December 4th 2018

Microbial Taxonomy as Seen Through a (Meta)genomic Lens

Jim Cole

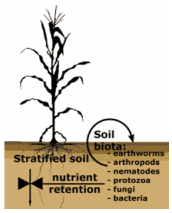
Center for Microbial Ecology
Dept. of Plant, Soil & Microbial Sciences
Michigan State University
East Lansing, Michigan U.S.A.

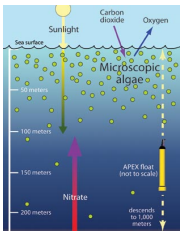
Michigan State University

- First land-grant university in USA, est. 1855
  - teaching of practical agriculture, science, and engineering
- 50,000 students (11,000 graduate students)

- College of Agriculture & Natural Resources
- Dept. of Plant Soil and Microbial Sciences
- Center for Microbial Ecology
  - Study **microbes** and their interactions in the environment

Microbial Communities: Important in many environments





http://rdp.cme.msu.edu 2

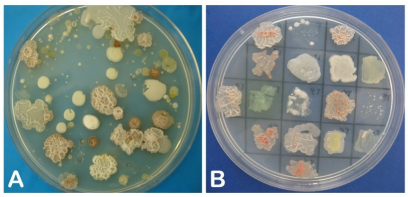
Microbes are Dominant in Biosphere

- Biomass:
  Prokaryotes 4-6 x 10^{30} cells, containing:
  - 350-550 Pg of C (0.6-1 X plants)
  - 85-130 Pg of N (10 X plants)
  - 9-14 Pg of P (10 X plants)
- Environments:
  - Open ocean 0.12 x 10^{30} cells
  - Soil 0.26 x 10^{30} cells
  - Oceanic subsurface 3.5 x 10^{30} cells
  - Terrestrial subsurface 0.25 – 2.5 x 10^{30} cells
- Biodiversity: 400,000 – 4,000,000 species

Whitman et al. (1998) *PNAS* 95:6578

http://rdp.cme.msu.edu 3

Can you name these bacteria?



Morphological properties are not phylogenetically coherent.

From: Ch. 2 -- Terrestrial Bacteria from Agricultural Soils: By Masoomeh Shams-Ghahfarokhi, Sanaz Kalantari and Mehdi Razzaghi-Abyaneh DOI: 10.5772/45918

http://rdp.cme.msu.edu 4

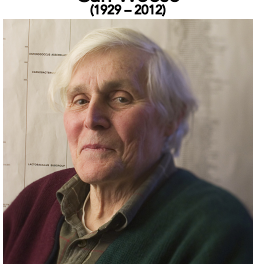
Challenges to Traditional Taxonomy

- Only a small number of phenotypic traits are easily tested.
- Phenotypic traits are often not congruent with (molecular) phylogeny.
- The vast majority of microbial species have not been studied in pure culture.

http://rdp.cme.msu.edu 5

Elucidation of the three domains of life

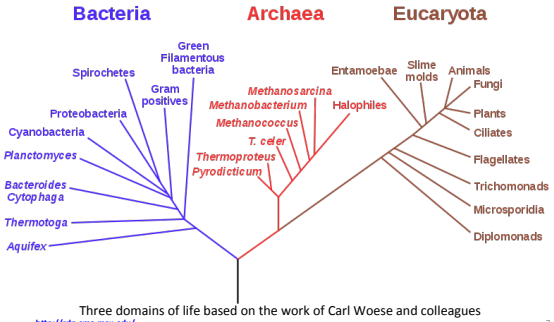
Carl Woese
(1929 – 2012)



Ribosomal RNA sequence as phylogenetic marker

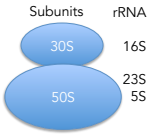
- Discovered "3rd kingdom"
- Archaea and Bacteria separate domains
- Contrast with former Prokaryote hypothesis

