 11

Additional software tools and resources for use in conjunction with Sequedex

In this chapter, we introduce the user to a variety of software tools and packages and how we have found them to be useful in the process of organizing data files and understanding the output. This is neither intended to be an exhaustive list of software packages nor a tutorial in using any of them; it is simply a short sketch of how we used widely available software to better understand the output of Sequedex.

Indeed, if the user has difficulties with the tools below, it is possible to choose from a wide array of open source analysis tools, such as those that have been described on Wikipedia and on seqanswer to identify an alternate method of acheiving the a similar result.

11.1 Graphing and sorting with Excel

Spreadsheets are popular tools for visualizing and comparing tabular outputs such as the phylogenetic and functional profiles produced by Sequedex. Excel is widely used on Windows and Mac operating systems, and can be run on Linux under the Windows Emulator (WINE, http://www.winehq.org/). WINE is also useful for running other Windows applications under Linux, such as *BioEdit and reference alignments*, but the open source spreadsheet programs Gnumeric and LibreOffice will also serve this function. Numerous books are available for Excel, and Gnumeric and LibreOffice both have extensive online documentation, here and here.

If you open (or 'data -> import') the *Output files - Who?*, *Output files - What?*, or *Output files - Who does what?* or *Output files - stats* files, it will be possible to see which categories are most-populated, to compare the different types of k-mer matches for the who file, and extract columns from different samples for direct comparison and statistical analysis. It is also possible to use the 'consolidate' function to perform *Phylogenetic rollups* and *Functional rollups*, although the user may want to further bundle the rollup categories or sort and discard them to simplify the plots. We supply programs, below, that perform the rollups shown in *Phylogenetic rollups*, which combines a number of multiphyla nodes into a single category.

Alternatively, the user can save output from multiple samples in a tab delimited text file to be compared with data from the Sequedex documentation or previously analyzed samples. When doing this, care must be taken to compare compatible samples, where the rows have the same meaning, and to ensure the user is able to provide column labels which adequately document which sample and analysis the data are from. The tree name is automatically used as part of the who, what, and wdw files, and the functional classifications are included in the what and wdw files, to help keep track of what analysis was performed.

11.2 How to install Cygwin and what to include.

Cygwin is a software package that provides a Unix-like environment under a native Windows operating system. It is available from http://cygwin.org/. During installation, you will be asked to select from about 30 categories of programs to install. While it is simple to rerun the setup.exe script at any time, if you plan to run a variety of bioinformatics tools it may be easiest to install Devel, Editors, Interpreters, Perl, and Python, since you often need to re-compile software developed for Linux from source code. Installing everything will require over 1 GB of disk space and typically require several hours of download time.

Once installed, Cygwin provides many of the functions of a command line with standard tools, discussed in the next section, including running the Sequedex executable, as described in the *Running Sequedex from the command line* section, although the s.install.cyg are distinct from eithr Linux or Windows without Cygwin. Note the the GUI version of Sequedex can be initiated by typing 'sequescan' on the command line of Cygwin, once Java and the path variables are correctly set.

Many programs, for example R, Python, and Matlab, can be installed and used either from within Cygwin or as a native Windows application in an integrated development environment (IDE). Typically the IDEs have numerous user-friendly features that you may find helpful. There is no difficult in mixing the different types of tools (native Windows and Cygwin), although navagation back and forth between the Windows home directory (My Documents) and the Cygwin home directory (typically c:/cygwin/home/username/) can be awkward. It is possible to put a shortcut (or link) to quickly navigate back and forth between these two frequented folders. Also, path names are treated differently in the two systems: see http://cygwin.com/cygwin-ug-net/using.html#using-pathnames.

11.3 The command line with standard tools (Linux, Mac, Cygwin for Windows)

The command line is not so much a software program as a flexible interface to the computer's operating system. It enables a wide variety of software to be run on a variety of input files, placing the output in well defined places. As described in *Directory structures for data, output files, and analysis*, appropriate choices in where to place files for projects and samples can greatly facilitate the automation of analysis of large numbers of files. If scripts are used to perform the analysis, a record will be created of exactly what was done to each data set. For example, we have already described in *Running Sequedex from the command line* how to run Sequedex from the command line.

Two of the most popular 'shells' for the command line are bash and tcsh, and if the user is new to its use, he or she is encouraged to locate an introductory book and a colleague who is willing to help you from time to time. We will only describe the features needed to perform the Sequedex analysis tasks described in this manual.

- The filesystems can be viewed and navigated with df -h, ls -lthr, and cd, for example, remembering that the 'tab' key normally auto-completes filenames or directory names.
- Files (such as the Sequedex outputfiles, which are tab-delimited, can be viewed with cat, head, or grep Percent.
- The functional categories can be sorted on the screen by decreasing prevalence with:

```
sort -n -k3 -r what-Life2550-32GB.0xseed_0911.m1.tsv | less
```

- The phylogenetic categories can be sorted on the screen by decreasing prevalence with:

```
sort -n -k3 -r who-Life2550-32GB.0.tsv |less
```

- A help page can be found for many commands by typing 'man command', such as 'man less' or 'man awk'
- Results from one command can be sent to another command with the pipe, or '|', and to a file with '>'.
- Commands can be iterated over a list with:

```
for i in file1 file2 file3;do
  awk '{print $3}' < $i.fas.sqdx/what-Life2550-40GB.0xseed_0911.m1.tsv > $i.out ; done ; paste *
done
```

- Commands can be saved in a file, and be run in batch by typing 'source filename'. A short introduction is available here. Note that variables can be passed to the script from the command line.
- The utilities ps2pdf, convert, and pdfunite are convenient tools for changing among image formats.

A simple text editor that is available in most terminal windows is emacs or xemacs, that can be used to edit the sequescan configuration file or modify any of the scripts and programs that are distributed with Sequedex. A short example script for creating a matrix of sequedex output from several runs is described in s.rollup. If the user knows how to program in Perl or Python already, it is easy to write short programs or scripts in these languages to process output files. For numerical analysis tools, many options are available, including gfortran, which we used to do simple numerical operations on groups of profiles in *Initial analysis with Sequedex: Phylogenetic and functional profiles*, using `pnorm.f90` or `fnorm.f90`. To use this code, you need to open the source code with a text editor (eg. emacs pnorm.f90 &), change the parameter in the 5th line to the appropriate number of parameters, and the input filename in the 9th line to the appropriate input filename, save the file, compile (gfortran pnorm.f90) and run the executable (./a.out). Several output files will be created, with normalized output, a rollup, and a similarity matrix of each sample compared to the others. The same objective could easily be accomplished in a variety of languages and analysis environments, such as Python, R, C, C++, Perl, or Matlab. It is a useful way to automate automate analysis and make use of functions written by others and distributed on the web.

11.4 Exploring trees with Archaeoptrix, FigTree, or NJPlot

Sequedex organizes its list of signature peptides with carefully constructed phylogenetic trees, ostensibly representing the ancestral history of the various organisms represented. Although Sequedex provides both html and pdf ways of looking at the trees in *Tree of Life, 2550 taxa*, specialized tree-viewing software will greatly facilitate the user's ability to both understand the reference database and the results. The variety of tree-viewers and file formats for defining trees and associated meta-data include:

- The most flexible is archaeopteryx, which defined it's own format, phyloxml.
- FigTree was designed to produce publication quality trees, and reads nexus and phylip format trees.
- NJplot is an easy-to-use program that enables viewing of phylip-style trees, such as is created by *Tree building with phyML or FastTree*.

11.5 Comparing and annotating phylogenetic and functional profiles with Gnuplot

Gnuplot is rapid method to view output files, perhaps as Sequedex is running, as a script-able method to make and annotate a series of plots of phylogenetic or functional profiles.

The example script `hi_res_HMB.gpt`, makes and annotates a multi-panel plot of the total reads matched in each phylogenetic bin of the tol1471. It can be run by typing *gnuplot hi_res_HMB.gpt* with the appropriate input files (plaque, stool, tongue, cheek, nose, ear, and bds) in the path, and will produce the output postscript file 'hmb.ps'. Much of the value of this method of plot-making lies in editing the example file ('emacs hi_res_HMB.gpt &), and changing the output filename, perhaps producing a pdf or png file instead of postscript, changing the scaling factor from the data, or changing the input files or line colors.

The advantage of this tool is that it is lightweight and easy to modify. It it, however, specific to a particular tree or functional set (eg. seed_0911), as the elements are identified by their position in the output table, not the label.

11.6 Statistical analysis and graphing with R and R Studio

R (http://www.r-project.org/ is a powerful open source software for statistical data analysis. Binaries and download instruction are found at http://cran.r-project.org/. In addition to the basic installation, many users will want to peruse the library of contributed packages that extend the functionality and methods available. Packages to consider include ape, ade4, cluster, and rgl.

Although R is available within Cygwin, the user may find it more convenient to utilize the Windows binary distributions for R, enabling installation of Rstudio and utilization of Windows-specific distribution of packages.

Sequestat (described in detail in *Using Sequestat*) utilizes three add-on packages distributed for use within the R environment:

- ape (Analysis of Phylogenetics and Evolution - APE) http://ape.mpl.ird.fr/
- gee (Generalized Estimation Equation solver - GEE) http://cran.r-project.org/web/packages/gee/
- lars (Least Angle Regression - LARS) http://cran.r-project.org/web/packages/lars/
- igraph http://cran.r-project.org/web/packages/igraph/

APE and GEE are used for displaying Sequedex output on the nodes of trees, while LARS is use by Sequestat for predicting which combination of reference genomes or metagenomes best-describes the phylogenetic profile of a metagenomic sample. Igraph is used for displaying functional profiles on a graph-layout of the hierarchical Seed subsystems. If the packages are downloaded, they can be installed from within R with the command:

```
install.packages(repos=NULL, pkgs="gee_4.13-18.tar.gz")
```

Packages and their dependencies can also be automatically downloades, but may require a proxy server be utilized, if your organization sits behind a firewall. This should already be set for sequedex-update and other network-using features of Sequedex, but can be done set from within R if needed by a comand such as:

```
Sys.setenv(http_proxy="http://proxyserver.org.org:8080")
```

Many users prefer to use an integrated development environment such as the freely distributed RStudio, from http://www.rstudio.com/ide/ instead of the command-line interface.

In either case, there are several basic functions the user needs to get started applying R to Sequedex data.

- Getting to the correct directory: Within RStudio, there is a 'set working directory option under 'session''. R can just be started in the desired directory, or the command 'setwd("~/sample.fas.sqdx")'.
- Sourcing an external script so the functions and variables can be used interactively: 'source("~/sequestat.r")'. In this case, the file *sequestat4.r* will be sourced, assuming it is in the R directory of user's home directory, and the external package, ape, described above will be loaded.
- Loading a nexus format phylogenetic tree, such as Life=Read.tree("~/Life2550.nexus"), into the data structure, tol.
- Reading tsv data into the data structure, dat: dat=read.table("~/env.tsv",sep="t"). The first column needs to have node names, and the first row needs to have sample names. Both the rows and columns can be referred to by either element number (position) or label.
- Attaching the tsv file just loaded to the tree structure: env$data=dat.
- Combining two data matrixes: all$data=c(env$data,new$data). (Don't forget to also update the column labels)
- Changing plot output to a png file instead of the screen: png("rplot.png") ; plot(x,y) ; dev.off()

Although a high-throughput genomics analysis package for R Bioconductor has been written, we have not yet incorporated features from this toolkit into our workflows.

11.7 Obtaining reference data from NCBI

NCBI has several interfaces with which data can be downloaded. We downloaded bacterial and archael genomes from the NCBI ftp site under the link for 'Genome annotation and assembly projects' for 'Bacteria' and *Bacteria_DRAFT*. Subdirectories for each of the genomes exist with various file-types and information for download. For the Bacteria, gzipped tar files exist to easily download a particular file-type for all of the organisms, such as amino acid sequences of called genes, in the faa file.

Data for eukaryotes was obtained from the 'RefSeq' link at the NCBI ftp site, and a fair amount of effort was required to identify and extract genes from each of the organisms.

11.8 Creating synthetic data

There are many reasons one might want to create synthetic data. In using this technique to evaluate analysis software, however, one needs to keep in mind that simulated data often derives directly from the same database of reference organisms that the analysis software uses to classify reads. Unfortunately, the reference genomes are not necessarily representative of observed natural diversity. For well-characterized pathogens, this problem is relatively minor. For most RNA viruses and environmental metagenomes, this problem can be quite severe.

MetaSim is a software package that can simulate metagenomic data from both reference genomes and evolutionary models that incorporates sequencer type-specific errors and distributions of read lengths. Since many metagenomics analysis methods, in particular those relying on an assembly step, are quite non-linear in that the results depend on both sequencing depth and the richness and diversity of the sequenced sample, such a flexible tool to generate synthetic data is probably essential to improving the method. For nucleotide-based methods of recognizing reads, such as Metaphlan and Phylophlan testing the analysis software with synthetic data derived from reference genomes is more problematic than with amino-acid-based recognition of reads, such as employed by Sequedex.

Beyond the need to characterize the impact of sequencing errors, database bias, and the impact of mixtures on analysis performance, it is also useful to characterize the phylogenetic dependence of the sensitivity and specificity of read-based annotation. For this purpose, synthetic data can be obtained by simply chopping the full nucleotide sequence of an organism into equal-length reads with, for example, the following two-line perl / bash script:

```
cat *.fna |grep -v ">"|tr -d '\n' |perl -ne 'while ($c = substr($_, 0, 100, "")) {print "$c\n"}' > sy
grep -n . syn |tr ':' ' '|awk '{print ">"$1"\n" $2}' > syn.fas;echo $i
```

11.9 Comparing functional profiles to KEGG

The Kyoto Encyclopedia of Genes and Genomes KEGG provides an extensive collection of how genes interact with one another in pathways. For highly conserved genes where paralogs performing a distinct function are absent, it is relatively straightforward to relate a gene fragment to a position in a KEGG pathway map. For most genes, however, this mapping can be error-prone, as described at in the SEED manifesto of the SEED project.

Consequently, we currently base our functional assignments on the *Definition of functional classifications*, and rely on further analysis of reads with reference genomes to confirm functional assignments, using the techniques described in *Annotated reads* and the reference data provided in *Annotated reads* to get the user started on this semi-infinite task. Anyone doubting the complexity of this task is invited to browse the Pfam families and assess the clarity of mapping onto the KEGG pathway maps.

11.10 The Ribosomal Database Project

In part because universally conserved regions permit PCR amplification ribosomal sequences, the database and tools for probing the microbial community structure are quite well-developed. The Ribosomal Database Project has systemetically collected and curated ribosomal sequence data, and developed and made available numerous tools for its analysis. Begining with the Life2550 tree, Sequedex has defined a functional category for ribosomal sequence, and, using the techniques described in *Annotated reads*, will generate a fasta or fastq file of all identified bacterial, archael, and eukaryotic ribosomal sequences, both the SSU and LSU.

11.11 Bergey's manual

When specific organisms are identified in a sequencing data set, they can be indicative of the ecological niche the sample came from. In addition to the peer reviewed literature, the references to Genbank and annotated pdf files provided in *Tree of Life, 2550 taxa*, and several online wiki sites, Bergey's manual of systematic microbiology proveds a systematic exposition of cultured bacterial and archaeal organisms.

11.12 Phylogeny vs. taxononomy - MEGAN and Krona plots

Several tools exist for visualizing phylogenetic output. Because Sequedex is defined by and utilizes a phylogenetic tree to appropriately capture abiguity in phylogenetic placement on a read-by-read basis, such as *Graphing and sorting with Excel*, *Tree of Life, 2550 taxa*, and *Using Sequestat*. Nevertheless, since phylogenetic rollups provide reasonably good correspondence to the NCBI taxonomy, and many users will be familliar with tools such as MEGAN and Krona to represent hierarchical placement of reads in a taxonomy, it may be worthwhile for the user to map the Sequedex output into these formats.

11.13 Removal of duplicate reads with CD-HIT

Software to remove duplicates from pyrosequencing data is available at http://weizhong-lab.ucsd.edu/cd-hit/.

11.14 BioPython, BioPerl, and EMBOSS utilities

In the flexible workflows described in the Sequedex documentation, especially regarding the phylogenetic analysis, the user will encounter numerous instances when sequences need to be translated, aligned, annotated, visualized on networks, or compared, or may want to bundle workflows together in a larger, more complex, program. While we typically provide robust examples of hot these tasks can be solved, several buldled packages of utilities are available, together with source code and documentation to enable modification. Examples of these packages include BioPython, BioPerl, and the EMBOSS utilities.

11.15 BioEdit and reference alignments

Multiple sequence alignments are a powerful tool for assessing the noise characteristics of sequencing errors and for placing novel sequences in the context of the existing database of genome sequences. Additionally, they provide the starting point for evolutionary analysis, such as that described in the section *Tree building with phyML or FastTree*. Such analysis was used to compute the reference tree for Sequedex's signature calculation, and the reference alignments are provided in *Phylogenetic placement of RNA Polymerase reads*. Additionally, when a dominant species is

found in a metagenome, it is often useful to compare the nucleotide sequence of a phylogenetic marker gene to that of an assembled unknown, such as provided in *Nucleotide Alignments for Strain Attribution* for *E. coli* and its neighbors.

If your web browser has a constant width font, the above-mentioned alignments will be rendered on your screen so that each column contains corresponding a corresponding amino acid or nucleotide. If the sequences are sufficiently similar to one another, numerous software packages (see *Muscle*, and *HMMER and Pfam*) are available to construct such alignments from appropriate unaligned sequence data. If, however, too much divergence is present in the input file or chimeric or inappropriate sequences are included in the input file, automated alignment programs can fail, sometimes quite badly. More divergent nucleotide sequences, such as from viral genomes, can often be more easily aligned as translated proteins, and then un-translated for the evolutionary analysis.

In this case, a sequence alignment viewer and editor can be quite useful. One of the most versatile is BioEdit, which is freely available, although it was written in Visual Basic, and will run only under a Windows operating system or Windows emulator, such as WINE.

When used for viewing alignments, sequencing errors, stray bar-codes, frame-shift mutations, chimeric reads or assemblies, mis-placed start codons, and un-alignable proteins are often glaringly obvious. Because many alignment programs place similar sequences next to one another in the alignment, it is also possible to distinguish orthologs from paralogs, and either select a sub-set of sequences for re-alignment, or delete problematic sequences and re-align the remainders.

Besides making it easier to spot inconsistencies and errors, sequence editors also enable manual editing of sequences (and sequence labels) in the alignment, by sliding residues or blocks of residues left or right, deleting residues judged to be errors, or, if the user so desires, sequences can be manually typed in or changed, hopefully with some justification. Although BioEdit does not provide a mechanism to access the quality scores, it is quite versitile, and numerous strategies exist to obtain alignments that are appropriate inputs to evolutionary analysis.

When downloading the reference alignments provided in the sections *Phylogenetic placement of RNA Polymerase reads* and *Nucleotide Alignments for Strain Attribution*, the user will need to change them into one of the supported file formats, such as fasta. It is a good practice to manually inspect all parts of any alignment which will be used for evolutionary analysis, perhaps consulting available structures and literature in cases where the sequence data alone is ambiguous.

For simply viewing alignments, ClustalX and SeaView both work well.

11.16 BLAST and nr / nt databases

Although much of the convenience of running Blast as a confirmatory analysis lies in the accessibility of the continuously updated databases and parallelized compuataional capacity availble through the web interface at NCBI, it is also quite valuable to create customized databases for use on local machines. Such databases can allow rapid searches for particular genes, over restricted phylogenetic search-space, or over a subset of organisms that are appropriately selected and annotated for the purposes of annotation.

A customized Blast database can be created from an unaligned fasta file with the command:

./makeblastdb -in RNAP.fas -dbtype ‘prot’ -out RNAP -name -RNAP

A Blast search can be run against a local database with the any of the commands:

```
blastn -db C:\blast\db\bact_complete.fna -query contigs_l.fa -out velveth_l.nt -num_alignments 1 blastn
-db ~/megan/nt -query contigs_l.fa -out velveth_l.nt -num_threads 4 -num_alignments 1 blastn -db
~/megan/nt -query rnap.fa -out rnap.nt -num_threads 4 -num_alignments 1 -outfmt 7
```

The top one is appropriate for a Windows machine running Cygwin, while the bottom two work on a Linux machine with four processors, and produce two different types of output file formats. Blast is an extremely flexible software package with extensive documentation.

11.17 HMMER and Pfam

HMMER constructs a statistical representation of the conserved elements in a multiple sequence alignment (a Hidden Markov Model) that can be used to rapidly align and score unknows against the original alignment. It is useful for rapidly aligning and classifying contigs or fragments against a family of proteins that can either be created from a curated reference or downloaded from (Pfam http://pfam.janelia.org/ is an exhaustive collection of annotated protein families). For the highest confidence in the match, it is suggested that a tree be built with reference genes and the annotations checked. This is a way to distinguish orthologs from paralogs.

11.18 Muscle

Muscle http://www.drive5.com/muscle/ is rapid enough to simply re-align both reference and unknown sequences, although it is also possible to use it to add sequences to a frozen alignment.

11.19 Velvet and de-novo assembly

Velvet is a robust assembler that does not require quality scores. It can be run with the following two commands:

```
velveth 701 21 701.fa
velvetg 701
```

These commands will create an assembly of the reads in the file 701.fa with several output files in the directory *701*. The *21* on the first line refers to the length of overlap identity required for two reads to be identified as overlapping. To sort the assembled contigs in order of decreasing length, from the directory created by velvet (701 in the above example), type:

```
tr '\n' ';' < contigs.fa |sed "s/>/\n>/g" |tr '_' ' '|sort -n -r -k4 |sed "s/[0-9];/\n/" |tr -d ';' >
```

The longest contigs are likely to be the most useful in identifying novel organisms, although simply selecting relevant reads for alignment will sometimes provide reads of nearly the same length, but with much greater choice in location (very important for comparing across samples, were it is useful to have the same region selected from all samples). It is also possible to estimate coverage from the assembly, and compare to that expected from examination of phylogenetic profiles.

11.20 Tree building with phyML or FastTree

Once an unknown read or contig has been placed into a multiple sequence alignment, and the alignment has been examined and / or curated, it is possible to compute a phylogenetic tree. Although neighbor-joining trees, such as can be easily run from ClustalX or BioEdit, can be a quick-and-dirty of screening sequences for appropriateness for inclusion of sequences in an alignment, anything of importance should be run through a maximum likelihood tree-building algorithm, such as phyML. Although maximum likelihood methods can be time consuming (they typically scale linearly with alignment length and cubically with the number of sequences, a variety of methods and options are available, and it is generally possible to curate the list of sequences to less than 100 taxa (which aids interpretation as well), so there isn't really a good reason for lengthy computations here.

11.21 p-values and Pplacer

Pplacer is designed to place metagenomic reads according to a reference alignment and tree, and enable downstream visualization and analysis. It requires a reference alignment, a potentially large number of reads aligned to this reference alignment, perhaps with *HMMER and Pfam*, and a tree-building algorithm, such as *Tree building with phyML or FastTree*.

11.22 BWA and BowTie2

When a suitable reference genome is within a couple of percent divergence in nucleotide sequence from an unknown in a metagenomic sample, read-mappers such BWA or BowTie2 can rapidly extract reads or contigs from large input files in a few minutes. Quality scores are preserved and a file format specially designed for such large alignments (SAM) as well as its binary version (BAM) are supported. A set of commands that creates such an alignment is:

```
bwa index all.na.fas
bwa aln -n 2 all.na.fas ../../../test008.fastq > aln.5.sai
bwa samse all.na.fas aln.sai ../../../test008.fastq > aln.sam
```

In the next section, *Samtools and Bamtools*, we will generate a consensus fasta file from the above.

11.23 Samtools and Bamtools

Samtools, Bamtools, and VCFtools are freely available utilities to perform a variety of useful manipulations on large alignments in the SAM or BAM files. VCF tools is an additional software suite associated with the *Variant Call Format*, designed to enable SNP analysis in the presence of sequencing errors. These files are best viewed with a specialized viewer, such as the Integrated Genomics Viewer, described in the next section.

The above example from BWA can be continued to create a consensus fasta file with the commands:

```
samtools faidx all.na.fas
samtools view -bt all.na.fas.fai -o aln.bam aln.sam
samtools sort aln.bam aln.sorted
samtools index aln.sorted.bam
samtools mpileup -uf all.na.fas rnap.bam | bcftools view -bvcg - > var.raw.bcf
bcftools view var.raw.bcf | vcfutils.pl varFilter -D100 > var.flt.vcf
samtools mpileup -uf all.na.fas rnap.sorted.bam | bcftools view -cg - | vcfutils.pl vcf2fq > cns.fq
```

The process of iteratively distinguishing sequence errors from true variants and correctly associating SNPs with one another is well-traveled territory; see, for example, vphaser.

11.24 The Integrated Genomics Viewer

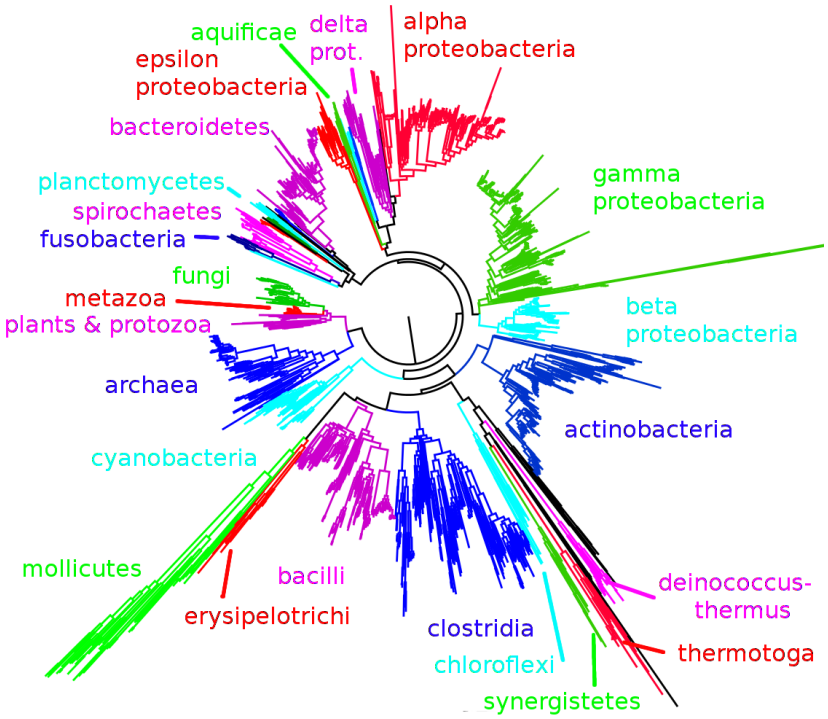
The Integrated Genomics Viewer from the Broad Institute is designed to compare large, disparate, data sets that can be mapped onto a reference genome. It requires reference genomes and can accommodate annotations in the gff format, available with completed and draft genomes from NCBI, described in *Obtaining reference data from NCBI*. It is primarily used for resequencing of large, eukaryotic, genomes, so the user will likely need to create local reference files for particular genomes of interest if environmental or microbiome data are being analyzed.

Since one valuable function the IGV enables is evaluating genome inventory, it is important to consider both the quality of the annotation and the completeness of the genome inventory, in addition to phylogenetic distance when choosing a reference genome.

CHAPTER 12

Tree of Life, 2550 taxa

One-per-species tree of life, made from the completed and draft genomes available on 1 April, 2013, in `phyloxml format`.



In order to provide correspondence to NCBI's taxonomy and the reference genomes, this chapter presents the list of reference genomes linked to NCBI on the basis of genus and species names. The complete directory name from the NCBI ftp site also appears below, and the combination allows the current release of each genome to be located.

Sequedex was developed with a signature list and phylogenetic classification based on a bacterial one-per-genus tree of 403 taxa. Since that time, we have increased both the breadth and resolution of this tree to include bacteria, archaea, and eukaryotes at the one-per-species resolution. The tree was based on an amino acid alignment of the concatenated RNA polymerase beta and beta prime subunits (for bacteria) or the analogous subunits of RNA pol II (for eukaryotes).

Phylogenetic information can be visualized in a variety of ways, and this chapter attempts to start the user with phylum-level, family-level, and species-level identifications of what is present in a sample. Although labeled as phylum and family, we took liberties in using the pair of classifications that best suited the density of reference genomes in each

region of the phylogeny, to obtain a labeling scheme that is parsimonious, spanning, and non-overlapping. By clicking on the taxon names, the user can easily compare to all levels of the taxonomy assigned by NCBI.

As described in the Introduction, Sequedex captures the uncertainty inherent in classifying fragments of DNA by making phylogenetic assignments to the nodes of a binary phylogenetic tree, and not to individual taxa. Consequently, some node numbers are indicative of phyla, some nodes are indicative of families, and some are specific to a pair of taxa. In the table below, we provide all three types of information. Because of the way the nodes on the trees were numbered, through a depth-first traversal, node numbers intermediate between these two extremes are intermediate nodes between the group name (family or phylum) and the leaves. PDF files are available with page-sized chunks of the tree throughout the tree, and are indicated at the beginning of each section, *Bacteroidetes, et al.*, *Alpha proteobacteria, et al.*, *Beta and Gamma Proteobacteria*, *Actinobacteria*, *Firmicutes, et al.*, *Cyanobacteria*, *Archaea*, and *Eukaryotes*.

There are undoubtably minor mistakes in this classification. The nature of the signature peptide method, however, is such that these mistakes serve to bluzrr out phylogenetic signals, not result in spurious assignments to specific nodes.

12.1 Bacteroidetes, et al.

This group contains Chlamydia, Verrucomicrobia, Planctomycetes, Spirochaetes, Fusobacteria, Elusimicrobia, Gemmatimonads, Salinibacter, Chlorobia, Cytophaga, Pedobacteria,Flavobacteria, and bacteroidetes. Annotated pdfs with node numbers and family names are available for:

- `spirochaetes + fusobacteria`,
- `bacteroidetes`,
- `chlamydia + verrucomicrobia + planctomycetes`,
- and `chlorobi`.

Phylum	Family	Organism
Elusimicrobia	Elusimicrobiales	uncultured Termite group 1 bacterium phylotype Rs D17 uid59059
(node 6)	(node 6)	Elusimicrobium minutum Pei191 uid58949
Fusobacteriales	Fusobacteriaceae	Ilyobacter polytropus DSM 2926 uid59769
(nodes 8-24)	(nodes 9-20)	Fusobacterium mortiferum ATCC 9817 uid55571
.		Fusobacterium varium ATCC 27725 uid55573
.		Fusobacterium 12 1B uid81787
.		Fusobacterium ulcerans ATCC 49185 uid55615
.		Fusobacterium gonidiaformans ATCC 25563 uid55569
.		Fusobacterium D12 uid55613
.		Fusobacterium necrophorum funduliforme 1 1 36S uid81603
.		Fusobacterium oral taxon 370 F0437 uid78149
.		Fusobacterium periodonticum ATCC 33693 uid55335
.		Fusobacterium D11 uid55627
.		Fusobacterium nucleatum ATCC 25586 uid57885
.		Fusobacterium 4 1 13 uid55609
.	Leptotrichiaceae	Sebaldella termitidis ATCC 33386 uid41865
.	(nodes 21-24)	Streptobacillus moniliformis DSM 12112 uid41863
.		Leptotrichia goodfellowii F0264 uid41359
.		Leptotrichia buccalis C 1013 b uid59211
.		Leptotrichia hofstadii F0254 uid55767
Spirochaetes	Leptospiraceae	Turneriella parva DSM 21527 uid168321
(nodes 25-74)	(nodes 26-40)	Leptonema illini DSM 21528 uid82715
.		Leptospira biflexa serovar Patoc Patoc 1 Ames uid58511

Continued on next page

Table 12.1 – continued from previous page

Phylum	Family	Organism
.		Leptospira meyeri serovar Hardjo Went 5 uid177325
.		Leptospira alexanderi serovar Manhao 3 L 60 uid171099
.		Leptospira kirschneri 200802841 uid180750
.		Leptospira noguchii 2006001870 uid78701
.		Leptospira kmetyi serovar Malaysia Bejo Iso9 uid171098
.		Leptospira interrogans serovar Copenhageni Fiocruz L1 130 uid58065
.		Leptospira borgpetersenii serovar Hardjo bovis JB197 uid58509
.		Leptospira weilii 2006001853 uid179809
.		Leptospira Fiocruz LV3954 uid178558
.		Leptospira santarosai 2000030832 uid78697
.		Leptospira licerasiae MMD4847 uid180102
.		Leptospira broomii 5399 uid171097
.		Leptospira inadai serovar Lyme 10 uid171096
.	Brachyspiraceae	Brachyspira pilosicoli 95 1000 uid50609
.	(nodes 42-46)	Brachyspira intermedia PWS A uid158369
.		Brachyspira hyodysenteriae WA1 uid59291
.		Brachyspira 30446 uid183377
.		Brachyspira hampsonii 30599 uid188379
.		Brachyspira murdochii DSM 12563 uid48819
.	Spirochaetacea	Borrelia valaisiana VS116 uid54823
.	(nodes 47-74)	Borrelia spielmanii A14S uid55069
.		Borrelia afzelii HLJ01 uid177930
.		Borrelia garinii BgVir uid162165
.		Borrelia bissettii DN127 uid71231
.		Borrelia burgdorferi JD1 uid161197
.		Borrelia SV1 uid55065
.		Borrelia crocidurae Achema uid162335
.		Borrelia recurrentis A1 uid58793
.		Borrelia duttonii Ly uid58791
.		Borrelia hermsii DAH uid59225
.		Borrelia turicatae 91E135 uid58311
.		Spirochaeta coccoides DSM 17374 uid66331
.		Sphaerochaeta pleomorpha Grapes uid82365
.		Spirochaeta Buddy uid63633
.		Spirochaeta thermophila DSM 6192 uid53037
.		Spirochaeta africana DSM 8902 uid81779
.		Spirochaeta caldaria DSM 7334 uid68753
.		Treponema azotonutricium ZAS 9 uid67365
.		Treponema primitia ZAS 2 uid67367
.		Treponema denticola ATCC 35405 uid57583
.		Treponema vincentii ATCC 35580 uid55865
.		Treponema phagedenis F0421 uid62291
.		Treponema pallidum Chicago uid159543
.		Treponema paraluiscuniculi Cuniculi A uid68447
.		Treponema brennaborense DSM 12168 uid66607
.		Treponema succinifaciens DSM 2489 uid65781
.		Treponema saccharophilum DSM 2985 uid156761
.		Treponema JC4 uid158669
Chlamydiae	Chlamydiaceae	Chlamydophila abortus S26 3 uid57963

Continued on next page

Table 12.1 – continued from previous page

Phylum	Family	Organism
(nodes 77-88)	(nodes 79-86)	Chlamydia psittaci 01DC12 uid179070
.		Chlamydophila psittaci 01DC11 uid159527
.		Chlamydophila caviae GPIC uid57783
.		Chlamydophila felis Fe C 56 uid57971
.		Chlamydia muridarum Nigg uid57785
.		Chlamydia trachomatis 434 Bu uid61633
.		Chlamydophila pneumoniae AR39 uid57809
.		Chlamydophila pecorum E58 uid66295
.	Waddliaceae	Waddlia chondrophila WSU 86 1044 uid49531
.	Parachlamydiaceae	Candidatus Protochlamydia amoebophila UWE25 uid58079
.	(node 88)	Parachlamydia acanthamoebae UV7 uid68335
.		Simkania negevensis Z uid68451
Verrucomicrobia	Lentisphaeraceae	Lentisphaera araneosa HTCC2155 uid54167
(nodes 90-98)	Methylacidiphilaceae	Methylacidiphilum fumariolicum uid159999
.	(node 92)	Methylacidiphilum infernorum V4 uid59161
.	Spartobacteria	Chthoniobacter flavus Ellin428 uid55037
.	Verrucomicrobia	Verrucomicrobium spinosum DSM 4136 uid54121
.	(node 94)	Akkermansia muciniphila ATCC BAA 835 uid58985
.		bacterium Ellin514 uid54925
.	Optitae	Coraliomargarita akajimensis DSM 45221 uid47079
.	(nodes 96-98)	Verrucomicrobiae bacterium DG1235 uid54739
.		Opitutaceae bacterium TAV1 uid82717
.		Opitutus terrae PB90 1 uid58965
Planctomycetes		planctomycete KSU 1 uid163683
(nodes 99-110)		Phycisphaera mikurensis NBRC 102666 uid157331
.		Gemmata obscuriglobus UQM 2246 uid54931
.		Isosphaera pallida ATCC 43644 uid62207
.		Singulisphaera acidiphila DSM 18658 uid81777
.		Planctomyces brasiliensis DSM 5305 uid60583
.		Planctomyces maris DSM 8797 uid54323
.		Planctomyces limnophilus DSM 3776 uid48643
.		Schlesneria paludicola DSM 18645 uid175442
.		Blastopirellula marina DSM 3645 uid54189
.		Pirellula staleyi DSM 6068 uid43209
.		Rhodopirellula baltica SH 1 uid61589
.		Rhodopirellula europaea 6C uid189169
Gemmatimonadetes	Gemmatimonadlaes	Gemmatimonas aurantiaca T 27 uid58813
WWE1	WWE1	Candidatus Cloacamonas acidaminovorans Evry uid62959
Fibrobacteres	Fibrobacterales	Fibrobacter succinogenes S85 uid161919
Chlorobi	Chlorobiales	Chloroherpeton thalassium ATCC 35110 uid59187
(nodes 115-124)	(nodes 115-124)	Chlorobium phaeobacteroides BS1 uid58131
.		Prosthecochloris aestuarii DSM 271 uid58151
.		Chlorobaculum parvum NCIB 8327 uid59185
.		Chlorobium tepidum TLS uid57897
.		Chlorobium limicola DSM 245 uid58127
.		Chlorobium luteolum DSM 273 uid58175
.		Chlorobium phaeovibrioides DSM 265 uid58129
.		Chlorobium chlorochromatii CaD3 uid58375
.		Chlorobium ferrooxidans DSM 13031 uid54401

Continued on next page

Table 12.1 – continued from previous page

Phylum	Family	Organism
.		Pelodictyon phaeoclathratiforme BU 1 uid58173
Ignavibacteria	Ignavibacteriales	Ignavibacterium album JCM 16511 uid162097
(node 126)	(node 126)	Melioribacter roseus P3M uid170941
Bacteroidetes	Order_II	Rhodothermus marinus DSM 4252 uid41729
(nodes 127-299)	(node 128)	Salinibacter ruber DSM 13855 uid58513
.	unclassified	Cardinium endosymbiont cEper1 of Encarsia pergandiella uid175524
.	unclassified	Candidatus Amoebophilus asiaticus 5a2 uid58963
.	Cytophagia	Emticicia oligotrophica DSM 17448 uid177079
.	(nodes 132-150)	Leadbetterella byssophila DSM 17132 uid60161
.		Fibrella aestuarina uid178352
.		Fibrisoma limi uid169430
.		Spirosoma linguale DSM 74 uid43413
.		Runella slithyformis DSM 19594 uid68317
.		Dyadobacter fermentans DSM 18053 uid59049
.		Cytophaga hutchinsonii ATCC 33406 uid57651
.		Flexibacter litoralis DSM 6794 uid168257
.		Microscilla marina ATCC 23134 uid54163
.		Marivirga tractuosa DSM 4126 uid60837
.		Fulvivirga imtechensis AK7 uid186536
.		Mariniradius saccharolyticus AK6 uid185795
.		Cecembia lonarensis LW9 uid175678
.		Marinilabilia AK2 uid176937
.		Indibacter alkaliphilus LW1 uid175140
.		Algoriphagus PR1 uid54611
.		Belliella baltica DSM 15883 uid168182
.		Cyclobacterium marinum DSM 745 uid71485
.		Echinicola vietnamensis DSM 17526 uid184076
.	Sphingobacteriia	Chitinophaga pinensis DSM 2588 uid59113
.	(nodes 152-162)	Niabella soli DSM 19437 uid82551
.		Niastella koreensis GR20 10 uid83125
.		Haliscomenobacter hydrossis DSM 1100 uid66777
.		Saprospira grandis Lewin uid89375
.		Solitalea canadensis DSM 3403 uid81783
.		Pedobacter saltans DSM 12145 uid61349
.		Pedobacter BAL39 uid54703
.		Pedobacter heparinus DSM 2366 uid59111
.		Mucilaginibacter paludis DSM 18603 uid60581
.		Sphingobacterium 21 uid64755
.		Sphingobacterium spiritivorum ATCC 33300 uid55527
.	Flavobacteriia	Cellulophaga lytica DSM 7489 uid63401
.	(nodes 164-218)	Cellulophaga algicola DSM 14237 uid62159
.		Flavobacteriales bacterium HTCC2170 uid51877
.		Zobellia galactanivorans uid70621
.		Robiginitalea biformata HTCC2501 uid58285
.		Mesoflavibacter zeaxanthinifaciens S86 uid81109
.		Muricauda ruestringensis DSM 13258 uid72479
.		Galbibacter ck I2 15 uid176601
.		Joostella marina DSM 19592 uid163687
.		Imtechella halotolerans K1 uid158663

Continued on next page

Table 12.1 – continued from previous page

Phylum	Family	Organism
.		Capnocytophaga CM59 uid174242
.		Capnocytophaga gingivalis ATCC 33624 uid55451
.		Capnocytophaga canimorsus Cc5 uid70727
.		Capnocytophaga sputigena ATCC 33612 uid55407
.		Capnocytophaga ochracea DSM 7271 uid59197
.		Capnocytophaga oral taxon 324 F0483 uid183775
.		Kordia algicida OT 1 uid54687
.		Bizionia argentinensis JUB59 uid72961
.		Lacinutrix 5H 3 7 4 uid68067
.		Psychroflexus torquis ATCC 700755 uid54205
.		Gillisia limnaea DSM 15749 uid82719
.		Gramella forsetii KT0803 uid58881
.		Zunongwangia profunda SM A87 uid48073
.		Krokinobacter 4H 3 7 5 uid66593
.		Leeuwenhoekiella blandensis MED217 uid54243
.		Croceibacter atlanticus HTCC2559 uid49661
.		Aquimarina agarilytica ZC1 uid173055
.		Persicivirga dokdonensis DSW 6 uid186842
.		Aequorivita sublithincola DSM 14238 uid168181
.		unidentified eubacterium SCB49 uid54737
.		Myroides odoratus DSM 2801 uid82721
.		Myroides injenensis M09 0166 uid171989
.		Myroides odoratimimus CCUG 10230 uid81601
.		Flavobacterium columnare ATCC 49512 uid80731
.		Flavobacterium indicum GPTSA100 9 uid157999
.		Flavobacteria bacterium BAL38 uid54617
.		Flavobacterium psychrophilum JIP02 86 uid61627
.		Flavobacterium branchiophilum FL 15 uid73421
.		Flavobacterium frigoris PS1 uid156765
.		Flavobacterium CF136 uid171678
.		Flavobacterium F52 uid170864
.		Flavobacterium johnsoniae UW101 uid58493
.		Polaribacter irgensii 23 P uid54179
.		Polaribacter MED152 uid54207
.		Riemerella anatipestifer ATCC 11845 DSM 15868 uid159857
.		Elizabethkingia anophelis Ag1 uid80705
.		Bergeyella zoohelcum ATCC 43767 uid181642
.		Flavobacteriaceae bacterium 3519 10 uid59413
.		Chryseobacterium CF314 uid171677
.		Chryseobacterium gleum ATCC 35910 uid55393
.		Ornithobacterium rhinotracheale DSM 15997 uid168256
.		Weeksella virosa DSM 16922 uid63627
.		Candidatus Uzinura diaspidicola ASNER uid186740
.		Candidatus Sulcia muelleri CARI uid52535
.		Owenweeksia hongkongensis DSM 17368 uid82951
.		Fluviicola taffensis DSM 16823 uid65271
.	Rikenellaceae	Alistipes indistinctus YIT 12060 uid75115
.	(nodes 220-222)	Alistipes putredinis DSM 17216 uid54803
.		Alistipes finegoldii DSM 17242 uid168180

Continued on next page

Table 12.1 – continued from previous page

Phylum	Family	Organism
.		Alistipes JC136 uid174622
.	Porphyromonadaceae	Odoribacter laneus YIT 12061 uid82557
.	(node 225)	Odoribacter splanchnicus DSM 20712 uid63397
.	Marinilabiliaceae	Anaerophaga HS1 uid80885
.	(node 226-227)	Marinilabilia salmonicolor JCM 21150 uid177805
.		Anaerophaga thermohalophila DSM 12881 uid72977
.	Porphyromonadacea	Porphyromonas endodontalis ATCC 35406 uid55449
.	(nodes 229-244)	Porphyromonas asaccharolytica DSM 20707 uid66603
.		Porphyromonas uenonis 60 3 uid55869
.		Porphyromonas catoniae F0037 uid183770
.		Porphyromonas oral taxon 279 F0450 uid174237
.		Porphyromonas gingivalis ATCC 33277 uid58879
.		Tannerella forsythia ATCC 43037 uid83157
.		Parabacteroides distasonis ATCC 8503 uid58301
.		Parabacteroides goldsteinii CL02T12C30 uid178274
.		Parabacteroides johnsonii CL02T12C29 uid181649
.		Parabacteroides merdae ATCC 43184 uid54545
.		Candidatus Azobacteroides pseudotrichonymphae genomovar CFP2 uid59163
.		Dysgonomonas gadei ATCC BAA 286 uid67099
.		Dysgonomonas mossii DSM 22836 uid67097
.		Barnesiella intestinihominis YIT 11860 uid175259
.		Tannerella 6 1 58FAA CT1 uid80413
.		Paludibacter propionicigenes WB4 uid
.	Bacteroidaceae	Bacteroides vulgatus ATCC 8482 uid58253
.	(nodes246-267)	Bacteroides plebeius DSM 17135 uid54991
.		Bacteroides coprophilus DSM 18228 uid55301
.		Bacteroides salanitronis DSM 18170 uid63269
.		Bacteroides coprocola DSM 17136 uid54879
.		Bacteroides coprosuis DSM 18011 uid66921
.		Bacteroides oleiciplenus YIT 12058 uid182882
.		Bacteroides cellulosilyticus CL02T12C19 uid181624
.		Bacteroides intestinalis DSM 17393 uid54881
.		Bacteroides helcogenes P 36 108 uid62135
.		Bacteroides fluxus YIT 12057 uid66157
.		Bacteroides D20 uid42369
.		Bacteroides clarus YIT 12056 uid66155
.		Bacteroides eggerthii 1 2 48FAA uid61869
.		Bacteroides stercoris ATCC 43183 uid54825
.		Bacteroides fragilis 638R uid84217
.		Bacteroides nordii CL02T12C05 uid170043
.		Bacteroides salyersiae CL02T12C01 uid170041
.		Bacteroides finegoldii CL09T03C10 uid181638
.		Bacteroides caccae ATCC 43185 uid54521
.		Bacteroides faecis MAJ27 uid86875
.		Bacteroides ovatus 3 8 47FAA uid68195
.		Bacteroides xylanisolvens CL03T12C04 uid181622
.	Prevotellaceae	Paraprevotella clara YIT 11840 uid76949
.	(nodes 268-299)	Paraprevotella xylaniphila YIT 11841 uid66381
.		Prevotella tannerae ATCC 51259 uid55769

Continued on next page

Table 12.1 – continued from previous page

Phylum	Family	Organism
.		Prevotella ruminicola 23 uid47507
.		Prevotella marshii DSM 16973 uid51709
.		Prevotella stercorea DSM 18206 uid78321
.		Prevotella saccharolytica F0055 uid183769
.		Prevotella buccae ATCC 33574 uid61457
.		Prevotella copri DSM 18205 uid55277
.		Prevotella bryantii B14 uid50549
.		Prevotella oralis ATCC 33269 uid61459
.		Prevotella buccalis ATCC 35310 uid42969
.		Prevotella timonensis CRIS 5C B1 uid42971
.		Prevotella oulorum F0390 uid72969
.		Prevotella maculosa OT 289 uid81767
.		Prevotella oris C735 uid49959
.		Prevotella salivae DSM 15606 uid61887
.		Prevotella multisaccharivorax DSM 17128 uid68187
.		Prevotella bergensis DSM 17361 uid55893
.		Prevotella dentalis DSM 3688 uid184818
.		Prevotella amnii CRIS 21A A uid52821
.		Prevotella bivia DSM 20514 uid182041
.		Prevotella histicola F0411 uid76947
.		Prevotella veroralis F0319 uid55991
.		Prevotella C561 uid72971
.		Prevotella melaninogenica ATCC 25845 uid51377
.		Prevotella denticola F0289 uid65091
.		Prevotella multiformis DSM 16608 uid63431
.		Prevotella micans F0438 uid81791
.		Prevotella nigrescens ATCC 33563 uid70557
.		Prevotella pallens ATCC 700821 uid70559
.		Prevotella disiens FB035 09AN uid51531
.		Prevotella intermedia 17 uid163151

Jump to: *Bacteroidetes, et al.*, *Alpha proteobacteria, et al.*, *Beta and Gamma Proteobacteria*, *Actinobacteria*, *Firmicutes, et al.*, *Cyanobacteria*, *Archaea*, and *Eukaryotes*.

12.2 Alpha proteobacteria, et al.

This group contains Helicobacter, Campylobacter, Acidobacteriales, Aquificae, Deferribacteria, Desulfovibriales, Desulfotaleales, Myxococcus, Gebacteriales, Rhizobia, Caulobacteriales, Rhodobacteria, Sphingomonads, Rickettesia, Rhodospirilliae. PDFs are available for:

- `Delta, Epsilon Proteobacteria, Aquificae, Acidobacteria, and others`.
- Alpha Proteobacteria, part 1, including `Rickettsiales and Rhodospirillales`.
- Alpha Proteobacteria, part 2, including `Sphingomonadales and Rhodobacteraceae`.

Phylum	Family	Organism
Epsilon_proteobacteria	Campylobacteraceae	Sulfurospirillum barnesii SES 3 uid168117
(nodes 302-340)	(nodes 306-319)	Sulfurospirillum deleyianum DSM 6946 uid41861

Continued on next page

Table 12.2 – continued from previous page

Phylum	Family	Organism
.		Campylobacter lari RM2100 uid58115
.		Campylobacter upsaliensis JV21 uid61485
.		Campylobacter coli 1091 uid180315
.		Campylobacter jejuni 81116 uid58771
.		Campylobacter gracilis RM3268 uid55421
.		Campylobacter fetus 82 40 uid58545
.		Campylobacter rectus RM3267 uid55417
.		Campylobacter FOBRC14 uid173871
.		Campylobacter 10 1 50 uid80421
.		Campylobacter concisus 13826 uid58667
.		Arcobacter nitrofigilis DSM 7299 uid49001
.		Arcobacter butzleri ED 1 uid158699
.		Arcobacter L uid158135
.	Helicobacteraceae	Wolinella succinogenes DSM 1740 uid61591
.	(nodes 320-337)	Helicobacter winghamensis ATCC BAA 430 uid55619
.		Helicobacter canadensis MIT 98 5491 uid55289
.		Helicobacter pullorum MIT 98 5489 uid55293
.		Helicobacter cinaedi PAGU611 uid162219
.		Helicobacter hepaticus ATCC 51449 uid57737
.		Helicobacter bilis ATCC 43879 uid55617
.		Helicobacter mustelae 12198 uid46647
.		Helicobacter cetorum MIT 00 7128 uid162217
.		Helicobacter acinonychis Sheeba uid58685
.		Helicobacter pylori 2017 uid161151
.		Helicobacter suis HS1 uid62531
.		Helicobacter bizzozeronii CIII 1 uid68141
.		Helicobacter felis ATCC 49179 uid61409
.		Thiovulum ES uid170493
.		Sulfuricurvum kujiense DSM 16994 uid60789
.		Sulfurimonas autotrophica DSM 16294 uid53043
.		Sulfurimonas denitrificans DSM 1251 uid58185
.		Sulfurimonas GD1 uid81843
.		Nitratifractor salsuginis DSM 16511 uid62183
.	unclassified	Sulfurovum AR uid162597
.	unclassified	Sulfurovum NBC37 1 uid58863
.	unclassified	Nitratiruptor SB155 2 uid58861
.	Nautiliales	Caminibacter mediatlanticus TB 2 uid54669
.	(node 340)	Nautilia profundicola AmH uid59345
Aquificae	Desulfurellaceae	Desulfurobacterium thermolithotrophum DSM 11699 uid63405
(nodes 342-352)	(node 343)	Thermovibrio ammonificans HB 1 uid62095
.	Hydrogenthermaceae	Persephonella marina EX H1 uid58119
.	(nodes 345-347)	Sulfurihydrogenibium azorense Az Fu1 uid58121
.		Sulfurihydrogenibium yellowstonense SS 5 uid54637
.		Sulfurihydrogenibium YO3AOP1 uid58855
.	Aquificaceae	Aquifex aeolicus VF5 uid57765
.	(nodes 348-352)	Hydrogenivirga 128 5 R1 1 uid54685
.		Hydrogenobacter thermophilus TK 6 uid159875
.		Thermocrinis albus DSM 14484 uid46231
.		Hydrogenobaculum HO uid190882

Continued on next page

Table 12.2 – continued from previous page

Phylum	Family	Organism
.		Hydrogenobaculum Y04AAS1 uid58857
Acidobacteria	Holophagae	Holophaga foetida DSM 6591 uid81781
(nodes 355-362)	unclassified	Candidatus Chloracidobacterium thermophilum B uid73587
.	Solibacteres	Candidatus Solibacter usitatus Ellin6076 uid58139
.	unclassified	Candidatus Koribacter versatilis Ellin345 uid58479
.	Acidobacteriaceae	Acidobacterium capsulatum ATCC 51196 uid59127
.	(nodes 359-362)	Acidobacterium MP5ACTX9 uid50551
.		Granulicella mallensis MP5ACTX8 uid49957
.		Terriglobus roseus DSM 18391 uid168183
.		Terriglobus saanensis SP1PR4 uid53251
Chrysiogenetes	Chrysiogenetes	Desulfurispirillum indicum S5 uid45897
Deferribacteres	Deferribacterales	Deferribacter desulfuricans SSM1 uid46653
(nodes 364-366)	(nodes 364-366)	Denitrovibrio acetiphilus DSM 12809 uid46657
.		Calditerrivibrio nitroreducens DSM 19672 uid60821
.		Flexistipes sinusarabici DSM 4947 uid68147
Delta_proteobacteria	Desulfovibrionales	Desulfovibrio fructosovorans JJ uid51537
(nodes 368-426)	(nodes 374-391)	Desulfovibrio magneticus RS 1 uid59309
.		Desulfovibrio FW1012B uid43335
.		Desulfovibrio U5L uid162937
.		Desulfovibrio aespoeensis Aspo 2 uid42613
.		Desulfovibrio hydrothermalis AM13 DSM 14728 uid184831
.		Desulfovibrio salexigens DSM 2638 uid59223
.		Desulfovibrio alaskensis G20 uid57941
.		Desulfovibrio A2 uid73581
.		Desulfovibrio vulgaris DP4 uid58679
.		Desulfovibrio piger ATCC 29098 uid54519
.		Desulfovibrio 6 1 46AFAA uid72975
.		Desulfovibrio desulfuricans ATCC 27774 uid59213
.		Bilophila wadsworthia 3 1 6 uid61875
.		Lawsonia intracellularis N343 uid186598
.		Desulfovibrio africanus Walvis Bay uid66847
.		Desulfohalobium retbaense DSM 5692 uid59183
.		Desulfonatronospira thiodismutans ASO3 1 uid55423
.		Desulfomicrobium baculatum DSM 4028 uid59217
.	Desulfobulbaceae	delta proteobacterium MLMS 1 uid54339
.	(nodes 392-395)	Desulfurivibrio alkaliphilus AHT2 uid49487
.		Desulfobulbus propionicus DSM 2032 uid62265
.		Desulfocapsa sulfexigens DSM 10523 uid189952
.		Desulfotalea psychrophila LSv54 uid58153
.	Desulfobacteraceae	Desulfatibacillum alkenivorans AK 01 uid58913
.	(nodes 397-400)	Desulfococcus oleovorans Hxd3 uid58777
.		Desulfobacterium autotrophicum HRM2 uid59061
.		Desulfobacter postgatei 2ac9 uid76943
.		Desulfobacula toluolica Tol2 uid175777
.	Syntrophobaceae	Desulfobacca acetoxidans DSM 11109 uid65785
.	Desulfarculales	Desulfarculus baarsii DSM 2075 uid51371
.	Syntrophobaceae	Syntrophobacter fumaroxidans MPOB uid58177
.	Syntrophobaceae	Syntrophus aciditrophicus SB uid58539
.	(node 403)	Desulfomonile tiedjei DSM 6799 uid168320

Continued on next page

Table 12.2 – continued from previous page

Phylum	Family	Organism
.	Desulfuromonadales	Geobacter metallireducens GS 15 uid57731
.	(nodes 405-414)	Geobacter sulfurreducens KN400 uid161977
.		Geobacter M18 uid55771
.		Geobacter bemidjiensis Bem uid58749
.		Geobacter M21 uid59037
.		Geobacter FRC 32 uid58543
.		Geobacter uraniireducens Rf4 uid58475
.		Geobacter lovleyi SZ uid58713
.		Pelobacter propionicus DSM 2379 uid58255
.		Desulfuromonas acetoxidans DSM 684 uid54145
.		Pelobacter carbinolicus DSM 2380 uid58241
.	Myxococcales	Stigmatella aurantiaca DW4 3 1 uid158509
.	(nodes 416-425)	Cystobacter fuscus DSM 2262 uid188352
.		Corallococcus coralloides DSM 2259 uid157997
.		Myxococcus stipitatus DSM 14675 uid186549
.		Chondromyces apiculatus DSM 436 uid175317
.		Myxococcus fulvus HW 1 uid68443
.		Anaeromyxobacter dehalogenans 2CP 1 uid58989
.		Anaeromyxobacter Fw109 5 uid58755
.		Sorangium cellulosum So ce 56 uid61629
.		Haliangium ochraceum DSM 14365 uid41425
.		Plesiocystis pacifica SIR 1 uid54707
.	Bdelovibrionales	Bacteriovorax marinus SJ uid82341
.	(node 426)	Bdellovibrio bacteriovorus HD100 uid61595
.	unclassified	SAR324 cluster bacterium JCVI SC AAA005 uid86871
.	Desulfurellales	Hippea maritima DSM 10411 uid65267
Zeta_proteobacteria	Mariprofundales	Mariprofundus ferrooxydans PV 1 uid54269
Alpha_proteobacteria	Magnetococcales	Magnetococcus MC 1 uid57833
(nodes 428-690	Rickettsiales	Neorickettsia risticii Illinois uid58889
.	(nodes 430-464)	Neorickettsia sennetsu Miyayama uid57965
.		Wolbachia wRi uid59371
.		Wolbachia endosymbiont of Culex quinquefasciatus Pel uid61645
.		Wolbachia pipientis wAlbB uid81759
.		Ehrlichia ruminantium Gardel uid58245
.		Ehrlichia canis Jake uid58071
.		Ehrlichia chaffeensis Arkansas uid57933
.		Anaplasma phagocytophilum HZ uid57951
.		Anaplasma centrale Israel uid42155
.		Anaplasma marginale Florida uid58577
.		Candidatus Midichloria mitochondrii IricVA uid68687
.		Orientia tsutsugamushi Boryong uid61621
.		Rickettsia bellii OSU 85 389 uid58681
.		Rickettsia helvetica C9P9 uid173054
.		Rickettsia heilongjiangensis 054 uid70839
.		Rickettsia japonica YH uid73963
.		Rickettsia slovaca 13 B uid82369
.		Rickettsia sibirica 246 uid54113
.		Rickettsia africae ESF 5 uid58799
.		Rickettsia conorii Malish 7 uid57633

Continued on next page

Table 12.2 – continued from previous page

Phylum	Family	Organism
.		Rickettsia parkeri Portsmouth uid158045
.		Rickettsia peacockii Rustic uid59301
.		Rickettsia philipii 364D uid89383
.		Rickettsia rickettsii Arizona uid86655
.		Candidatus Rickettsia amblyommii GAT 30V uid156845
.		Rickettsia montanensis OSU 85 930 uid158043
.		Rickettsia massiliae AZT80 uid86751
.		Rickettsia rhipicephali 3 7 female6 CWPP uid156977
.		Rickettsia akari Hartford uid58161
.		Rickettsia australis Cutlack uid158039
.		Rickettsia felis URRWXCal2 uid58331
.		Rickettsia endosymbiont of Ixodes scapularis uid55851
.		Rickettsia canadensis CA410 uid88063
.		Rickettsia prowazekii BuV67 CWPP uid158063
.		Rickettsia typhi B9991CWPP uid158357
.	SAR_11	Candidatus Pelagibacter ubique HTCC1062 uid58401
.	(nodes 466-468)	Candidatus Pelagibacter IMCC9063 uid66305
.		alpha proteobacterium HIMB59 uid175778
.	unclassified	Candidatus Hodgkinia cicadicola Dsem uid59311
.	Rhodospirillales	Tistrella mobilis KA081020 065 uid167486
.	Rickettsiales	Candidatus Odyssella thessalonicensis L13 uid72365
.	Rhodospirillales	Azospirillum brasilense Sp245 uid162161
.	(nodes 472-507)	Azospirillum B510 uid46085
.		Azospirillum lipoferum 4B uid82343
.		Azospirillum amazonense Y2 uid73583
.		Rhodospirillum centenum SW uid58805
.		Oceanibaculum indicum P24 uid176351
.		Thalassobaculum L2 uid182483
.		SAR 116 cluster alpha proteobacterium HIMB100 uid78325
.		Candidatus Puniceispirillum marinum IMCC1322 uid47081
.		Thalassospira profundimaris WP0211 uid176349
.		Thalassospira xiamenensis M 5 DSM 17429 uid176348
.		Magnetospirillum magneticum AMB 1 uid58527
.		Phaeospirillum molischianum uid156755
.		Caenispirillum salinarum AK4 uid182892
.		Rhodospirillum photometricum uid159003
.		Rhodospirillum rubrum ATCC 11170 uid57655
.		Micavibrio aeruginosavorus ARL 13 uid73585
.		Acetobacteraceae bacterium AT 5844 uid80697
.		Roseomonas cervicalis ATCC 49957 uid49155
.		Acidiphilium cryptum JF 5 uid58447
.		Acidiphilium multivorum AIU301 uid63345
.		Granulibacter bethesdensis CGDNIH1 uid58661
.		Commensalibacter intestini A911 uid75109
.		Gluconacetobacter diazotrophicus PAl 5 uid59075
.		Gluconacetobacter hansenii ATCC 23769 uid48811
.		Gluconacetobacter europaeus LMG 18494 uid73763
.		Gluconacetobacter oboediens 174Bp2 uid73765
.		Gluconacetobacter SXCC 1 uid66379

Continued on next page

Table 12.2 – continued from previous page

Phylum	Family	Organism
.		Gluconacetobacter xylinus NBRC 3288 uid46523
.		Gluconobacter frateurii NBRC 101659 uid178735
.		Gluconobacter thailandicus NBRC 3255 uid191942
.		Gluconobacter morbifer G707 uid76941
.		Gluconobacter oxydans 621H uid58239
.		Acetobacter aceti NBRC 14818 uid70715
.		Acetobacter tropicalis NBRC 101654 uid68643
.		Acetobacter pasteurianus IFO 3283 01 42C uid158377
.		Acetobacter pomorum DM001 uid65823
.	Sphingomonadales	Sphingopyxis alaskensis RB2256 uid58351
.	(nodes 509-534)	Sphingobium SYK 6 uid73353
.		Sphingomonas SKA58 uid54251
.		Sphingobium chlorophenolicum L 1 uid52597
.		Sphingobium indicum B90A uid167985
.		Sphingobium japonicum UT26S uid47077
.		Sphingobium AP49 uid171561
.		Sphingobium yanoikuyae ATCC 51230 uid182888
.		Zymomonas mobilis ATCC 10988 uid55403
.		Sphingomonas wittichii RW1 uid58691
.		Sphingomonas KC8 uid77733
.		Sphingomonas elodea ATCC 31461 uid157063
.		Sphingomonas echinoides ATCC 14820 uid169543
.		Sphingomonas PAMC 26605 uid169544
.		Sphingomonas S17 uid66923
.		Novosphingobium aromaticivorans DSM 12444 uid57747
.		Novosphingobium nitrogenifigens DSM 19370 uid64475
.		Novosphingobium pentaromativorans US6 1 uid78315
.		Novosphingobium PP1Y uid67383
.		Sphingomonas LH128 uid174245
.		Novosphingobium AP12 uid171681
.		Novosphingobium Rr 2 17 uid170038
.		Erythrobacter NAP1 uid54197
.		Erythrobacter litoralis HTCC2594 uid58299
.		Erythrobacter SD 21 uid54677
.		Citromicrobium bathyomarinum JL354 uid48817
.		Citromicrobium JLT1363 uid70837
.	Rhodobacteraceae	Rhodobacter sphaeroides 2 4 1 uid57653
.	(nodes 536-576)	Rhodobacter SW2 uid40865
.		Rhodobacter capsulatus SB 1003 uid47509
.		Paracoccus denitrificans PD1222 uid58187
.		Paracoccus TRP uid70151
.		Ketogulonicigenium vulgare Y25 uid59581
.		Ketogulonigenium vulgarum WSH 001 uid161161
.		Oceaniovalibus guishaninsula JLT2003 uid176356
.		Oceanicola granulosus HTCC2516 uid54177
.		Thalassiobium R2A62 uid55955
.		Octadecabacter antarcticus 238 uid54699
.		Loktanella vestfoldensis SKA53 uid54169
.		Roseobacter CCS2 uid54627

Continued on next page

Table 12.2 – continued from previous page

Phylum	Family	Organism
.		Celeribacter baekdonensis B30 uid176352
.		Maritimibacter alkaliphilus HTCC2654 uid53325
.		Dinoroseobacter shibae DFL 12 uid58707
.		Jannaschia CCS1 uid58147
.		Rhodobacterales bacterium HTCC2083 uid54195
.		Roseobacter denitrificans OCh 114 uid58597
.		Roseobacter litoralis Och 149 uid54719
.		Oceanibulbus indolifex HEL 45 uid54697
.		Roseobacter GAI101 uid55361
.		Sulfitobacter EE 36 uid54191
.		Oceanicola S124 uid81119
.		Sagittula stellata E 37 uid54621
.		Pelagibaca bermudensis HTCC2601 uid53327
.		Citreicella SE45 uid55957
.		Citreicella 357 uid160001
.		Ruegeria TM1040 uid58193
.		Silicibacter TrichCH4B uid55971
.		Phaeobacter gallaeciensis 2 10 uid54715
.		Ruegeria R11 uid54725
.		Roseobacter MED193 uid54257
.		Roseobacter SK209 2 6 uid54629
.		Ruegeria pomeroyi DSS 3 uid57863
.		Silicibacter lacuscaerulensis ITI 1157 uid55959
.		Rhodobacteraceae bacterium KLH11 uid55441
.		Ruegeria TW15 uid73135
.		Roseovarius nubinhibens ISM uid54183
.		Roseovarius TM1035 uid54723
.		Roseobacter AzwK 3b uid54721
.		Oceanicola batsensis HTCC2597 uid54175
.	Parvularculales	Parvularcula bermudensis HTCC2503 uid51641
.	Caulobacterales	Phenylobacterium zucineum HLK1 uid58959
.	(nodes 580-588)	Caulobacter AP07 uid171682
.		Caulobacter K31 uid58551
.		Caulobacter crescentus CB15 uid57891
.		Caulobacter segnis ATCC 21756 uid41709
.		Brevundimonas diminuta 470 4 uid183780
.		Brevundimonas subvibrioides ATCC 15264 uid42117
.		Brevundimonas BAL3 uid54665
.		Asticcacaulis biprosthecum C19 uid66137
.		Asticcacaulis excentricus CB 48 uid55641
.	Hyphomonadaceae	Hirschia baltica ATCC 49814 uid59365
.	(nodes 589-591)	Hyphomonas neptunium ATCC 15444 uid58433
.		Maricaulis maris MCS10 uid58689
.		Oceanicaulis alexandrii HTCC2633 uid54173
.	Phyllobacteriaceae	Parvibaculum lavamentivorans DS 1 uid58739
.	Hyphomicrobiaceae	Rhodomicrobium vannielii ATCC 17100 uid43247
.	(nodes 594-595)	Hyphomicrobium denitrificans ATCC 51888 uid50325
.		Hyphomicrobium MC1 uid68453
.	Methylobacteriaceae	Microvirga WSM3557 uid167863

Continued on next page

Table 12.2 – continued from previous page

Phylum	Family	Organism
.	(nodes 599-604)	Methylobacterium 4 46 uid58843
.		Methylobacterium nodulans ORS 2060 uid59023
.		Methylobacterium extorquens AM1 uid57605
.		Methylobacterium populi BJ001 uid58937
.		Methylobacterium GXF4 uid170037
.		Methylobacterium radiotolerans JCM 2831 uid58845
.	Beijerinckiaceae	Beijerinckia indica ATCC 9039 uid59057
.	(node 606)	Methylocella silvestris BL2 uid59433
.	Methylocystaceae	Methylosinus trichosporium OB3b uid49119
.	(node 607-608)	Methylocystis ATCC 49242 uid62743
.		Methylocystis SC2 uid174072
.	Xanthobacteraceae	Starkeya novella DSM 506 uid48815
.	(nodes 610-611)	Azorhizobium caulinodans ORS 571 uid58905
.		Xanthobacter autotrophicus Py2 uid58453
.	Rhodobacteraceae	Rhodovulum PH10 uid174236
.	Bradyrhizobiaceae	Bradyrhizobiaceae bacterium SG 6C uid68635
.	(Nodes 613-625)	Afipia 1NLS2 uid49955
.		Oligotropha carboxidovorans OM4 uid162135
.		Rhodopseudomonas palustris BisA53 uid58445
.		Nitrobacter hamburgensis X14 uid58293
.		Nitrobacter winogradskyi Nb 255 uid58295
.		Nitrobacter Nb 311A uid54203
.		Bradyrhizobium WSM471 uid82711
.		Bradyrhizobium japonicum USDA 110 uid57599
.		Bradyrhizobium S23321 uid158167
.		Bradyrhizobium YR681 uid171652
.		Bradyrhizobium STM 3843 uid80711
.		Bradyrhizobium BTAi1 uid58505
.		Bradyrhizobium uid80709
.	Hyphomicrobiaceae	Pelagibacterium halotolerans B2 uid74393
.	Rhodobacteraceae	Pseudovibrio JE062 uid54711
.	(nodes 628-631)	Polymorphum gilvum SL003B 26A1 uid65447
.		Roseibium TrichSKD4 uid53349
.		Labrenzia aggregata IAM 12614 uid54575
.		Labrenzia alexandrii DFL 11 uid54727
.		Ahrensia R2A130 uid51717
.	Aurantimonadaceae	Aurantimonas manganoxydans SI85 9A1 uid54267
.	(node 634)	Fulvimarina pelagi HTCC2506 uid54193
.	Phyllobacteriaceae	Hoeflea phototrophica DFL 43 uid54683
.	Rhizobiaceae	Sinorhizobium medicae WSM419 uid58549
.	(nodes 637-654)	Sinorhizobium meliloti 1021 uid57603
.		Sinorhizobium fredii HH103 uid86865
.		Candidatus Liberibacter asiaticus psy62 uid59227
.		bacterium BT 1 uid184079
.		Rhizobium CF122 uid171679
.		Rhizobium mesoamericanum STM3625 uid178067
.		Rhizobium CF142 uid175318
.		Rhizobium etli CFN 42 uid58377
.		Rhizobium CCGE 510 uid173872

Continued on next page

Table 12.2 – continued from previous page

Phylum	Family	Organism
.		Rhizobium leguminosarum bv trifolii WSM1325 uid58991
.		Rhizobium AP16 uid171562
.		Rhizobium tropici CIAT 899 uid185179
.		Agrobacterium H13 3 uid63403
.		Agrobacterium tumefaciens C58 uid57865
.		Rhizobium CF080 uid180531
.		Agrobacterium vitis S4 uid58249
.		Agrobacterium albertimagni AOL15 uid176603
.		Rhizobium PDO1 076 uid180152
.	Phyllobacteriaceae	Mesorhizobium alhagi CCNWXJ12 2 uid78825
.	(nodes 656-664)	Mesorhizobium loti MAFF303099 uid57601
.		Mesorhizobium amorphae CCNWGS0123 uid76939
.		Mesorhizobium opportunistum WSM2075 uid40861
.		Mesorhizobium ciceri biovar biserrulae WSM1271 uid62101
.		Mesorhizobium australicum WSM2073 uid75101
.		Chelativorans BNC1 uid58069
.		Nitratireductor indicus C115 uid176490
.		Nitratireductor aquibiodomus RA22 uid168057
.		Nitratireductor pacificus pht 3B uid176350
.	Bartonellaceae	Bartonella tamiae Th239 uid170859
.	(nodes 666-680)	Bartonella bacilliformis KC583 uid58533
.		Bartonella melophagi K 2C uid170858
.		Bartonella clarridgeiae 73 uid62131
.		Bartonella doshiae NCTC 12862 uid170852
.		Bartonella alsatica IBS 382 uid170862
.		Bartonella DB5 6 uid170850
.		Bartonella taylorii 8TBB uid170861
.		Bartonella elizabethae F9251 uid181718
.		Bartonella tribocorum CIP 105476 uid59129
.		Bartonella grahamii as4aup uid59405
.		Bartonella rattimassiliensis 15908 uid170856
.		Bartonella birtlesii LL WM9 uid170860
.		Bartonella washoensis 085 0475 uid170851
.		Bartonella quintana RM 11 uid174512
.		Bartonella henselae Houston 1 uid57745
.	Phyllobacteriaceae	Phyllobacterium YR531 uid171653
.	Brucellaceae	Ochrobactrum anthropi ATCC 49188 uid58921
.	(nodes 682-690)	Brucella melitensis ATCC 23457 uid59241
.		Brucella pinnipedialis B2 94 uid71131
.		Brucella ovis ATCC 25840 uid58113
.		Brucella ceti B1 94 uid55807
.		Brucella microti CCM 4915 uid59319
.		Brucella neotomae 5K33 uid55801
.		Brucella abortus A13334 uid83615
.		Brucella canis ATCC 23365 uid59009
.		Brucella suis 1330 uid159871

Jump to: *Bacteroidetes, et al.*, *Alpha proteobacteria, et al.*, *Beta and Gamma Proteobacteria*, *Actinobacteria*, *Firmicutes, et al.*, *Cyanobacteria*, *Archaea*, and *Eukaryotes*.

12.3 Beta and Gamma Proteobacteria

This group contains Enteric, Haemophilus, Vibrio, Shewanella, Pseudomonads, Coxiella, Thiomicrospira, Nitrosococcus, Xanthomonads, Methylobacteria, Acidothiobacillus, Francisella, Neisseria, Burkholderiales, Azoarcus, Thiobacillus.

- Gamma proteobacteria 1, including `Enterobacteriales, Vibrionales, Alteromonadales, Shewanellaceae, and others`
- Gamma proteobacteria 2, including `Pseudomonadales, Moraxellaceae, Oceanospirillales, Xanthomonadales, and others`
- `Beta proteobacteria`.

Phylum	Family	Organism
Gamma_proteobacteria	Enterobacteriales	Serratia marcescens FGI94 uid185180
(nodes 692-1082)	(nodes 705-792)	Serratia symbiotica Cinara cedri uid82363
.		Serratia odorifera 4Rx13 uid42253
.		Serratia plymuthica AS9 uid67313
.		Serratia proteamaculans 568 uid58725
.		Rahnella aquatilis CIP 78 65 ATCC 33071 uid86855
.		Serratia M24T3 uid158323
.		Candidatus Hamiltonella defensa 5AT Acyrthosiphon pisum uid59289
.		Candidatus Regiella insecticola LSR1 uid51727
.		Yersinia frederiksenii ATCC 33641 uid54347
.		Yersinia enterocolitica 8081 uid57741
.		Yersinia aldovae ATCC 35236 uid55243
.		Yersinia intermedia ATCC 29909 uid54349
.		Yersinia bercovieri ATCC 43970 uid54343
.		Yersinia mollaretii ATCC 43969 uid54345
.		Yersinia ruckeri ATCC 29473 uid55249
.		Yersinia rohdei ATCC 43380 uid55247
.		Yersinia kristensenii ATCC 33638 uid55245
.		Yersinia pestis KIM 10 uid57875
.		Yersinia pseudotuberculosis IP 31758 uid58487
.		Brenneria EniD312 uid75111
.		Pectobacterium carotovorum PC1 uid59295
.		Pectobacterium atrosepticum SCRI1043 uid57957
.		Pectobacterium wasabiae WPP163 uid41297
.		Hafnia alvei ATCC 51873 uid80687
.		Edwardsiella ictaluri 93 146 uid59403
.		Edwardsiella tarda EIB202 uid41819
.		Photorhabdus asymbiotica uid59243
.		Photorhabdus luminescens laumondii TTO1 uid61593
.		Xenorhabdus bovienii SS 2004 uid46345
.		Xenorhabdus nematophila ATCC 19061 uid49133
.		Proteus mirabilis HI4320 uid61599
.		Morganella morganii KT uid180867
.		Providencia stuartii MRSN 2154 uid162193
.		Providencia burhodogranariea DSM 19968 uid181280
.		Providencia rettgeri DSM 1131 uid55121
.		Providencia rustigianii DSM 4541 uid55071

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Providencia alcalifaciens Dmel2 uid181281
.		Erwinia billingiae Eb661 uid50547
.		Erwinia tasmaniensis Et1 99 uid59029
.		Erwinia amylovora ATCC 49946 uid46943
.		Erwinia pyrifoliae DSM 12163 uid159693
.		Pantoea At 9b uid55845
.		Pantoea GM01 uid171648
.		Pantoea stewartii DC283 uid86627
.		Pantoea ananatis AJ13355 uid162073
.		Pantoea Sc1 uid157059
.		Pantoea agglomerans 299R uid185797
.		Pantoea vagans C9 1 uid49871
.		Cronobacter sakazakii ATCC BAA 894 uid58145
.		Cronobacter turicensis z3032 uid40821
.		Enterobacter Ag1 uid170516
.		Escherichia hermannii NBRC 105704 uid84453
.		Enterobacter cloacae ATCC 13047 uid48363
.		Enterobacter hormaechei ATCC 49162 uid67395
.		Enterobacter cancerogenus ATCC 35316 uid55079
.		Enterobacter mori LMG 25706 uid75365
.		Enterobacter 638 uid58727
.		Enterobacter asburiae LF7a uid72793
.		Yokenella regensburgei ATCC 43003 uid80693
.		Citrobacter freundii 4 7 47CFAA uid80411
.		Citrobacter youngae ATCC 29220 uid55081
.		Enterobacteriaceae bacterium FGI 57 uid185181
.		Klebsiella oxytoca E718 uid170256
.		Enterobacter aerogenes EA1509E uid187411
.		Klebsiella pneumoniae 1084 uid174151
.		Klebsiella variicola At 22 uid42113
.		Citrobacter koseri ATCC BAA 895 uid58143
.		Citrobacter rodentium ICC168 uid43089
.		Salmonella bongori NCTC 12419 uid70155
.		Salmonella enterica arizonae serovar 62 z4 z23 uid58191
.		Escherichia blattae DSM 4481 uid165043
.		Escherichia coli K 12 substr MG1655 uid57779
.		Shigella flexneri 2002017 uid159233
.		Shigella sonnei 53G uid84383
.		Shigella boydii CDC 3083 94 uid58415
.		Shigella dysenteriae Sd197 uid58213
.		Escherichia fergusonii ATCC 35469 uid59375
.		Escherichia albertii TW07627 uid55089
.		Dickeya dadantii 3937 uid52537
.		Dickeya zeae Ech1591 uid59297
.		Sodalis glossinidius morsitans uid58553
.		Candidatus Moranella endobia PCIT uid68739
.		secondary endosymbiont of Ctenarytaina eucalypti uid172737
.		Baumannia cicadellinicola Hc Homalodisca coagulata uid58111
.		Buchnera aphidicola 5A Acyrthosiphon pisum uid59285

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis uid57853
.		Candidatus Blochmannia chromaiodes 640 uid185308
.		Candidatus Riesia pediculicola USDA uid46841
.	Pasteurellales	Gallibacterium anatis UMN179 uid66567
.	(nodes 793-817)	Haemophilus parasuis SH0165 uid59273
.		Mannheimia haemolytica PHL213 uid54093
.		Actinobacillus suis H91 0380 uid176363
.		Actinobacillus ureae ATCC 25976 uid62663
.		Haemophilus ducreyi 35000HP uid57625
.		Actinobacillus pleuropneumoniae serovar 3 JL03 uid58891
.		Actinobacillus minor 202 uid55633
.		Haemophilus sputorum HK 2154 uid173855
.		Haemophilus parahaemolyticus HK385 uid165733
.		Haemophilus paraphrohaemolyticus HK411 uid162935
.		Pasteurella dagmatis ATCC 43325 uid41077
.		Pasteurella multocida 3480 uid161955
.		Haemophilus somnus 129PT uid57929
.		Actinobacillus succinogenes 130Z uid58247
.		Mannheimia succiniciproducens MBEL55E uid58197
.		Pasteurella bettyae CCUG 2042 uid165735
.		Haemophilus haemolyticus HK386 uid180431
.		Haemophilus oral taxon 851 F0397 uid81771
.		Haemophilus aegyptius ATCC 11116 uid65831
.		Haemophilus influenzae 10810 uid86647
.		Haemophilus parainfluenzae T3T1 uid72801
.		Haemophilus pittmaniae HK 85 uid72189
.		Aggregatibacter actinomycetemcomitans ANH9381 uid80743
.		Aggregatibacter segnis ATCC 33393 uid61481
.		Aggregatibacter aphrophilus NJ8700 uid59407
.	Vibrionales	Vibrio parahaemolyticus BB22OP uid184822
.	(nodes 818-851)	Vibrio EJY3 uid83161
.		Vibrio alginolyticus 12G01 uid54241
.		Vibrio ichthyoenteri ATCC 700023 uid72183
.		Vibrio scophthalmi LMG 19158 uid72185
.		Vibrio rotiferianus DAT722 uid73893
.		Vibrio AND4 uid54233
.		Vibrio campbellii CAIM 519 uid188135
.		Vibrio harveyi ATCC BAA 1116 uid58957
.		Vibrio sinaloensis DSM 21326 uid63103
.		Vibrio brasiliensis LMG 20546 uid63101
.		Vibrio orientalis CIP 102891 ATCC 33934 uid179869
.		Vibrio coralliilyticus ATCC BAA 450 uid41031
.		Vibrio tubiashii ATCC 19109 uid72181
.		Vibrio caribbenthicus ATCC BAA 2122 uid60579
.		Vibrio nigripulchritudo ATCC 27043 uid72179
.		Vibrio shilonii AK1 uid54743
.		Vibrionales bacterium SWAT 3 uid54745
.		Vibrio cyclitrophicus ZF14 uid175805
.		Vibrio splendidus LGP32 uid59353