Phylogenetic Tree of Life



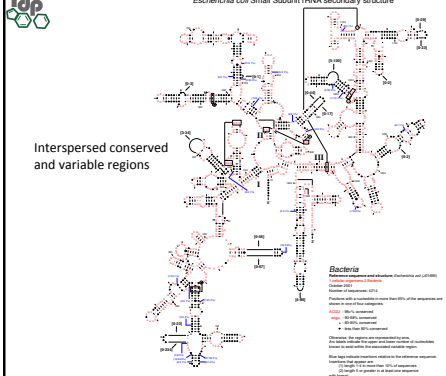
Three domains of life based on the work of Carl Woese and colleagues

Ribosomes – Universal Marker



- Protein synthesis factory.
- Core function present in all cellular organisms.
- Very little evidence of horizontal gene transfer.
- Historically easy to work with.
  - Purify by centrifugation and extract rRNA.
- Now we use PCR to amplify from genomic DNA
  - rRNA genes have conserved regions interspersed with highly variable regions.
  - Conserved regions used for both PCR primers and sequencing primers.

Phylogenetic conservation superimposed onto the *Escherichia coli* Small Subunit rRNA secondary structure



Interspersed conserved and variable regions

Diversity of uncultured organisms explored by rRNA sequencing

David A. Stahl, David J. Lane, Gary J. Olsen and Norman R. Pace
Science, New Series, Vol. 224, No. 4647 (Apr. 27, 1984), pp. 409-411
Published by: American Association for the Advancement of Science

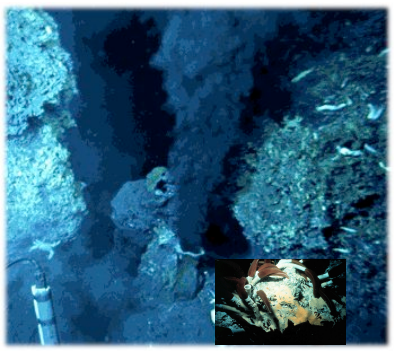
Analysis of Hydrothermal Vent–Associated Symbionts by Ribosomal RNA Sequences

Abstract. *Ribosomal RNA (rRNA) sequences were used to establish the phylogenetic affiliations of symbioses in which prokaryotes appear to confer sulfur-based chemoautotrophy on their invertebrate hosts. Two submarine hydrothermal vent animals, the vestimentiferan tube worm* Riftia pachyptila *and the clam* Calyptogena magnifica, *and a tidal-flat bivalve,* Solemya velum, *were inspected. 5S rRNA's were extracted from symbiont-bearing tissues, separated into unique forms, and their nucleotide sequences determined and related to other 5S rRNA's in a phylogenetic tree analysis. The prokaryotic symbionts are related to one another and affiliated with the same narrow phylogenetic grouping as* Escherichia coli *and* Pseudomonas aeruginosa. *The sequence comparisons suggest that* Riftia *is more closely related to the bivalves than their current taxonomic status would suggest.*

Evidence has accumulated that sulfur-oxidizing microbes can establish symbiotic relationships with certain invertebrates, producing "chemoautotrophic animals" (*1*). The putative symbionts were identified histologically and by the presence of high levels of certain Calvin cycle and sulfur-oxidative enzymes in the hydrothermal vent tube worm *Riftia pachyptila* (*2*), in which the bacteria fill a specialized organ, the trophosome. Similar (trunk wall and trophosome) and *Calyptogena magnifica* (gill tissue) and live specimens of *Solemya velum* were obtained (*3*); gill and foot tissues were excised and frozen immediately upon receipt. Total RNA was isolated from homogenized tissues extracted with hot phenol and sodium dodecyl sulfate and fractionated by polyacrylamide gel electrophoresis (Fig. 1A). After elution, the mixtures of 5S rRNA's (host and symbiont) were labeled at their 5′ termini with [γ-^{32}P]ATP (adenosine triphosphate) and polynucleotide kinase or at their 3′ termini with [5′-^{32}P]pCp (C, cytosine) and RNA ligase and were resolved by electrophoresis on 8 percent polyacrylamide sequencing gels (Fig. 1B). All 5S rRNA's were sequenced from both termini by enzymatic and chemical partial digestions (Fig. 1C). The derived sequences and the alignments used for phylogenetic analysis are shown in Fig. 2.