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Vibrio anguillarum 775 uid68057
.		Vibrio ordalii ATCC 33509 uid80489
.		Vibrio vulnificus CMCP6 uid62909
.		Vibrio furnissii NCTC 11218 uid82347
.		Vibrio cholerae IEC224 uid89389
.		Vibrio RC586 uid41619
.		Aliivibrio salmonicida LFI1238 uid59251
.		Vibrio fischeri ES114 uid58163
.		Enterovibrio AK16 uid188104
.		Grimontia hollisae CIP 101886 uid41501
.		Photobacterium profundum SS9 uid62923
.		Photobacterium AK15 uid186537
.		Photobacterium damselae CIP 102761 uid41873
.		Photobacterium leiognathi mandapamensis svers 1 1 uid66161
.		Photobacterium angustum S14 uid54237
.	Aeromonadales	Oceanimonas GK1 uid81627
.	(nodes 852-859)	Aeromonas veronii B565 uid66323
.		Aeromonas salmonicida A449 uid58631
.		Aeromonas media WS uid180868
.		Aeromonas caviae Ae398 uid67585
.		Aeromonas hydrophila ATCC 7966 uid58617
.		Aeromonas aquariorum AAK1 uid178353
.		Tolumonas auensis DSM 9187 uid59395
.		Succinatimonas hippei YIT 12066 uid62747
.	Alteromonadales_1	Moritella PE36 uid54201
.	(nodes 960-878)	Psychromonas CNPT3 uid54249
.		Psychromonas ingrahamii 37 uid58521
.		Ferrimonas balearica DSM 9799 uid53371
.		Shewanella amazonensis SB2B uid58257
.		Shewanella MR 4 uid58345
.		Shewanella oneidensis MR 1 uid57949
.		Shewanella baltica BA175 uid52601
.		Shewanella HN 41 uid68137
.		Shewanella putrefaciens 200 uid161927
.		Shewanella denitrificans OS217 uid58263
.		Shewanella frigidimarina NCIMB 400 uid58265
.		Shewanella loihica PV 4 uid58349
.		Shewanella piezotolerans WP3 uid58745
.		Shewanella halifaxensis HAW EB4 uid59007
.		Shewanella pealeana ATCC 700345 uid58705
.		Shewanella woodyi ATCC 51908 uid58721
.		Shewanella sediminis HAW EB3 uid58835
.		Shewanella benthica KT99 uid54155
.		Shewanella violacea DSS12 uid47085
.	unclassified	Gallaecimonas xiamenensis 3 C 1 uid176353
.	Alteromonadales_2	Glaciecola HTCC2999 uid54867
.	(nodes 881-891)	Glaciecola lipolytica E3 uid178730
.		Glaciecola arctica BSs20135 uid178729
.		Glaciecola psychrophila 170 uid178725

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Glaciecola chathamensis S18K6 uid178731
.		Glaciecola polaris LMG 21857 uid178726
.		Glaciecola mesophila KMM 241 uid178728
.		Alteromonas macleodii AltDE1 uid179068
.		Alteromonas SN2 uid67349
.		Glaciecola punicea DSM 14233 ACAM 611 uid157053
.		Glaciecola nitratireducens FR1064 uid73759
.		Glaciecola pallidula DSM 14239 ACAM 615 uid178727
.	Chromatiales	Rheinheimera A13L uid68139
.	(nodes 893-894)	Alishewanella aestuarii B11 uid171146
.		Alishewanella agri BL06 uid170036
.	Pseudoalteromonadaceae	Pseudoalteromonas tunicata D2 uid54181
.	(nodes 895-907)	Pseudoalteromonas spongiae UST010723 006 uid168330
.		Pseudoalteromonas luteoviolacea B ATCC 29581 uid186644
.		Pseudoalteromonas rubra ATCC 29570 uid168329
.		Pseudoalteromonas flavipulchra JG1 uid177806
.		Pseudoalteromonas piscicida JCM 20779 uid168328
.		Pseudoalteromonas citrea NCIMB 1889 uid168326
.		Pseudoalteromonas SM9913 uid61247
.		Pseudoalteromonas undina NCIMB 2128 uid168331
.		Pseudoalteromonas marina mano4 uid168327
.		Pseudoalteromonas haloplanktis TAC125 uid58431
.		Pseudoalteromonas arctica A 37 1 2 uid168325
.		Pseudoalteromonas Bsw20308 uid179221
.		Pseudoalteromonas BSi20652 uid78645
.	Alteromonadales	Colwellia psychrerythraea 34H uid57855
.	Idiomarinaceae	Idiomarina A28L uid68223
.	(nodes 909-911)	Idiomarina xiamenensis 10 D 4 uid176354
.		Idiomarina baltica OS145 uid54159
.		Idiomarina loihiensis L2TR uid58087
.	Oceanospirillales	Kangiella koreensis DSM 16069 uid59209
.	Pseudomonadales	Pseudomonas GM18 uid171673
.	(nodes 915-944)	Pseudomonas mandelii JR 1 uid175806
.		Pseudomonas GM41 2012 uid171667
.		Pseudomonas chlororaphis aureofaciens 30 84 uid181709
.		Pseudomonas GM80 uid171657
.		Pseudomonas GM30 uid171669
.		Pseudomonas GM55 uid171663
.		Pseudomonas fragi A22 uid173061
.		Pseudomonas viridiflava UASWS0038 uid178273
.		Pseudomonas avellanae BPIC 631 uid177065
.		Pseudomonas syringae B728a uid57931
.		Pseudomonas brassicacearum NFM421 uid66303
.		Pseudomonas PAMC 25886 uid170201
.		Pseudomonas fluorescens A506 uid165185
.		Pseudomonas synxantha BG33R uid167316
.		Pseudomonas extremaustralis 14 3 substr 14 3b uid170171
.		Pseudomonas poae RE 1 1 14 uid188480
.		Pseudomonas fuscovaginae UPB0736 uid174621

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Pseudomonas M47T1 uid167375
.		Pseudomonas HYS uid177807
.		Pseudomonas entomophila L48 uid58639
.		Pseudomonas GM84 uid171656
.		Pseudomonas putida BIRD 1 uid162055
.		Pseudomonas fulva 12 X uid67351
.		Pseudomonas pseudoalcaligenes CECT 5344 uid171984
.		Pseudomonas mendocina NK 01 uid66299
.		Pseudomonas Chol1 uid178064
.		Pseudomonas stutzeri A1501 uid58641
.		Azotobacter vinelandii DJ uid57597
.		Pseudomonas psychrotolerans L19 uid78819
.		Pseudomonas aeruginosa DK2 uid168996
.	OMG_group	Congregibacter litoralis KT71 uid54253
.	(node 946)	marine gamma proteobacterium HTCC2080 uid54583
.	Alteromonadales_3	Alteromonas S89 uid81111
.	(nodes 947-951)	Saccharophagus degradans 2 40 uid57921
.		Teredinibacter turnerae T7901 uid59267
.		Simiduia agarivorans SA1 uid177713
.		Cellvibrio BR uid167161
.		Cellvibrio japonicus Ueda107 uid59139
.	Oceanospirillales	Hahella chejuensis KCTC 2396 uid58483
.	Alteromonadales_4	Marinobacter aquaeolei VT8 uid59419
.	(nodes 956-959)	Marinobacter adhaerens HP15 uid162009
.		Marinobacter algicola DG893 uid54691
.		Marinobacter BSs20148 uid171995
.		Marinobacter ELB17 uid54619
.	Oceanospirillales_1	Marinomonas MED121 uid54255
.	(nodes 960-966)	Marinomonas mediterranea MMB 1 uid64753
.		Marinomonas MWYL1 uid58715
.		Marinomonas posidonica IVIA Po 181 uid67323
.		Marinobacterium stanieri S30 uid81115
.		Neptuniibacter caesariensis uid54227
.		Bermanella marisrubri uid54229
.		Reinekea blandensis MED297 uid54199
.	unclassified	SAR86 cluster bacterium SAR86C uid173062
.	Oceanospirillales_2	Candidatus Portiera aleyrodidarum BT B uid173859
.	(nodes 967-974)	Chromohalobacter salexigens DSM 3043 uid62921
.		Halomonas elongata DSM 2581 uid52781
.		Halomonas KM 1 uid171986
.		Halomonas boliviensis LC1 uid78313
.		Halomonas titanicae BH1 uid188651
.		Halomonas GFAJ 1 uid78821
.		Halomonas TD01 uid68639
.	Oceanospirillales_3	gamma proteobacterium HdN1 uid51635
.	(nodes 975-979)	Alcanivorax W11 5 uid176219
.		Alcanivorax dieselolei B5 uid176364
.		Alcanivorax hongdengensis A 11 3 uid176602
.		Alcanivorax borkumensis SK2 uid58169

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Alcanivorax DG881 uid54659
.	Moraxellaceae	Enhydrobacter aerosaccus SK60 uid55439
.	(nodes 980-1004)	Moraxella macacae 0408225 uid183957
.		Moraxella catarrhalis BBH18 uid48809
.		Psychrobacter PAMC 21119 uid172255
.		Psychrobacter arcticus 273 4 uid58021
.		Psychrobacter cryohalolentis K5 uid58373
.		Psychrobacter 1501 2011 uid67091
.		Psychrobacter PRwf 1 uid58459
.		Acinetobacter radioresistens DSM 6976 NBRC 102413 uid180968
.		Acinetobacter ADP1 uid61597
.		Acinetobacter ursingii DSM 16037 CIP 107286 uid173056
.		Acinetobacter bereziniae LMG 1003 uid173060
.		Acinetobacter P8 3 8 uid77731
.		Acinetobacter johnsonii SH046 uid41583
.		Acinetobacter lwoffii SH145 uid41587
.		Acinetobacter HA uid167861
.		Acinetobacter RUH2624 uid40849
.		Acinetobacter nosocomialis Ab22222 uid181746
.		Acinetobacter baumannii 1656 2 uid158677
.		Acinetobacter calcoaceticus PHEA 2 uid83123
.		Acinetobacter oleivorans DR1 uid50119
.		Acinetobacter NBRC 100985 uid78547
.		Acinetobacter parvus DSM 16617 CIP 108168 uid173057
.		Acinetobacter junii SH205 uid41585
.		Acinetobacter ATCC 27244 uid55389
.		Acinetobacter haemolyticus ATCC 19194 uid47361
.	Legionellales	Legionella pneumophila 2300 99 Alcoy uid48801
.	(nodes 1005-1010)	Legionella drancourtii LLAP12 uid56003
.		Legionella longbeachae NSW150 uid46099
.		Fluoribacter dumoffii Tex KL uid167163
.		Coxiella burnetii CbuG Q212 uid58893
.		Diplorickettsia massiliensis 20B uid176383
.		Rickettsiella grylli uid54407
.	Thiotrichales	Beggiatoa alba B18LD uid163695
.	(nodes 1012-1015)	Methylophaga aminisulfidivorans MP uid67817
.		Methylophaga thiooxidans DMS010 uid54693
.		Methylophaga JAM1 uid162947
.		Methylophaga JAM7 uid162949
.	Chromatiales	Nitrosococcus halophilus Nc4 uid46803
.	(nodes 1018-1028)	Nitrosococcus oceani ATCC 19707 uid58403
.		Nitrosococcus watsonii C 113 uid50331
.		Thiothrix nivea DSM 5205 uid163685
.		Thioalkalivibrio K90mix uid46181
.		Thioalkalivibrio nitratireducens DSM 14787 uid184011
.		Thioalkalivibrio thiocyanoxidans ARh 4 uid74027
.		Thiorhodospira sibirica ATCC 700588 uid74023
.		Ectothiorhodospira PHS 1 uid82731
.		Nitrococcus mobilis Nb 231 uid54185

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Alkalilimnicola ehrlichii MLHE 1 uid58467
.		Halorhodospira halophila SL1 uid58473
.	Xanthomonadales	Hydrocarboniphaga effusa AP103 uid169224
.	(nodes 1029-1050)	Salinisphaera shabanensis E1L3A uid67923
.		Frateuria aurantia DSM 6220 uid81775
.		Rhodanobacter fulvus Jip2 uid167858
.		Rhodanobacter spathiphylli B39 uid167859
.		Rhodanobacter 2APBS1 uid74431
.		Rhodanobacter thiooxydans LCS2 uid167856
.		Pseudoxanthomonas suwonensis 11 1 uid62105
.		Xylella fastidiosa 9a5c uid57849
.		Xanthomonas campestris 8004 uid57595
.		Xanthomonas oryzae KACC 10331 uid58155
.		Xanthomonas perforans 91 118 uid63619
.		Xanthomonas fuscans aurantifolii ICPB 10535 uid47495
.		Xanthomonas axonopodis citri 306 uid57889
.		Xanthomonas gardneri ATCC 19865 uid63615
.		Xanthomonas vesicatoria ATCC 35937 uid63613
.		Stenotrophomonas maltophilia D457 uid162199
.		Pseudomonas geniculata N1 uid176938
.		Stenotrophomonas SKA14 uid54729
.		Pseudoxanthomonas spadix BD a59 uid75113
.		Xanthomonas translucens DAR61454 uid185791
.		Xanthomonas albilineans GPE PC73 uid43163
.		Xanthomonas sacchari NCPPB 4393 uid86877
.	Chromatiales	Thiorhodovibrio 970 uid74037
.	(nodes 1052-1057)	Thioflavicoccus mobilis 8321 uid184343
.		Thiocapsa marina 5811 uid72573
.		Marichromatium purpuratum 984 uid72575
.		Allochromatium vinosum DSM 180 uid46083
.		Thiorhodococcus drewsii AZ1 uid72963
.		Thiocystis violascens DSM 198 uid74025
.	Thiotrichales	Cycloclasticus P1 uid176368
.	Methylococcales	Methylococcus capsulatus Bath uid57607
.	(nodes 1059-1062)	Methylomonas methanica MC09 uid67363
.		Methylomicrobium alcaliphilum uid77119
.		Methylobacter tundripaludum SV96 uid53253
.		Methylomicrobium album BG8 uid67385
.	Cardiobacteriales	Dichelobacter nodosus VCS1703A uid57643
.	(nodes 1064-1065)	Cardiobacterium hominis ATCC 15826 uid55949
.		Cardiobacterium valvarum F0432 uid80727
.	Xanthanomonadales	Wohlfahrtiimonas chitiniclastica SH04 uid188353
.	Chromatiales	Halothiobacillus neapolitanus c2 uid41317
.	sulfur-oxid_symb	Candidatus Ruthia magnifica Cm Calyptogena magnifica uid58645
.	(node 1069)	Candidatus Vesicomyosocius okutanii HA uid59427
.	Piscirickettsiaceae	Thiomicrospira crunogena XCL 2 uid58183
.	(nodes 1070-1071)	Thioalkalimicrobium aerophilum AL3 uid74029
.		Thioalkalimicrobium cyclicum ALM1 uid67391
.	unclassified	Candidatus Tremblaya princeps PCIT uid68741

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.	Burkholderiales	Candidatus Zinderia insecticola CARI uid52459
.	unclassified	Candidatus Carsonella ruddii CE isolate Thao2000 uid172732
.	Francisellaceae	Francisella cf novicida 3523 uid162107
.	(nodes 1075-1079)	Francisella novicida U112 uid58499
.		Francisella tularensis FSC198 uid58693
.		Francisella noatunensis orientalis Toba 04 uid164779
.		Francisella philomiragia ATCC 25017 uid59105
.		Francisella TX077308 uid68321
.	Acidithiobacillales	Acidithiobacillus caldus SM 1 uid70791
.	(nodes 1080-1082)	Acidithiobacillus thiooxidans ATCC 19377 uid157459
.		Acidithiobacillus ferrivorans SS3 uid67387
.		Acidithiobacillus ferrooxidans ATCC 23270 uid57649
Beta_proteobacteria	Burkholderiaceae	Candidatus Burkholderia kirkii UZHbot1 uid74017
(nodes 1083-1218)	(nodes 1088-1133)	Burkholderia SJ98 uid160003
.		Burkholderia YI23 uid81081
.		Burkholderia CCGE1002 uid42523
.		Burkholderia H160 uid55101
.		Burkholderia phytofirmans PsJN uid58729
.		Burkholderia xenovorans LB400 uid57823
.		Burkholderia CCGE1001 uid42975
.		Burkholderia phenoliruptrix BR3459a uid176370
.		Burkholderia phymatum STM815 uid58699
.		Burkholderia terrae BS001 uid168186
.		Burkholderia gladioli BSR3 uid66301
.		Burkholderia glumae BGR1 uid59397
.		Burkholderia oklahomensis C6786 uid54789
.		Burkholderia thailandensis E264 uid58081
.		Burkholderia mallei NCTC 10247 uid58385
.		Burkholderia pseudomallei K96243 uid57733
.		Burkholderia multivorans ATCC 17616 uid58697
.		Burkholderia 383 uid58073
.		Burkholderia ubonensis Bu uid54793
.		Burkholderia ambifaria AMMD uid58303
.		Burkholderia TJI49 uid179699
.		Burkholderia cepacia GG4 uid173858
.		Burkholderia vietnamiensis G4 uid58075
.		Burkholderia cenocepacia AU 1054 uid58371
.		Burkholderia rhizoxinica HKI 454 uid60487
.		Candidatus Glomeribacter gigasporarum uid73781
.		Polynucleobacter necessarius asymbioticus QLW P1DMWA 1 uid58611
.		Ralstonia PBA uid170039
.		Ralstonia pickettii 12D uid58859
.		Ralstonia solanacearum CFBP2957 uid50545
.		Cupriavidus metallidurans CH34 uid57815
.		Cupriavidus taiwanensis LMG 19424 uid61615
.		Cupriavidus necator N 1 uid68689
.		Ralstonia eutropha H16 uid62925
.		Oxalobacter formigenes HOxBLS uid55623
.		Herbaspirillum CF444 uid171676

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Herbaspirillum YR522 uid171654
.		Herbaspirillum frisingense GSF30 uid179598
.		Herbaspirillum seropedicae SmR1 uid50427
.		Collimonas fungivorans Ter331 uid70793
.		Herminiimonas arsenicoxydans uid58291
.		Janthinobacterium Marseille uid58603
.		Massilia timonae CCUG 45783 uid182886
.		Janthinobacterium HH01 uid188444
.		Lautropia mirabilis ATCC 51599 uid62167
.		Limnobacter MED105 uid54689
.	unclassified_Burk	Thiomonas 3As uid178369
.	(node 1135)	Thiomonas intermedia K12 uid48825
.	Sutterellaceae	Parasutterella excrementihominis YIT 11859 uid66383
.	(nodes 1137-1138)	Sutterella parvirubra YIT 11816 uid86919
.		Sutterella wadsworthensis 2 1 59BFAA uid181408
.	unclassified_Burk	Leptothrix cholodnii SP 6 uid58971
.	(nodes 1140-1142)	Methylibium petroleiphilum PM1 uid58085
.		Rubrivivax benzoatilyticus JA2 uid66743
.		Rubrivivax gelatinosus IL144 uid158163
.	Comamonadaceae	Hydrogenophaga PBC uid167376
.	(nodes 1143-1164)	Hylemonella gracilis ATCC 19624 uid66745
.		Rhodoferax ferrireducens T118 uid58353
.		Polaromonas naphthalenivorans CJ2 uid58273
.		Polaromonas CF318 uid171649
.		Polaromonas JS666 uid58207
.		Ramlibacter tataouinensis TTB310 uid68279
.		Variovorax CF313 uid171650
.		Variovorax paradoxus EPS uid62107
.		Acidovorax KKS102 uid176500
.		Acidovorax NO 1 uid80409
.		Acidovorax delafieldii 2AN uid55645
.		Acidovorax radicis N35 uid74319
.		Acidovorax CF316 uid170384
.		Verminephrobacter At4 uid73529
.		Verminephrobacter eiseniae EF01 2 uid58675
.		Acidovorax avenae ATCC 19860 uid42497
.		Acidovorax citrulli AAC00 1 uid58429
.		Alicycliphilus denitrificans BC uid49953
.		Acidovorax ebreus TPSY uid59233
.		Comamonas testosteroni CNB 2 uid62961
.		Delftia acidovorans SPH 1 uid58703
.		Delftia Cs1 4 uid67319
.	Alcaligenaceae	Bordetella parapertussis 12822 uid57615
.	(nodes 1165-1176)	Bordetella pertussis CS uid158859
.		Bordetella bronchiseptica 253 uid178913
.		Bordetella avium 197N uid61563
.		Achromobacter SY8 uid78829
.		Achromobacter xylosoxidans A8 uid59899
.		Achromobacter piechaudii ATCC 43553 uid47029

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Bordetella petrii DSM 12804 uid61631
.		Alcaligenes faecalis NCIB 8687 uid170035
.		Pusillimonas T7 7 uid66391
.		Taylorella asinigenitalis MCE3 uid73771
.		Taylorella equigenitalis ATCC 35865 uid170255
.		Candidatus Kinetoplastibacterium blastocrithidii ex Strigomonas culicis uid183
.	Rhodocyclaceae	Candidatus Accumulibacter phosphatis clade IIA UW 1 uid59207
.	(nodes 1177-1182)	Dechloromonas aromatica RCB uid58025
.		Dechlorosoma suillum PS uid81439
.		Methyloversatilis universalis FAM5 uid67399
.		Aromatoleum aromaticum EbN1 uid58231
.		Azoarcus BH72 uid61603
.		Thauera MZ1T uid58987
.	Hydrogenophilales	Thiobacillus denitrificans ATCC 25259 uid58189
.	Hydrogenophilales	Sulfuricella denitrificans skB26 uid170240
.	Methylophilales	Methylobacillus flagellatus KT uid58049
.	(nodes 1186-1190)	Methylovorus glucosetrophus SIP3 4 uid59367
.		Methylovorus MP688 uid60723
.		Methylophilales bacterium HTCC2181 uid54577
.		Methylotenera 301 uid49469
.		Methylotenera mobilis JLW8 uid59373
.	Nitrosomonadales	Nitrosospira multiformis ATCC 25196 uid58361
.	(nodes 1192-1195)	Nitrosomonas europaea ATCC 19718 uid57647
.		Nitrosomonas eutropha C91 uid58363
.		Nitrosomonas AL212 uid55727
.		Nitrosomonas Is79A3 uid68745
.	Gallionellales	Gallionella capsiferriformans ES 2 uid51505
.	(node 1196)	Sideroxydans lithotrophicus ES 1 uid46801
.	Neisseriaceae	Neisseria oral taxon 014 F0314 uid49701
.	(nodes 1197-1218)	Neisseria sicca ATCC 29256 uid55323
.		Neisseria macacae ATCC 33926 uid70617
.		Neisseria mucosa ATCC 25996 uid55321
.		Neisseria lactamica 020 06 uid60851
.		Neisseria gonorrhoeae FA 1090 uid57611
.		Neisseria meningitidis 053442 uid58587
.		Neisseria cinerea ATCC 14685 uid55315
.		Neisseria flavescens NRL30031 H210 uid55317
.		Neisseria subflava NJ9703 uid55325
.		Neisseria weaveri ATCC 51223 uid179870
.		Neisseria wadsworthii 9715 uid74035
.		Eikenella corrodens ATCC 23834 uid55333
.		Neisseria shayeganii 871 uid73821
.		Neisseria bacilliformis ATCC BAA 1200 uid65827
.		Kingella denitrificans ATCC 33394 uid63429
.		Kingella oralis ATCC 51147 uid55763
.		Simonsiella muelleri ATCC 29453 uid47807
.		Kingella kingae ATCC 23330 uid67095
.		Laribacter hongkongensis HLHK9 uid59265
.		Chromobacterium violaceum ATCC 12472 uid58001

Continued on next p

Table 12.3 – continued from previous page

Phylum	Family	Organism
.		Lutiella nitroferrum 2002 uid55377
.		Pseudogulbenkiania NH8B uid73423

Jump to: *Bacteroidetes, et al.*, *Alpha proteobacteria, et al.*, *Beta and Gamma Proteobacteria*, *Actinobacteria*, *Firmicutes, et al.*, *Cyanobacteria*, *Archaea*, and *Eukaryotes*.

12.4 Actinobacteria

This group contains Rubrobacteria, Slackia, Acidothermus, Streptomyces, and Nocardia

- Actinobacteria group 1, including `Coriobacteriales`, `Micrococcinea`, and `Bifidobacteriaea`,
- Actinobacteria group 2, including `Corynebacterineae`, `Nocardiaceae`, and `Frankiaceae`,

Phylum	Family	Organism
Actinobacteria	Coriobacteriales	Atopobium parvulum DSM 20469 uid59195
(nodes 1221-1510)	(nodes 1224-1238)	Atopobium rimae ATCC 49626 uid55409
.		Atopobium vaginae DSM 15829 uid53885
.		Olsenella oral taxon 809 F0356 uid76573
.		Olsenella uli DSM 7084 uid51367
.		Coriobacterium glomerans PW2 uid65787
.		Collinsella tanakaei YIT 12063 uid73035
.		Collinsella aerofaciens ATCC 25986 uid54525
.		Collinsella intestinalis DSM 13280 uid55125
.		Collinsella stercoris DSM 13279 uid54813
.		Cryptobacterium curtum DSM 15641 uid59041
.		Eggerthella lenta DSM 2243 uid59079
.		Eggerthella YY7918 uid68707
.		Slackia piriformis YIT 12062 uid175495
.		Slackia heliotrinireducens DSM 20476 uid59051
.		Slackia exigua ATCC 700122 uid55993
.	Rubrobacteridae	Rubrobacter xylanophilus DSM 9941 uid58057
.	(nodes 1239-1240)	Conexibacter woesei DSM 14684 uid43467
.		Patulibacter I11 uid80835
.	Acidimicrobiales	Acidimicrobium ferrooxidans DSM 10331 uid59215
.	Bifidobacteriales	Bifidobacterium angulatum DSM 20098 uid55113
.	(nodes 1249-1263)	Bifidobacterium breve ACS 071 V Sch8b uid158863
.		Bifidobacterium 12 1 47BFAA uid61873
.		Bifidobacterium longum BBMN68 uid60163
.		Bifidobacterium bifidum BGN4 uid167988
.		Bifidobacterium dentium Bd1 uid43091
.		Bifidobacterium adolescentis ATCC 15703 uid58559
.		Bifidobacterium catenulatum DSM 16992 uid55369
.		Bifidobacterium pseudocatenulatum DSM 20438 uid55303
.		Gardnerella vaginalis 409 05 uid43211
.		Bifidobacterium animalis ATCC 25527 uid162513
.		Bifidobacterium gallicum DSM 20093 uid55371

Continued on next page

Table 12.4 – continued from previous page

Phylum	Family	Organism
.		Bifidobacterium asteroides PRL2011 uid176921
.		Scardovia wiggsiae F0424 uid181651
.		Parascardovia denticolens DSM 10105 uid61281
.		Scardovia inopinata F0304 uid47829
.	Microbacteriaceae	Candidatus Aquiluna IMCC13023 uid158317
.	(nodes 1264-1271)	Tropheryma whipplei TW08 27 uid57961
.		Leucobacter chromiiresistens JG 31 uid162357
.		Microbacterium laevaniformans OR221 uid157055
.		Microbacterium testaceum StLB037 uid62789
.		Leifsonia xyli CTCB07 uid57759
.		Clavibacter michiganensis NCPPB 382 uid61625
.		marine actinobacterium PHSC20C1 uid54153
.		Salinibacterium PAMC 21357 uid172846
.	Actinomycetaceae	Actinobaculum massiliae ACS 171 V Col2 uid182883
.	(nodes 1272-1287)	Arcanobacterium haemolyticum DSM 20595 uid49489
.		Actinomyces graevenitzii C83 uid80681
.		Actinomyces oral taxon 170 F0386 uid66153
.		Actinomyces oris K20 uid63275
.		Actinomyces viscosus C505 uid51663
.		Actinomyces urogenitalis DSM 15434 uid55459
.		Actinomyces massiliensis F0489 uid180496
.		Mobiluncus curtisii ATCC 43063 uid49695
.		Mobiluncus mulieris 28 1 uid42169
.		Actinomyces coleocanis DSM 15436 uid55457
.		Actinomyces turicensis ACS 279 V Col4 uid175493
.		Actinomyces georgiae F0490 uid180495
.		Actinomyces ICM39 uid171968
.		Actinomyces odontolyticus ATCC 17982 uid54529
.		Actinomyces ICM47 uid170984
.		Atopobium ICM58 uid171969
.	Brevibacteriaceae	Brevibacterium linens BL2 uid54109
.	(node 1291)	Brevibacterium mcbrellneri ATCC 49030 uid48125
.	Micrococcaceae_1	Rothia mucilaginosa uid43093
.	(nodes 1292-1306)	Rothia aeria F0474 uid158319
.		Rothia dentocariosa ATCC 17931 uid49331
.		Kocuria palustris TAGA27 uid182893
.		Kocuria rhizophila DC2201 uid59099
.		Arthrobacter crystallopoietes BAB 32 uid184765
.		Renibacterium salmoninarum ATCC 33209 uid58899
.		Arthrobacter aurescens TC1 uid58109
.		Arthrobacter FB24 uid58141
.		Arthrobacter globiformis NBRC 12137 uid78817
.		Arthrobacter chlorophenolicus A6 uid58969
.		Arthrobacter phenanthrenivorans Sphe3 uid63629
.		Arthrobacter arilaitensis Re117 uid53509
.		Nesterenkonia F uid81597
.		Citricoccus CH26A uid85573
.		Micrococcus luteus NCTC 2665 uid59033
.	Dermabacteraceae	Brachybacterium squillarum M 6 3 uid86873

Continued on next page

Table 12.4 – continued from previous page

Phylum	Family	Organism
.	(nodes 1307-1308)	Brachybacterium faecium DSM 4810 uid58649
.		Brachybacterium paraconglomeratum LC44 uid162599
.	Microccineae	Beutenbergia cavernae DSM 12333 uid59047
.	Celluomonadaceae	Cellulomonas fimi ATCC 484 uid66779
.	(node 1311)	Cellulomonas flavigena DSM 20109 uid48821
.	Promicromonosporaceae	Isoptericola variabilis 225 uid67501
.	(node 1313)	Xylanimonas cellulosilytica DSM 15894 uid41935
.	Micrococcineae_2	Jonesia denitrificans DSM 20603 uid59053
.	(node 1314)	Sanguibacter keddieii DSM 10542 uid40845
.	Kineosporiaceae	Kineococcus radiotolerans SRS30216 uid58067
.	Intrasporangiaceae	Intrasporangium calvum DSM 43043 uid61729
.	(nodes 1317-1318)	Janibacter HTCC2649 uid54213
.		Janibacter hoylei PVAS 1 uid175562
.	Dermacoccaceae	Dermacoccus Ellin185 uid59483
.	Dermatophilaceae	Austwickia chelonae NBRC 105200 uid174617
.	(nodes 1320-1322)	Kineosphaera limosa NBRC 100340 uid174615
.		Mobilicoccus pelagius NBRC 104925 uid84451
.	Dermatophilaceae	Kytococcus sedentarius DSM 20547 uid59071
.	Intrasporangiaceae	Serinicoccus profundi MCCC 1A05965 uid86869
.	Propionibacterineae	Propionibacterium avidum ATCC 25577 uid74033
.	(nodes 1324-1333)	Propionibacterium acnes 266 uid162059
.		Propionibacterium humerusii P08 uid170491
.		Propionibacterium acidipropionici ATCC 4875 uid179069
.		Propionibacterium freudenreichii shermanii CIRM BIA1 uid49535
.		Propionibacterium propionicum F0230a uid170533
.		Microlunatus phosphovorus NM 1 uid68055
.		Kribbella flavida DSM 17836 uid43465
.		Aeromicrobium marinum DSM 15272 uid55873
.		Nocardioidaceae bacterium Broad 1 uid64407
.		Nocardioides JS614 uid58149
.	unclassified	actinobacterium SCGC AAA027 L06 uid174471
.	Catenulisporineae	Catenulispora acidiphila DSM 44928 uid59077
.	Streptomyceniaea	Kitasatospora setae KM 6054 uid77027
.	(nodes 1336-1373)	Streptomyces cattleya NRRL 8057 DSM 46488 uid162187
.		Streptomyces auratus AGR0001 uid171646
.		Streptomyces rimosus ATCC 10970 uid185996
.		Streptomyces bingchenggensis BCW 1 uid82931
.		Streptomyces violaceusniger Tu 4113 uid52609
.		Streptomyces somaliensis DSM 40738 uid176939
.		Streptomyces clavuligerus ATCC 27064 uid47867
.		Streptomyces tsukubaensis NRRL18488 uid162933
.		Streptomyces pristinaespiralis ATCC 25486 uid59511
.		Streptomyces W007 uid80699
.		Streptomyces griseus NBRC 13350 uid58983
.		Streptomyces roseosporus NRRL 11379 uid55565
.		Streptomyces globisporus C 1027 uid178734
.		Streptomyces flavogriseus ATCC 33331 uid40839
.		Streptomyces SirexAA E uid72627
.		Streptomyces venezuelae ATCC 10712 uid177080

Continued on next page

Table 12.4 – continued from previous page

Phylum	Family	Organism
.		Streptomyces C uid55823
.		Streptomyces Mg1 uid55043
.		Streptomyces albus J1074 uid55547
.		Streptomyces S4 uid78151
.		Streptomyces SPB74 uid48415
.		Streptomyces Tu6071 uid66919
.		Streptomyces turgidiscabies Car8 uid185785
.		Streptomyces avermitilis MA 4680 uid57739
.		Streptomyces griseoaurantiacus M045 uid66149
.		Streptomyces hygroscopicus jinggangensis 5008 uid89409
.		Streptomyces e14 uid47353
.		Streptomyces chartreusis NRRL 12338 uid156769
.		Streptomyces ghanaensis ATCC 14672 uid55543
.		Streptomyces sviceus ATCC 29083 uid59513
.		Streptomyces acidiscabies 84 104 uid170242
.		Streptomyces scabiei 87 22 uid46531
.		Streptomyces viridochromogenes DSM 40736 uid55829
.		Streptomyces zinciresistens K42 uid72955
.		Streptomyces griseoflavus Tu4000 uid55831
.		Streptomyces coelicoflavus ZG0656 uid180030
.		Streptomyces coelicolor A3 2 uid57801
.		Streptomyces lividans TK24 uid55825
.	Acidothermaceae	Acidothermus cellulolyticus 11B uid58501
.	Streptosporangineae	Nocardiopsis alba ATCC BAA 2165 uid174334
.	(nodes 1377-1379)	Thermobifida fusca YX uid57703
.		Thermomonospora curvata DSM 43183 uid41885
.		Streptosporangium roseum DSM 43021 uid42521
.		Thermobispora bispora DSM 43833 uid48999
.	Frankiaea	Frankia CN3 uid78145
.	(nodes 1381-1386)	Frankia EuI1c uid42615
.		Frankia symbiont of Datisca glomerata uid46257
.		Frankia EAN1pec uid58367
.		Frankia EUN1f uid46069
.		Frankia CcI3 uid58397
.		Frankia alni ACN14a uid58695
.		Frankia QA3 uid169226
.	Glycomycetaceae	Stackebrandtia nassauensis DSM 44728 uid46663
.	Micromonosporineae	Actinoplanes missouriensis 431 uid158169
.	(nodes 1390-1395)	Actinoplanes SE50 110 uid162333
.		Micromonospora lupini Lupac 08 uid163145
.		Verrucosispora maris AB 18 032 uid66297
.		Salinispora arenicola CNS 205 uid58659
.		Salinispora tropica CNB 440 uid58565
.		Micromonospora ATCC 39149 uid55817
.		Micromonospora aurantiaca ATCC 27029 uid42501
.	Geodermatophilaceae	Modestobacter marinus uid167487
.	(nodes 1397-1398)	Geodermatophilus obscurus DSM 43160 uid43725
.		Blastococcus saxobsidens DD2 uid89391
.	Nakamurellaceae	Nakamurella multipartita DSM 44233 uid59221

Continued on next page

Table 12.4 – continued from previous page

Phylum	Family	Organism
.	Pseudonocardiaceae	Pseudonocardia dioxanivorans CB1190 uid65087
.	(node 1401-1417)	Pseudonocardia P1 uid63251
.		Saccharopolyspora erythraea NRRL 2338 uid62947
.		Saccharopolyspora spinosa NRRL 18395 uid73653
.		Actinosynnema mirum DSM 43827 uid58951
.		Saccharothrix espanaensis DSM 44229 uid184826
.		Amycolatopsis azurea DSM 43854 uid189461
.		Amycolatopsis decaplanina DSM 44594 uid191935
.		Amycolatopsis mediterranei S699 uid158689
.		Streptomyces AA4 uid55821
.		Amycolatopsis ATCC 39116 uid159503
.		Saccharomonospora marina XMU15 uid83009
.		Saccharomonospora paurometabolica YIM 90007 uid75123
.		Saccharomonospora viridis DSM 43017 uid59055
.		Saccharomonospora glauca K62 uid82727
.		Saccharomonospora cyanea NA 134 uid83011
.		Saccharomonospora azurea NA 128 uid75103
.		Saccharomonospora xinjiangensis XJ 54 uid158329
.	Nocardiaceae	Nocardia cyriacigeorgica GUH 2 uid89395
.	(nodes 1421-1429)	Nocardia farcinica IFM 10152 uid58203
.		Nocardia brasiliensis ATCC 700358 uid86913
.		Rhodococcus P14 uid175804
.		Rhodococcus pyridinivorans AK37 uid79055
.		Rhodococcus equi 103S uid60171
.		Rhodococcus opacus B4 uid13791
.		Rhodococcus imtechensis RKJ300 uid158667
.		Rhodococcus jostii RHA1 uid58325
.		Rhodococcus erythropolis PR4 uid59019
.		Rhodococcus AW25M09 uid187352
.	Segniliparaceae	Segniliparus rotundus DSM 44985 uid49049
.	(node 1432)	Segniliparus rugosus ATCC BAA 974 uid61889
.	Tsukamurellaceae	Tsukamurella paurometabola DSM 20162 uid48829
.	Gordoniaceae	Gordonia soli NBRC 108243 uid188136
.	(nodes 1434-1453)	Gordonia rhizosphera NBRC 16068 uid174616
.		Gordonia bronchialis DSM 43247 uid41403
.		Gordonia KTR9 uid174812
.		Gordonia terrae NBRC 100016 uid84449
.		Gordonia amarae NBRC 15530 uid78545
.		Gordonia polyisoprenivorans VH2 uid86651
.		Gordonia otitidis NBRC 100426 uid84445
.		Gordonia aichiensis NBRC 108223 uid186538
.		Gordonia sputi NBRC 100414 uid84447
.		Gordonia paraffinivorans NBRC 108238 uid191944
.		Gordonia alkanivorans NBRC 16433 uid72353
.		Gordonia amicalis NBRC 100051 uid186539
.		Gordonia namibiensis NBRC 108229 uid174798
.		Gordonia rubripertincta NBRC 101908 uid186606
.		Gordonia effusa NBRC 100432 uid78815
.		Gordonia araii NBRC 100433 uid78543

Continued on next page

Table 12.4 – continued from previous page

Phylum	Family	Organism
.		Gordonia hirsuta DSM 44140 NBRC 16056 uid186540
.		Gordonia malaquae NBRC 108250 uid191943
.		Gordonia neofelifaecis NRRL B 59395 uid64473
.		Gordonia sihwensis NBRC 108236 uid186541
.	Mycobacteriaceae	Mycobacterium abscessus uid61613
.	(nodes 1454-1483)	Mycobacterium massiliense GO 06 uid170732
.		Mycobacterium fortuitum DSM 46621 uid175127
.		Mycobacterium JLS uid58489
.		Mycobacterium KMS uid58491
.		Mycobacterium phlei RIVM601174 uid158325
.		Mycobacterium hassiacum DSM 44199 uid176492
.		Mycobacterium thermoresistibile ATCC 19527 uid76937
.		Mycobacterium chubuense NBB4 uid168322
.		Mycobacterium vanbaalenii PYR 1 uid58463
.		Mycobacterium gilvum PYR GCK uid59421
.		Mycobacterium vaccae ATCC 25954 uid175128
.		Mycobacterium smegmatis JS623 uid184820
.		Mycobacterium rhodesiae NBB3 uid75107
.		Mycobacterium tusciae JS617 uid82725
.		Mycobacterium JDM601 uid67369
.		Mycobacterium xenopi RIVM700367 uid158327
.		Mycobacterium africanum GM041182 uid68839
.		Mycobacterium tuberculosis CCDC5079 uid161943
.		Mycobacterium bovis AF2122 97 uid57695
.		Mycobacterium canettii CIPT 140010059 uid70731
.		Mycobacterium leprae Br4923 uid59293
.		Mycobacterium ulcerans Agy99 uid62939
.		Mycobacterium marinum M uid59423
.		Mycobacterium liflandii 128FXT uid59005
.		Mycobacterium parascrofulaceum ATCC BAA 614 uid48977
.		Mycobacterium avium 104 uid57693
.		Mycobacterium colombiense CECT 3035 uid71463
.		Mycobacterium MOTT36Y uid164001
.		Mycobacterium indicus pranii MTCC 9506 uid175523
.		Mycobacterium intracellulare ATCC 13950 uid167994
.	Corynebacteriaceae	Amycolicicoccus subflavus DQS3 9A1 uid67253
.	(nodes 1484-1510)	Dietzia alimentaria 72 uid177888
.		Dietzia cinnamea P4 uid62173
.		Corynebacterium amycolatum SK46 uid55411
.		Corynebacterium kroppenstedtii DSM 44385 uid59411
.		Corynebacterium urealyticum DSM 7109 uid61639
.		Corynebacterium jeikeium K411 uid58399
.		Corynebacterium resistens DSM 45100 uid50555
.		Corynebacterium bovis DSM 20582 uid67345
.		Corynebacterium nuruki S6 4 uid77677
.		Corynebacterium variabile DSM 44702 uid62003
.		Corynebacterium glucuronolyticum ATCC 51866 uid55539
.		Corynebacterium matruchotii ATCC 14266 uid51885
.		Corynebacterium diphtheriae 241 uid83607

Continued on next page

Table 12.4 – continued from previous page

Phylum	Family	Organism
.		Corynebacterium pseudotuberculosis 1002 uid159677
.		Corynebacterium ulcerans 0102 uid169879
.		Corynebacterium durum F0235 uid183766
.		Corynebacterium efficiens YS 314 uid62905
.		Corynebacterium glutamicum ATCC 13032 uid57905
.		Corynebacterium ammoniagenes DSM 20306 uid48813
.		Corynebacterium casei uid78139
.		Corynebacterium accolens ATCC 49725 uid55467
.		Corynebacterium pseudogenitalium ATCC 33035 uid55395
.		Corynebacterium tuberculostearicum SK141 uid55413
.		Corynebacterium aurimucosum ATCC 700975 uid59409
.		Corynebacterium striatum ATCC 6940 uid55471
.		Corynebacterium genitalium ATCC 33030 uid52785
.		Corynebacterium lipophiloflavum DSM 44291 uid55469

Jump to: *Bacteroidetes, et al.*, *Alpha proteobacteria, et al.*, *Beta and Gamma Proteobacteria*, *Actinobacteria*, *Firmicutes, et al.*, *Cyanobacteria*, *Archaea*, and *Eukaryotes*.

12.5 Firmicutes, et al.

This group contains Thermotogae, Dictyoglomus, Thermoanaerovibrio, Deinococcus/Thermus, Chloroflexi, Syntrophomonads, Selenomonads, Bacilli, Lactobacilli, Mycoplasmas, and Clostridia. Annotated pdfs with node numbers are available for:

- `Deinococcus+Thermotogae+Synergistetes+Chloroflexi`,
- Clostridia part 1 `(Clostridiaceae, Ruminococcaceae, Eubacteriaceae)`,
- Clostridia part 2 `(Negativicutes, Peptococcaceae, Halanaerobales)`,
- `Lactobacillales`,
- `Bacillales`, and
- `Eryspelotrichi+Mollicutes`.

Family names and links to Genbank are below.

Phylum	Family	Organism
.		Candidatus Methylomirabilis oxyfera uid161981
Nitrospirae	Nitrospirales	Leptospirillum ferriphilum ML 04 uid175904
(nodes 1517-1520)	(node 1517)	Leptospirillum ferrooxidans C2 3 uid158171
.	Nitrospiraceae	Candidatus Nitrospira defluvii uid51175
.	(nodes 1519-1520)	Candidatus Poribacteria WGA A3 uid43489
.		Thermodesulfovibrio yellowstonii DSM 11347 uid59257
Thermodesulfobacteria	Thermodesulfobacteria	Thermodesulfatator indicus DSM 15286 uid68285
.		Thermodesulfobacterium OPB45 uid68283
.		Plautia stali symbiont uid65033
.		Candidatus Cloacamonas acidaminovorans Evry uid62959
.		Caldithrix abyssi DSM 13497 uid81765
Deinococcus-Thermus	Thermales	Thermus aquaticus Y51MC23 uid55053

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
(nodes 1524-1540)	(nodes 1525-1533)	Thermus CCB US3 UF1 uid81197
.		Thermus thermophilus HB27 uid58033
.		Thermus scotoductus SA 01 uid62273
.		Thermus RLM uid156767
.		Thermus oshimai JL 2 uid178948
.		Meiothermus ruber DSM 1279 uid46661
.		Meiothermus silvanus DSM 9946 uid49485
.		Marinithermus hydrothermalis DSM 14884 uid65783
.		Oceanithermus profundus DSM 14977 uid60855
.	Deinococcus	Truepera radiovictrix DSM 17093 uid49533
.	(nodes 1534-1540)	Deinococcus peraridilitoris DSM 19664 uid183485
.		Deinococcus maricopensis DSM 21211 uid62225
.		Deinococcus proteolyticus MRP uid63399
.		Deinococcus geothermalis DSM 11300 uid58275
.		Deinococcus gobiensis I 0 uid162509
.		Deinococcus deserti VCD115 uid58615
.		Deinococcus radiodurans R1 uid57665
Dictyoglomi	Dictyoglomi	Dictyoglomus thermophilum H 6 12 uid59439
(node 1544)	(node 1544)	Dictyoglomus turgidum DSM 6724 uid59177
.		Thermodesulfobium narugense DSM 14796 uid66601
.		Coprothermobacter proteolyticus DSM 5265 uid59253
.		Caldisericum exile AZM16c01 uid158173
Thermotogae	Thermotogae	Kosmotoga olearia TBF 19 5 1 uid59205
(nodes 1547-1561)	(nodes 1547-1561)	Mesotoga prima MesG1 Ag 4 2 uid52599
.		Marinitoga piezophila KA3 uid81629
.		Petrotoga mobilis SJ95 uid58747
.		Thermotoga lettingae TMO uid58419
.		Thermotoga thermarum DSM 5069 uid68449
.		Thermosipho africanus TCF52B uid59095
.		Thermosipho melanesiensis BI429 uid58683
.		Fervidobacterium nodosum Rt17 B1 uid58625
.		Fervidobacterium pennivorans DSM 9078 uid78143
.		Thermotoga petrophila RKU 1 uid58655
.		Thermotoga maritima MSB8 uid57723
.		Thermotoga naphthophila RKU 10 uid42777
.		Thermotoga neapolitana DSM 4359 uid59065
.		Thermotoga RQ2 uid58935
.		Thermotoga EMP uid174473
Synergistetes	Synergistetes	Anaerobaculum hydrogeniformans ATCC BAA 1850 uid55759
(nodes 1562-1571)	(nodes 1562-1571)	Anaerobaculum mobile DSM 13181 uid168323
.		Thermovirga lienii DSM 17291 uid77129
.		Synergistes 3 1 syn1 uid80429
.		Aminomonas paucivorans DSM 12260 uid60577
.		Thermanaerovibrio acidaminovorans DSM 6589 uid41925
.		Thermanaerovibrio velox DSM 12556 uid80703
.		Aminobacterium colombiense DSM 12261 uid47083
.		Dethiosulfovibrio peptidovorans DSM 11002 uid54917
.		Jonquetella anthropi DSM 22815 uid182031
.		Pyramidobacter piscolens W5455 uid42959

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
Chloroflexi	Chloroflexi	Herpetosiphon aurantiacus DSM 785 uid58599
(nodes 1572-1589)	Chloroflexales	Roseiflexus castenholzii DSM 13941 uid58287
.	(nodes 1575-1578	Roseiflexus RS 1 uid58523
.		Oscillochloris trichoides DG6 uid59465
.		Chloroflexus aggregans DSM 9485 uid58621
.		Chloroflexus aurantiacus J 10 fl uid57657
.		Thermobaculum terrenum ATCC BAA 798 uid42011
.		Thermomicrobium roseum DSM 5159 uid59341
.	Sphaerobacterales	Nitrolancetus hollandicus Lb uid168185
.	(node 1581)	Sphaerobacter thermophilus DSM 20745 uid41997
.		Ktedonobacter racemifer DSM 44963 uid49671
.	Dehalococcoidetes	Dehalogenimonas lykanthroporepellens BL DC 9 uid48131
.	(nodes 1584-1588)	Dehalococcoides VS uid42393
.		Dehalococcoides mccartyi BTF08 uid190183
.		Dehalococcoides CBDB1 uid58413
.		Dehalococcoides BAV1 uid58477
.		Dehalococcoides ethenogenes 195 uid57763
.		Anaerolinea thermophila UNI 1 uid62245
.		Caldilinea aerophila DSM 14535 NBRC 104270 uid158165
Clostridia	Clostridiaceae	Clostridium nexile DSM 1787 uid55077
(nodes 1591-1819)	(nodes 1602-1649)	Clostridium hylemonae DSM 15053 uid55299
.		Clostridium scindens ATCC 35704 uid54533
.		Dorea formicigenerans 4 6 53AFAA uid73033
.		Dorea longicatena DSM 13814 uid54515
.		Clostridium D5 uid63427
.		Ruminococcus gnavus ATCC 29149 uid54537
.		Ruminococcus torques ATCC 27756 uid54511
.		Ruminococcus lactaris ATCC 29176 uid54903
.		Bryantella formatexigens DSM 14469 uid54943
.		Blautia hansenii DSM 20583 uid55275
.		Blautia hydrogenotrophica DSM 10507 uid54939
.		Ruminococcus 5 1 39BFAA uid55629
.		Ruminococcus obeum ATCC 29174 uid54509
.		Roseburia hominis A2 183 uid73419
.		Roseburia intestinalis L1 82 uid55267
.		Shuttleworthia satelles DSM 14600 uid55775
.		Eubacterium rectale ATCC 33656 uid59169
.		Clostridium SY8519 uid68705
.		Butyrivibrio proteoclasticus B316 uid51489
.		Eubacterium cellulosolvens 6 uid61055
.		Eubacterium eligens ATCC 27750 uid59171
.		Butyrivibrio crossotus DSM 2876 uid55091
.		Bacteroides pectinophilus ATCC 43243 uid54987
.		Johnsonella ignava ATCC 51276 uid77897
.		Eubacterium saburreum DSM 3986 uid61491
.		Lachnospiraceae oral taxon 107 F0167 uid66385
.		Clostridium M62 1 uid54557
.		Clostridium symbiosum WAL 14163 uid63097
.		Clostridium bolteae ATCC BAA 613 uid54523

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Clostridium clostridioforme 2 1 49FAA uid76955
.		Clostridium citroniae WAL 17108 uid76581
.		Clostridiales bacterium 1 7 47FAA uid55287
.		Clostridium hathewayi DSM 13479 uid55373
.		Clostridium saccharolyticum WM1 uid51419
.		Oribacterium oral taxon 078 F0262 uid55773
.		Oribacterium sinus F0268 uid55891
.		Oribacterium ACB7 uid79229
.		Oribacterium ACB1 uid79231
.		Oribacterium ACB8 uid170518
.		Eubacterium ventriosum ATCC 27560 uid54517
.		Clostridium phytofermentans ISDg uid58519
.		Clostridium L2 50 uid54559
.		Coprococcus eutactus ATCC 27759 uid54541
.		Eubacterium hallii DSM 3353 uid54535
.		Anaerostipes 3 2 56FAA uid61867
.		Anaerostipes caccae DSM 14662 uid54561
.		Clostridium SS2 1 uid54553
.		Eubacterium hadrum DSM 3319 uid183778
.		Clostridium lentocellum DSM 5427 uid49117
.		Oscillibacter valericigenes uid73895
.		Ruminococcaceae bacterium D16 uid52825
.		Bacteroides capillosus ATCC 29799 uid54531
.		Flavonifractor plautii ATCC 29863 uid80691
.	Ruminococcaceae	Clostridium methylpentosum DSM 5476 uid55281
.	(nodes 1654-1664)	Eubacterium siraeum DSM 15702 uid54603
.		Ruminococcus albus 7 uid51721
.		Ruminococcus flavefaciens FD 1 uid55965
.		Clostridium leptum DSM 753 uid54605
.		Clostridium MSTE9 uid180497
.		Anaerotruncus colihominis DSM 17241 uid54807
.		Ethanoligenens harbinense YUAN 3 uid46255
.		Subdoligranulum 4 3 54A2FAA uid80415
.		Subdoligranulum variabile DSM 15176 uid54539
.		Faecalibacterium cf prausnitzii KLE1255 uid60645
.		Faecalibacterium prausnitzii A2 165 uid54551
.		Clostridiales genomosp BVAB3 UPII9 5 uid46219
.	Clostridiaceae	Clostridium stercorarium DSM 8532 uid186819
.	(nodes 1665-1670	Clostridium cellulolyticum H10 uid58709
.		Clostridium BNL1100 uid84307
.		Clostridium papyrosolvens DSM 2782 uid55815
.		Clostridium thermocellum ATCC 27405 uid57917
.		Acetivibrio cellulolyticus CD2 uid51533
.		Clostridium clariflavum DSM 19732 uid82345
.	Family III	Caldicellulosiruptor saccharolyticus DSM 8903 uid58289
.	(nodes 1672-1677)	Caldicellulosiruptor bescii DSM 6725 uid59201
.		Caldicellulosiruptor kronotskyensis 2002 uid60491
.		Caldicellulosiruptor hydrothermalis 108 uid60157
.		Caldicellulosiruptor kristjanssonii 177R1B uid60393

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Caldicellulosiruptor lactoaceticus 6A uid60575
.		Caldicellulosiruptor owensensis OL uid60165
.		Caldicellulosiruptor obsidiansis OB47 uid51501
.		Mahella australiensis 50 1 BON uid66917
.		Thermoanaerobacterium thermosaccharolyticum DSM 571 uid51639
.		Thermoanaerobacterium xylanolyticum LX 11 uid63163
.		Thermoanaerobacterium saccharolyticum JW SL YS485 uid167781
.	Thermoanaerobacteraceae	Thermoanaerobacter tengcongensis MB4 uid57813
.	(nodes 1683-1690)	Thermoanaerobacter pseudethanolicus ATCC 33223 uid58339
.		Thermoanaerobacter wiegelii Rt8 B1 uid52581
.		Thermoanaerobacter X514 uid58589
.		Thermoanaerobacter X561 uid55835
.		Thermoanaerobacter siderophilus SR4 uid169227
.		Thermoanaerobacter brockii finnii Ako 1 uid55639
.		Thermoanaerobacter italicus Ab9 uid46241
.		Thermoanaerobacter ethanolicus CCSD1 uid55837
.	Clostridiaceae	Caloramator australicus RC3 uid171387
.	(nodes 1691-1710)	Candidatus Arthromitus SFB mouse Japan uid71379
.		Clostridium perfringens 13 uid57681
.		Clostridium butyricum 5521 uid54843
.		Clostridium DL VIII uid78327
.		Clostridium beijerinckii NCIMB 8052 uid58137
.		Clostridium Maddingley MBC34 26 uid178945
.		Clostridium 7 2 43FAA uid55597
.		Clostridium celatum DSM 1785 uid183779
.		Clostridium JC122 uid174326
.		Clostridium tetani E88 uid57683
.		Clostridium botulinum A2 Kyoto uid59229
.		Clostridium sporogenes ATCC 15579 uid54895
.		Clostridium carboxidivorans P7 uid48985
.		Clostridium kluyveri DSM 555 uid58885
.		Clostridium ljungdahlii DSM 13528 uid50583
.		Clostridium novyi NT uid58643
.		Clostridium cellulovorans 743B uid51503
.		Clostridium acetobutylicum ATCC 824 uid57677
.		Clostridium arbusti SL206 uid171987
.		Clostridium pasteurianum DSM 525 uid185802
.		Anaerofustis stercorihominis DSM 17244 uid54805
.		Acetobacterium woodii DSM 1030 uid88073
.		Pseudoramibacter alactolyticus ATCC 23263 uid61507
.		Eubacterium limosum KIST612 uid59777
.		Eubacterium saphenum ATCC 49989 uid55765
.	Family XIII	Eubacterium infirmum F0142 uid81785
.	(node 1718)	Mogibacterium CM50 uid174240
.	Family XI	Peptoniphilus oral taxon 375 F0436 uid71189
.	(nodes 1719-1732)	Peptoniphilus duerdenii ATCC BAA 1640 uid51739
.		Peptoniphilus indolicus ATCC 29427 uid74031
.		Peptoniphilus lacrimalis 315 B uid42973
.		Peptoniphilus harei ACS 146 V Sch2b uid61041

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Peptoniphilus rhinitidis 1 13 uid171988
.		Parvimonas micra ATCC 33270 uid54527
.		Parvimonas oral taxon 110 F0139 uid67821
.		Finegoldia magna ATCC 29328 uid58867
.		Helcococcus kunzii ATCC 51366 uid83005
.		Anaerococcus hydrogenalis ACS 025 V Sch4 uid63589
.		Anaerococcus vaginalis ATCC 51170 uid56001
.		Anaerococcus lactolyticus ATCC 51172 uid55391
.		Anaerococcus prevotii DSM 20548 uid59219
.		Anaerococcus tetradius ATCC 35098 uid55461
.		Clostridium acidurici 9a uid176126
.	Peptostreptococcaceae	Peptostreptococcus anaerobius 653 L uid46195
.	(nodes 1736-1739)	Peptostreptococcus stomatis DSM 17678 uid52563
.		Clostridium hiranonis DSM 13275 uid55075
.		Clostridium bartlettii DSM 16795 uid54809
.		Clostridium difficile 630 uid57679
.		Filifactor alocis ATCC 35896 uid46625
.	Eubacteriaceae	Eubacteriaceae bacterium ACC19a uid79227
.	(nodes 1741-1742)	Eubacterium AS15 uid173856
.		Eubacterium yurii margaretiae ATCC 43715 uid52347
.		Alkaliphilus metalliredigens QYMF uid58171
.		Alkaliphilus oremlandii OhILAs uid58495
.	Thermoanaerobacterales	Thermosediminibacter oceani DSM 16646 uid51421
.	(node 1744-1745)	Tepidanaerobacter acetatoxydans Re1 uid184827
.		Tepidanaerobacter Re1 uid66873
.	Family XVII	Thermaerobacter marianensis DSM 12885 uid61727
.	(node 1747)	Thermaerobacter subterraneus DSM 13965 uid61053
.	Syntrophomonadaceae	Syntrophomonas wolfei Goettingen uid58179
.	(node 1748)	Syntrophothermus lipocalidus DSM 12680 uid49527
.	Halanaerobales	Halothermothrix orenii H 168 uid58585
.	(nodes 1749-1752)	Halanaerobium hydrogeniformans uid60191
.		Halanaerobium praevalens DSM 2228 uid161959
.		Acetohalobium arabaticum DSM 5501 uid51423
.		Halobacteroides halobius DSM 5150 uid184862
.		Dethiobacter alkaliphilus AHT 1 uid55401
.		Natranaerobius thermophilus JW NM WN LF uid59001
.		Sulfobacillus acidophilus DSM 10332 uid88061
.		Symbiobacterium thermophilum IAM 14863 uid58165
.	Thermoanaerobacterales	Moorella thermoacetica ATCC 39073 uid58051
.	(node 1759)	Thermacetogenium phaeum DSM 12270 uid177811
.	Peptococcaceae	Thermincola potens JR uid48823
.	(nodes 1760-1783)	Desulfotomaculum kuznetsovii DSM 6115 uid67357
.		Desulfotomaculum gibsoniae DSM 7213 uid76945
.		Pelotomaculum thermopropionicum SI uid58877
.		Desulfotomaculum acetoxidans DSM 771 uid59109
.		Desulfotomaculum carboxydivorans CO 1 SRB uid67317
.		Desulfotomaculum nigrificans DSM 574 uid63161
.		Desulfotomaculum ruminis DSM 2154 uid67507
.		Desulfotomaculum hydrothermale Lam5 DSM 18033 uid179385

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Desulfotomaculum reducens MI 1 uid58277
.		Candidatus Desulforudis audaxviator MP104C uid59067
.		Ammonifex degensii KC4 uid41053
.		Carboxydothermus hydrogenoformans Z 2901 uid57821
.		Heliobacterium modesticaldum Ice1 uid58279
.		Dehalobacter 11DCA uid177715
.		Syntrophobotulus glycolicus DSM 8271 uid63343
.		Desulfosporosinus acidiphilus SJ4 uid156759
.		Desulfosporosinus OT uid72957
.		Desulfosporosinus orientis DSM 765 uid82939
.		Desulfosporosinus meridiei DSM 13257 uid75097
.		Desulfosporosinus youngiae DSM 17734 uid82713
.		Desulfitobacterium metallireducens DSM 15288 uid75095
.		Desulfitobacterium dichloroeliminans LMG P 21439 uid82555
.		Desulfitobacterium dehalogenans ATCC 51507 uid82553
.		Desulfitobacterium hafniense DCB 2 uid57749
.	Negativicutes	Veillonella dispar ATCC 17748 uid55331
.	(nodes 1784-1819)	Veillonella oral taxon 158 F0412 uid61047
.		Veillonella parvula DSM 2008 uid41927
.		Veillonella 3 1 44 uid47845
.		Veillonella 6 1 27 uid47835
.		Veillonella ACP1 uid172974
.		Veillonella atypica ACS 049 V Sch6 uid51525
.		Veillonella ratti ACS 216 V Col6b uid182889
.		Anaeroglobus geminatus F0357 uid80689
.		Megasphaera micronuciformis F0359 uid60585
.		Megasphaera elsdenii DSM 20460 uid71135
.		Megasphaera genomosp type 1 28L uid46567
.		Megasphaera UPII 135 E uid71191
.		Dialister invisus DSM 15470 uid55761
.		Dialister succinatiphilus YIT 11850 uid81763
.		Dialister micraerophilus DSM 19965 uid65829
.		Dialister microaerophilus UPII 345 E uid61045
.		Phascolarctobacterium YIT 12067 uid62745
.		Acidaminococcus fermentans DSM 20731 uid43471
.		Acidaminococcus D21 uid55871
.		Acidaminococcus intestini RyC MR95 uid74445
.		Megamonas funiformis YIT 11815 uid82999
.		Selenomonas CM52 uid174241
.		Selenomonas sputigena ATCC 35185 uid55329
.		Selenomonas infelix ATCC 43532 uid76957
.		Selenomonas flueggei ATCC 43531 uid55953
.		Selenomonas F0473 uid182887
.		Selenomonas FOBRC9 uid173853
.		Selenomonas artemidis F0399 uid62283
.		Selenomonas oral taxon 137 F0430 uid61049
.		Selenomonas noxia ATCC 43541 uid55895
.		Centipeda periodontii DSM 2778 uid67397
.		Mitsuokella multacida DSM 20544 uid55073

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Selenomonas ruminantium lactilytica TAM6421 uid157247
.		Pelosinus fermentans A11 uid180746
.		Thermosinus carboxydivorans Nor1 uid54463
.		Acetonema longum DSM 6540 uid68557
Lactobacillales	Streptococcaceae	Streptococcus SK643 uid161589
(nodes 1822-1956)	(nodes 1826-1874)	Streptococcus mitis B6 uid46097
.		Streptococcus pneumoniae 670 6B uid52533
.		Streptococcus pseudopneumoniae IS7493 uid71153
.		Streptococcus GMD4S uid180425
.		Streptococcus C300 uid62527
.		Streptococcus oralis Uo5 uid65449
.		Streptococcus BS35b uid172972
.		Streptococcus M143 uid42367
.		Streptococcus SK140 uid161587
.		Streptococcus peroris ATCC 700780 uid62543
.		Streptococcus infantis ATCC 700779 uid180433
.		Streptococcus cristatus ATCC 51100 uid179856
.		Streptococcus sanguinis SK36 uid58381
.		Streptococcus AS14 uid172973
.		Streptococcus gordonii Challis substr CH1 uid57667
.		Streptococcus 2 1 36FAA uid41507
.		Streptococcus australis ATCC 700641 uid179855
.		Streptococcus parasanguinis ATCC 15912 uid49313
.		Streptococcus intermedius JTH08 uid168614
.		Streptococcus anginosus 1 2 62CV uid62163
.		Streptococcus constellatus pharyngis SK1060 uid72195
.		Streptococcus suis 05ZYH33 uid58663
.		Streptococcus macacae NCTC 11558 uid71359
.		Streptococcus mutans GS 5 uid169223
.		Streptococcus ratti FA 1 uid172632
.		Streptococcus criceti uid71271
.		Streptococcus downei F0415 uid60561
.		Streptococcus thermophilus CNRZ1066 uid58221
.		Streptococcus salivarius 57 I uid162151
.		Streptococcus vestibularis ATCC 49124 uid62665
.		Streptococcus C150 uid62525
.		Streptococcus equinus ATCC 9812 uid62297
.		Streptococcus infantarius CJ18 uid87033
.		Streptococcus bovis ATCC 700338 uid52359
.		Streptococcus macedonicus ACA DC 198 uid81631
.		Streptococcus agalactiae 2603V R uid57943
.		Streptococcus urinalis 2285 97 uid71277
.		Streptococcus ictaluri 707 05 uid71275
.		Streptococcus equi 4047 uid59259
.		Streptococcus canis FSL Z3 227 uid168663
.		Streptococcus pyogenes A20 uid178106
.		Streptococcus dysgalactiae equisimilis AC 2713 uid178644
.		Streptococcus iniae 9117 uid175677
.		Streptococcus parauberis KCTC 11537 uid67355

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Streptococcus uberis 0140J uid57959
.		Streptococcus porcinus Jelinkova 176 uid66741
.		Streptococcus pseudoporcinus LQ 940 04 uid179694
.		Lactococcus garvieae ATCC 49156 uid73413
.		Lactococcus lactis cremoris A76 uid160937
.	Leuconostocaceae	Leuconostoc mesenteroides ATCC 8293 uid57919
.	(nodes 1878-1893)	Leuconostoc pseudomesenteroides 4882 uid171985
.		Leuconostoc argentinum KCTC 3773 uid64589
.		Leuconostoc citreum KM20 uid58481
.		Leuconostoc kimchii IMSNU 11154 uid48589
.		Leuconostoc gasicomitatum LMG 18811 uid50385
.		Leuconostoc gelidum JB7 uid175682
.		Leuconostoc carnosum JB16 uid176371
.		Fructobacillus fructosus KCTC 3544 uid68703
.		Leuconostoc fallax KCTC 3537 uid66211
.		Oenococcus kitaharae DSM 17330 uid81361
.		Oenococcus oeni PSU 1 uid59417
.		Weissella koreensis KACC 15510 uid68837
.		Weissella ceti NC36 uid184763
.		Weissella paramesenteroides ATCC 33313 uid55901
.		Weissella cibaria KACC 11862 uid66733
.		Weissella confusa LBAE C39 2 uid168254
.	Lactobacillaceae_3	Lactobacillus pentosus KCA1 uid169225
.	(nodes 1894-1914)	Lactobacillus plantarum JDM1 uid59361
.		Pediococcus claussenii ATCC BAA 344 uid81103
.		Pediococcus acidilactici 7 4 uid42365
.		Pediococcus pentosaceus ATCC 25745 uid57981
.		Lactobacillus suebicus KCTC 3549 uid80667
.		Lactobacillus vaginalis ATCC 49540 uid55515
.		Lactobacillus reuteri DSM 20016 uid58471
.		Lactobacillus antri DSM 16041 uid55491
.		Lactobacillus oris F0423 uid179841
.		Lactobacillus mucosae LM1 uid86029
.		Lactobacillus coleohominis 101 4 CHN uid55977
.		Lactobacillus gastricus PS3 uid84443
.		Lactobacillus fermentum CECT 5716 uid162003
.		Lactobacillus brevis ATCC 367 uid57989
.		Lactobacillus malefermentans KCTC 3548 uid80665
.		Lactobacillus hilgardii ATCC 8290 uid55501
.		Lactobacillus buchneri NRRL B 30929 uid66205
.		Lactobacillus kisonensis F0435 uid81769
.		Lactobacillus fructivorans KCTC 3543 uid68677
.		Lactobacillus florum 2F uid177324
.		Lactobacillus sanfranciscensis TMW 1 1304 uid72937
.	Lactobacillaceae_2	Lactobacillus mali KCTC 3596 DSM 20444 uid180535
.	(nodes 1915-1919)	Lactobacillus vini DSM 20605 uid175441
.		Lactobacillus animalis KCTC 3501 uid67869
.		Lactobacillus ruminis ATCC 27782 uid73417
.		Lactobacillus acidipiscis KCTC 13900 uid80675

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Lactobacillus salivarius CECT 5713 uid162005
.	Lactobacillaceae_1	Lactobacillus coryniformis CECT 5711 uid180492
.	(nodes 1921-1940)	Lactobacillus curvatus CRL 705 uid76157
.		Lactobacillus sakei 23K uid58281
.		Lactobacillus rhamnosus ATCC 8530 uid162169
.		Lactobacillus paracasei 8700 2 uid55295
.		Lactobacillus zeae KCTC 3804 uid80671
.		Lactobacillus versmoldensis KCTC 3814 uid80673
.		Lactobacillus iners AB 1 uid51205
.		Lactobacillus hominis CRBIP 24 179 uid170497
.		Lactobacillus gasseri ATCC 33323 uid57687
.		Lactobacillus johnsonii DPC 6026 uid162057
.		Lactobacillus jensenii 115 3 CHN uid40897
.		Lactobacillus delbrueckii bulgaricus 2038 uid161929
.		Lactobacillus equicursoris CIP 110162 uid177398
.		Lactobacillus amylolyticus DSM 11664 uid48293
.		Lactobacillus ultunensis DSM 16047 uid55513
.		Lactobacillus acidophilus 30SC uid63605
.		Lactobacillus helveticus DPC 4571 uid58761
.		Lactobacillus kefiranofaciens ZW3 uid67985
.		Lactobacillus crispatus ST1 uid48359
.		Lactobacillus gigeriorum CRBIP 24 85 uid170495
.		Lactobacillus pasteurii CRBIP 24 76 uid170496
.	Enterococcaceae	Melissococcus plutonius ATCC 35311 uid66803
.	(nodes 1941-1946)	Enterococcus faecalis 62 uid159663
.		Enterococcus faecium Aus0004 uid87025
.		Enterococcus casseliflavus ATCC 12755 uid63559
.		Enterococcus saccharolyticus 30 1 uid76953
.		Tetragenococcus halophilus uid74441
.		Enterococcus italicus DSM 15952 uid61487
.	Carnobacteriaceae	Carnobacterium 17 4 uid65789
.	(nodes 1947-1951)	Carnobacterium AT7 uid54673
.		Granulicatella adiacens ATCC 49175 uid55951
.		Granulicatella elegans ATCC 700633 uid40873
.		Alloiococcus otitis ATCC 51267 uid182884
.		Dolosigranulum pigrum ATCC 51524 uid83001
.	Aerococcaceae	Facklamia hominis CCUG 36813 uid175684
.	(nodes 1952-1956)	Facklamia languida
.		Eremococcus coleocola ACS 139 V Col8 uid61037
.		Facklamia ignava CCUG 37419 uid175683
.		Aerococcus urinae ACS 120 V Col10a uid64757
.		Aerococcus viridans ATCC 11563 uid48397
Bacillales	Family XI	Gemella sanguinis M325 uid66133
(nodes 1957-2064)	(nodes 1961-1962)	Gemella haemolysans ATCC 10379 uid55327
.	Staphylococcaceae	Gemella moribillum M424 uid61881
.	(nodes 1963-1975)	Macrococcus caseolyticus JCSC5402 uid59003
.		Staphylococcus carnosus TM300 uid59401
.		Staphylococcus pseudintermedius ED99 uid162109
.		Staphylococcus lugdunensis HKU09 01 uid46233

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Staphylococcus hominis C80 uid61127
.		Staphylococcus haemolyticus JCSC1435 uid62919
.		Staphylococcus caprae C87 uid61125
.		Staphylococcus epidermidis ATCC 12228 uid57861
.		Staphylococcus warneri SG1 uid187059
.		Staphylococcus aureus 04 02981 uid161969
.		Staphylococcus pettenkoferi VCU012 uid179999
.		Staphylococcus arlettae CVD059 uid175126
.		Staphylococcus equorum Mu2 uid169178
.		Staphylococcus saprophyticus ATCC 15305 uid58411
.	Planococcaceae	Sporosarcina newyorkensis uid70561
.	(nodes 1976-1982)	Planococcus antarcticus DSM 14505 uid167855
.		Planococcus donghaensis MPA1U2 uid63099
.		Solibacillus silvestris StLB046 uid168516
.		Bacillus B14905 uid54613
.		Lysinibacillus sphaericus C3 41 uid58945
.		Kurthia JC30 uid174325
.		Kurthia JC8E uid174327
.	Listeria	Listeria grayi DSM 20601 uid55523
.	(nodes 1983-1986)	Listeria ivanovii PAM 55 uid73473
.		Listeria seeligeri serovar 1 2b SLCC3954 uid46215
.		Listeria monocytogenes 07PF0776 uid162185
.		Listeria welshimeri serovar 6b SLCC5334 uid61605
.	Bacillaceae	Anoxybacillus flavithermus WK1 uid59135
.	(nodes 1988-2039)	Geobacillus thermoleovorans CCB US3 UF5 uid82949
.		Geobacillus G11MC16 uid55035
.		Geobacillus thermodenitrificans NG80 2 uid58829
.		Geobacillus thermoglucosidans TNO 09 020 uid181720
.		Geobacillus WCH70 uid59045
.		Bacillus methanolicus MGA3 uid179596
.		Bacillus 2 A 57 CT2 uid62059
.		Bacillus NRRL B 14911 uid54211
.		Bacillus 1NLA3E uid81841
.		Bacillus bataviensis LMG 21833 uid178560
.		Bacillus coahuilensis m4 4 uid54797
.		Bacillus SG 1 uid54663
.		Bacillus m3 13 uid59767
.		Bacillus 10403023 uid174324
.		Bacillus megaterium DSM319 uid48371
.		Bacillus pseudomycoides DSM 12442 uid55213
.		Bacillus mycoides DSM 2048 uid55207
.		Bacillus weihenstephanensis KBAB4 uid58315
.		Bacillus 7 6 55CFAA CT2 uid80427
.		Bacillus cereus 03BB102 uid59299
.		Bacillus anthracis A0248 uid59385
.		Bacillus thuringiensis Al Hakam uid58795
.		Bacillus licheniformis ATCC 14580 uid58097
.		Bacillus HYC 10 uid176491
.		Bacillus pumilus SAFR 032 uid59017

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Bacillus 5B6 uid161591
.		Bacillus amyloliquefaciens DSM 7 uid53535
.		Bacillus atrophaeus 1942 uid59887
.		Bacillus mojavensis RO H 1 uid171173
.		Bacillus subtilis 168 uid57675
.		Bacillus coagulans 2 6 uid68053
.		Bacillus smithii 7 3 47FAA uid80423
.		Bacillus azotoformans LMG 9581 uid178561
.		Ornithinibacillus TW25 uid72367
.		Lentibacillus Grbi uid86863
.		Oceanobacillus iheyensis HTE831 uid57867
.		Oceanobacillus Ndiop uid174322
.		Salimicrobium MJ3 uid176218
.		Halobacillus BAB 2008 uid184766
.		Halobacillus halophilus DSM 2266 uid162033
.		Amphibacillus xylanus NBRC 15112 uid176453
.		Exiguobacterium AT1b uid59093
.		Exiguobacterium antarcticum B7 uid176125
.		Exiguobacterium sibiricum 255 15 uid58053
.		Bacillus macauensis ZFHKF 1 uid169222
.		Bacillus halodurans C 125 uid57791
.		Bacillus clausii KSM K16 uid58237
.		Bacillus alcalophilus ATCC 27647 uid173850
.		Bacillus pseudofirmus OF4 uid45847
.		Bacillus selenitireducens MLS10 uid49513
.		Bacillus cellulosilyticus DSM 2522 uid43329
.		Caldalkalibacillus thermarum TA2 A1 uid67815
.	Sporolactobacillus	Sporolactobacillus inulinus CASD uid82635
.	(node 2041)	Sporolactobacillus vineae DSM 21990 SL153 uid171990
.	Paenibacillaceae	Brevibacillus laterosporus GI 9 uid180972
.	(node 2043-2062	Brevibacillus agri BAB 2500 uid184804
.		Brevibacillus brevis NBRC 100599 uid59175
.		Desmospora 8437 uid67093
.		Paenibacillus alvei DSM 29 uid174248
.		Paenibacillus dendritiformis C454 uid82723
.		Paenibacillus popilliae ATCC 14706 uid180969
.		Paenibacillus oral taxon 786 D14 uid55903
.		Paenibacillus peoriae KCTC 3763 uid167853
.		Paenibacillus terrae HPL 003 uid82371
.		Paenibacillus polymyxa E681 uid53477
.		Paenibacillus lactis 154 uid75105
.		Paenibacillus vortex V453 uid61489
.		Paenibacillus Y412MC10 uid41127
.		Thermobacillus composti KWC4 uid74021
.		Paenibacillus curdlanolyticus YK9 uid51723
.		Paenibacillus JDR 2 uid59021
.		Paenibacillus JC66 uid174323
.		Paenibacillus HGF7 uid67401
.		Paenibacillus elgii B69 uid76799

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Paenibacillus mucilaginosus 3016 uid89377
.	Alicylobacillaceae	Kyrpidia tusciae DSM 2912 uid48361
.	(node 2063-2064)	Alicyclobacillus acidocaldarius DSM 446 uid59199
.		Alicyclobacillus hesperidum URH17 3 68 uid174618
Erysipelotrichi	Erysipelotrichi	Clostridium spiroforme DSM 1552 uid54607
(nodes 2066-2080)	(nodes 2066-2080)	Clostridium ramosum DSM 1402 uid54811
.		Coprobacillus 3 3 56FAA uid80685
.		Catenibacterium mitsuokai DSM 15897 uid54829
.		Coprobacillus 29 1 uid62161
.		Erysipelothrix rhusiopathiae Fujisawa uid68021
.		Eubacterium biforme DSM 3989 uid55117
.		Eubacterium 3 1 31 uid81761
.		Eubacterium dolichum DSM 3991 uid54609
.		Clostridium HGF2 uid61051
.		Erysipelotrichaceae bacterium 21 3 uid81609
.		Holdemania filiformis DSM 12042 uid55297
.		Bulleidia extructa W1219 uid43197
.		Solobacterium moorei F0204 uid62281
.		Haloplasma contractile SSD 17B uid67925
.		Turicibacter PC909 uid46977
Mollicutes	Acholeplasmataceae	Acholeplasma laidlawii PG 8A uid58901
(nodes 2081-2119)	(nodes 2082-2084)	Candidatus Phytoplasma australiense uid61641
.		Aster yellows witches broom phytoplasma AYWB uid58297
.		Onion yellows phytoplasma OY M uid58015
.	Micoplasmataceae	Spiroplasma melliferum IPMB4A uid185130
.	(nodes 2085-2119)	Mesoplasma florum L1 uid58055
.		Mycoplasma putrefaciens KS1 uid72481
.		Mycoplasma G5847 uid184764
.		Mycoplasma mycoides capri LC 95010 uid66189
.		Mycoplasma leachii 99 014 6 uid162031
.		Mycoplasma capricolum ATCC 27343 uid58525
.		Ureaplasma parvum serovar 3 ATCC 27815 uid58887
.		Ureaplasma urealyticum serovar 10 ATCC 33699 uid59011
.		Mycoplasma iowae 695 uid74019
.		Mycoplasma penetrans HF 2 uid57729
.		Mycoplasma gallisepticum CA06 2006 052 5 2P uid172630
.		Mycoplasma genitalium G37 uid57707
.		Mycoplasma pneumoniae 309 uid85495
.		Mycoplasma suis Illinois uid61897
.		Candidatus Mycoplasma haemolamae Purdue uid171259
.		Mycoplasma wenyonii Massachusetts uid170731
.		Mycoplasma haemocanis Illinois uid82367
.		Mycoplasma haemofelis Langford 1 uid62461
.		Mycoplasma arthritidis 158L3 1 uid58005
.		Mycoplasma hominis ATCC 23114 uid41875
.		Mycoplasma hyorhinis GDL 1 uid87003
.		Mycoplasma conjunctivae uid59325
.		Mycoplasma hyopneumoniae 168 uid162053
.		Mycoplasma ovipneumoniae SC01 uid79051

Continued on next page

Table 12.5 – continued from previous page

Phylum	Family	Organism
.		Mycoplasma mobile 163K uid58077
.		Mycoplasma pulmonis UAB CTIP uid61569
.		Mycoplasma columbinum SF7 uid72177
.		Mycoplasma fermentans JER uid53543
.		Mycoplasma agalactiae PG2 uid61619
.		Mycoplasma bovis HB0801 uid168665
.		Mycoplasma alligatoris A21JP2 uid46947
.		Mycoplasma anatis 1340 uid71183
.		Mycoplasma synoviae 53 uid58061
.		Mycoplasma canis PG 14 uid180388
.		Mycoplasma cynos C142 uid184824

Jump to: *Bacteroidetes, et al.*, *Alpha proteobacteria, et al.*, *Beta and Gamma Proteobacteria*, *Actinobacteria*, *Firmicutes, et al.*, *Cyanobacteria*, *Archaea*, and *Eukaryotes*.

12.6 Cyanobacteria

This group contains Cyanobacteria. An annotated pdf with node numbers is available for

- `Cyanobacteria`.

  For a recent discussion of the cyanobacterial phylogeny, please see Shih, et al., Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing.