The relation of the symbiont 5S rRNA's to those of better-known organisms is best understood as a phylogenetic tree (Fig. 3). The branch lengths

One approach to characterizing uncultivable organisms is to establish their phylogenetic relationships to better-known organisms by appropriate macromolecular sequence comparisons (*5*). Ribosomal RNA's (rRNA) seem well-suited among cellular macromolecules for such analyses because of their ubiquitous distribution, functional constancy, high conservation of primary structure, and apparent freedom from artifacts of



Hydrothermal Vent Black Smoker

Prokaryotic Taxonomy

- Current taxonomy (nearly) coherent with phylogeneny.
- Taxonomy is informed by phenotype.
- Taxon boundaries are circumscribed by experts to attempt to give groups meaningful to the practitioners
- Uncultivated organisms not included

Phylogenetic Analysis vs. Classification

- Classification is conceptually easier to interpret.
- Often preferred when the groups are well understood.
- Phylogenetic methods are preferred for new groups or when the placement is not clear.

Simple classifiers from our work

Taxonomic Classifiers

- Use concepts from Machine Learning
- Many algorithms can be applied

- Sequence Match
  - Nearest-Neighbor (Last Common Ancestor)
- RDP Classifier
  - Naïve Bayesian

Cole et al. (2005) Nucleic Acids Research
33, Database issue: D294–D296
10.1186/s40168-015-0093-6

OXFORD JOURNALS
Nucleic Acids Research

The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis

ABSTRACT

The Ribosomal Database Project (RDP-II) provides the research community with aligned and annotated rRNA gene sequences, along with analysis services and a phylogenetically consistent taxonomic framework for these data. Updated monthly, these services are made available through the RDP-II website (http://rdp.cme.msu.edu/). RDP-II release 9.21 (August 2004) contains 101 632 bacterial small subunit rRNA gene sequences in aligned and annotated format. High-throughput tools for initial taxonomic placement, identification of related sequences, probe and primer testing, data navigation and subalignment download are provided. The RDP-II email address for questions or comments is rdpstaff@msu.edu.

DESCRIPTION

Release 9 introduces substantial changes to the Ribosomal Database Project (RDP). These changes are in response to the rapidly increasing number of available ribosomal RNA gene sequences (rRNA sequences) and the trend toward high-throughput rRNA sequencing with the concomitant need for high volume rRNA analysis tools. This paper describes changes since the 2003 description (1). Details about the data and analysis services can be found at the RDP-II website (http://

into its internal model. This aligner is trained on a set of high-quality hand-aligned sequences and incorporates the conserved bacterial secondary structure model of Gutell and co-workers (3). As of release 9.21 (August 2004), the database contained 101 632 total small subunit bacterial rRNA sequences. Of these, 39 772 were near full-length (≥1200 bases), 54 316 came from uncultured organisms and 4431 were from type strains of validly named bacterial species.

Taxonomy. All Release 9 tools use a new hierarchical framework (RDP Hierarchy) differing significantly from the hierarchy provided with previous RDP releases. The RDP Hierarchy is based on the new phylogenetically consistent higher-order bacterial taxonomy proposed by Garrity *et al.* (4) (http://dx.doi.org/10.1007/bergeysoutline). This hierarchy provides order to the collection. It provides a phylogenetic framework into which to place results of the RDP analysis functions, and it provides an entry point for users looking for sequences from specific groups of organisms. New sequences are placed into the RDP Hierarchy using the RDP Classifier (see below).