Phylum	Family	Organism
Cyanobacteria	Prochlorales	Synechococcus RCC307 uid61609
(nodes 2120-2212)	(nodes 2121-2145)	Cyanobium gracile PCC 6307 uid182931
.		Cyanobium PCC 7001 uid54675
.		Synechococcus WH 5701 uid54219
.		Synechococcus CB0205 uid61893
.		Prochlorococcus marinus MIT 9303 uid58305
.		Prochlorococcus marinus MIT 9313 uid57773
.		Synechococcus WH 8102 uid61581
.		Synechococcus BL107 uid54225
.		Synechococcus CC9605 uid58319
.		Synechococcus RS9917 uid54221
.		Synechococcus RS9916 uid54223
.		Synechococcus WH 7803 uid61607
.		Synechococcus WH 7805 uid54217
.		Synechococcus CC9311 uid58123
.		Synechococcus WH 8016 uid74433
.		Prochlorococcus marinus NATL1A uid58423
.		Prochlorococcus marinus CCMP1375 uid57995
.		Prochlorococcus marinus MIT 9211 uid58309
.		Prochlorococcus marinus MIT 9515 uid58313
.		Prochlorococcus marinus pastoris CCMP1986 uid57761
.		Prochlorococcus marinus MIT 9312 uid58357
.		Prochlorococcus marinus AS9601 uid58307

Continued on next page

Table 12.6 – continued from previous page

Phylum	Family	Organism
.		Prochlorococcus marinus MIT 9301 uid58437
.		Prochlorococcus marinus MIT 9202 uid54709
.		Prochlorococcus marinus MIT 9215 uid58819
.	Gloeobacteriales	Synechococcus elongatus PCC 6301 uid58235
.	(nodes 2147,2149-2157)	Synechococcus elongatus PCC 7942 uid58045
.		Acaryochloris marina MBIC11017 uid58167
.		Cyanothece PCC 7425 uid59435
.		Synechococcus PCC 6312 uid182934
.		Thermosynechococcus elongatus BP 1 uid57907
.		Synechococcus PCC 7502 uid183008
.		Pseudanabaena biceps PCC 7429 uid187056
.		Pseudanabaena PCC 7367 uid183004
.		Gloeobacter violaceus PCC 7421 uid58011
.		Synechococcus JA 2 3B a 2 13 uid58537
.		Synechococcus JA 3 3Ab uid58535
.	Oscillatoriales	Leptolyngbya PCC 6406 uid186936
.	(nodes 2159/62, 2164/70)	Synechococcus PCC 7335 uid54731
.		Leptolyngbya PCC 7375 uid182890
.		Geitlerinema PCC 7407 uid183007
.		Oscillatoriales cyanobacterium JSC 12 uid179140
.		Oscillatoria acuminata PCC 6304 uid183003
.		Oscillatoria PCC 6506 uid50611
.		Microcoleus vaginatus FGP 2 uid67389
.		Oscillatoria PCC 7112 uid183110
.		Trichodesmium erythraeum IMS101 uid57925
.		Lyngbya PCC 8106 uid54161
.		Arthrospira platensis C1 uid181584
.		Arthrospira platensis Paraca uid55907
.	Nostocales	Cylindrospermopsis raciborskii CS 505 uid42983
.	(nodes 2172-2189)	Raphidiopsis brookii D9 uid42981
.		Anabaena cylindrica PCC 7122 uid183339
.		Nostoc azollae 0708 uid49725
.		Anabaena 90 uid179383
.		Nodularia spumigena CCY9414 uid54171
.		Nostoc punctiforme PCC 73102 uid57767
.		Cylindrospermum stagnale PCC 7417 uid183111
.		Calothrix PCC 7507 uid182930
.		Nostoc PCC 7107 uid182932
.		Nostoc PCC 7524 uid182933
.		Anabaena variabilis ATCC 29413 uid58043
.		Nostoc PCC 7120 uid57803
.		Calothrix PCC 6303 uid183109
.		Rivularia PCC 7116 uid182929
.		Fischerella JSC 11 uid75099
.		Chroococcidiopsis thermalis PCC 7203 uid183002
.		Gloeocapsa PCC 7428 uid183112
.		Synechocystis PCC 7509 uid186935
.	Croococcales	Chamaesiphon PCC 6605 uid183005
.	(nodes 2190-2212)	Crinalium epipsammum PCC 9333 uid183113

Continued on next page

Table 12.6 – continued from previous page

Phylum	Family	Organism
.		Microcoleus chthonoplastes PCC 7420 uid54695
.		Lyngbya majuscula 3L uid66849
.		Dactylococcopsis salina PCC 8305 uid183341
.		Halothece PCC 7418 uid183338
.		Leptolyngbya PCC 7376 uid182928
.		Synechococcus PCC 7002 uid59137
.		Stanieria cyanosphaera PCC 7437 uid183115
.		Xenococcus PCC 7305 uid186938
.		Gloeocapsa PCC 73106 uid186937
.		Cyanobacterium PCC 10605 uid183340
.		Cyanobacterium stanieri PCC 7202 uid183337
.		Pleurocapsa PCC 7327 uid183006
.		Synechocystis PCC 6803 substr GT I uid158059
.		Cyanothece PCC 8801 uid59027
.		Cyanothece PCC 8802 uid59143
.		cyanobacterium UCYN A uid43697
.		Crocosphaera watsonii WH 0003 uid179643
.		Cyanothece ATCC 51142 uid59013
.		Cyanothece PCC 7424 uid59025
.		Cyanothece PCC 7822 uid52547
.		Microcystis aeruginosa NIES 843 uid59101
.		Microcystis aeruginosa TAIHU98 uid185796

Jump to: *Bacteroidetes, et al.*, *Alpha proteobacteria, et al.*, *Beta and Gamma Proteobacteria*, *Actinobacteria*, *Firmicutes, et al.*, *Cyanobacteria*, *Archaea*, and *Eukaryotes*.

12.7 Archaea

This group contains the Archaea. Annotated pdfs with node numbers and family names are available for

- `crenarchaeota` and
- `euryarchaeota`.

Phylum	Family	Organism
Nanoarchaeota		Nanoarchaeum equitans Kin4 M uid58009
Euryarachaeota		Methanopyrus kandleri AV19 uid57883
Korarchaeota		Candidatus Korarchaeum cryptofilum OPF8 uid58601
Crenarchaeota		Sulfolobus acidocaldarius DSM 639 uid58379
(nodes 2219-2252)		Sulfolobus tokodaii 7 uid57807
.		Sulfolobus islandicus HVE10 4 uid162067
.		Sulfolobus solfataricus 98 2 uid167998
.		Acidianus hospitalis W1 uid66875
.		Metallosphaera yellowstonensis MK1 uid82737
.		Metallosphaera cuprina Ar 4 uid66329
.		Metallosphaera sedula DSM 5348 uid58717
.		Staphylothermus hellenicus DSM 12710 uid45893
.		Thermogladius 1633 uid167488

Continued on next page

Table 12.7 – continued from previous page

Phylum	Family	Organism
.		Thermosphaera aggregans DSM 11486 uid48993
.		Desulfurococcus mucosus DSM 2162 uid62227
.		Desulfurococcus fermentans DSM 16532 uid75119
.		Desulfurococcus kamchatkensis 1221n uid59133
.		Hyperthermus butylicus DSM 5456 uid57755
.		Pyrolobus fumarii 1A uid73415
.		Aeropyrum pernix K1 uid57757
.		Acidilobus saccharovorans 345 15 uid51395
.		Caldisphaera lagunensis DSM 15908 uid183486
.		Ignicoccus hospitalis KIN4 I uid58365
.		Fervidicoccus fontis Kam940 uid162201
.		Ignisphaera aggregans DSM 17230 uid51875
.		Thermofilum pendens Hrk 5 uid58563
.		Pyrobaculum calidifontis JCM 11548 uid58787
.		Pyrobaculum islandicum DSM 4184 uid58635
.		Pyrobaculum neutrophilum V24Sta uid58421
.		Pyrobaculum arsenaticum DSM 13514 uid58409
.		Pyrobaculum oguniense TE7 uid84411
.		Pyrobaculum 1860 uid82379
.		Pyrobaculum aerophilum IM2 uid57727
.		Thermoproteus tenax Kra 1 uid74443
.		Thermoproteus uzoniensis 768 20 uid65089
.		Caldivirga maquilingensis IC 167 uid58711
.		Vulcanisaeta distributa DSM 14429 uid52827
.		Vulcanisaeta moutnovskia 768 28 uid63631
Thaumarchaeota		Candidatus Nitrososphaera gargensis Ga9 2 uid176707
(nodes 2253-2258)		Cenarchaeum symbiosum A uid61411
.		Nitrosopumilus maritimus SCM1 uid58903
.		marine archaeal group 1 BD31 uid81845
.		Candidatus Nitrosopumilus AR2 uid176130
.		Candidatus Nitrosoarchaeum limnia BG20 uid169874
.		Nitrosopumilus MY1 uid70553
Euryarchaeota	Thermococcales	Pyrococcus yayanosii CH1 uid68281
(nodes 2259-2400)	(nodes 2260-2274)	Pyrococcus furiosus COM1 uid169620
.		Pyrococcus ST04 uid167261
.		Pyrococcus abyssi GE5 uid62903
.		Pyrococcus NA2 uid66551
.		Pyrococcus horikoshii OT3 uid57753
.		Thermococcus barophilus MP uid54733
.		Thermococcus litoralis DSM 5473 uid82997
.		Thermococcus sibiricus MM 739 uid59399
.		Thermococcus onnurineus NA1 uid59043
.		Thermococcus 4557 uid70841
.		Thermococcus CL1 uid168259
.		Thermococcus kodakarensis KOD1 uid58225
.		Thermococcus zilligii AN1 uid176935
.		Thermococcus AM4 uid54735
.		Thermococcus gammatolerans EJ3 uid59389
.	Methanococci	Methanocaldococcus infernus ME uid48803

Continued on next page

Table 12.7 – continued from previous page

Phylum	Family	Organism
.	(nodes 2277-2287)	Methanocaldococcus vulcanius M7 uid41131
.		Methanocaldococcus jannaschii DSM 2661 uid57713
.		Methanocaldococcus fervens AG86 uid59347
.		Methanocaldococcus FS406 22 uid42499
.		Methanococcus voltae A3 uid49529
.		Methanococcus maripaludis C5 uid58741
.		Methanococcus vannielii SB uid58767
.		Methanococcus aeolicus Nankai 3 uid58823
.		Methanothermococcus okinawensis IH1 uid51535
.		Methanotorris formicicus Mc S 70 uid82739
.		Methanotorris igneus Kol 5 uid67321
.		Methanothermus fervidus DSM 2088 uid60167
.		Methanosphaera stadtmanae DSM 3091 uid58407
.		Methanobacterium formicicum DSM 3637 uid176604
.		Methanobacterium Maddingley MBC34 uid178946
.		Methanobacterium AL 21 uid63623
.		Methanobacterium SWAN 1 uid67359
.		Methanobrevibacter ruminantium M1 uid45857
.		Methanobrevibacter smithii ATCC 35061 uid58827
.		Methanothermobacter marburgensis Marburg uid51637
.		Methanothermobacter thermautotrophicus Delta H uid57877
.		Aciduliprofundum boonei T469 uid43333
.		Aciduliprofundum MAR08 339 uid184407
.		Thermoplasma acidophilum DSM 1728 uid61573
.		Thermoplasma volcanium GSS1 uid57751
.		Ferroplasma acidarmanus fer1 uid54095
.		Picrophilus torridus DSM 9790 uid58041
.		Ferroglobus placidus DSM 10642 uid40863
.		Archaeoglobus veneficus SNP6 uid65269
.		Archaeoglobus profundus DSM 5631 uid43493
.		Archaeoglobus fulgidus DSM 4304 uid57717
.		Methanocella arvoryzae MRE50 uid61623
.		Methanocella conradii HZ254 uid157911
.		Methanocella paludicola SANAE uid42887
.		Methanosaeta thermophila PT uid58469
.		Methanosaeta harundinacea 6Ac uid81199
.		Methanosaeta concilii GP6 uid66207
.		Methanosarcina barkeri Fusaro uid57715
.		Methanosarcina acetivorans C2A uid57879
.		Methanosarcina mazei Go1 uid57893
.		Methanohalobium evestigatum Z 7303 uid49857
.		Methanosalsum zhilinae DSM 4017 uid68249
.		Methanohalophilus mahii DSM 5219 uid47313
.		Methanococcoides burtonii DSM 6242 uid58023
.		Methanomethylovorans hollandica DSM 15978 uid184864
.		Methanolobus psychrophilus R15 uid177925
.		Methanofollis liminatans DSM 4140 uid170729
.		Methanoplanus limicola DSM 2279 uid82735
.		Methanoplanus petrolearius DSM 11571 uid52695

Continued on next page

Table 12.7 – continued from previous page

Phylum	Family	Organism
.		Methanospirillum hungatei JF 1 uid58181
.		Methanocorpusculum labreanum Z uid58785
.		Methanoculleus bourgensis MS2 uid171377
.		Methanoculleus marisnigri JR1 uid58561
.		Methanosphaerula palustris E1 9c uid59193
.		Methanolinea tarda NOBI 1 uid76575
.		Methanoregula boonei 6A8 uid58815
.		Methanoregula formicicum SMSP uid184406
.		Halorubrum hochstenium ATCC 700873 uid188622
.		Halorubrum tebenquichense DSM 14210 uid188611
.		Halorubrum coriense DSM 10284 uid188619
.		Halorubrum californiensis DSM 19288 uid188618
.		Halorubrum distributum JCM 10118 uid188621
.		Halorubrum terrestre JCM 10247 uid188610
.		Halorubrum arcis JCM 13916 uid188617
.		Halorubrum litoreum JCM 13561 uid188613
.		Halorubrum aidingense JCM 13560 uid188616
.		Halorubrum lacusprofundi ATCC 49239 uid58807
.		Halorubrum saccharovorum DSM 1137 uid188612
.		Halorubrum kocurii JCM 14978 uid188615
.		Halorubrum lipolyticum DSM 21995 uid188614
.		halophilic archaeon DL31 uid72619
.		Halogranum salarium B 1 uid172007
.		Haloquadratum walsbyi C23 uid162019
.		Halogeometricum borinquense DSM 11551 uid54919
.		Halosarcina pallida JCM 14848 uid188609
.		Haloferax elongans ATCC BAA 1513 uid188630
.		Haloferax larsenii JCM 13917 uid188628
.		Haloferax alexandrinus JCM 10717 uid188631
.		Haloferax volcanii DS2 uid46845
.		Haloferax lucentense DSM 14919 uid188627
.		Haloferax prahovense DSM 18310 uid188626
.		Haloferax gibbonsii ATCC 33959 uid188629
.		Haloferax denitrificans ATCC 35960 uid188635
.		Haloferax sulfurifontis ATCC BAA 897 uid188632
.		Haloferax mediterranei ATCC 33500 uid167315
.		Haloferax mucosum ATCC BAA 1512 uid188633
.		Halalkalicoccus jeotgali B3 uid50305
.		Haladaptatus paucihalophilus DX253 uid62523
.		Halococcus hamelinensis 100A6 uid177808
.		Halococcus morrhuae DSM 1307 uid188645
.		Halomicrobium mukohataei DSM 12286 uid59107
.		Natronomonas moolapensis 8 8 11 uid190182
.		Natronomonas pharaonis DSM 2160 uid58435
.		Halosimplex carlsbadense 2 9 1 uid188608
.		Halorhabdus tiamatea SARL4B uid67927
.		Halorhabdus utahensis DSM 12940 uid59189
.		Halobacterium DL1 uid75121
.		Halobacterium NRC 1 uid57769

Continued on next page

Table 12.7 – continued from previous page

Phylum	Family	Organism
.		Halobacterium salinarum R1 uid61571
.		Halovivax ruber XH 70 uid184819
.		Natronococcus occultus SP4 uid184863
.		Natronococcus jeotgali DSM 18795 uid188591
.		Natronococcus amylolyticus DSM 10524 uid188592
.		Natrialba aegyptia DSM 13077 uid188603
.		Natrialba asiatica DSM 12278 uid188602
.		Natrialba taiwanensis DSM 12281 uid188599
.		Natrialba hulunbeirensis JCM 10989 uid188600
.		Natrialba magadii ATCC 43099 uid46245
.		Natrialba chahannaoensis JCM 10990 uid188601
.		Natronolimnobius innermongolicus JCM 12255 uid188590
.		Haloterrigena salina JCM 13891 uid188606
.		Haloterrigena turkmenica DSM 5511 uid43501
.		Halopiger xanaduensis SH 6 uid68105
.		Natronobacterium gregoryi SP2 uid74439
.		Halobiforma lacisalsi AJ5 uid157065
.		Halobiforma nitratireducens JCM 10879 uid188643
.		Haloterrigena limicola JCM 13563 uid188607
.		Natrinema pallidum DSM 3751 uid188596
.		Natrinema gari JCM 14663 uid188597
.		Natrinema altunense JCM 12890 uid188598
.		Natrinema versiforme JCM 10478 uid188594
.		Haloterrigena thermotolerans DSM 11522 uid188605
.		Natrinema pellirubrum DSM 15624 uid74437
.		Natronorubrum tibetense GA33 uid188587
.		Natronorubrum bangense JCM 10635 uid188589
.		Natronorubrum sulfidifaciens JCM 14089 uid188588

Jump to: *Bacteroidetes, et al.*, *Alpha proteobacteria, et al.*, *Beta and Gamma Proteobacteria*, *Actinobacteria*, *Firmicutes, et al.*, *Cyanobacteria*, *Archaea*, and *Eukaryotes*.

12.8 Eukaryotes

This group contains the the Eukaryotes. Annotated pdfs with node numbers and family names are available for download for

- `plants and protozoa` and
- `Metazoa and fungi`.

The phylogeny of lower eukaryotes is significantly more difficult to understand than the rest of this tree, and we relied heavily on Ginger, et al., Koonin, and Adl, et al. in assembling this phylogeny.

Phylum	Family	Organism
Protozoa	Alveolata	Cryptosporidium parvum
(nodes 2401-2442)	(nodes 2404-2416)	Cryptosporidium muris
.		Theileria parva
.		Theileria annulata

Continued on next page

Table 12.8 – continued from previous page

Phylum	Family	Organism
.		Babesia bovis
.		Plasmodium vivax
.		Plasmodium knowlesi
.		Plasmodium falciparum
.		Plasmodium yoelii
.		Toxoplasma gondii
.		Neospora caninum
.		Perkinsus marinus
.		Cilophora
.		Paramecium tetraurelia
.		Tetrahymena thermophila
.	Stramenopiles	Thalassiosira pseudonana
.	(nodes 2416-2417)	Phaeodactylum tricornutum
.		Phytophthora infestans
.	Cryptophyta	Guillardia theta
.		Trypanosoma brucei
.		Leishmania donovani
.		Leishmania braziliensis
.		Naegleria gruberi
.		Giardia lamblia
.		Trichomonas vaginalis
.		Glycine max
.		Medicago truncatula
.		Cucumis sativus
.		Ricinus communis
.		Populus trichocarpa
.		Arabidopsis thaliana
.		Solanum lycopersicum
.		Vitis vinifera
.		Oryza sativa
.		Brachypodium distachyon
.		Sorghum bicolor
.		Selaginella moellendorffii
.		Physcomitrella patens
.		Micromonas sp.
.		Ostreococcus tauri
.		Ostreococcus lucimarinus
.		Chlorella unnamed
.		Chlamydomonas reinhardtii
Amoebozoa	Mycetozoa	Dictyostelium purpureum
(nodes 2444-2445)	(node 2445)	Dictyostelium discoideum
.		Entamoeba dispar
.	Choanoflagellida	Monosiga brevicollis
Metazoa	Cnidaria	Nematostella vectensis
(nodes 2448-2481)	(node 2450	Hydra magnipapillata
.	Porifera	Amphimedon queenslandica
.	Placozoa	Trichoplax adhaerens
.	Nematodes	Trichinella spiralis
.	(nodes 2453-2455)	Caenorhabditis elegans

Continued on next page

Table 12.8 – continued from previous page

Phylum	Family	Organism
.		Loa loa
.		Brugia malayi
.	Trematode	Schistosoma mansoni
.	Hemichordate	Saccoglossus kowalevskii
.	Echinodermata	Strongylocentrotus prupuratus
.	Arthropods	Metaseiulus occidentalis
.	(nodes 2458-2469)	Ixodes unnamed
.		Tribolium castaneum
.		Pediculus humanus
.		Culex quinquefasciatus
.		Anopheles gambiae
.		Drosophila melanogaster
.		Megachile rotundata
.		Bombus impatiens
.		Apis mellifera
.		Nasonia vitripennis
.	Chordates	Ciona intestinalis
.	(nodes 2470-2481)	Oreochromis niloticus
.		Oryzias latipes
.		Danio rerio
.		Salmo salar
.		Takifugu rubripes
.		Xenopus tropicalis
.		Anolis carolinensis
.		Taeniopygia guttata
.		Gallus gallus
.		Sarcophilus harrisii
.		Mus musculus
.		Homo sapiens
Fungi	Microsporidia	Enterocytozoon bieneusi
(nodes 2482-2548)	(nodes 2483-2486)	Nosema ceranae
.		Encephalitozoon cuniculi
.		Encephalitozoon intestinalis
.		Encephalitozoon hellem
.	Basidiomycota	Puccinia graminis
.	(nodes 2488-2493)	Malassezia globosa
.		Schizophyllum commune
.		Coprinopsis cinerea
.		Laccaria bicolor
.		Ustilago maydis
.		Cryptococcus neoformans
.	Basidiomycota	Schizosaccharomyces pombe
.	(node 2495)	Schizosaccharomyces japonicus
.	Saccharomycotina	Yarrowia lipolytica
.	(nodes 2497-2517)	Lodderomyces elongisporus
.		Candida tropicalis
.		Candida dubliniensis
.		Debaryomyces hansenii
.		Meyerozyma guilliermondii

Continued on next page

Table 12.8 – continued from previous page

Phylum	Family	Organism
.		Scheffersomyces stipitis
.		Clavispora lusitaniae
.		Komagataella pastoris
.		Naumovozyma dairenensis
.		Naumovozyma castellii
.		Kazachstania africana
.		Saccharomyces cerevisiae
.		Candida glabrata
.		Torulaspora delbrueckii
.		Zygosaccharomyces rouxii
.		Vanderwaltozyma polyspora
.		Tetrapisispora phaffii
.		Lachancea thermotolerans
.		Eremothecium cymbalariae
.		Ashbya gossypii
.		Kluyveromyces lactis
.	Pezizomycotina	Zymoseptoria tritici
.	(nodes 2518-2548)	Pyrenophora tritici-repentis
.		Leptosphaeria maculans
.		Phaeosphaeria nodorum
.		Tuber melanosporum
.		Neosartorya fischeri
.		Aspergillus fumigatus
.		Aspergillus clavatus
.		Aspergillus niger
.		Aspergillus terreus
.		Aspergillus oryzae
.		Aspergillus flavus
.		Aspergillus nidulans
.		Penicillium chrysogenum
.		Talaromyces stipitatus
.		Uncinocarpus reesii
.		Coccidioides posadasii
.		Trichophyton rubrum
.		Arthroderma otae
.		Paracoccidioides sp.
.		Ajellomyces dermatitidis
.		Sclerotinia sclerotiorum
.		Botryotinia fuckeliana
.		Myceliophthora thermophila
.		Thielavia terrestris
.		Chaetomium globosum
.		Podospora anserina
.		Sordaria macrospora
.		Neurospora crassa
.		Magnaporthe oryzae
.		Verticillium albo-atrum
.		Nectria haematococca

CHAPTER 13

Definition of functional classifications

While the SEED project is a living annotation system, Sequedex, by necessity, used a snapshot of this system that was frozen in September of 2011 in its functional module seed_0911, distributed by default in its first data modules. In this chapter, we define the various subsystems. The 'si' number is used only by Sequedex, and does not have meaning at theSEED.org beyond that presented in this table. They are grouped by high-level rollup category. Most of the subsystems should be available at http://pubseed.theseed.org/seedviewer.cgi?page=SubsystemSelect.

13.1 Amino Acids and Derivatives

si_#	category	subsystem
0000	Alanine, serine,	Alanine_biosynthesis
.	and glycine	
0001		Glycine_Biosynthesis
0002		Glycine_and_Serine_Utilization
0003		Glycine_cleavage_system
0004		Serine_Biosynthesis
0005	Arginine; urea cycle,	Anaerobic_Oxidative_Degradation_of_L-Ornithine
.	polyamines	.
0006		Arginine_Biosynthesis_extended
0007		Arginine_Deiminase_Pathway
0008		Arginine_and_Ornithine_Degradation
0009		Cyanophycin_Metabolism
0010		Polyamine_Metabolism
0011		Putrescine_utilization_pathways
0012		Urea_decomposition
0013	Aromatic amino acids	Aromatic_amino_acid_degradation
.	and derivatives	
0014		Aromatic_amino_acid_interconversions_with_aryl_acids
0015		Bacilysin_biosynthesis
0016		Chorismate:_Intermediate_for_synthesis_of_PAPA_antibiotics,_PABA,
0017		Chorismate_Synthesis
0018		Common_Pathway_For_Synthesis_of_Aromatic_Compounds_(DAHP_synthase
0019		Indole-pyruvate_oxidoreductase_complex
0020		Phenylalanine_and_Tyrosine_Branches_from_Chorismate
0021		Tryptophan_catabolism
0022		Tryptophan_synthesis

Table 13.1 – continued from previous page

si_#	category	subsystem
0023	Branched-chain	Branched-Chain_Amino_Acid_Biosynthesis
.	amino acids	
0024		Branched_chain_amino_acid_degradation_regulons
0025		HMG_CoA_Synthesis
0026		Isoleucine_degradation
0027		Ketoisovalerate_oxidoreductase
0028		Leucine_Biosynthesis
0029		Leucine_Degradation_and_HMG-CoA_Metabolism
0030		Valine_degradation
0031	Glutamine, glutamate,	Glutamate_dehydrogenases
.	aspartate, asparagine;	
.	ammonia assimilation	
0032		Glutamine,_Glutamate,_Aspartate_and_Asparagine_Biosynthesis
0033		Glutamine_synthetases
0034		Poly-gamma-glutamate_biosynthesis
0035	Histidine Metabolism	Histidine_Biosynthesis
0036		Histidine_Degradation
0037	Lysine, threonine,	Cysteine_Biosynthesis
.	methionine, and cysteine	
0038		Lysine_Biosynthesis_DAP_Pathway
0039		Lysine_biosynthesis_AAA_pathway_2
0040		Lysine_degradation
0041		Lysine_fermentation
0042		Methionine_Biosynthesis
0043		Methionine_Degradation
0044		Methionine_Salvage
0045		Threonine_anaerobic_catabolism_gene_cluster
0046		Threonine_and_Homoserine_Biosynthesis
0047		Threonine_degradation
0048	Proline and	A_Hypothetical_Protein_Related_to_Proline_Metabolism
.	4-hydroxyproline	
0049		Proline,_4-hydroxyproline_uptake_and_utilization
0050		Proline_Synthesis
0051	Unclassified amino	Creatine_and_Creatinine_Degradation
.	acids and derivatives	
0052		L-2-amino-thiazoline-4-carboxylic_acid-Lcysteine_conversion
0053		Phosphonoalanine_utilization

13.2 Carbohydrates

si_#	category	subsystem
0054	Aminosugars	(GlcNAc)2_Catabolic_Operon
0055		Chitin_and_N-acetylglucosamine_utilization
0056		Fructoselysine_(Amadori_product)_utilization_pathway
0057		N-Acetyl-Galactosamine_and_Galactosamine_Utilization
0058		Neotrehalosadiamine_(NTD)_Biosynthesis_Operon
0059	CO2 fixation	CO2_uptake,_carboxysome

Table 13.2 – continued from previous page

si_#	category	subsystem
0060		Calvin-Benson_cycle
0061		Carboxysome
0062		Photorespiration_(oxidative_C2_cycle)
0063	Central carbohydrate	Dehydrogenase_complexes
.	metabolism	
0064		Dihydroxyacetone_kinases
0065		Entner-Doudoroff_Pathway
0066		Ethylmalonyl-CoA_pathway_of_C2_assimilation
0067		Glycolate,_glyoxylate_interconversions
0068		Glycolysis_and_Gluconeogenesis
0069		Glycolysis_and_Gluconeogenesis,_including_Archaeal_enzymes
0070		Glyoxylate_bypass
0071		Methylglyoxal_Metabolism
0072		Particulate_methane_monooxygenase_(pMMO)
0073		Pentose_phosphate_pathway
0074		Peripheral_Glucose_Catabolism_Pathways
0075		Pyruvate:ferredoxin_oxidoreductase
0076		Pyruvate_Alanine_Serine_Interconversions
0077		Pyruvate_metabolism_I:_anaplerotic_reactions,_PEP
0078		Pyruvate_metabolism_II:_acetyl-CoA,_acetogenesis_from_pyruvate
0079		Soluble_methane_monooxygenase_(sMMO)
0080		TCA_Cycle
0081	Di- and oligosaccharides	Beta-Glucoside_Metabolism
0082		Fructooligosaccharides(FOS)_and_Raffinose_Utilization
0083		Lactose_and_Galactose_Uptake_and_Utilization
0084		Lactose_utilization
0085		Maltose_and_Maltodextrin_Utilization
0086		Melibiose_Utilization
0087		Sucrose_utilization
0088		Sucrose_utilization_Shewanella
0089		Trehalose_Biosynthesis
0090		Trehalose_Uptake_and_Utilization
0091		Unknown_oligosaccharide_utilization_Sde_1396
0092	Fermentation	Acetoin,_butanediol_metabolism
0093		Acetone_Butanol_Ethanol_Synthesis
0094		Acetyl-CoA_fermentation_to_Butyrate
0095		Butanol_Biosynthesis
0096		Fermentations:_Lactate
0097		Fermentations:_Mixed_acid
0098	Glycoside hydrolases	Predicted_carbohydrate_hydrolases
0099	Monosaccharides	2-Ketogluconate_Utilization
0100		D-Galacturonate_and_D-Glucuronate_Utilization
0101		D-Sorbitol(D-Glucitol)_and_L-Sorbose_Utilization
0102		D-Tagatose_and_Galactitol_Utilization
0103		D-allose_utilization
0104		D-galactarate,_D-glucarate_and_D-glycerate_catabolism
0105		D-galactonate_catabolism
0106		D-gluconate_and_ketogluconates_metabolism
0107		D-ribose_utilization

Table 13.2 – continued from previous page

si_#	category	subsystem
0108		Deoxyribose_and_Deoxynucleoside_Catabolism
0109		Fructose_utilization
0110		Hexose_Phosphate_Uptake_System
0111		L-Arabinose_utilization
0112		L-ascorbate_utilization_(and_related_gene_clusters)
0113		L-fucose_utilization
0114		L-fucose_utilization_temp
0115		L-rhamnose_utilization
0116		Mannose_Metabolism
0117		Xylose_utilization
0118	One-carbon Metabolism	Formaldehyde_assimilation:_Ribulose_monophosphate_pathway
0119		Methanogenesis
0120		Methanogenesis_from_methylated_compounds
0121		One-carbon_metabolism_by_tetrahydropterines
0122		Serine-glyoxylate_cycle
0123	Organic acids	2-methylcitrate_to_2-methylaconitate_metabolism_cluster
0124		Alpha-acetolactate_operon
0125		Glycerate_metabolism
0126		Isobutyryl-CoA_to_Propionyl-CoA_Module
0127		Lactate_utilization
0128		Malonate_decarboxylase
0129		Methylcitrate_cycle
0130		Propionate-CoA_to_Succinate_Module
0131		Propionyl-CoA_to_Succinyl-CoA_Module
0132		Tricarballylate_Utilization
0133	Polysaccharides	Alpha-Amylase_locus_in_Streptococococcus
0134		Cellulosome
0135		Glycogen_metabolism
0136		Unknown_carbohydrate_utilization_containing_Fructose-bisphosphat
0137		Xyloglucan_Utilization
0138	Sugar alcohols	Erythritol_utilization
0139		Ethanolamine_utilization
0140		Glycerol_and_Glycerol-3-phosphate_Uptake_and_Utilization
0141		Glycerol_fermenation_to_1,3-propanediol
0142		Inositol_catabolism
0143		Mannitol_Utilization
0144		Propanediol_utilization
0145	Unclassified carbohydrates	Conserved_cluster_around_inner_membrane_protein_gene_yghQ,_proba
0146		Lacto-N-Biose_I_and_Galacto-N-Biose_Metabolic_Pathway
0147		Sugar_utilization_in_Thermotogales
0148		Unknown_carbohydrate_utilization_(_cluster_Ydj_)
0149		Unknown_carbohydrate_utilization_(_cluster_Yeg_)
0150		Unknown_sugar_utilization_(cluster_yphABCDEFG)
0151		VC0266
0152		beta-glucuronide_utilization

13.3 Cell Division and Cell Cycle

si_#	category	subsystem
0153	Unclassified cell division	Bacterial_Cytoskeleton
0154		Control_of_cell_elongation__division_cycle_in_Bacilli
0155		Cyanobacterial_Circadian_Clock
0156		Heterocyst_formation_in_cyanobacteria
0157		Intracellular_septation_in_Enterobacteria
0158		Macromolecular_synthesis_operon
0159		MukBEF_Chromosome_Condensation
0160		Two_cell_division_clusters_relating_to_chromosome_partitioning
0161		YgjD_and_YeaZ

13.4 Cell Wall and Capsule

si_#	category	subsystem
0162	Capsular and extracellular	
.	polysacchrides	Alginate_metabolism
0163		CMP-N-acetylneuraminate_Biosynthesis
0164		Capsular_Polysaccharide_(CPS)_of_Campylobacter
0165		Capsular_Polysaccharides_Biosynthesis_and_Assembly
0166		Capsular_heptose_biosynthesis
0167		Capsular_surface_virulence_antigen_loci
0168		Colanic_acid_biosynthesis
0169		Exopolysaccharide_Biosynthesis
0170		Extracellular_Polysaccharide_Biosynthesis_of_Streptococci
0171		Legionaminic_Acid_Biosynthesis
0172		O-Methyl_Phosphoramidate_Capsule_Modification_in_Campylobacter
0173		Phosphorylcholine_incorporation_in_LPS
0174		Polysaccharide_deacetylases
0175		Pseudaminic_Acid_Biosynthesis
0176		Rhamnose_containing_glycans
0177		Serotype_determining_Capsular_polysaccharide_biosynthesis_in_Sta
0178		Sialic_Acid_Metabolism
0179		Streptococcal_Hyaluronic_Acid_Capsule
0180		Vibrio_Polysaccharide_(VPS)_Biosynthesis
0181		Xanthan_Exopolysaccharide_Biosynthesis_and_Export
0182		YjbEFGH_Locus_Involved_in_Exopolysaccharide_Production
0183		dTDP-rhamnose_synthesis
0184	Cell wall of Mycobacteria	mycolic_acid_synthesis
0185	Gram-Negative cell	
.	wall components	Core_Oligosaccharide_Glycosylation_in_Pseudomonas
0186		Inner_membrane_protein_YhjD_and_conserved_cluster_involved_in_LP
0187		KDO2-Lipid_A_biosynthesis
0188		LOS_core_oligosaccharide_biosynthesis
0189		Lipid_A-Ara4N_pathway_(_Polymyxin_resistance_)
0190		Lipid_A_modifications

Conti

Table 13.3 – continued from previous page

si_#	category	subsystem
0191		Lipopolysaccharide-related_cluster_in_Alphaproteobacteria
0192		Lipopolysaccharide_assembly
0193		Lipoprotein_sorting_system
0194		Major_Outer_Membrane_Proteins
0195		Outer_membrane
0196		Perosamine_Synthesis_Vibrio
0197		Vibrio_Core_Oligosaccharide_Biosynthesis
0198	Gram-Positive cell	
.	wall components	Anthrose_Biosynthesis
0199		D-Alanyl_Lipoteichoic_Acid_Biosynthesis
0200		Polyglycerolphosphate_lipoteichoic_acid_biosynthesis
0201		Sortase
0202		Teichoic_and_lipoteichoic_acids_biosynthesis
0203		Teichuronic_acid_biosynthesis
0204	Unclassified cell wall	Murein_Hydrolases
.	and capsule	
0205		Peptidoglycan_Biosynthesis
0206		Peptidoglycan_Crosslinking_of_Peptide_Stems
0207		Peptidoglycan_biosynthesis--gjo
0208		Recycling_of_Peptidoglycan_Amino_Acids
0209		Recycling_of_Peptidoglycan_Amino_Sugars
0210		UDP-N-acetylmuramate_from_Fructose-6-phosphate_Biosynthesis
0211		YjeE
0212		tRNA-dependent_amino_acid_transfers

13.5 Clustering-based subsystems

si_#	category	subsystem
0213	Biosynthesis of galacto-	CBSS-258594.1.peg.3339
.	glycans and related	
.	lipopolysacharides	
0214		CBSS-376686.6.peg.291
0215	CRISPRs and associated	CBSS-216592.1.peg.3534
.	hypotheticals	
0216	Carbohydrates	Beta-lactamase_cluster_in_Streptococcus
0217		Cluster_Ytf_and_putative_sugar_transporter
0218		Predicted_mycobacterial_monooxygenase
0219		Putative_sugar_ABC_transporter_(ytf_cluster)
0220	Carotenoid biosynthesis	CBSS-320388.3.peg.3759
0221	Catabolism of an	CBSS-262316.1.peg.2929
.	unknown compound	
0222	Cell Division	CBSS-393130.3.peg.794
0223		Cell_Division_Cluster
0224	Chemotaxis, response	CBSS-323850.3.peg.3142
.	regulators	
0225	Choline bitartrate	CBSS-344610.3.peg.2335
.	degradation, putative	

Table 13.4 – continued from previous pag

si_#	category	subsystem
0226	Chromosome Replication	SeqA_and_Co-occurring_Genes
0227	Clustering-based	CBSS-262719.3.peg.410
.	subsystems	
0228		CBSS-280355.3.peg.2835
0229		CBSS-292415.3.peg.2341
0230		Putative_diaminopropionate_ammonia-lyase_cluster
0231		Sporulation-related_Hypotheticals
0232	Cytochrome biogenesis	CBSS-196164.1.peg.1690
0233		CBSS-196164.1.peg.461
0234		D-tyrosyl-tRNA(Tyr) deacylase (EC 3.1.-.-) cluster CBSS-342610.3.
0235		DNA metabolism CBSS-269801.1.peg.2186
0236		DNA polymerase III epsilon cluster CBSS-342610.3.peg.1536
0237		Fatty acid metabolic cluster CBSS-246196.1.peg.364
0238		Flagella protein? CBSS-323098.3.peg.2823
0239		Hypothetical Related to Dihydroorate Dehydrogenase Hypothetical_R
0240		Hypothetical associated with RecF Hypothetical_Coupled_to_RecF
0241		Hypothetical in Lysine biosynthetic cluster CBSS-323850.3.peg.326
0242		Hypothetical lipase related to Phosphatidate metabolism CBSS-3164
0243		Hypothetical protein possible functionally linked with Alanyl-tRN
0244		Isoprenoid/cell wall biosynthesis: PREDICTED UNDECAPRENYL DIPHOS
0245		Lysine Biosynthesis A_Glutathione-dependent_Thiol_Reductase_Assoc
0246	Lysine, threonine,	CBSS-84588.1.peg.1247
.	methionine, and	
.	cysteine	
0247		YeiH
0248		Membrane-bound hydrogenase CBSS-69014.3.peg.2094
0249		Methylamine utilization CBSS-265072.7.peg.546
0250		Molybdopterin oxidoreductase CBSS-269799.3.peg.2220
0251		Monosaccharides Unspecified_monosaccharide_transport_cluster
0252		Nucleotidyl-phosphate metabolic cluster CBSS-222523.1.peg.1311
0253		Pigment biosynthesis CBSS-176299.3.peg.235
0254		Probably GTP or GMP signaling related CBSS-176299.4.peg.1292
0255		Probably Pyrimidine biosynthesis-related CBSS-306254.1.peg.1508
0256		Probably Ybbk-related hypothetical membrane proteins CBSS-316057.
0257		Probably organic hydroperoxide resistance related hypothetical pr
0258		Proteasome related clusters Proteasome_subunit_alpha_archaeal_clu
0259		Protein export? CBSS-393121.3.peg.2760
0260		Putative GGDEF domain protein related to agglutinin secretion CBS
0261		Putative Isoquinoline 1-oxidoreductase subunit CBSS-314267.3.peg.
0262		Putative asociate of RNA polymerase sigma-54 factor rpoN CBSS-316
0263		Putrescine/GABA utilization cluster-temporal,to add to SSs GABA_a
0264		Pyruvate kinase associated cluster CBSS-288000.5.peg.1793
0265		Recombination related cluster CBSS-198094.1.peg.4426
0266		Related to Menaquinone-cytochrome C reductase CBSS-393130.3.peg.1
0267		Ribosomal Protein L28P ... A_Gram-positive_cluster_that_relates_
0268		Ribosome-related cluster A_Gammaproteobacteria_Cluster_Relating_t
0269		Sarcosine oxidase CBSS-188.1.peg.6170
0270		Shiga toxin cluster CBSS-194948.1.peg.143
0271		Sulfatases and sulfatase modifying factor 1 (and a hypothetical)

Table 13.4 – continued from previous pag

si_#	category	subsystem
0272		Three hypotheticals linked to lipoprotein biosynthesis CBSS-188.1
0273		TldD cluster CBSS-354.1.peg.2917
0274	Translation	CBSS-243265.1.peg.198
0275		CBSS-326442.4.peg.1852
0276		Tricarboxylate transporter CBSS-49338.1.peg.459
0277		Two related proteases CBSS-257314.1.peg.676
0278		Type III secretion system, extended Type_III_secretion_systems,_e
0279	Unclassified	Bacterial_Cell_Division
0280		Bacterial_RNA-metabolizing_Zn-dependent_hydrolases
0281		CBSS-138119.3.peg.2719
0282		CBSS-159087.4.peg.2189
0283		CBSS-160492.1.peg.550
0284		CBSS-176279.3.peg.1262
0285		CBSS-176279.3.peg.868
0286		CBSS-176280.1.peg.1561
0287		CBSS-196620.1.peg.2477
0288		CBSS-214092.1.peg.3450
0289		CBSS-224911.1.peg.435
0290		CBSS-228410.1.peg.134
0291		CBSS-235.1.peg.567
0292		CBSS-243277.1.peg.4359
0293		CBSS-251221.1.peg.1863
0294		CBSS-257314.1.peg.752
0295		CBSS-261594.1.peg.2640
0296		CBSS-266117.6.peg.2476
0297		CBSS-269801.1.peg.1715
0298		CBSS-269801.1.peg.1725
0299		CBSS-281090.3.peg.464
0300		CBSS-288681.3.peg.1039
0301		CBSS-290633.1.peg.1906
0302		CBSS-291331.3.peg.3674
0303		CBSS-296591.1.peg.2330
0304		CBSS-312309.3.peg.1965
0305		CBSS-314269.3.peg.1840
0306		CBSS-316057.3.peg.3521
0307		CBSS-316057.3.peg.563
0308		CBSS-316273.3.peg.227
0309		CBSS-316273.3.peg.2378
0310		CBSS-316273.3.peg.2709
0311		CBSS-316273.3.peg.448
0312		CBSS-316273.3.peg.922
0313		CBSS-316279.3.peg.746
0314		CBSS-316407.3.peg.2816
0315		CBSS-320372.3.peg.6046
0316		CBSS-323097.3.peg.2594
0317		CBSS-342610.3.peg.1794
0318		CBSS-345072.3.peg.1318
0319		CBSS-350688.3.peg.1509
0320		CBSS-370552.3.peg.1240

Table 13.4 – continued from previous pag

si_#	category	subsystem
0321		CBSS-393121.3.peg.1913
0322		CBSS-393124.3.peg.2657
0323		CBSS-393131.3.peg.612
0324		CBSS-393133.3.peg.2787
0325		CBSS-562.2.peg.5158_SK3_including
0326		CBSS-56780.10.peg.1536
0327		CBSS-630.2.peg.3360
0328		CBSS-83332.1.peg.3803
0329		CBSS-83333.1.peg.946
0330		CBSS-87626.3.peg.3639
0331		Cell_division-ribosomal_stress_proteins_cluster
0332		Conserved_cluster_around_acetyltransferase_YpeA_in_Enterobacteria
0333		Conserved_cluster_in_Enterobacteriaceae_downstream_from_YqjA,_a_D
0334		Conserved_gene_cluster_associated_with_Met-tRNA_formyltransferase
0335		EC49-61
0336		EC699-706
0337		KH_domain_RNA_binding_protein_YlqC
0338		LMPTP_YfkJ_cluster
0339		LMPTP_YwlE_cluster
0340		NusA-TFII_Cluster
0341		PA0057_cluster
0342		Putative_hemin_transporter
0343		Putative_sulfate_assimilation_cluster
0344		Spore_Coat
0345		Staphylococcus_aureus_hypothetical_repetitive_gene_loci
0346		USS-DB-1
0347		USS-DB-2
0348		USS-DB-4
0349		USS-DB-6
0350		USS-DB-7
0351		Yfa_cluster
0352		Urate degradation CBSS-205922.3.peg.1809
0353		alpha-proteobacterial cluster of hypotheticals CBSS-52598.3.peg.2
0354		proteosome related Cluster-based_Subsystem_Grouping_Hypotheticals
0355		recX and regulatory cluster CBSS-261594.1.peg.788
0356		tRNA sulfuration CBSS-89187.3.peg.2957

13.6 Cofactors, Vitamins, Prosthetic Groups, Pigments

si_#	category	subsystem
0357	Biotin	Biotin_biosynthesis
0358	Coenzyme A	Coenzyme_A_Biosynthesis
0359	Coenzyme B	Coenzyme_B_synthesis
0360	Coenzyme F420	Coenzyme_F420_synthesis
0361	Coenzyme M	coenzyme_M_biosynthesis
0362	Folate and pterines	5-FCL-like_protein
0363		Folate_Biosynthesis

Continued on next page

Table 13.5 – continued from previous page

si_#	category	subsystem
0364		Methanopterin_biosynthesis2
0365		Molybdenum_cofactor_biosynthesis
0366		Pterin_biosynthesis
0367		Pterin_carbinolamine_dehydratase
0368		Pterin_metabolism
0369		Pterin_metabolism_3
0370		methanopterin_biosynthesis
0371		p-Aminobenzoyl-Glutamate_Utilization
0372	Lipoic acid	Lipoic_acid_metabolism
0373	NAD and NADP	NAD_and_NADP_cofactor_biosynthesis_global
0374		NAD_consumption
0375		NAD_regulation
0376	Pyridoxine	Pyridoxin(Vitamin_B6)_Degradation_Pathway
0377		Pyridoxin_(Vitamin_B6)_Biosynthesis
0378	Quinone cofactors	Coenzyme_PQQ_synthesis
0379		Menaquinone_Biosynthesis_via_Futalosine_--_gjo
0380		Menaquinone_and_Phylloquinone_Biosynthesis
0381		Plastoquinone_Biosynthesis
0382		Pyrroloquinoline_Quinone_biosynthesis
0383		Tocopherol_Biosynthesis
0384		Ubiquinone_Biosynthesis
0385	Riboflavin, FMN, FAD	Flavodoxin
0386		Riboflavin,_FMN_and_FAD_metabolism
0387		riboflavin_to_FAD
0388	Tetrapyrroles	Bilin_Biosynthesis
0389		Chlorophyll_Biosynthesis
0390		Chlorophyll_Degradation
0391		Cobalamin_synthesis
0392		Coenzyme_B12_biosynthesis
0393		Heme_and_Siroheme_Biosynthesis
0394		Heme_biosynthesis_orphans
0395	Unclassified	Molybdopterin_cytosine_dinucleotide
0396		Thiamin_biosynthesis

13.7 DNA Metabolism

si_#	category	subsystem
0397	CRISPs	CRISPRs
0398		CRISP_Cmr_Cluster
0399	DNA recombination	DNA_recombination,_archaeal
0400		RuvABC_plus_a_hypothetical
0401	DNA repair	2-phosphoglycolate_salvage
0402		ATP-dependent_Nuclease
0403		DNA_Repair_Base_Excision
0404		DNA_repair,_UvrABC_system
0405		DNA_repair,_bacterial
0406		DNA_repair,_bacterial_DinG_and_relatives

Continued on next page

Table 13.6 – continued from previous page

si_#	category	subsystem
0407		DNA_repair,_bacterial_MutL-MutS_system
0408		DNA_repair,_bacterial_RecBCD_pathway
0409		DNA_repair,_bacterial_RecFOR_pathway
0410		DNA_repair,_bacterial_UmuCD_system
0411		DNA_repair,_bacterial_UvrD_and_related_helicases
0412		DNA_repair,_bacterial_photolyase
0413		Nonhomologous_End-Joining_in_Bacteria
0414		Uracil-DNA_glycosylase
0415	DNA replication	DNA-replication
0416		DNA_Helicase_of_Unknown_Function
0417		DNA_replication,_archaeal
0418		DNA_topoisomerases,_Type_I,_ATP-independent
0419		DNA_topoisomerases,_Type_II,_ATP-dependent
0420		Plasmid_replication
0421	DNA uptake, competence	Competence_in_Streptococci
0422		DNA_processing_cluster
0423		Gram_Positive_Competence
0424		Natural_DNA_Transformation_in_Vibrio
0425	Unclassified	DNA_phosphorothioation
0426		DNA_structural_proteins,_bacterial
0427		Nucleoid-associated_proteins_in_Bacteria
0428		Restriction-Modification_System
0429		Type_I_Restriction-Modification
0430		YcfH

13.8 Dormancy and Sporulation

si_#	category	subsystem
0431	Spore DNA protection	Dipicolinate_Synthesis
0432		Small_acid-soluble_spore_proteins
0433	Unclassified	Bacillus_Sporulation_Killing_Factor_A_Biosynthetic_Cluster
0434		Bacillus_biofilm_matrix_protein_component_TasA_and_homologs
0435		Exosporium
0436		Persister_Cells
0437		SpoVS_protein_family
0438		Spore_Core_Dehydration
0439		Spore_germination
0440		Spore_pigment_biosynthetic_cluster_in_Actinomycetes
0441		Sporulation-associated_proteins_with_broader_functions
0442		Sporulation_Cluster_III_A
0443		Sporulation_draft
0444		Sporulation_gene_orphans

13.9 Fatty Acids, Lipids, and Isoprenoids

si_#	category	subsystem
0445	Fatty acids	Acyl-CoA_thioesterase_II
0446		Carnitine_Metabolism_in_Microorganisms
0447		Fatty_Acid_Biosynthesis_FASI
0448		Fatty_Acid_Biosynthesis_FASII
0449		Fatty_acid_degradation_regulons
0450		Phospholipid_and_Fatty_acid_biosynthesis_related_cluster
0451		Polyunsaturated_Fatty_Acids_synthesis
0452	Isoprenoids	Acyclic_terpenes_utilization
0453		Archaeal_lipids
0454		Carotenoids
0455		Isoprenoid_Biosynthesis
0456		Myxoxanthophyll_biosynthesis_in_Cyanobacteria
0457		Polyprenyl_Diphosphate_Biosynthesis
0458		polyprenyl_synthesis
0459	Phospholipids	Glycerolipid_and_Glycerophospholipid_Metabolism_in_Bacteria
0460		Sphingolipid_biosynthesis
0461	Triacylglycerols	Triacylglycerol_metabolism
0462	Unclassified	Cholesterol_catabolic_operon_in_Mycobacteria
0463		Polyhydroxybutyrate_metabolism

13.10 Iron acquisition and metabolism

si_#	category	subsystem
0464	Siderophores	Alcaligin_Siderophore
0465		Bacillibactin_Siderophore
0466		Iron_siderophore_sensor_&_receptor_system
0467		Petrobactin-mediated_iron_uptake_system
0468		Salmochelin-mediated_Iron_Acquisition
0469		Siderophore_Achromobactin
0470		Siderophore_Aerobactin
0471		Siderophore_Desferrioxamine_E
0472		Siderophore_Enterobactin
0473		Siderophore_Pyoverdine
0474		Siderophore_Staphylobactin
0475		Siderophore_Yersiniabactin_Biosynthesis
0476		Siderophore_[Alcaligin-like]
0477		Siderophore_assembly_kit
0478		Siderophore_pyochelin
0479		Vibrioferrin_synthesis
0480	Unclassified	Campylobacter_Iron_Metabolism
0481		Ferrous_iron_transporter_EfeUOB,_low-pH-induced
0482		Heme,_hemin_uptake_and_utilization_systems_in_GramPositives
0483		Hemin_transport_system
0484		Iron_Scavenging_cluster_in_Thermus
0485		Iron_acquisition_in_Streptococcus
0486		Iron_acquisition_in_Vibrio
0487		Transport_of_Iron

13.11 Membrane Transport

si_#	category	subsystem
0488	ABC transporters	ABC_transporter_alkylphosphonate_(TC_3.A.1.9.1)
0489		ABC_transporter_branched-chain_amino_acid_(TC_3.A.1.4.1)
0490		ABC_transporter_dipeptide_(TC_3.A.1.5.2)
0491		ABC_transporter_of_unknown_substrate_X
0492		ABC_transporter_oligopeptide_(TC_3.A.1.5.1)
0493		ABC_transporter_peptide_(TC_3.A.1.5.5)
0494		ABC_transporter_tungstate_(TC_3.A.1.6.2)
0495		ATP-dependent_efflux_pump_transporter_Ybh
0496		Periplasmic-Binding-Protein-Dependent_Transport_System_for_&#94
0497	Protein and nucleoprotein	Dot-Icm_type_IV_secretion_system
.	secretion system, Type IV	
0498		Mannose-sensitive_hemagglutinin_type_4_pilus
0499		Toxin_co-regulated_pilus
0500		Type_4_conjugative_transfer_system,_IncI1_type
0501		Type_IV_pilus
0502		pVir_Plasmid_of_Campylobacter
0503	Protein secretion system,	General_Secretion_Pathway

Contin

Table 13.7 – continued from previous page

si_#	category	subsystem
.	Type II	
0504		Widespread_colonization_island
0505	Protein secretion system,	Type_III_secretion_system
.	Type III	
0506		Type_III_secretion_system_orphans
0507	Protein secretion system,	Type_VI_secretion_systems
.	Type VI	
0508	Protein secretion system,	Colonization_factor_antigen_I_fimbriae
.	Type VII (Chaperone/Usher	
.	pathway, CU)	
0509		The_fimbrial_Sfm_cluster
0510		The_fimbrial_Stf_cluster
0511		The_usher_protein_HtrE_fimbrial_cluster
0512	Protein secretion system,	Curli_production
.	Type VIII (Extracellular	
.	nucleation/precipitation	
.	pathway, ENP)	
0513	Protein translocation	ESAT-6_proteins_secretion_system_in_Actinobacteria
.	across cytoplasmic	
.	membrane	
0514		ESAT-6_proteins_secretion_system_in_Firmicutes
0515		HtrA_and_Sec_secretion
0516		SecY2-SecA2_Specialized_Transport_System
0517		Twin-arginine_translocation_system
0518	Sugar Phosphotransferase	Fructose_and_Mannose_Inducible_PTS
.	Systems, PTS	
0519		Galactose-inducible_PTS
0520		Sucrose-specific_PTS
0521	Unclassified	Agrobacterium_opine_transport
0522		Choline_Transport
0523		Citrate_Utilization_System_(CitAB,_CitH,_and_tctABC)
0524		ECF_class_transporters
0525		Phosphoglycerate_transport_system
0526		Ton_and_Tol_transport_systems
0527		Transport_of_Manganese
0528		Transport_of_Molybdenum
0529		Transport_of_Nickel_and_Cobalt
0530		Transport_of_Zinc
0531	Uni- Sym- and Antiporters	Na(+)_H(+)_antiporter
0532		Proton-dependent_Peptide_Transporters
0533		Sodium_Hydrogen_Antiporter

13.12 Metabolism of Aromatic Compounds

si_#	category	subsystem
0534	Anaerobic degradation	Anaerobic_benzoate_metabolism
.	of aromatic compounds	

Continued on next page

Table 13.8 – continued from previous page

si_#	category	subsystem
0535		Hydroxyaromatic_decarboxylase_family
0536	Metabolism of central	4-Hydroxyphenylacetic_acid_catabolic_pathway
.	aromatic intermediates	
0537		Catechol_branch_of_beta-ketoadipate_pathway
0538		Central_meta-cleavage_pathway_of_aromatic_compound_degradation
0539		Homogentisate_pathway_of_aromatic_compound_degradation
0540		N-heterocyclic_aromatic_compound_degradation
0541		Protocatechuate_branch_of_beta-ketoadipate_pathway
0542		Salicylate_and_gentisate_catabolism
0543	Peripheral pathways for	Benzoate_catabolism
.	catabolism of aromatic	
.	compounds	
0544		Benzoate_degradation
0545		Biphenyl_Degradation
0546		Chloroaromatic_degradation_pathway
0547		Chlorobenzoate_degradation
0548		Naphtalene_and_antracene_degradation
0549		Phenol_hydroxylase
0550		Phenylpropanoid_compound_degradation
0551		Quinate_degradation
0552		Salicylate_ester_degradation
0553		Toluene_degradation
0554		n-Phenylalkanoic_acid_degradation
0555		p-Hydroxybenzoate_degradation
0556	Unclassified	Aromatic_Amin_Catabolism
0557		Benzoate_transport_and_degradation_cluster
0558		Cresol_degradation
0559		Gentisare_degradation
0560		Phenylacetyl-CoA_catabolic_pathway_(core)
0561		Toluene_4-monooxygenase_(T4MO)
0562		carbazol_degradation_cluster
0563		p-cymene_degradation

13.13 Miscellaneous

si_#	category	subsystem
0564	Plant-Prokaryote DOE	At2g33980_At1g28960
.	project	
0565		At3g21300
0566		COG2363
0567		Conserved_gene_cluster_possibly_involved_in_RNA_metabolism
0568		Iojap
0569		Synechocystis_experimental
0570	Unclassified	Archease
0571		Archease2
0572		Broadly_distributed_proteins_not_in_subsystems
0573		Luciferases
0574		Magnetosome_Biomineralization_and_Function
0575		Muconate_lactonizing_enzyme_family
0576		YaaA
0577		YbbK
0578		ZZ_gjo_need_homes

13.14 Motility and Chemotaxis

si_#	category	subsystem
0579	Flagellar motility in	Additional_flagellar_genes_in_Vibrionales
.	Prokaryota	
0580		Archaeal_Flagellum
0581		Flagellar_motility
0582		Flagellum
0583		Flagellum_in_Campylobacter
0584	Social motility and	Control_of_Swarming_in_Vibrio_and_Shewanella_species
.	nonflagellar swimming	
.	in bacteria	
0585		Rhamnolipids_in_Pseudomonas
0586	Unclassified	Bacterial_Chemotaxis

13.15 Nitrogen Metabolism

si_#	category	subsystem
0587	Unclassified	Allantoin_Utilization
0588		Amidase_clustered_with_urea_and_nitrile_hydratase_functions
0589		Ammonia_assimilation
0590		Cyanate_hydrolysis
0591		Denitrification
0592		Dissimilatory_nitrite_reductase
0593		Nitrate_and_nitrite_ammonification
0594		Nitric_oxide_synthase
0595		Nitrogen_fixation
0596		Nitrosative_stress

13.16 Nucleosides and Nucleotides

si_#	category	subsystem
0597	Detoxification	Housecleaning_nucleoside_triphosphate_pyrophosphatases
0598		Nucleoside_triphosphate_pyrophosphohydrolase_MazG
0599		Nudix_proteins_(nucleoside_triphosphate_hydrolases)
0600	Purines	De_Novo_Purine_Biosynthesis
0601		Purine_Utilization
0602		Purine_conversions
0603		Purine_nucleotide_synthesis_regulator
0604		Xanthine_Metabolism_in_Bacteria
0605		Xanthosine_utilization_(xap_region)
0606	Pyrimidines	De_Novo_Pyrimidine_Synthesis
0607		Novel_non-oxidative_pathway_of_Uracil_catabolism
0608		Pyrimidine_utilization
0609		pyrimidine_conversions
0610	Unclassified	AMP_to_3-phosphoglycerate
0611		Adenosyl_nucleosidases
0612		Hydantoin_metabolism
0613		Pseudouridine_Metabolism
0614		Ribonucleotide_reduction

13.17 Phages, Prophages, Transposable elements, Plasmids

si_#	category	subsystem
0615	Bacteriophage integration	Phage_integration_and_excision
.	/excision/lysogeny	
0616	Bacteriophage structural	Phage_capsid_proteins
.	proteins	
0617	Superinfection Exclusion	Phage_Dual_Exonuclease_Exclusion
0618	Pathogenicity islands	Listeria_Pathogenicity_Island_LIPI-1_extended
0619		Staphylococcal_pathogenicity_islands_SaPI
0620		Vibrio_pathogenicity_island
0621	Phages, Prophages	Listeria_phi-A118-like_prophages
0622		Phage_baseplate_proteins
0623		Phage_neck_proteins
0624		Phage_nin_genes_-_N-independent_survival
0625		Phage_packaging_machinery
0626		Phage_tail_fiber_proteins
0627		Phage_tail_proteins_2
0628		Prophage-encoded_Rst_operon
0629		Prophage_lysogenic_conversion_modules
0630		Staphylococcal_phi-Mu50B-like_prophages
0631	Plasmid related functions	Plasmid-encoded_T-DNA_transfer
0632		Rolling-circle_replication
0633	Transposable elements	CBSS-203122.12.peg.188
0634		Conjugative_transposon,_Bacteroidales
0635		Tn552
0636	Unclassified	Integrons

13.18 Phosphorous Metabolism

si_#	category	subsystem
0637	Unclassified	Alkylphosphonate_utilization
0638		High_affinity_phosphate_transporter_and_control_of_PHO_regulon
0639		P_uptake_(cyanobacteria)
0640		Phosphate_metabolism
0641		Phosphoenolpyruvate_phosphomutase
0642		Phosphonate_metabolism

13.19 Photosynthesis

si_#	category	subsystem
0643	Electron transport and	Photosystem_I
.	photophosphorylation	
0644		Photosystem_I-type_photosynthetic_reaction_center
0645		Photosystem_II
0646		Photosystem_II-type_photosynthetic_reaction_center
0647	Light-harvesting complexes	Bacterial_light-harvesting_proteins
0648		Chlorosome
0649		Phycobilisome
0650	Unclassified	Bacteriorhodopsin
0651		Proteorhodopsin

13.20 Potassium metabolism

si_#	category	subsystem
0652	Unclassified	Glutathione-regulated_potassium-efflux_system_and_associated_functions
0653		Hyperosmotic_potassium_uptake
0654		Potassium_homeostasis
0655		pH_adaptation_potassium_efflux_system

13.21 Protein Metabolism

si_#	category	subsystem
0656	Protein biosynthesis	Nucleolar_protein_complex
0657		Programmed_frameshift
0658		Pyrrolysine
0659		Ribosomal_protein_paralogs
0660		Ribosome_LSU_bacterial
0661		Ribosome_LSU_eukaryotic_and_archaeal
0662		Ribosome_SSU_bacterial
0663		Ribosome_SSU_eukaryotic_and_archaeal
0664		Ribosome_activity_modulation
0665		Ribosome_biogenesis_bacterial
0666		Trans-translation_by_stalled_ribosomes
0667		Translation_elongation_factor_G_family
0668		Translation_elongation_factors_bacterial
0669		Translation_elongation_factors_eukaryotic_and_archaeal
0670		Translation_initiation_factors_bacterial
0671		Translation_initiation_factors_eukaryotic_and_archaeal
0672		Translation_termination_factors_bacterial
0673		Universal_GTPases
0674		tRNA_aminoacylation,_Ala
0675		tRNA_aminoacylation,_Arg
0676		tRNA_aminoacylation,_Asp_and_Asn

Continued on next page

Table 13.9 – continued from previous page

si_#	category	subsystem
0677		tRNA_aminoacylation,_Cys
0678		tRNA_aminoacylation,_Glu_and_Gln
0679		tRNA_aminoacylation,_Gly
0680		tRNA_aminoacylation,_His
0681		tRNA_aminoacylation,_Ile
0682		tRNA_aminoacylation,_Leu
0683		tRNA_aminoacylation,_Lys
0684		tRNA_aminoacylation,_Met
0685		tRNA_aminoacylation,_Phe
0686		tRNA_aminoacylation,_Pro
0687		tRNA_aminoacylation,_Pyr
0688		tRNA_aminoacylation,_Ser
0689		tRNA_aminoacylation,_Thr
0690		tRNA_aminoacylation,_Trp
0691		tRNA_aminoacylation,_Tyr
0692		tRNA_aminoacylation,_Val
0693	Protein degradation	Aminopeptidases_(EC_3.4.11.-)
0694		Dipeptidases_(EC_3.4.13.-)
0695		Metallocarboxypeptidases_(EC_3.4.17.-)
0696		Metalloendopeptidases_(EC_3.4.24.-)
0697		Omega_peptidases_(EC_3.4.19.-)
0698		Proteasome_archaeal
0699		Proteasome_bacterial
0700		Protein_degradation
0701		Proteolysis_in_bacteria,_ATP-dependent
0702		Putative_TldE-TldD_proteolytic_complex
0703		Serine_endopeptidase_(EC_3.4.21.-)
0704	Protein folding	GroEL_GroES
0705		Peptidyl-prolyl_cis-trans_isomerase
0706		Periplasmic_disulfide_interchange
0707		Protein_chaperones
0708		Thermosome,_archaeal
0709	Protein processing and	Inteins
.	modification	
0710		Lipoprotein_Biosynthesis
0711		N-linked_Glycosylation_in_Bacteria
0712		Peptide_methionine_sulfoxide_reductase
0713		Protein_Acetylation_and_Deacetylation_in_Bacteria
0714		Ribosomal_protein_S12p_Asp_methylthiotransferase
0715		Ribosomal_protein_S5p_acylation
0716		Signal_peptidase
0717		Ubiquitin-like_archaeal_modifier_proteins_(SAMPs)
0718	Secretion	Protein_secretion_by_ABC-type_exporters
0719	Selenoproteins	Glycine_reductase,_sarcosine_reductase_and_betaine_reductase
0720		Selenocysteine_metabolism
0721		Selenoprotein_O

13.22 RNA Metabolism

si_#	category	subsystem
0722	RNA processing and	ATP-dependent_RNA_helicases,_bacterial
.	modification	
0723		Methylthiotransferases
0724		Polyadenylation_bacterial
0725		Queuosine-Archaeosine_Biosynthesis
0726		RNA_3'-terminal_phosphate_cyclase
0727		RNA_processing_and_degradation,_bacterial
0728		RNA_processing_orphans
0729		RNA_pseudouridine_syntheses
0730		Ribonuclease_H
0731		Ribonuclease_P_archaeal_and_eukaryal
0732		Ribosome_biogenesis_archaeal
0733		Wyeosine-MimG_Biosynthesis
0734		YrdC-YciO
0735		eukaryotic_rRNA_modification_and_related_functions
0736		rRNA_modification_Archaea
0737		rRNA_modification_Bacteria
0738		tRNA_modification_Archaea
0739		tRNA_modification_Bacteria
0740		tRNA_nucleotidyltransferase
0741		tRNA_processing
0742		tRNA_splicing
0743	Transcription	RNA_polymerase_III
0744		RNA_polymerase_archaeal
0745		RNA_polymerase_archaeal_initiation_factors
0746		RNA_polymerase_bacterial
0747		Rrf2_family_transcriptional_regulators
0748		Transcription_elongation_factors,_archaeal
0749		Transcription_factors_bacterial
0750		Transcription_factors_cyanobacterial_RpoD-like_sigma_factors
0751		Transcription_initiation,_bacterial_sigma_factors
0752	Unclassified	Group_II_intron-associated_genes

13.23 Regulation and Cell signaling

si_#	category	subsystem
0753	Programmed Cell Death and	A_toxin-antitoxin_module_cotranscribed_with_DinB
.	Toxin-antitoxin Systems	
0754		Bacterial_Caspases
0755		Fratricide_in_Streptococcus
0756		MazEF_toxin-antitoxing_(programmed_cell_death)_system
0757		Murein_hydrolase_regulation_and_cell_death
0758		Phd-Doc,_YdcE-YdcD_toxin-antitoxin_(programmed_cell_death)_syst
0759		Toxin-antitoxin_replicon_stabilization_systems
0760		Toxin-antitoxin_systems_(other_than_RelBE_and_MazEF)
0761	Quorum sensing and	Autoinducer_2_(AI-2)_transport_and_processing_(lsrACDBFGE_operc
.	biofilm formation	
0762		Biofilm_Adhesin_Biosynthesis

Continued

Table 13.11 – continued from previous page

si_#	category	subsystem
0763		Biofilm_formation_in_Staphylococcus
0764		Quorum-sensing_in_Vibrio
0765		Quorum_sensing_in_Yersinia
0766		Quorum_sensing_regulation_in_Pseudomonas
0767		Symbiotic_colonization_and_sigma-dependent_biofilm_formation_ge
0768	Regulation of virulence	A_conserved_operon_linked_to_TyrR_and_possibly_involved_in_viru
0769		SpeB-SpeF_extended_regulon
0770		Streptococcal_Mga_Regulon
0771		Streptococcus_pyogenes_virulence_regulators
0772		Two-component_Response_Regulator_of_Virulence_ResDE
0773		VieSAB_signal_transduction_system_of_Vibrio
0774	Unclassified	Cell_envelope-associated_LytR-CpsA-Psr_transcriptional_attenuat
0775		CytR_regulation
0776		DNA-binding_regulatory_proteins,_strays
0777		Global_Two-component_Regulator_PrrBA_in_Proteobacteria
0778		HPr_catabolite_repression_system
0779		Orphan_regulatory_proteins
0780		Oxygen_and_light_sensor_PpaA-PpsR
0781		Pseudomonas_quinolone_signal_PQS
0782		Rcs_phosphorelay_signal_transduction_pathway
0783		Sex_pheromones_in_Enterococcus_faecalis_and_other_Firmicutes
0784		Stringent_Response,_(p)ppGpp_metabolism
0785		The_Chv_regulatory_system_of_Alphaproteobacteria
0786		Trans-envelope_signaling_system_VreARI_in_Pseudomonas
0787		Two-component_regulatory_systems_in_Campylobacter
0788		WhiB_and_WhiB-type_regulatory_proteins_
0789		Zinc_regulated_enzymes
0790		cAMP_signaling_in_bacteria

13.24 Respiration

si_#	category	subsystem
0791	ATP synthases	F0F1-type_ATP_synthase
0792		V-Type_ATP_synthase
0793	Electron accepting reactions	Anaerobic_respiratory_reductases
0794		Terminal_AA3-600_quinol_oxidase
0795		Terminal_cytochrome_C_oxidases
0796		Terminal_cytochrome_O_ubiquinol_oxidase
0797		Terminal_cytochrome_d_ubiquinol_oxidases
0798		Terminal_cytochrome_oxidases
0799		Tetrathionate_respiration
0800		Ubiquinone_Menaquinone-cytochrome_c_reductase_complexes
0801		trimethylamine_N-oxide_(TMAO)_reductase
0802	Electron donating reactions	CO_Dehydrogenase
0803		Coenzyme_F420-H2_dehydrogenase_(methanophenazine)
0804		Coenzyme_F420_hydrogenase
0805		Energy-conserving_hydrogenase_(ferredoxin)

Table 13.12 – continued from previous page

si_#	category	subsystem
0806		Formate_dehydrogenase
0807		H2:CoM-S-S-HTP_oxidoreductase
0808		Hydrogenases
0809		Methanophenazine_hydrogenase
0810		Na(+)-translocating_NADH-quinone_oxidoreductase_and_rnf-like_gr
0811		NiFe_hydrogenase_maturation
0812		Respiratory_Complex_I
0813		Respiratory_dehydrogenases_1
0814		Succinate_dehydrogenase
0815	Sodium Ion-Coupled	Na+_translocating_decarboxylases_and_related_biotin-dependent_e
.	Energetics	
0816	Unclassified	Biogenesis_of_c-type_cytochromes
0817		Biogenesis_of_cbb3-type_cytochrome_c_oxidases
0818		Biogenesis_of_cytochrome_c_oxidases
0819		Carbon_monoxide_dehydrogenase_maturation_factors
0820		Carbon_monoxide_induced_hydrogenase
0821		Cytochrome_B6-F_complex
0822		Flavocytochrome_C
0823		Formate_hydrogenase
0824		Quinone_oxidoreductase_family
0825		Reductive_Dechlorination
0826		Soluble_cytochromes_and_functionally_related_electron_carriers

13.25 Secondary Metabolism

si_#	category	subsystem
0827	Aromatic amino acids and	Cinnamic_Acid_Degradation
.	derivatives	
0828	Bacterial cytostatics,	2-isocapryloyl-3R-hydroxymethyl-gamma-butyrolactone_and_other_bacteri
.	differentiation factors and	
.	antibiotic	
0829		Clavulanic_acid_biosynthesis
0830		Nonribosomal_peptide_synthetases_(NRPS)_in_Frankia_sp._Ccl3
0831		Paerucumarin_Biosynthesis
0832		Phenazine_biosynthesis
0833	Biologically active compounds	Steroid_sulfates
.	in metazoan cell defence and	
.	differentiation	
0834	Biosynthesis of	Apigenin_derivatives
.	phenylpropanoids	
0835		Biflavanoid_biosynthesis
0836		Caffeic_acid_derivatives
0837		Flavanone_biosynthesis
0838		Phenylpropionate_Degradation
0839		Tannin_biosynthesis
0840	Plant Alkaloids	Alkaloid_biosynthesis_from_L-lysine
0841	Plant Hormones	Auxin_biosynthesis
0842		Auxin_degradation

13.26 Stress Response

si_#	category	subsystem
0843	Acid stress	Acid_resistance_mechanisms
0844		Glutamate_transporter_involved_in_acid_tolerance_in_Streptococcus
0845	Cold shock	Cold_shock,_CspA_family_of_proteins
0846	Dessication stress	O-antigen_capsule_important_for_environmental_persistence
0847	Detoxification	D-tyrosyl-tRNA(Tyr)_deacylase
0848		Glutathione-dependent_pathway_of_formaldehyde_detoxification
0849		Tellurite_resistance:_Chromosomal_determinants
0850		Uptake_of_selenate_and_selenite
0851	Heat shock	Heat_shock_dnaK_gene_cluster_extended
0852	Osmotic stress	Betaine_biosynthesis_from_glycine
0853		Choline_and_Betaine_Uptake_and_Betaine_Biosynthesis
0854		Ectoine_biosynthesis_and_regulation
0855		Osmoprotectant_ABC_transporter_YehZYXW_of_Enterobacteriales
0856		Osmoregulation
0857		Synthesis_of_osmoregulated_periplasmic_glucans

Continued on next page

Table 13.13 – continued from previous page

si_#	category	subsystem
0858		Glutaredoxins
0859		Glutathione:_Biosynthesis_and_gamma-glutamyl_cycle
0860		Glutathione:_Non-redox_reactions
0861		Glutathione:_Redox_cycle
0862		Glutathione_analogs:_mycothiol
0863		Glutathionylspermidine_and_Trypanothione
0864		Oxidative_stress
0865		Protection_from_Reactive_Oxygen_Species
0866		Redox-dependent_regulation_of_nucleus_processes
0867		Regulation_of_Oxidative_Stress_Response
0868		Rubrerythrin
0869	Periplasmic Stress	Periplasmic_Acid_Stress_Response_in_Enterobacteria
0870		Periplasmic_Stress_Response
0871	Unclassified	Bacterial_hemoglobins
0872		Carbon_Starvation
0873		Commensurate_regulon_activation
0874		Dimethylarginine_metabolism
0875		Flavohaemoglobin
0876		Hfl_operon
0877		Phage_shock_protein_(psp)_operon
0878		SigmaB_stress_responce_regulation
0879		Sugar-phosphate_stress_regulation
0880		Universal_stress_protein_family

13.27 Sulfur Metabolism

si_#	category	subsystem
0881	Inorganic sulfur	Inorganic_Sulfur_Assimilation
.	assimilation	
0882	Organic sulfur assimilation	Alkanesulfonate_assimilation
0883		Alkanesulfonates_Utilization
0884		DMSP_breakdown
0885		L-Cystine_Uptake_and_Metabolism
0886		Taurine_Utilization
0887		Utilization_of_glutathione_as_a_sulphur_source
0888	Unclassified	Dimethylsulfoniopropionate_(DMSP)_mineralization
0889		Galactosylceramide_and_Sulfatide_metabolism
0890		Release_of_Dimethyl_Sulfide_(DMS)_from_Dimethylsulfoniopropionate_(DM
0891		Sulfate_reduction-associated_complexes
0892		Sulfur_oxidation
0893		Thioredoxin-disulfide_reductase

13.28 Virulence, Disease, and Defense

si_#	category	subsystem
0894	Adhesion	Type_1_pili_(mannose-sensitive_fimbriae)
0895	Fimbriae of the Chaperone	α-Fimbriae
.	/Usher Assembly Pathway	
0896		α-related_Fimbriae_in_Yersinia
0897		β-Fimbriae
0898		κ-Fimbriae
0899		σ-Fimbriae
0900	Quorum sensing and	Acyl_Homoserine_Lactone_(AHL)_Autoinducer_Quorum_Sensing_
.	biofilm formation	
0901	Resistance to antibiotics	BlaR1_Family_Regulatory_Sensor-transducer_Disambiguation
.	and toxic compounds	
0902		Erythromycin_resistance
0903		Polymyxin_Synthetase_Gene_Cluster_in_Bacillus
0904	Toxins and superantigens	Diphtheria_toxin
0905		Streptococcus_agalactiae_hemolysin_operon
0906	Type III, Type IV, Type	Type_4_secretion_and_conjugative_transfer
.	VI, ESAT secretion systems	
0907	Unclassified virulence	Streptococcus_agalactiae_virulome
0908	Adhesion	Accessory_colonization_factor
0909		Adhesins_in_Staphylococcus
0910		Adhesion_of_Campylobacter
0911		Decorin_binding_proteins_of_Borrelia
0912		Mediator_of_hyperadherence_YidE_in_Enterobacteria_and_its_conse
0913		Streptococcus_pyogenes_recombinatorial_zone
0914	Bacteriocins, ribosomally	Bacitracin_Stress_Response
.	synthesized antibacterial	
.	peptides	
0915		Bacteriocin-like_peptides_Blp
0916		Marinocine,_a_broad-spectrum_antibacterial_protein
0917		Tolerance_to_colicin_E2
0918	Detection	MLST
0919	Invasion and	Cytolysin_and_Lipase_operon_in_Vibrio
.	intracellular resistance	
0920		Gram-Positive_Extracellular_Nucleases
0921		Listeria_surface_proteins:_Internalin-like_proteins
0922		Listeria_surface_proteins:_LPXTG_motif
0923		Salmonella_invasion_locus
0924	Resistance to antibiotics	Adaptation_to_d-cysteine
.	and toxic compounds	
0925		Aminoglycoside_adenylyltransferases
0926		Arsenic_resistance
0927		Beta-lactamase
0928		Cadmium_resistance
0929		Cobalt-zinc-cadmium_resistance
0930		Copper_homeostasis
0931		Fosfomycin_resistance
0932		Lysozyme_inhibitors
0933		Mercuric_reductase
0934		Mercury_resistance_operon
0935		Methicillin_resistance_in_Staphylococci

Continue

Table 13.14 – continued from previous page

si_#	category	subsystem
0936		MexA-MexB-OprM_Multidrug_Efflux_System
0937		MexC-MexD-OprJ_Multidrug_Efflux_System
0938		MexE-MexF-OprN_Multidrug_Efflux_System
0939		Multidrug_Resistance_Efflux_Pumps
0940		Multidrug_Resistance_Operon_mdtRP_of_Bacillus
0941		Multidrug_efflux_pump_in_Campylobacter_jejuni_(CmeABC_operon)
0942		Multiple_Antibiotic_Resistance_MAR_locus
0943		Resistance_to_Vancomycin
0944		Resistance_to_chromium_compounds
0945		Resistance_to_fluoroquinolones
0946		Streptococcus_pneumoniae_Vancomycin_Tolerance_Locus
0947		Streptothricin_resistance
0948		Teicoplanin-resistance_in_Staphylococcus
0949		The_mdtABCD_multidrug_resistance_cluster
0950		Zinc_resistance
0951	Toxins and superantigens	Cholera_toxin
0952		Cytolethal_distending_toxin_of_Campylobacter_jejuni
0953		Cytolethal_distending_toxins
0954		Pertussis_toxin
0955		Prophage-encoded_Exotoxins
0956		SLO-NADGH_Locus
0957		Staphylococcus_Two-component_and_Pore-forming_Cytolysins
0958		Streptolysin_S_Biosynthesis_and_Transport
0959	Unclassified	Bacterial_cyanide_production_and_tolerance_mechanisms
0960		C_jejuni_colonization_of_chick_caeca
0961		Streptococcus_pyogenes_Virulome

13.29 Ribosome

si_#	category	subsystem
0962	Ribosome	Ribosome

13.30 0963 No Function Match

CHAPTER 14

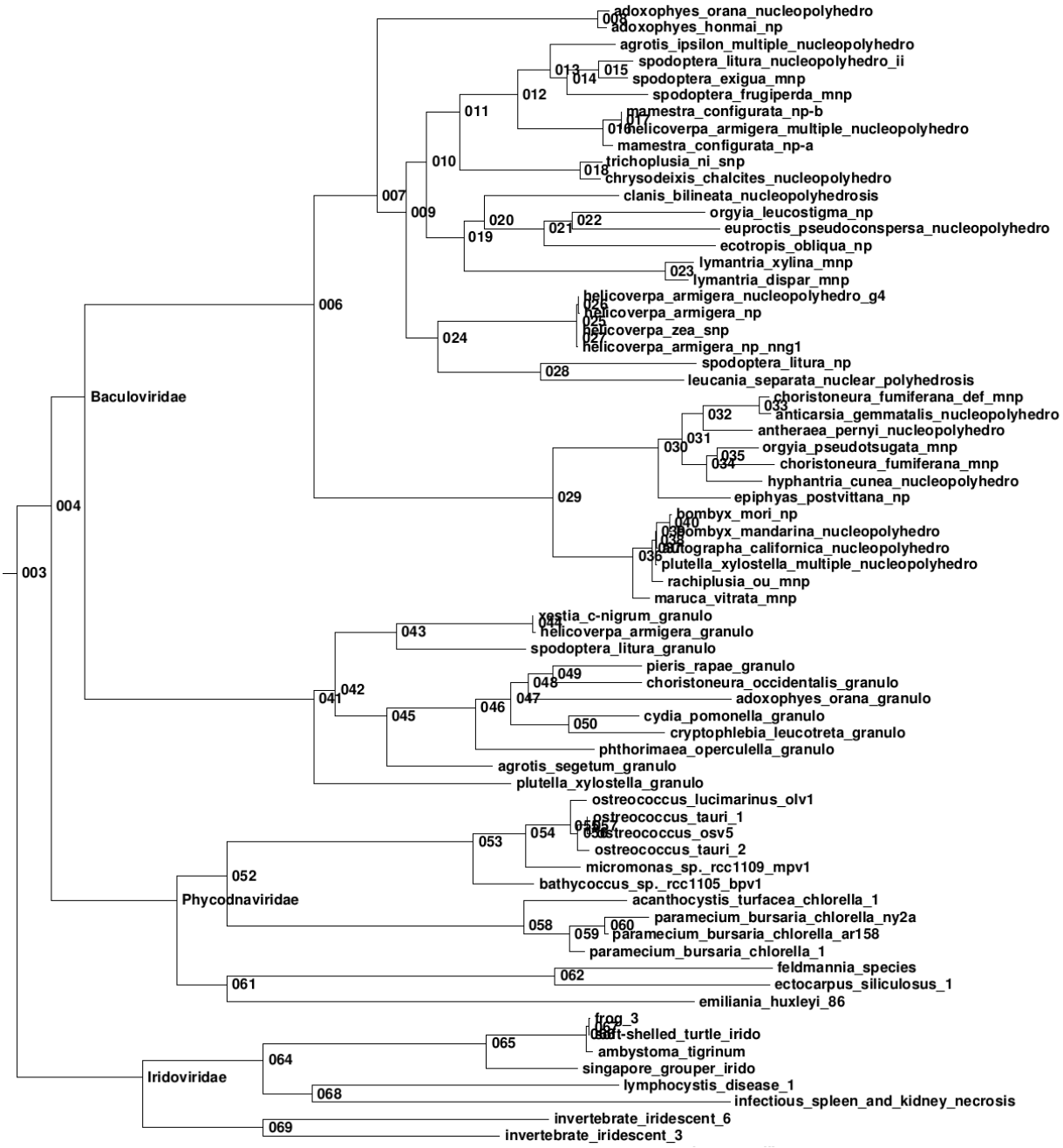
Viral tree, 1252 taxa

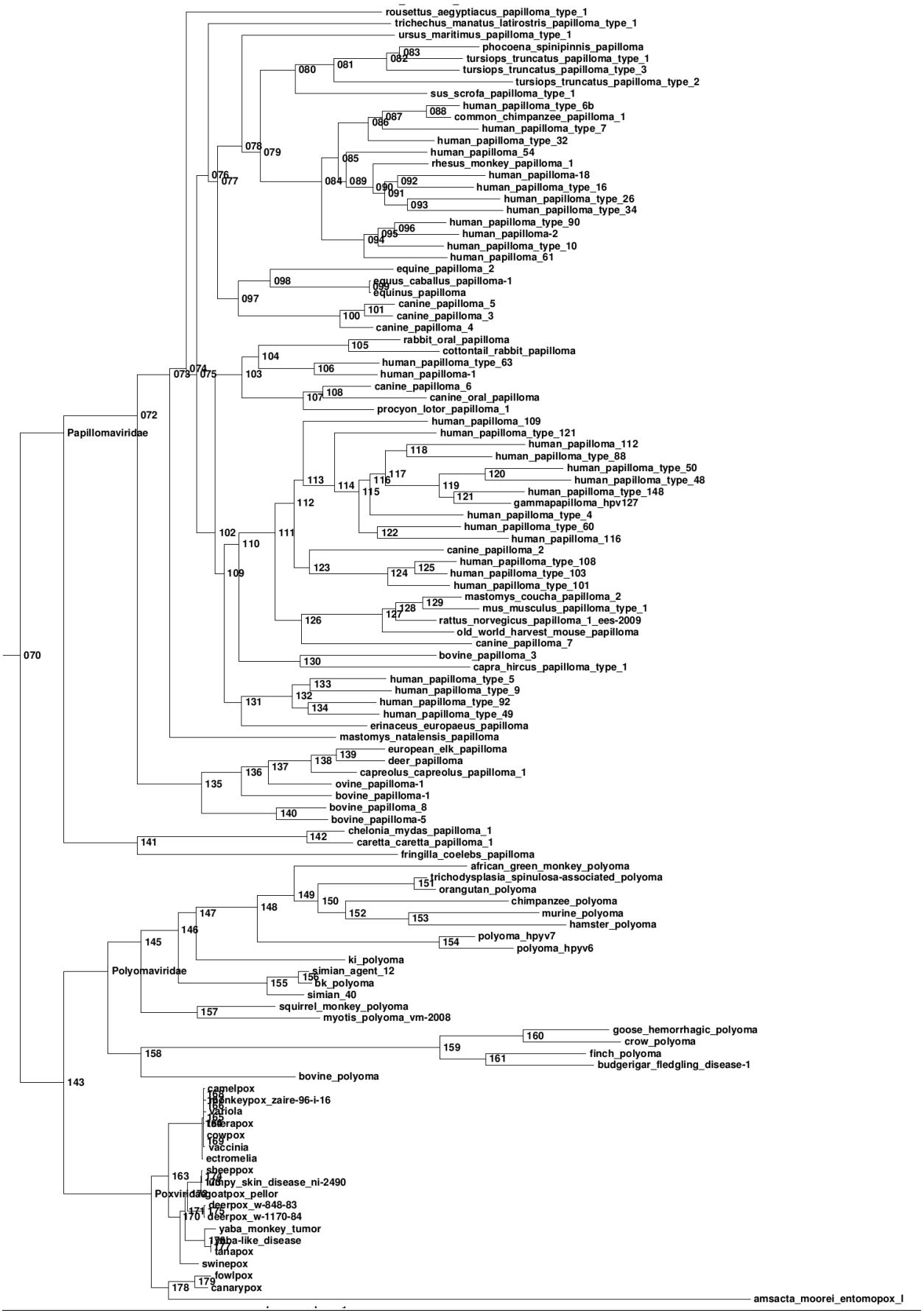
One-per-species tree of viruses, made from the completed genomes available in RefSeq.