Analysis services

The RDP analysis services have been completely revised to support the emerging trend toward high-throughput rRNA sequence analysis in microbial ecology and related disciplines. Three of the tools listed below incorporate the concept of data filters. The user can choose to apply up to three data filters on

Cole et al. (2005) Nucleic Acids Research
33, Database issue: D294–D296
10.1186/s40168-015-0093-6

OXFORD JOURNALS
Nucleic Acids Research

The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis

Sequence Match is a complete re-implementation of the original Sequence Match method (1). Sequence Match finds sequences similar to a user's query sequences using a word matching strategy not requiring prior alignment. Sequence Match is more accurate than BLAST (6) at finding closely related rRNA sequences (Table 1). The related sequences returned by Sequence Match serve as a good starting point for more detailed examination of relatedness by classical phylogenetic or other methods. The initial result page presents a *k*-nearest neighbor (*k*-NN) classifier assignment of the query sequences. A query is assigned to the lowest taxonomic rank that includes the *k* highest scoring database sequences. The value of *k*, as well as the three data filters can be changed at will in this view. The user can switch from the summary *k*-NN view to a detailed results view for any query sequence. In this view, the top *k* database matches to the query are displayed in the RDP Hierarchy. In this mode, any subset of the matches can be selected for transfer to the Hierarchy Browser and later

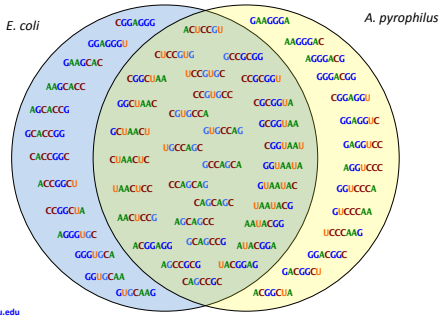
SeqMatch

- NN – Nearest-neighbor classifier
- k-NN – k nearest-neighbors classifier
- LCA – Last common ancestor
  (Lowest common ancestor)

Can represent a sequence as a set of overlapping k-mers

GAAGCACCGGCUAACUCCGUGCCAGCAGCCGCGGUAAUACGGAGGGUGCAAG

SeqMatch Comparison of 16S rRNA genes from two species



SeqMatch Math

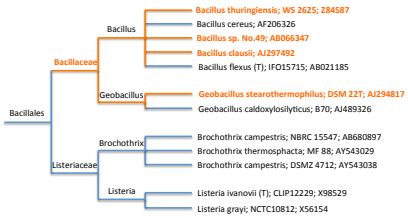
Given query sequence A and training sequence B, the k-mer similarity between A and B is defined as:

$$S_{AB} \equiv \frac{|k\text{-}mers \text{ in common}|}{\min(\,|k\text{-}mers \text{ in } A|, |k\text{-}mers \text{ in } B|)}$$

The training sequences with the highest S_{AB} are the *nearest-neighbors* of query A

SeqMatch classifies query A into the same taxa as its nearest-neighbors

LCA – Last Common Ancestor



Wang et al. (2007) AEM 73(16): 5261-5267
doi:10.1128/AEM.00062-07

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology

Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.

The Ribosomal Database Project (RDP) Classifier, a naïve Bayesian classifier, can rapidly and accurately classify bacterial 16S rRNA sequences into the new higher-order taxonomy proposed in Bergey's *Taxonomic Outline of the Prokaryotes* (2nd ed., release 5.0, Springer-Verlag, New York, NY, 2004). It provides taxonomic assignments from domain to genus, with confidence estimates for each assignment. The majority of classifications (98%) were of high estimated confidence (≥95%) and high accuracy (98%). In addition to being tested with the corpus of 5,014 type strain sequences from Bergey's outline, the RDP Classifier was tested with a corpus of 23,095 rRNA sequences as assigned by the NCBI into their alternative higher-order taxonomy. The results from leave-one-out testing on both corpora show that the overall accuracies at all levels of confidence for near-full-length and 400-base segments were 89% or above down to the genus level, and the majority of the classification errors appear to be due to anomalies in the current taxonomies. For shorter rRNA segments, such as those that might be generated by pyrosequencing, the error rate varied greatly over the length of the 16S rRNA gene, with segments around the V2 and V4 variable regions giving the lowest error rates. The RDP Classifier is suitable both for the analysis of single rRNA sequences and for the analysis of libraries of thousands of sequences. Another related tool, RDP Library Compare, was developed to facilitate microbial-community comparison based on 16S rRNA gene sequence libraries. It combines the RDP Classifier with a statistical test to flag taxa differentially represented between samples. The RDP Classifier and RDP Library Compare are available online at http://rdp.cme.msu.edu/.