In order to provide node definitions for use with the virus1252 data module, we provide the reference phylogeny with numbers, divided into the sections: *dsDNA viruses 1: Baculoviridae, Phycodnaviridae, and Irdoviridae*, *dsDNA viruses 2: Papillomaviridae and Polyomaviridae, and Poxviridae*, *dsDNA viruses 3: Adenoviridae and Herpesviridae*, *Reverse transcriptase viruses: Caulimoviridae, Retroviridae, and Hepadnaviridae*, *ssRNA+ 1: Caliciviridae, Nidovirales*, *ssRNA+ 2: Flaviviridae*, *ssRNA+ 3: Tombusviridae and Virgaviridae*, *ssRNA+ 4: Tymoviridae*, *ssRNA+ 5: Picornaviridae, Togaviridae*, *ssRNA+ 6: Potyviridae*, *ssRNA, segmented: Arenaviridae, Bunyaviridae*, *Mononegavirales*, *Orthomyxoviridae*, *ssDNA 1: Parvoviridae*, *ssDNA 2: Geminiviridae_1*, *ssDNA 2: Geminiviridae_2*, *dsDNA: Reoviridae*

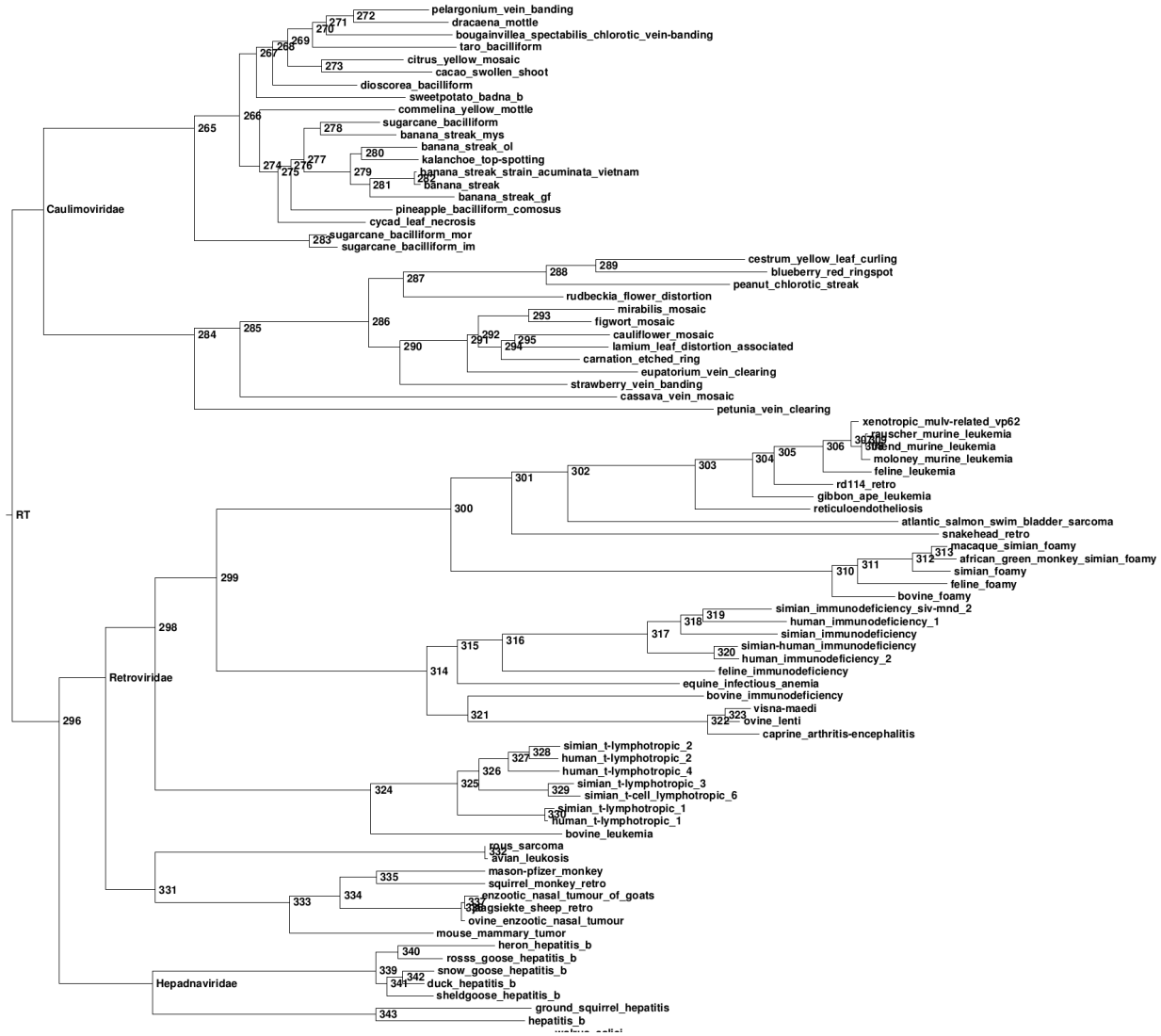
Tree files can be downloaded in * `nexus` * `phyloxml` * `pdf` formats.

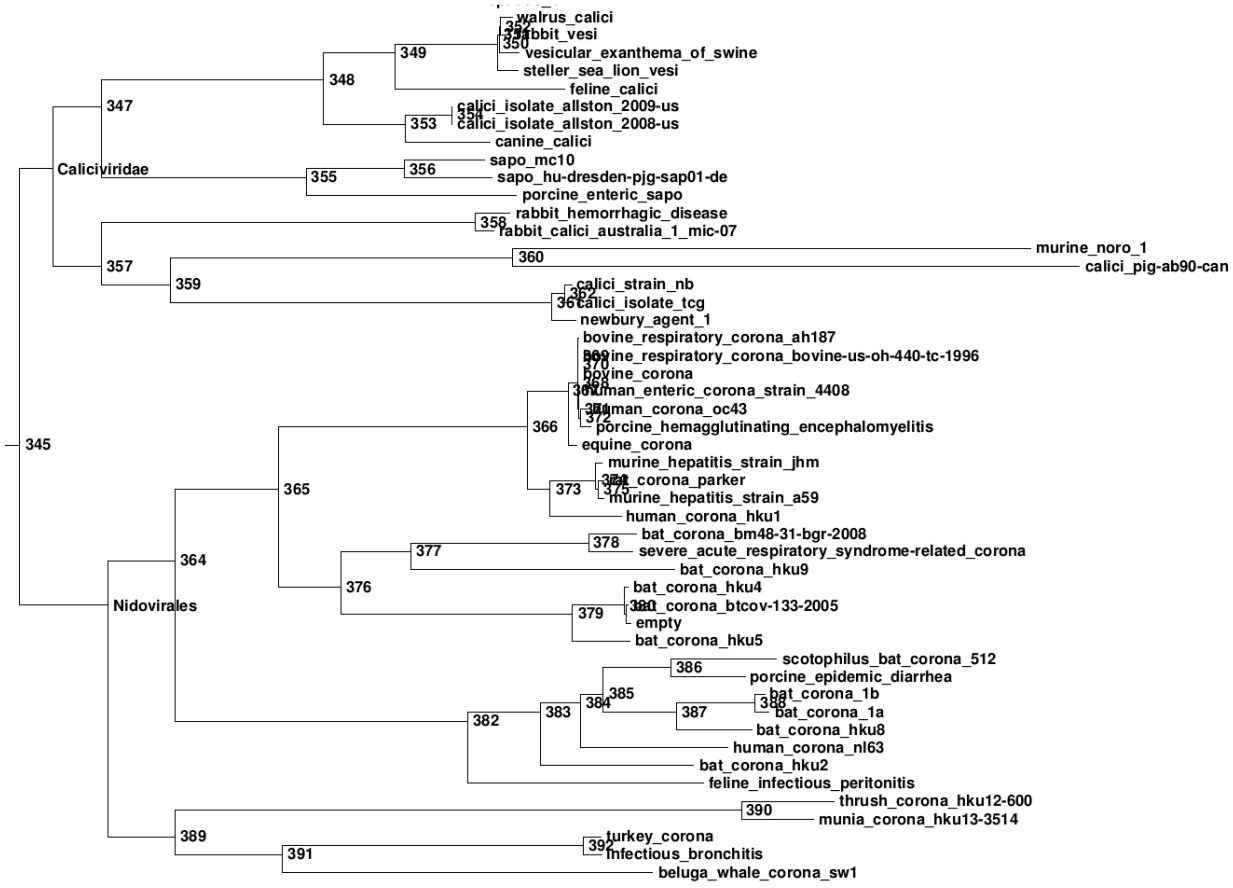
The viral sequence data module virus1252.1 is annotated with the functional classifications present in Pfam release 27.

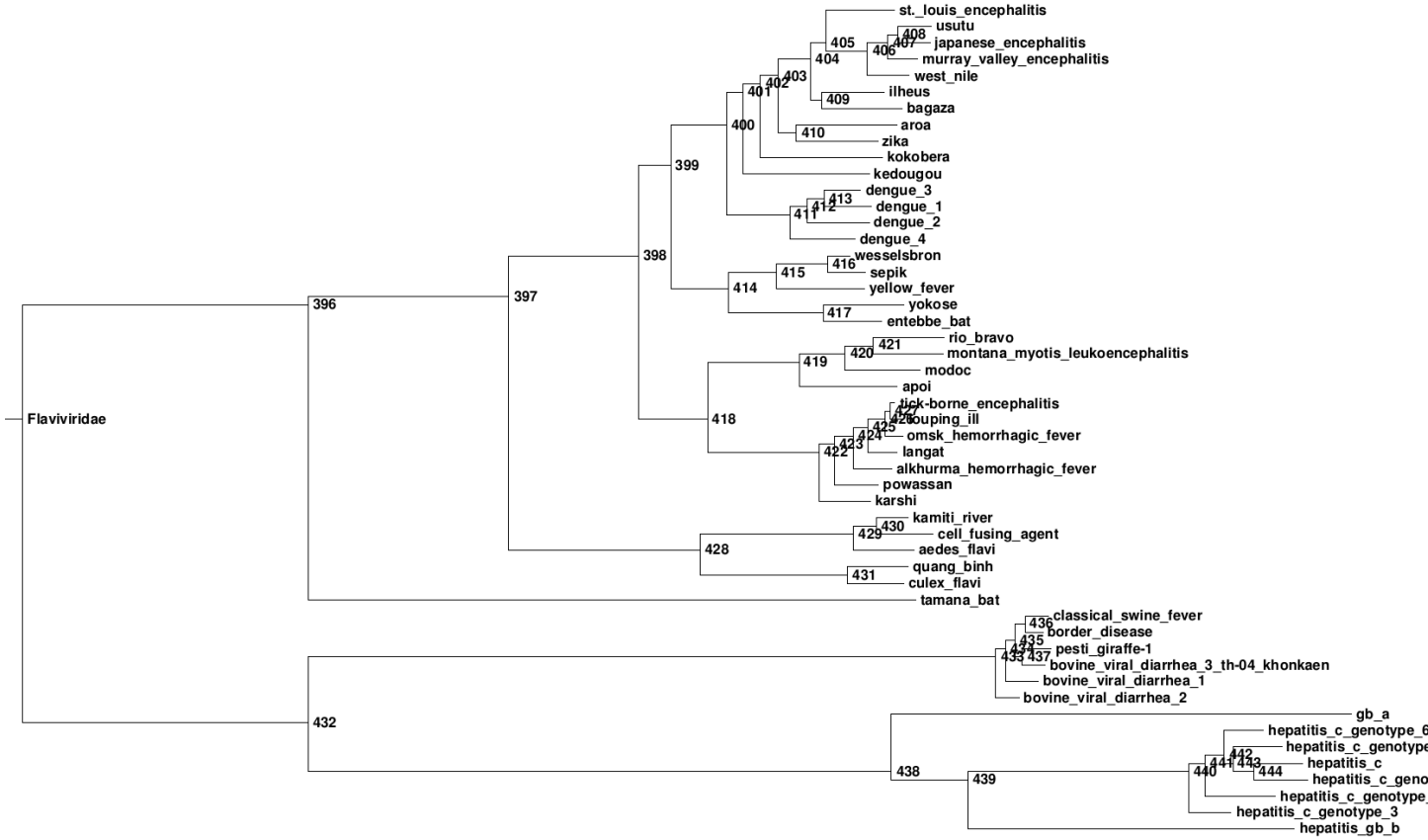

adoxophyes_orana_nucleopolyhedro
adoxophyes_honmai_np
agrotis_ipsilon_multiple_nucleopolyhedro
spodoptera_litura_nucleopolyhedro_ii
spodoptera_exigua_mnp
spodoptera_frugiperda_mnp
mamestra_configurata_np-b
helicoverpa_armigera_multiple_nucleopolyhedro
mamestra_configurata_np-a
trichoplusia_ni_snp
chrysodeixis_chalcites_nucleopolyhedro
clanis_bilineata_nucleopolyhedrosis
orgyia_leucostigma_np
euproctis_pseudoconspersa_nucleopolyhedro
ecotropis_obliqua_np
lymantria_xylina_mnp
lymantria_dispar_mnp
helicoverpa_armigera_nucleopolyhedro_g4
helicoverpa_armigera_np
helicoverpa_zea_snp
helicoverpa_armigera_np_nng1
spodoptera_litura_np
leucania_separata_nuclear_polyhedrosis
Baculoviridae
choristoneura_fumiferana_def_mnp
anticarsia_gemmatalis_nucleopolyhedro
antheraea_pernyi_nucleopolyhedro
orgyia_pseudotsugata_mnp
choristoneura_fumiferana_mnp
hyphantria_cunea_nucleopolyhedro
epiphyas_postvittana_np
bombyx_mori_np
bombyx_mandarina_nucleopolyhedro
autographa_californica_nucleopolyhedro
plutella_xylostella_multiple_nucleopolyhedro
rachiplusia_ou_mnp
maruca_vitrata_mnp
xestia_c-nigrum_granulo
helicoverpa_armigera_granulo
spodoptera_litura_granulo
pieris_rapae_granulo
choristoneura_occidentalis_granulo
adoxophyes_orana_granulo
cydia_pomonella_granulo
cryptophlebia_leucotreta_granulo
phthorimaea_operculella_granulo
agrotis_segetum_granulo
plutella_xylostella_granulo
ostreococcus_lucimarinus_olv1
ostreococcus_tauri_1
ostreococcus_osv5
ostreococcus_tauri_2
micromonas_sp._rcc1109_mpv1
bathycoccus_sp._rcc1105_bpv1
Phycodnaviridae
acanthocystis_turfacea_chlorella_1
paramecium_bursaria_chlorella_ny2a
paramecium_bursaria_chlorella_ar158
paramecium_bursaria_chlorella_1
feldmannia_species
ectocarpus_siliculosus_1
emiliania_huxleyi_86
frog_3
soft-shelled_turtle_irido
ambystoma_tigrinum
singapore_grouper_irido
Iridoviridae
lymphocystis_disease_1
infectious_spleen_and_kidney_necrosis
invertebrate_iridescent_6
invertebrate_iridescent_3

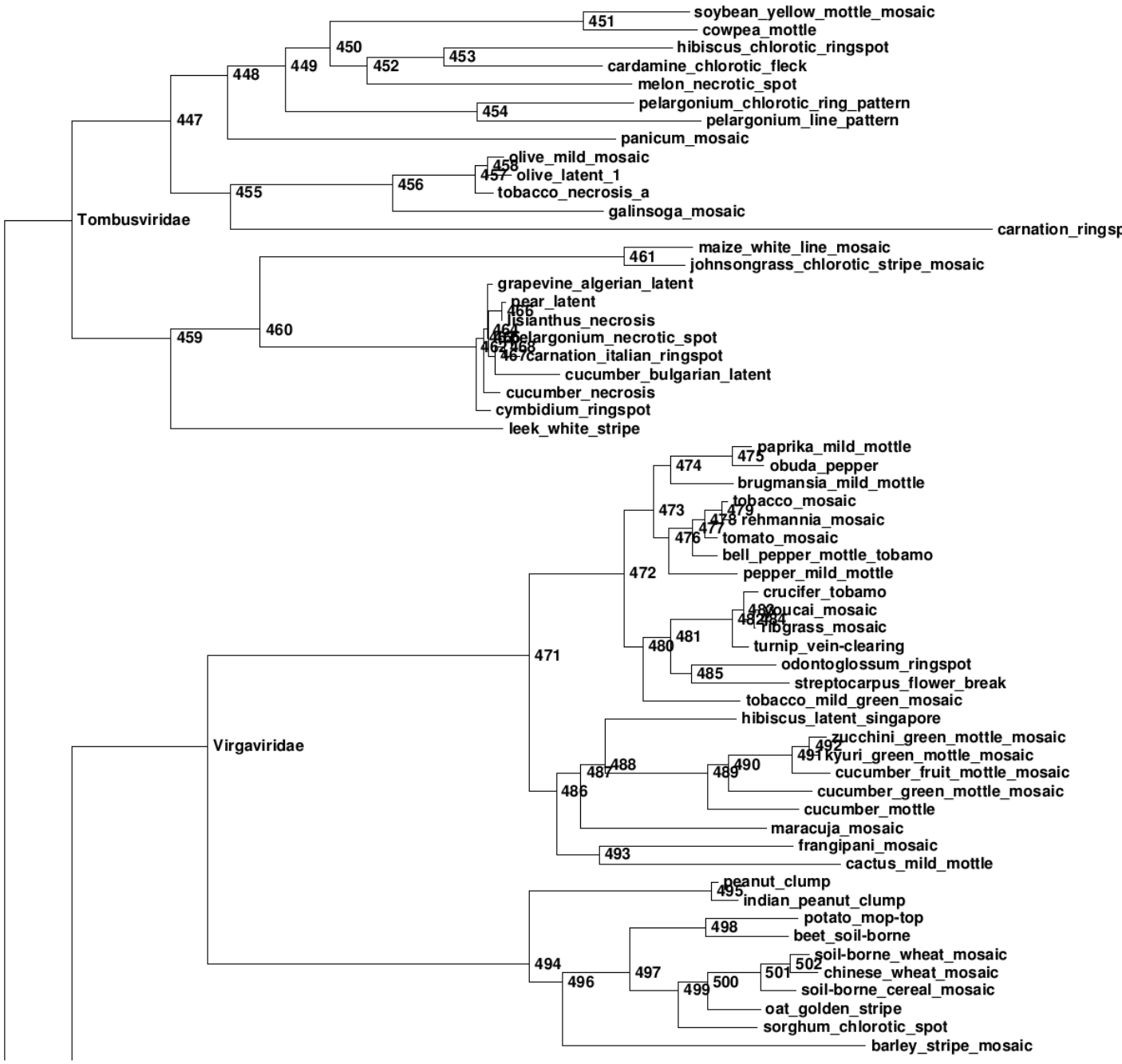

rousettus_aegyptiacus_papilloma_type_1
trichechus_manatus_latirostris_papilloma_type_1
ursus_maritimus_papilloma_type_1
phocoena_spinipinnis_papilloma
tursiops_truncatus_papilloma_type_1
tursiops_truncatus_papilloma_type_3
tursiops_truncatus_papilloma_type_2
sus_scrofa_papilloma_type_1
human_papilloma_type_6b
common_chimpanzee_papilloma_1
human_papilloma_type_7
human_papilloma_type_32
human_papilloma_54
rhesus_monkey_papilloma_1
human_papilloma-18
human_papilloma_type_16
human_papilloma_type_26
human_papilloma_type_34
human_papilloma_type_90
human_papilloma-2
human_papilloma_type_10
human_papilloma_61
equine_papilloma_2
equus_caballus_papilloma-1
equinus_papilloma
canine_papilloma_5
canine_papilloma_3
canine_papilloma_4
rabbit_oral_papilloma
cottontail_rabbit_papilloma
human_papilloma_type_63
human_papilloma-1
canine_papilloma_6
canine_oral_papilloma
procyon_lotor_papilloma_1
Papillomaviridae
human_papilloma_109
human_papilloma_type_121
human_papilloma_112
human_papilloma_type_88
human_papilloma_type_50
human_papilloma_type_48
human_papilloma_type_148
gammapapilloma_hpv127
human_papilloma_type_4
human_papilloma_type_60
human_papilloma_116
canine_papilloma_2
human_papilloma_type_108
human_papilloma_type_103
human_papilloma_type_101
mastomys_coucha_papilloma_2
mus_musculus_papilloma_type_1
rattus_norvegicus_papilloma_1_ees-2009
old_world_harvest_mouse_papilloma
canine_papilloma_7
bovine_papilloma_3
capra_hircus_papilloma_type_1
human_papilloma_type_5
human_papilloma_type_9
human_papilloma_type_92
human_papilloma_type_49
erinaceus_europaeus_papilloma
mastomys_natalensis_papilloma
european_elk_papilloma
deer_papilloma
capreolus_capreolus_papilloma_1
ovine_papilloma-1
bovine_papilloma-1
bovine_papilloma_8
bovine_papilloma-5
chelonia_mydas_papilloma_1
caretta_caretta_papilloma_1
fringilla_coelebs_papilloma
african_green_monkey_polyoma
trichodysplasia_spinulosa-associated_polyoma
orangutan_polyoma
chimpanzee_polyoma
murine_polyoma
hamster_polyoma
polyoma_hpyv7
polyoma_hpyv6
ki_polyoma
simian_agent_12
bk_polyoma
simian_40
squirrel_monkey_polyoma
myotis_polyoma_vm-2008
Polyomaviridae
goose_hemorrhagic_polyoma
crow_polyoma
finch_polyoma
budgerigar_fledgling_disease-1
bovine_polyoma
camelpox
monkeypox_zaire-96-i-16
variola
taterapox
cowpox
vaccinia
ectromelia
sheeppox
lumpy_skin_disease_ni-2490
goatpox_pellor
deerpox_w-848-83
deerpox_w-1170-84
yaba_monkey_tumor
yaba-like_disease
tanapox
swinepox
fowlpox
canarypox
Poxviridae
amsacta_moorei_entomopox_l

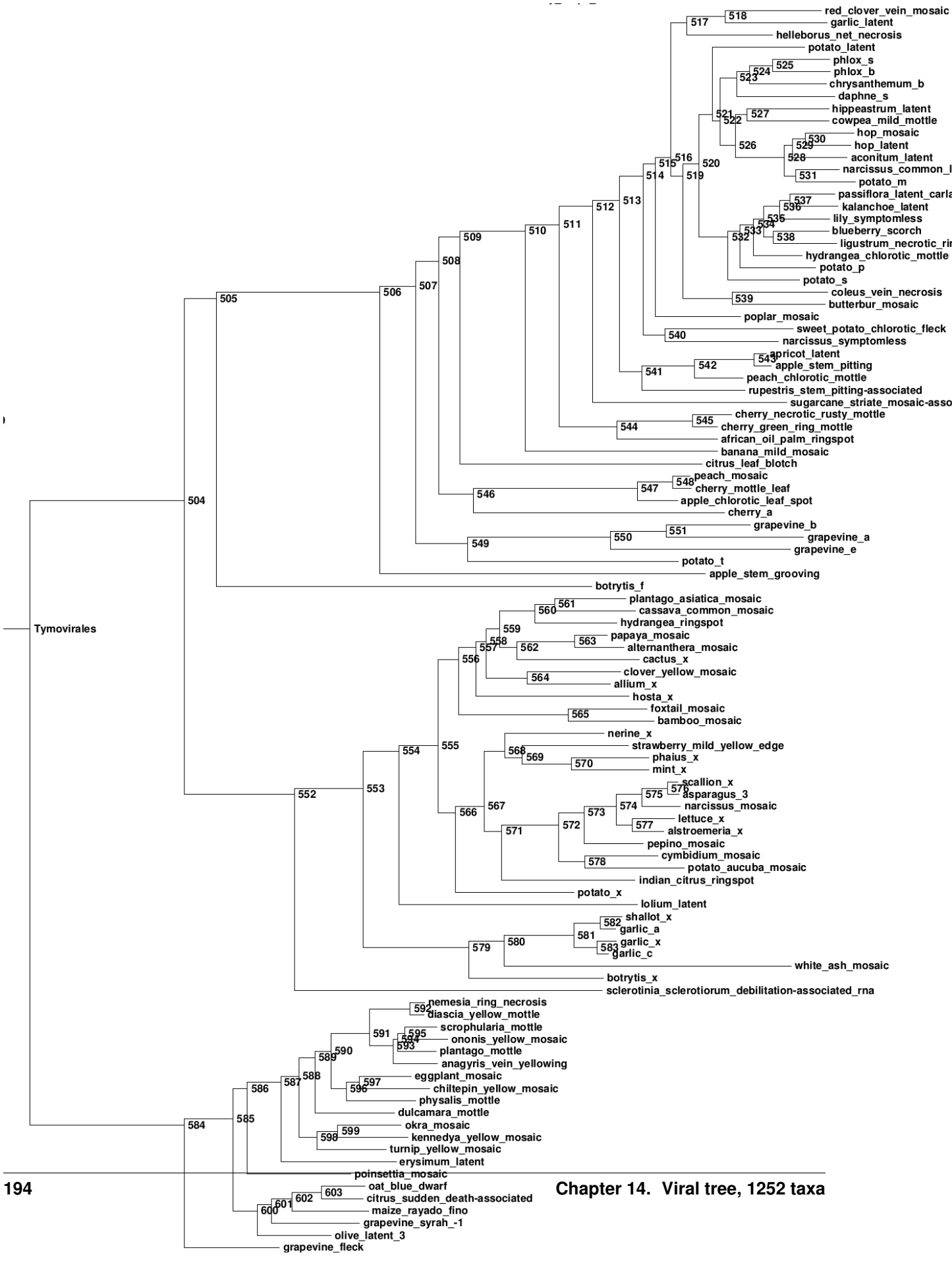

human_adeno_1
human_adeno_c
human_adeno_2
human_adeno_5
simian_adeno_21
human_adeno_b
human_adeno_35
human_adeno_7
simian_adeno_25
human_adeno_e
human_adeno_d
human_adeno_54
human_adeno_a
simian_adeno_1
human_adeno_f
simian_adeno_3
tree_shrew_adeno
ovine_adeno_a
bovine_adeno_a
porcine_adeno_c
canine_adeno
canine_adeno_1
canine_adeno_2
porcine_adeno_a
bovine_adeno_b
murine_adeno_a
murine_adeno_3
Adenoviridae
fowl_adeno_d
fowl_adeno_a
turkey_adeno_1
turkey_adeno_a
frog_adeno_1
bovine_adeno_d
duck_adeno_a
snake_adeno
gallid_herpes_3
gallid_herpes_2
meleagrid_herpes_1
anatid_herpes_1
bovine_herpes_5
bovine_herpes_1
suid_herpes_1
equid_herpes_9
equid_herpes_1
equid_herpes_4
felid_herpes_1
human_herpes_3
cercopithecine_herpes_9
human_herpes_2
human_herpes_1
papiine_herpes_2
cercopithecine_herpes_2
macacine_herpes_1
saimiriine_herpes_1
psittacid_herpes_1
gallid_herpes_1
anguillid_herpes_1
ranid_herpes_1
ranid_herpes_2
ictalurid_herpes_1
murid_herpes_4
bovine_herpes_4
macacine_herpes_5
human_herpes_8
saimiriine_herpes_2
ateline_herpes_3
equid_herpes_2
ovine_herpes_2
alcelaphine_herpes_1
human_herpes_4_type_2
human_herpes_4
macacine_herpes_4
callitrichine_herpes_3
panine_herpes_2
human_herpes_5
macacine_herpes_3
tupaiid_herpes_1
murid_herpes_2
caviid_herpes_2
human_herpes_7
human_herpes_6
Herpesviridae
180
182
183
184
185
186
187
188
189
190
194
195
201
202
203
204
205
206
207
208
209
210
211
212
213
214
216
217
218
219
220
221
222
223
224
225
226
228
229
231
232
233
234
237
238
239
240
241
242
243
244
245
246
247
248
249
250
253
254
255
256
257
258
259

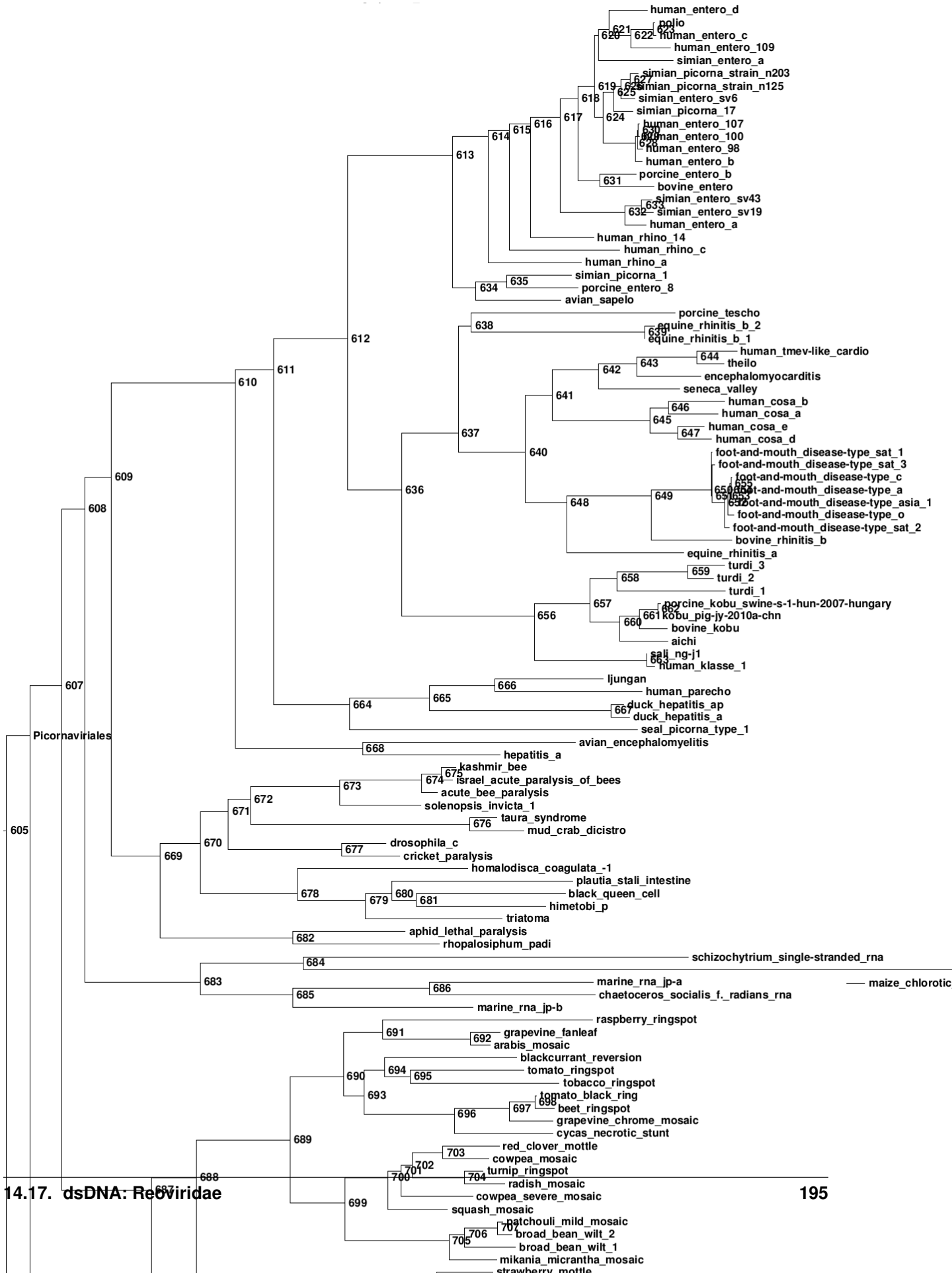

RT
Caulimoviridae
Retroviridae
Hepadnaviridae
pelargonium_vein_banding
dracaena_mottle
bougainvillea_spectabilis_chlorotic_vein-banding
taro_bacilliform
citrus_yellow_mosaic
cacao_swollen_shoot
dioscorea_bacilliform
sweetpotato_badna_b
commelina_yellow_mottle
sugarcane_bacilliform
banana_streak_mys
banana_streak_ol
kalanchoe_top-spotting
banana_streak_strain_acuminata_vietnam
banana_streak
banana_streak_gf
pineapple_bacilliform_comosus
cycad_leaf_necrosis
sugarcane_bacilliform_mor
sugarcane_bacilliform_im
cestrum_yellow_leaf_curling
blueberry_red_ringspot
peanut_chlorotic_streak
rudbeckia_flower_distortion
mirabilis_mosaic
figwort_mosaic
cauliflower_mosaic
lamium_leaf_distortion_associated
carnation_etched_ring
eupatorium_vein_clearing
strawberry_vein_banding
cassava_vein_mosaic
petunia_vein_clearing
xenotropic_mulv-related_vp62
moloney_murine_leukemia
feline_leukemia
rd114_retro
gibbon_ape_leukemia
reticuloendotheliosis
atlantic_salmon_swim_bladder_sarcoma
snakehead_retro
macaque_simian_foamy
african_green_monkey_simian_foamy
simian_foamy
feline_foamy
bovine_foamy
simian_immunodeficiency_siv-mnd_2
human_immunodeficiency_1
simian_immunodeficiency
simian-human_immunodeficiency
human_immunodeficiency_2
feline_immunodeficiency
equine_infectious_anemia
bovine_immunodeficiency
visna-maedi
ovine_lenti
caprine_arthritis-encephalitis
simian_t-lymphotropic_2
human_t-lymphotropic_2
human_t-lymphotropic_4
simian_t-lymphotropic_3
simian_t-cell_lymphotropic_6
simian_t-lymphotropic_1
human_t-lymphotropic_1
bovine_leukemia
rous_sarcoma
avian_leukosis
mason-pfizer_monkey
squirrel_monkey_retro
enzootic_nasal_tumour_of_goats
jaagsiekte_sheep_retro
ovine_enzootic_nasal_tumour
mouse_mammary_tumor
heron_hepatitis_b
rosss_goose_hepatitis_b
snow_goose_hepatitis_b
duck_hepatitis_b
sheldgoose_hepatitis_b
ground_squirrel_hepatitis
hepatitis_b

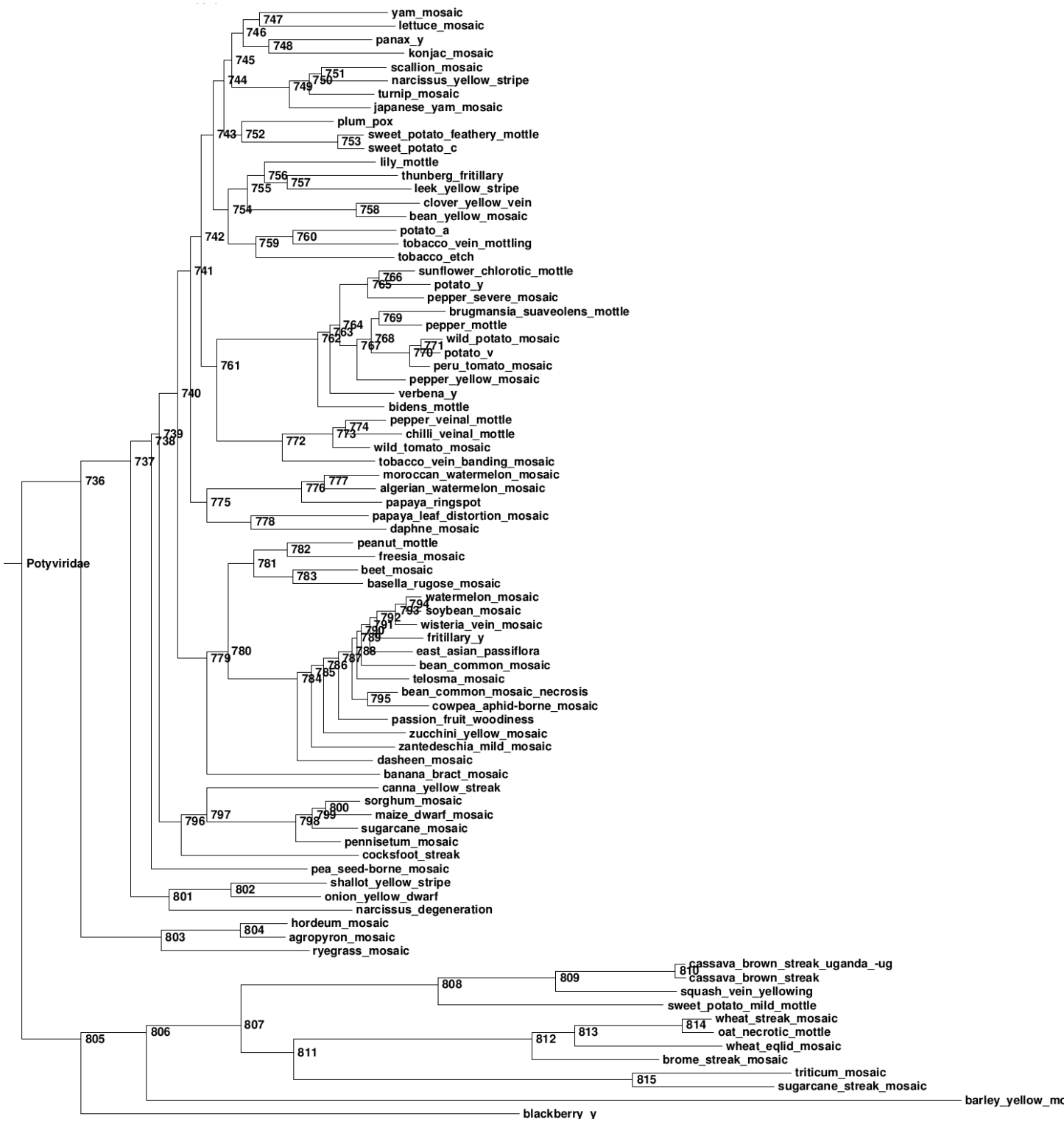

walrus_calici
rabbit_vesi
vesicular_exanthema_of_swine
steller_sea_lion_vesi
feline_calici
calici_isolate_allston_2009-us
calici_isolate_allston_2008-us
canine_calici
sapo_mc10
sapo_hu-dresden-pjg-sap01-de
porcine_enteric_sapo
rabbit_hemorrhagic_disease
rabbit_calici_australia_1_mic-07
murine_noro_1
calici_pig-ab90-can
calici_strain_nb
calici_isolate_tcg
newbury_agent_1
bovine_respiratory_corona_ah187
bovine_respiratory_corona_bovine-us-oh-440-tc-1996
bovine_corona
human_enteric_corona_strain_4408
human_corona_oc43
porcine_hemagglutinating_encephalomyelitis
equine_corona
murine_hepatitis_strain_jhm
murine_hepatitis_strain_a59
human_corona_hku1
bat_corona_bm48-31-bgr-2008
severe_acute_respiratory_syndrome-related_corona
bat_corona_hku9
bat_corona_hku4
bat_corona_btcov-133-2005
empty
bat_corona_hku5
scotophilus_bat_corona_512
porcine_epidemic_diarrhea
bat_corona_1b
bat_corona_1a
bat_corona_hku8
human_corona_nl63
bat_corona_hku2
feline_infectious_peritonitis
thrush_corona_hku12-600
munia_corona_hku13-3514
turkey_corona
infectious_bronchitis
beluga_whale_corona_sw1
Coronaviridae
Nidovirales
345
347
348
349
350
353
354
355
356
357
358
359
360
362
364
365
366
370
372
373
375
376
377
378
379
382
383
384
385
386
387
388
389
390
391
392