Starting in the mid-1980s, Carl Woese revolutionized the field of microbiology with his rRNA-based phylogenetic comparisons delineating the three main branches of life (28). Today, rRNA-based analysis remains a central method in microbiology, used not only to explore microbial diversity but also as a day-to-day method for bacterial identification. Identification methods are conceptually easier to interpret than molecular phylogenetic analyses and are often preferred when the groups are well understood. Most rRNA identification (classification) showed that the Bayesian method can still be optimal even when this attribute independency is violated. The method has also been reported to perform well on problems similar to the classification of sequence data, such as the classification of text documents, that have a high-dimensional feature space and sparse data (16).

The Ribosomal Database Project II (RDP) provides data, tools, and services related to rRNA sequences to the research community. As of January 2007, the RDP maintains over 300,000 bacterial sequences and averages over 5,000 new sequences

Wang *et al.* (2007) *AEM* 73(16): 5261-5267
10.1128/AEM.00062-07
Applied and Environmental Microbiology

Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.

The Ribosomal Database Project (RDP) Classifier, a naïve Bayesian classifier, can rapidly and accurately classify bacterial 16S rRNA sequences into the new higher-order taxonomy proposed in Bergey's *Taxonomic Outline of the Prokaryotes* (2nd ed., release 5.0, Springer-Verlag, New York, NY, 2004). It

- Rapidly assigns sequences into bacterial, archaeal and fungal taxonomies
- Works well on partial or full-length sequences
- Bootstrap confidence estimates
- Still in use as a standard microbial classification method

RDP Classifier

- Naïve Bayesian classifier
- Uses training data to estimate k-mer likelihood
- Assignment based on genus for which k-mers are most common

http://rdp.cme.msu.edu 25

Estimating P(V | G)

	k-mer Set V			
	CGGCUAA	GUAAUAC	ACGGAGG	GCCGCGG
B. subtilis	+	+	–	–
B. clausii	+	–	+	–
B. smithii	+	+	–	–
B. flexus	+	+	+	–
B. ruris	+	+	–	–
Probability:	**99%**	**80%**	**40%**	**01%**

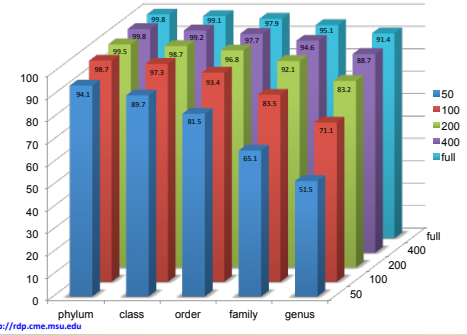
P(V|*Bacillus*) = (0.99 × 0.80 × 0.40 × 0.01) = 0.003168 or ≈ 1 in 316
P(V|*Sinobaca*) = (0.99 × 0.01 × 0.20 × 0.01) = 0.000019 or ≈ 1 in 50505

When you have eliminated the impossible, whatever remains, however improbable, must be the truth.

Arthur Conan Doyle

http://rdp.cme.msu.edu 26

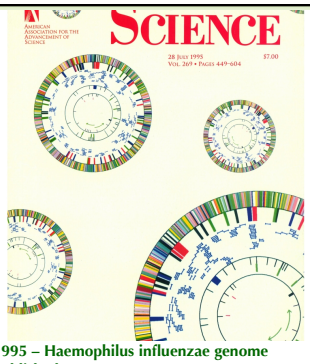
RDP Classifier: Leave-one-out Testing



http://rdp.cme.msu.edu 27

Taxonomy in the Genome Era

http://rdp.cme.msu.edu 28



1995 – Haemophilus influenzae genome published

Bacterial and Archaeal Genomes Available at NCBI (9/3/2018) (Compare to >3 million rRNA genes)

11403	Complete Genomes
2025	Complete Chromosome
70950	Scaffolds
72311	Contigs only
156689	Total Genomes

Majority from a small number of species

Metagenomics

Metagenomics is the study of genetic material recovered directly from environmental samples…

Because of its ability to reveal the previously hidden diversity of microscopic life, metagenomics offers a powerful lens for viewing the microbial world that has the potential to revolutionize understanding of the entire living world.