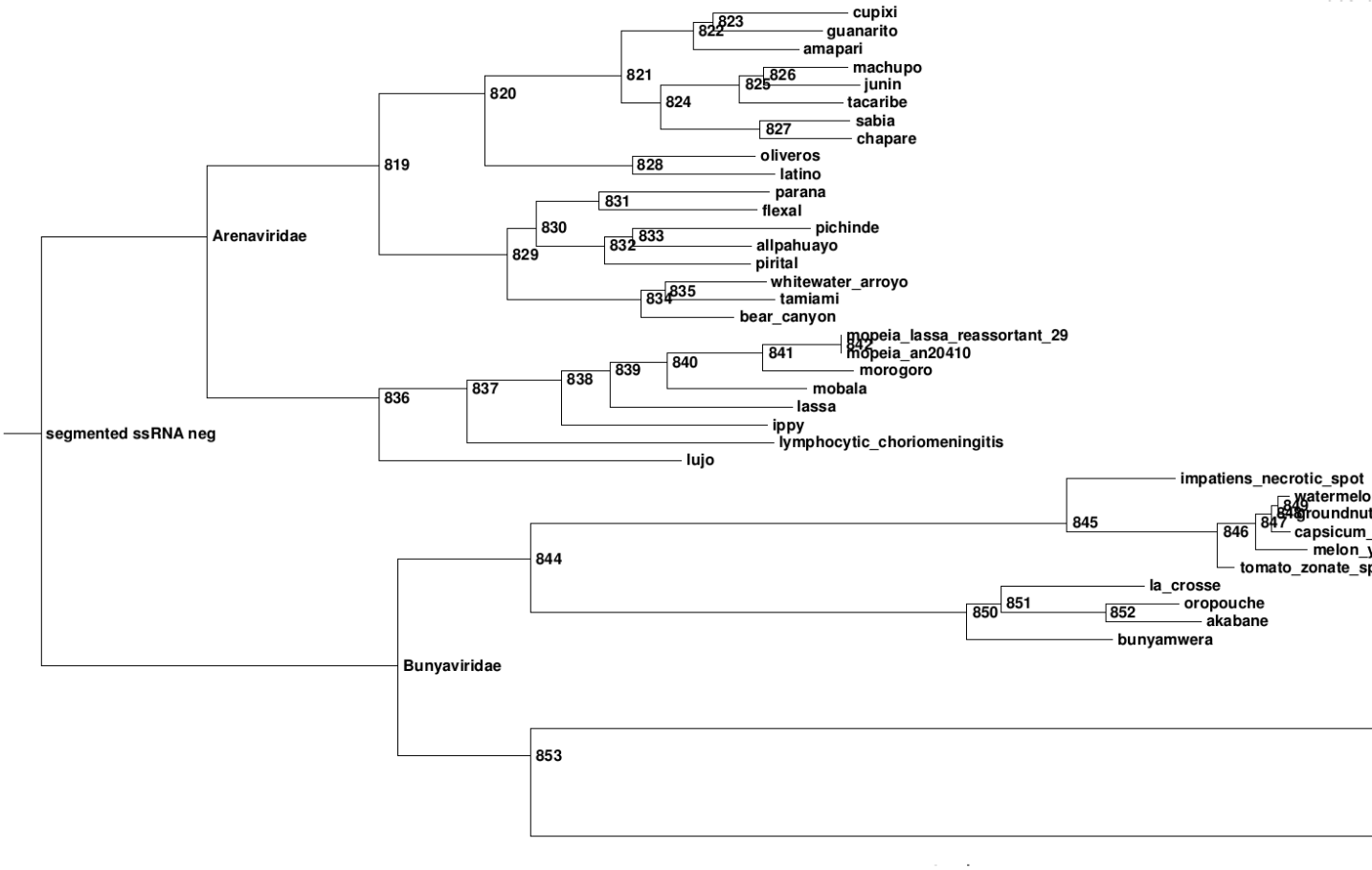

st._louis_encephalitis
usutu
japanese_encephalitis
murray_valley_encephalitis
west_nile
ilheus
bagaza
aroa
zika
kokobera
kedougou
dengue_3
dengue_1
dengue_2
dengue_4
wesselsbron
sepik
yellow_fever
yokose
entebbe_bat
rio_bravo
montana_myotis_leukoencephalitis
modoc
apoi
tick-borne_encephalitis
louping_ill
omsk_hemorrhagic_fever
langat
alkhurma_hemorrhagic_fever
powassan
karshi
kamiti_river
cell_fusing_agent
aedes_flavi
quang_binh
culex_flavi
tamana_bat
classical_swine_fever
border_disease
pesti_giraffe-1
bovine_viral_diarrhea_3_th-04_khonkaen
bovine_viral_diarrhea_1
bovine_viral_diarrhea_2
gb_a
hepatitis_c_genotype_6
hepatitis_c_genotype
hepatitis_c
hepatitis_c_genotype_3
hepatitis_gb_b
Flaviviridae
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444

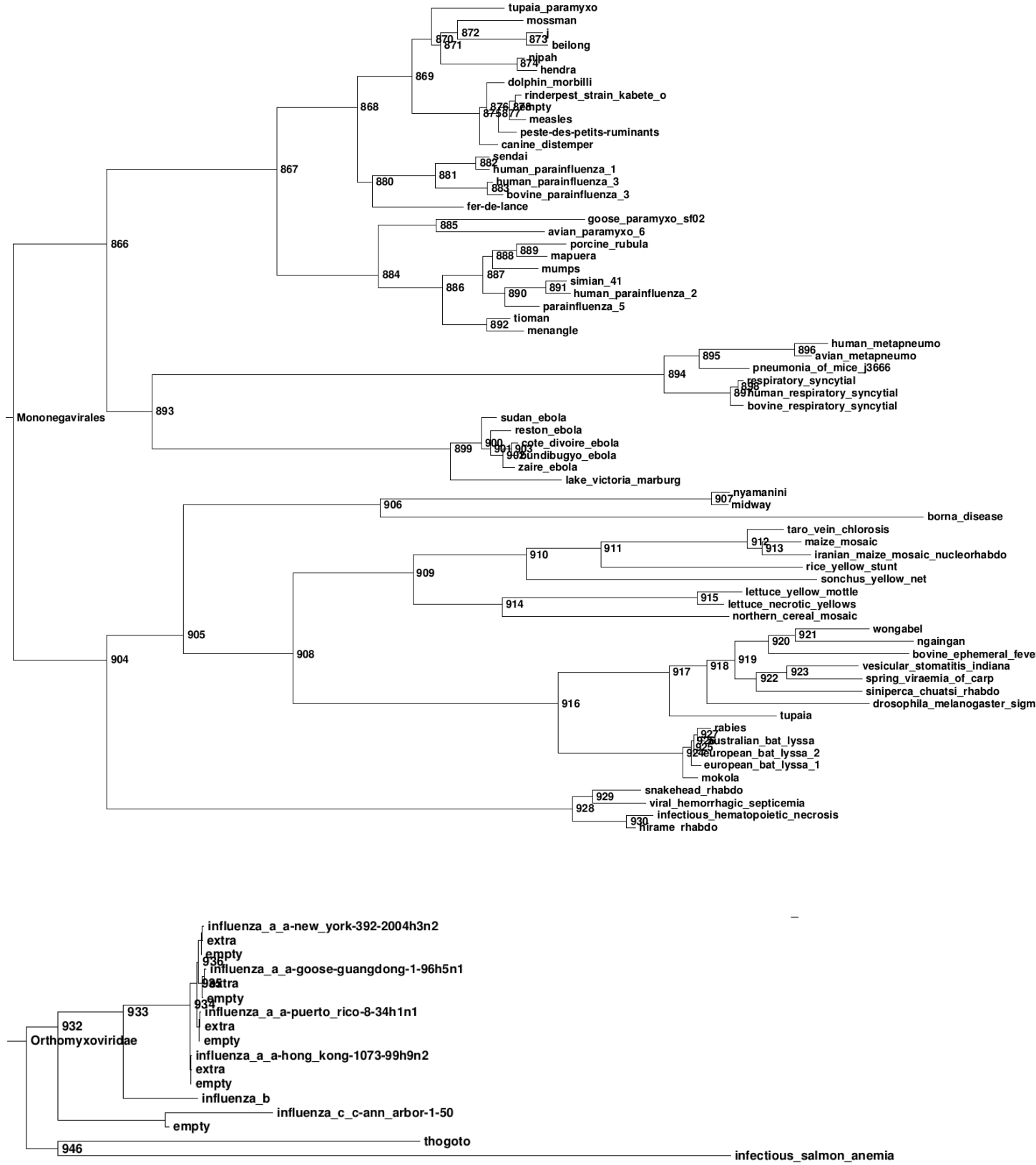

soybean_yellow_mottle_mosaic
451
cowpea_mottle
450
hibiscus_chlorotic_ringspot
453
452
cardamine_chlorotic_fleck
449
448
melon_necrotic_spot
pelargonium_chlorotic_ring_pattern
454
pelargonium_line_pattern
447
panicum_mosaic
olive_mild_mosaic
458
457
olive_latent_1
456
tobacco_necrosis_a
455
galinsoga_mosaic
Tombusviridae
carnation_ringsp
maize_white_line_mosaic
461
johnsongrass_chlorotic_stripe_mosaic
grapevine_algerian_latent
pear_latent
466
lisianthus_necrosis
460
459
pelargonium_necrotic_spot
carnation_italian_ringspot
cucumber_bulgarian_latent
cucumber_necrosis
cymbidium_ringspot
leek_white_stripe
paprika_mild_mottle
475
474
obuda_pepper
brugmansia_mild_mottle
tobacco_mosaic
473
479
478
rehmannia_mosaic
477
476
tomato_mosaic
bell_pepper_mottle_tobamo
472
pepper_mild_mottle
crucifer_tobamo
ribgrass_mosaic
481
480
turnip_vein-clearing
odontoglossum_ringspot
485
streptocarpus_flower_break
tobacco_mild_green_mosaic
471
hibiscus_latent_singapore
zucchini_green_mottle_mosaic
492
491
kyuri_green_mottle_mosaic
Virgaviridae
488
487
490
489
cucumber_fruit_mottle_mosaic
cucumber_green_mottle_mosaic
486
cucumber_mottle
maracuja_mosaic
frangipani_mosaic
493
cactus_mild_mottle
peanut_clump
495
indian_peanut_clump
potato_mop-top
498
beet_soil-borne
soil-borne_wheat_mosaic
494
497
502
501
chinese_wheat_mosaic
496
500
soil-borne_cereal_mosaic
499
oat_golden_stripe
sorghum_chlorotic_spot
barley_stripe_mosaic

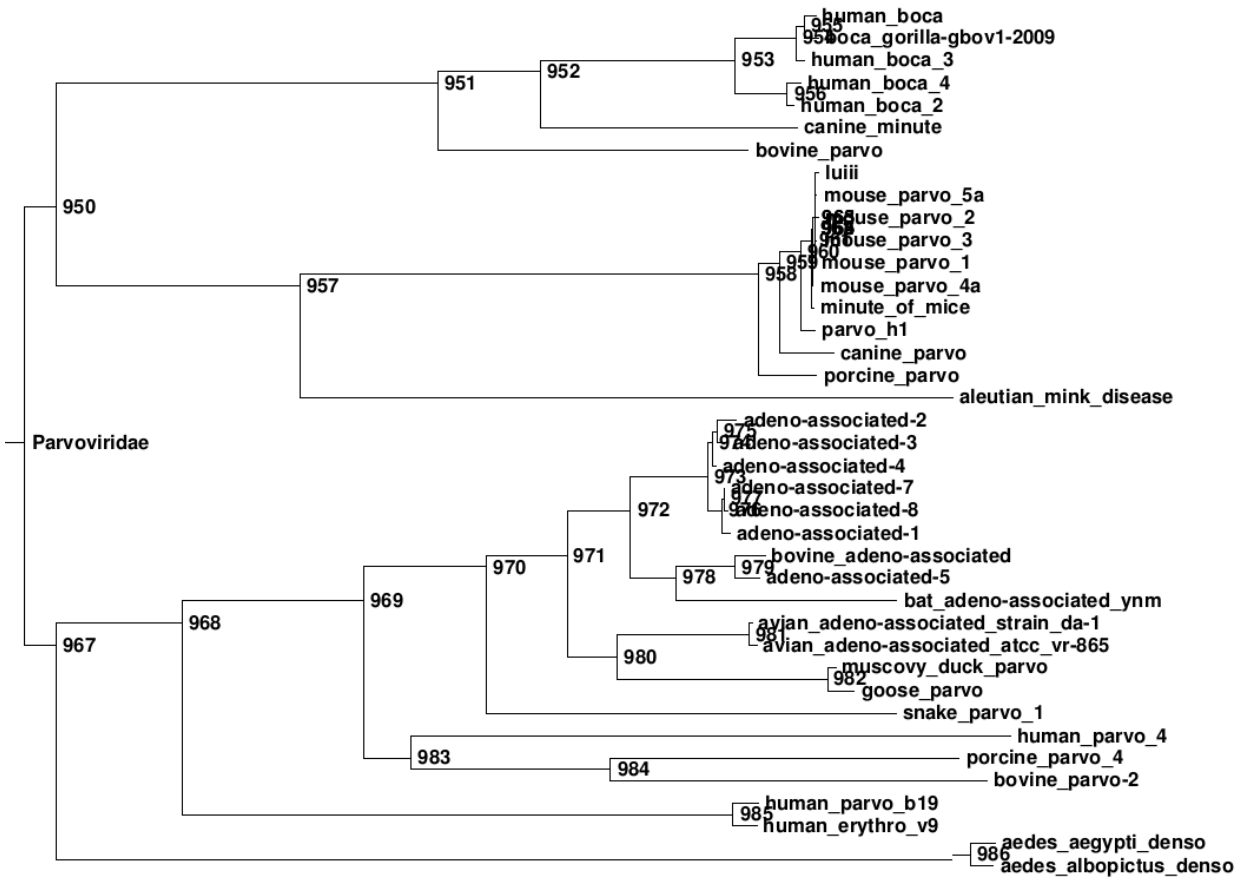

red_clover_vein_mosaic
garlic_latent
helleborus_net_necrosis
potato_latent
phlox_s
phlox_b
chrysanthemum_b
daphne_s
hippeastrum_latent
cowpea_mild_mottle
hop_mosaic
hop_latent
aconitum_latent
narcissus_common_l
potato_m
passiflora_latent_carla
kalanchoe_latent
lily_symptomless
blueberry_scorch
ligustrum_necrotic_ri
hydrangea_chlorotic_mottle
potato_p
potato_s
coleus_vein_necrosis
butterbur_mosaic
poplar_mosaic
sweet_potato_chlorotic_fleck
narcissus_symptomless
apricot_latent
apple_stem_pitting
peach_chlorotic_mottle
rupestris_stem_pitting-associated
sugarcane_striate_mosaic-asso
cherry_necrotic_rusty_mottle
cherry_green_ring_mottle
african_oil_palm_ringspot
banana_mild_mosaic
citrus_leaf_blotch
peach_mosaic
cherry_mottle_leaf
apple_chlorotic_leaf_spot
cherry_a
grapevine_b
grapevine_a
grapevine_e
potato_t
apple_stem_grooving
botrytis_f
plantago_asiatica_mosaic
cassava_common_mosaic
hydrangea_ringspot
papaya_mosaic
alternanthera_mosaic
cactus_x
clover_yellow_mosaic
allium_x
hosta_x
foxtail_mosaic
bamboo_mosaic
nerine_x
strawberry_mild_yellow_edge
phaius_x
mint_x
scallion_x
asparagus_3
narcissus_mosaic
lettuce_x
alstroemeria_x
pepino_mosaic
cymbidium_mosaic
potato_aucuba_mosaic
indian_citrus_ringspot
potato_x
lolium_latent
shallot_x
garlic_a
garlic_x
garlic_c
white_ash_mosaic
botrytis_x
sclerotinia_sclerotiorum_debilitation-associated_rna
nemesia_ring_necrosis
diascia_yellow_mottle
scrophularia_mottle
ononis_yellow_mosaic
plantago_mottle
anagyris_vein_yellowing
eggplant_mosaic
chiltepin_yellow_mosaic
physalis_mottle
dulcamara_mottle
okra_mosaic
kennedya_yellow_mosaic
turnip_yellow_mosaic
erysimum_latent
poinsettia_mosaic
oat_blue_dwarf
citrus_sudden_death-associated
maize_rayado_fino
grapevine_syrah_-1
olive_latent_3
grapevine_fleck
Tymovirales

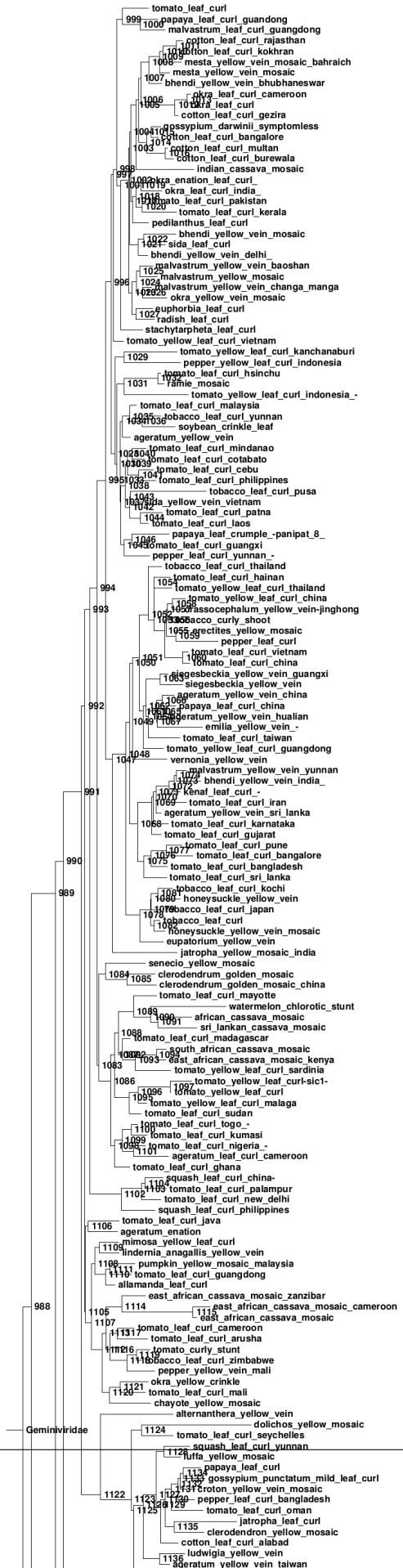

human_entero_d
polio
human_entero_c
human_entero_109
simian_entero_a
simian_picorna_strain_n203
simian_picorna_strain_n125
simian_entero_sv6
simian_picorna_17
human_entero_107
human_entero_100
human_entero_98
human_entero_b
porcine_entero_b
bovine_entero
simian_entero_sv43
simian_entero_sv19
human_entero_a
human_rhino_14
human_rhino_c
human_rhino_a
simian_picorna_1
porcine_entero_8
avian_sapelo
porcine_tescho
equine_rhinitis_b_2
equine_rhinitis_b_1
human_tmev-like_cardio
theilo
encephalomyocarditis
seneca_valley
human_cosa_b
human_cosa_a
human_cosa_e
human_cosa_d
foot-and-mouth_disease-type_sat_1
foot-and-mouth_disease-type_sat_3
foot-and-mouth_disease-type_c
foot-and-mouth_disease-type_asia_1
foot-and-mouth_disease-type_o
foot-and-mouth_disease-type_sat_2
bovine_rhinitis_b
equine_rhinitis_a
turdi_3
turdi_2
turdi_1
porcine_kobu_swine-s-1-hun-2007-hungary
kobu_pig-jy-2010a-chn
bovine_kobu
aichi
saji_ng-j1
human_klasse_1
ljungan
human_parecho
duck_hepatitis_ap
duck_hepatitis_a
seal_picorna_type_1
avian_encephalomyelitis
hepatitis_a
kashmir_bee
israel_acute_paralysis_of_bees
acute_bee_paralysis
solenopsis_invicta_1
taura_syndrome
mud_crab_dicistro
drosophila_c
cricket_paralysis
homalodisca_coagulata_-1
plautia_stali_intestine
black_queen_cell
himetobi_p
triatoma
aphid_lethal_paralysis
rhopalosiphum_padi
schizochytrium_single-stranded_rna
marine_rna_jp-a
chaetoceros_socialis_f._radians_rna
marine_rna_jp-b
maize_chlorotic_
raspberry_ringspot
grapevine_fanleaf
arabis_mosaic
blackcurrant_reversion
tomato_ringspot
tobacco_ringspot
tomato_black_ring
beet_ringspot
grapevine_chrome_mosaic
cycas_necrotic_stunt
red_clover_mottle
cowpea_mosaic
turnip_ringspot
radish_mosaic
cowpea_severe_mosaic
squash_mosaic
patchouli_mild_mosaic
broad_bean_wilt_2
broad_bean_wilt_1
mikania_micrantha_mosaic
Picornaviriales

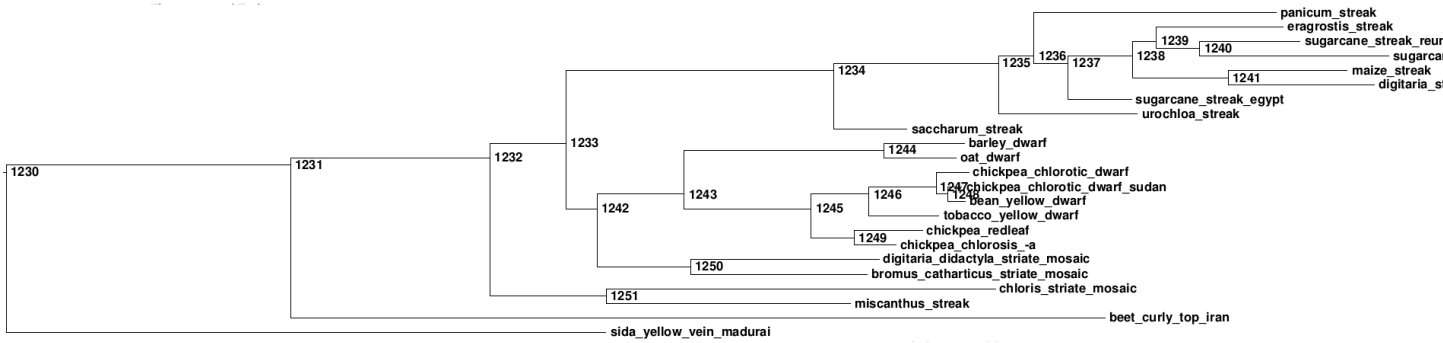

yam_mosaic
747
lettuce_mosaic
746
panax_y
748
konjac_mosaic
745
scallion_mosaic
744
751
narcissus_yellow_stripe
750
749
turnip_mosaic
japanese_yam_mosaic
plum_pox
743
752
sweet_potato_feathery_mottle
753
sweet_potato_c
lily_mottle
thunberg_fritillary
756
755
757
leek_yellow_stripe
clover_yellow_vein
754
758
bean_yellow_mosaic
potato_a
742
760
759
tobacco_vein_mottling
tobacco_etch
741
sunflower_chlorotic_mottle
766
765
potato_y
pepper_severe_mosaic
brugmansia_suaveolens_mottle
769
764
pepper_mottle
763
762
768
767
wild_potato_mosaic
771
770
potato_v
761
peru_tomato_mosaic
pepper_yellow_mosaic
740
verbena_y
bidens_mottle
pepper_veinal_mottle
774
739
773
chilli_veinal_mottle
738
772
wild_tomato_mosaic
tobacco_vein_banding_mosaic
737
moroccan_watermelon_mosaic
777
736
776
algerian_watermelon_mosaic
papaya_ringspot
775
papaya_leaf_distortion_mosaic
778
daphne_mosaic
peanut_mottle
782
freesia_mosaic
781
Potyviridae
beet_mosaic
783
basella_rugose_mosaic
watermelon_mosaic
794
793
soybean_mosaic
792
791
wisteria_vein_mosaic
790
789
fritillary_y
780
788
east_asian_passiflora
787
779
786
bean_common_mosaic
785
telosma_mosaic
784
bean_common_mosaic_necrosis
795
cowpea_aphid-borne_mosaic
passion_fruit_woodiness
zucchini_yellow_mosaic
zantedeschia_mild_mosaic
dasheen_mosaic
banana_bract_mosaic
canna_yellow_streak
sorghum_mosaic
800
maize_dwarf_mosaic
799
797
796
798
sugarcane_mosaic
pennisetum_mosaic
cocksfoot_streak
pea_seed-borne_mosaic
shallot_yellow_stripe
802
801
onion_yellow_dwarf
narcissus_degeneration
hordeum_mosaic
804
803
agropyron_mosaic
ryegrass_mosaic
cassava_brown_streak_uganda_-ug
810
809
cassava_brown_streak
808
squash_vein_yellowing
sweet_potato_mild_mottle
wheat_streak_mosaic
814
807
oat_necrotic_mottle
806
813
805
812
wheat_eqlid_mosaic
811
brome_streak_mosaic
triticum_mosaic
815
sugarcane_streak_mosaic
barley_yellow_mo
blackberry_y

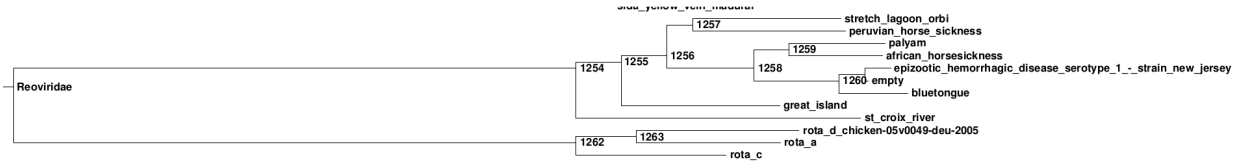

segmented ssRNA neg
Arenaviridae
Bunyaviridae
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
844
845
846
847
848
849
850
851
852
853
cupixi
guanarito
amapari
machupo
junin
tacaribe
sabia
chapare
oliveros
latino
parana
flexal
pichinde
allpahuayo
pirital
whitewater_arroyo
tamiami
bear_canyon
mopeia_lassa_reassortant_29
mopeia_an20410
morogoro
mobala
lassa
ippy
lymphocytic_choriomeningitis
lujo
impatiens_necrotic_spot
watermelo
groundnut
capsicum_
melon_y
tomato_zonate_sp
la_crosse
oropouche
akabane
bunyamwera


tupaia_paramyxo
mossman
872
870
873
j
871
beilong
nipah
874
hendra
869
dolphin_morbilli
rinderpest_strain_kabete_o
868
empty
measles
peste-des-petits-ruminants
canine_distemper
sendai
882
867
human_parainfluenza_1
881
880
human_parainfluenza_3
883
bovine_parainfluenza_3
fer-de-lance
goose_paramyxo_sf02
885
avian_paramyxo_6
866
porcine_rubula
889
888
mapuera
mumps
887
884
simian_41
886
891
890
human_parainfluenza_2
parainfluenza_5
tioman
892
menangle
human_metapneumo
896
895
avian_metapneumo
pneumonia_of_mice_j3666
894
respiratory_syncytial
human_respiratory_syncytial
bovine_respiratory_syncytial
Mononegavirales
893
sudan_ebola
reston_ebola
900
cote_divoire_ebola
899
zaire_ebola
lake_victoria_marburg
nyamanini
907
midway
906
borna_disease
taro_vein_chlorosis
912
maize_mosaic
913
910
911
iranian_maize_mosaic_nucleorhabdo
rice_yellow_stunt
909
sonchus_yellow_net
lettuce_yellow_mottle
915
914
lettuce_necrotic_yellows
northern_cereal_mosaic
wongabel
921
920
ngaingan
bovine_ephemeral_feve
905
919
vesicular_stomatitis_indiana
918
917
923
922
spring_viraemia_of_carp
904
908
siniperca_chuatsi_rhabdo
drosophila_melanogaster_sigm
916
tupaia
rabies
australian_bat_lyssa
european_bat_lyssa_2
european_bat_lyssa_1
mokola
snakehead_rhabdo
929
928
viral_hemorrhagic_septicemia
infectious_hematopoietic_necrosis
930
hirame_rhabdo
influenza_a_a-new_york-392-2004h3n2
extra
empty
influenza_a_a-goose-guangdong-1-96h5n1
extra
empty
933
influenza_a_a-puerto_rico-8-34h1n1
932
extra
Orthomyxoviridae
empty
influenza_a_a-hong_kong-1073-99h9n2
extra
empty
influenza_b
influenza_c_c-ann_arbor-1-50
empty
thogoto
946
infectious_salmon_anemia


human_boca
boca_gorilla-gbov1-2009
953
human_boca_3
951
952
human_boca_4
956
human_boca_2
canine_minute
bovine_parvo
luiii
mouse_parvo_5a
950
mouse_parvo_2
mouse_parvo_3
mouse_parvo_1
958
957
mouse_parvo_4a
minute_of_mice
parvo_h1
canine_parvo
porcine_parvo
aleutian_mink_disease
adeno-associated-2
Parvoviridae
adeno-associated-3
adeno-associated-4
adeno-associated-7
972
adeno-associated-8
adeno-associated-1
971
bovine_adeno-associated
970
978
979
adeno-associated-5
969
bat_adeno-associated_ynm
968
avian_adeno-associated_strain_da-1
967
avian_adeno-associated_atcc_vr-865
980
muscovy_duck_parvo
982
goose_parvo
snake_parvo_1
human_parvo_4
983
porcine_parvo_4
984
bovine_parvo-2
human_parvo_b19
985
human_erythro_v9
aedes_aegypti_denso
986
aedes_albopictus_denso


tomato_leaf_curl
papaya_leaf_curl_guandong
malvastrum_leaf_curl_guangdong
cotton_leaf_curl_rajasthan
cotton_leaf_curl_kokhran
mesta_yellow_vein_mosaic_bahraich
mesta_yellow_vein_mosaic
bhendi_yellow_vein_bhubhaneswar
okra_leaf_curl_cameroon
okra_leaf_curl
cotton_leaf_curl_gezira
gossypium_darwinii_symptomless
cotton_leaf_curl_bangalore
cotton_leaf_curl_multan
cotton_leaf_curl_burewala
indian_cassava_mosaic
okra_enation_leaf_curl_
okra_leaf_curl_india_
tomato_leaf_curl_pakistan
tomato_leaf_curl_kerala
pedilanthus_leaf_curl
bhendi_yellow_vein_mosaic
sida_leaf_curl
bhendi_yellow_vein_delhi_
malvastrum_yellow_vein_baoshan
malvastrum_yellow_mosaic
malvastrum_yellow_vein_changa_manga
okra_yellow_vein_mosaic
euphorbia_leaf_curl
radish_leaf_curl
stachytarpheta_leaf_curl
tomato_yellow_leaf_curl_vietnam
tomato_yellow_leaf_curl_kanchanaburi
pepper_yellow_leaf_curl_indonesia
tomato_leaf_curl_hsinchu
ramie_mosaic
tomato_yellow_leaf_curl_indonesia_-
tomato_leaf_curl_malaysia
tobacco_leaf_curl_yunnan
soybean_crinkle_leaf
ageratum_yellow_vein
tomato_leaf_curl_mindanao
tomato_leaf_curl_cotabato
tomato_leaf_curl_cebu
tomato_leaf_curl_philippines
tobacco_leaf_curl_pusa
sida_yellow_vein_vietnam
tomato_leaf_curl_patna
tomato_leaf_curl_laos
papaya_leaf_crumple_-panipat_8_
tomato_leaf_curl_guangxi
pepper_leaf_curl_yunnan_-
tobacco_leaf_curl_thailand
tomato_leaf_curl_hainan
tomato_yellow_leaf_curl_thailand
tomato_yellow_leaf_curl_china
cassocephalum_yellow_vein-jinghong
tobacco_curly_shoot
spectites_yellow_mosaic
pepper_leaf_curl
tomato_leaf_curl_vietnam
tomato_leaf_curl_china
siegesbeckia_yellow_vein_guangxi
siegesbeckia_yellow_vein
ageratum_yellow_vein_china
papaya_leaf_curl_china
ageratum_yellow_vein_hualian
emilia_yellow_vein_-
tomato_leaf_curl_taiwan
tomato_yellow_leaf_curl_guangdong
vernonia_yellow_vein
malvastrum_yellow_vein_yunnan
bhendi_yellow_vein_india_
kenaf_leaf_curl_-
tomato_leaf_curl_iran
ageratum_yellow_vein_sri_lanka
tomato_leaf_curl_karnataka
tomato_leaf_curl_gujarat
tomato_leaf_curl_pune
tomato_leaf_curl_bangalore
tomato_leaf_curl_bangladesh
tomato_leaf_curl_sri_lanka
tobacco_leaf_curl_kochi
honeysuckle_yellow_vein
tobacco_leaf_curl_japan
tobacco_leaf_curl
honeysuckle_yellow_vein_mosaic
eupatorium_yellow_vein
jatropha_yellow_mosaic_india
senecio_yellow_mosaic
clerodendrum_golden_mosaic
clerodendrum_golden_mosaic_china
tomato_leaf_curl_mayotte
watermelon_chlorotic_stunt
african_cassava_mosaic
sri_lankan_cassava_mosaic
tomato_leaf_curl_madagascar
south_african_cassava_mosaic
east_african_cassava_mosaic_kenya
tomato_yellow_leaf_curl_sardinia
tomato_yellow_leaf_curl-sic1-
tomato_yellow_leaf_curl
tomato_yellow_leaf_curl_malaga
tomato_leaf_curl_sudan
tomato_leaf_curl_togo_-
tomato_leaf_curl_kumasi
tomato_leaf_curl_nigeria_-
ageratum_leaf_curl_cameroon
tomato_leaf_curl_ghana
squash_leaf_curl_china-
tomato_leaf_curl_palampur
tomato_leaf_curl_new_delhi
squash_leaf_curl_philippines
tomato_leaf_curl_java
ageratum_enation
mimosa_yellow_leaf_curl
lindernia_anagallis_yellow_vein
pumpkin_yellow_mosaic_malaysia
tomato_leaf_curl_guangdong
allamanda_leaf_curl
east_african_cassava_mosaic_zanzibar
east_african_cassava_mosaic_cameroon
east_african_cassava_mosaic
tomato_leaf_curl_cameroon
tomato_leaf_curl_arusha
tomato_curly_stunt
tobacco_leaf_curl_zimbabwe
pepper_yellow_vein_mali
okra_yellow_crinkle
tomato_leaf_curl_mali
chayote_yellow_mosaic
alternanthera_yellow_vein
dolichos_yellow_mosaic
Geminiviridae
tomato_leaf_curl_seychelles
squash_leaf_curl_yunnan
luffa_yellow_mosaic
papaya_leaf_curl
gossypium_punctatum_mild_leaf_curl
croton_yellow_vein_mosaic
pepper_leaf_curl_bangladesh
tomato_leaf_curl_oman
jatropha_leaf_curl
clerodendron_yellow_mosaic
cotton_leaf_curl_alabad
ludwigia_yellow_vein
ageratum_yellow_vein_taiwan


panicum_streak
eragrostis_streak
sugarcane_streak_reun
sugarca
maize_streak
digitaria_s
sugarcane_streak_egypt
urochloa_streak
saccharum_streak
barley_dwarf
oat_dwarf
chickpea_chlorotic_dwarf
chickpea_chlorotic_dwarf_sudan
bean_yellow_dwarf
tobacco_yellow_dwarf
chickpea_redleaf
chickpea_chlorosis_-a
digitaria_didactyla_striate_mosaic
bromus_catharticus_striate_mosaic
chloris_striate_mosaic
miscanthus_streak
beet_curly_top_iran
sida_yellow_vein_madurai
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251


stretch_lagoon_orbi
peruvian_horse_sickness
palyam
african_horsesickness
epizootic_hemorrhagic_disease_serotype_1_-_strain_new_jersey
empty
bluetongue
great_island
st_croix_river
rota_d_chicken-05v0049-deu-2005
rota_a
rota_c
Reoviridae
1254
1255
1256
1257
1258
1259
1260
1262
1263

CHAPTER 15

Reference data

As explained in the *How is Sequedex used with other software?* and described in some detail in *Annotated reads*, further analysis of closely related sequences can occur with reference alignments and tree-building. Creation and annotation of reference sequence alignments is a laborious process that consumes a significant portion of effort at such specialized efforts as the HIV sequece database or the Ribosomal Database project. We are not attempting to be comprehensive (for this, see Pfam or The Hemorrhagic Fever Viruses (HFV) Database Project) or even particularly accurate.

Instead, we aim to enable the user to build off of our efforts at characterizing one important phylogenetic marker (RNA Polymerase) to illustrate the techniques and enable others to replicate and trouble-shoot our phylogenetic work in creating our reference tree. Additionally, we provide reference alignments for a variety of other problems, such as strain-ID of pathogens and distinguishing orthologs from paralogs in ascribing functional roles to proteins such as virulence factors, enzymes, or regulatory proteins.

Another advantage to the two-step process for prodcing phylogenetic or functional assignments is that the most recent information can be used in the latter step. Because Sequedex uses a database of orthogenomic signature peptides to identify reads for further analysis, improvements in the reference database (Sequedex data module) affect only the overall sensitivity, and not the resolution of placement. Depending on the question being asked, the sensitivity may already be quite high (> 90%).

These alignments can be used in two ways to characterize an unknown; both require an alignment of metagenomic reads or assembled contigs to the reference alignment. For a relatively small number of aligned reads or contigs, the tree can simply be re-calculated, and the relative position of reference and sample can be compared. For a large number of aligned reads or contigs, *p-values and Pplacer* can be used to visualize which leaves or internal branches provide the most appropriate placement of the reads.

15.1 Phylogenetic placement of RNA Polymerase reads

RNAaPol alignment files, including beta and beta prime subunits for bacterial groups 1-5 of *Tree of Life, 2550 taxa*, beta, gamma, and beta prime subunits of the cyanobacteria, and the B1 and B2 subunits of eukaryotic RNA Pol II (care was taken to exclude RNA Pols I & III). The files are in Phylip format, because that is what is needed for *Tree building with phyML or FastTree*. It has the additional advantage that sub-alignments can easily be extracted by simply using the grep command:

```
grep Bacteroides data.1 > Bacteroides.RNAP.fas
```

will extract the 23 Bacteroides polymerase sequences into a separate file. This provides a convenient alignment to which assembled genes from metagenomes can be added, for the purpose of generating trees showing how organisms from the metagenomic samples relate both to the reference species and one another.

Sequence file formats can be converted online at the HIV database.

- Sequence alignments for `Group 1`, `Group 2`, `Group 3`, `Group 4`, `Group 5`, `Group 6`, `Group 7`.

Subsets of the above files generated the bacterial reference trees used by Sequedex. The above seven groups cover the bacteria and archaea - for the eukaryotes, RNAPol II must be separated from RNAPols I and III (data not provided here) to create a tree based on RNAPol II, such as this:

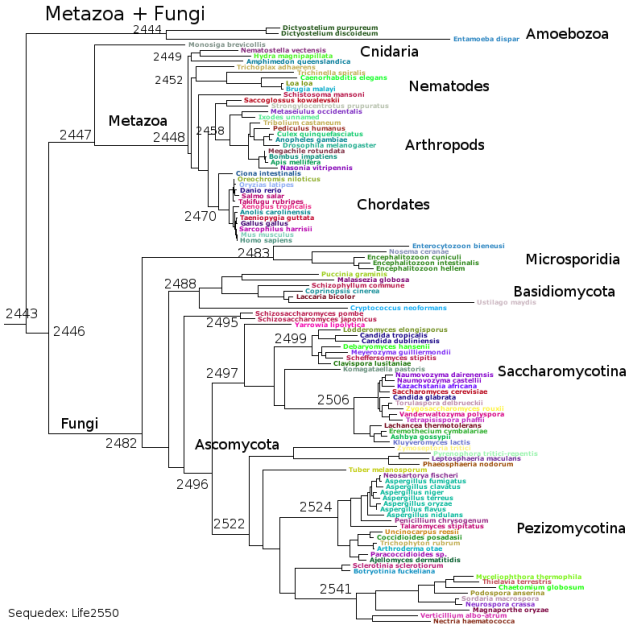


Fig. 15.1: Maximum likelihood tree of the fungi and metazoa computed with PhyML from an amino acid alignment of the B1 and B2 genes of RNA Polymerase II.

15.2 Nucleotide Alignments for Strain Attribution

By using nucleotide alignments and all available strains from NCBI's completed and draft genomes instead of amino acid alignments and one representative per species, it is possible to see considerable structure in the various strains of different species, as well as interspersing of strains from two different species. The figure below was made from near-neighbors of *E. coli*.

The sequence alignments to make this type of tree are available for `E.coli`.

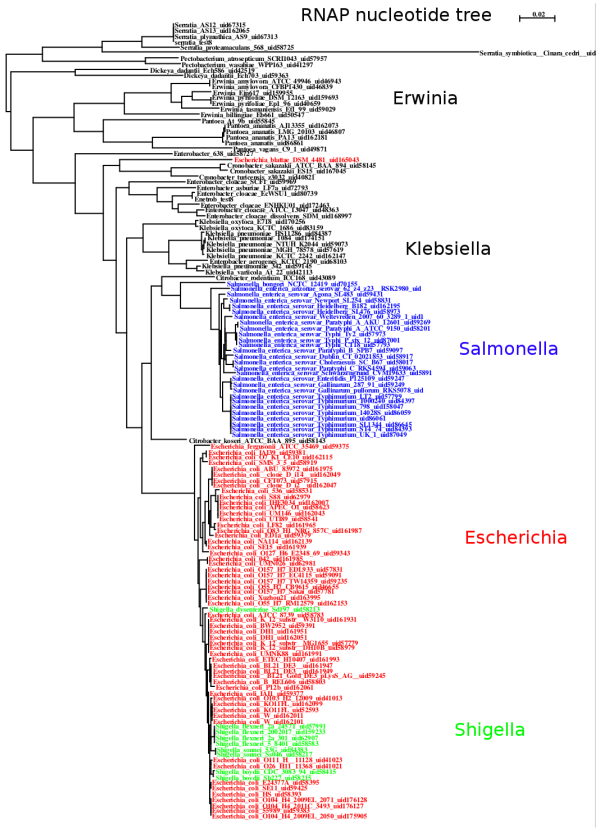


Fig. 15.2: Maximum likelihood tree of near-neighbors of *E. coli* computed with PhyML from a nucleotide alignment of the beta and beta prime subunits of the RNA Polymerase genes.