— Wikipedia

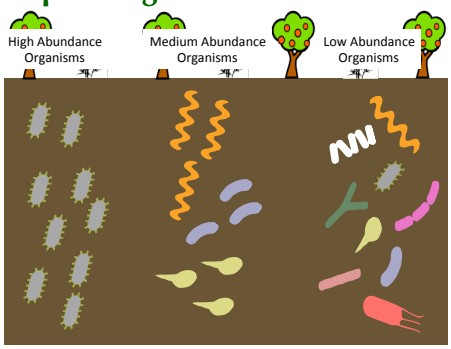
http://rdp.cme.msu.edu

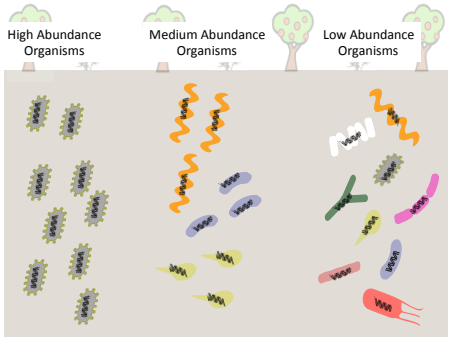
31

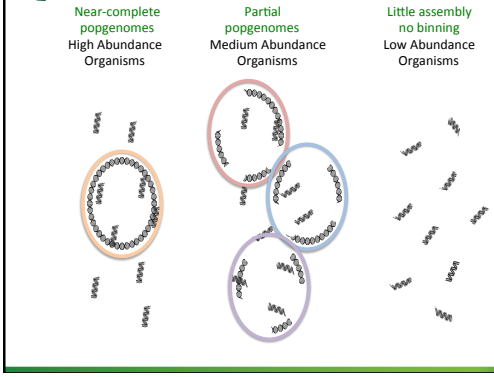
Microbial Genomes from Uncultured Organisms

- **Single Cell Genomes:** Single microbial cells are separated before sequencing
  - **Issues:** Incomplete genomes, enzymatic DNA amplification causes artifacts
- **Metagenomic Binning:** Grouped from metagenomic assemblies
  - **Issues:** Incomplete genomes, may mix allelic variants, contamination an issue

Sequencing a model microbiome







mSystems®
RESEARCH ARTICLE
Novel Systems Biology Techniques

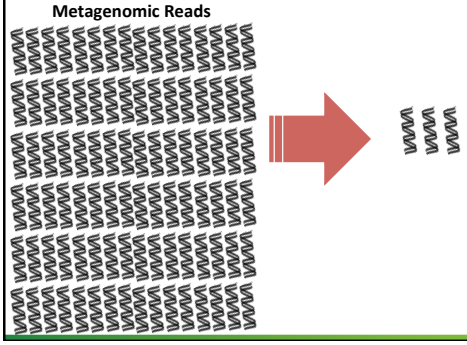
Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity

Luis M. Rodriguez-R,[a] Santosh Gunturu,[c] James M. Tiedje,[c,d,e] James R. Cole,[c,e] Konstantinos T. Konstantinidis[a,b]

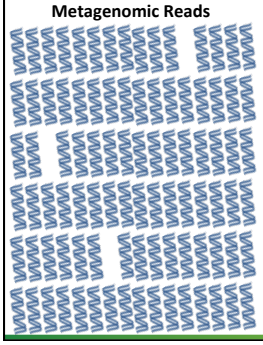
ABSTRACT Estimations of microbial community diversity based on metagenomic data sets are affected, often to an unknown degree, by biases derived from insufficient coverage and reference database-dependent estimations of diversity. For instance, the completeness of reference databases cannot be generally estimated since it depends on the extant diversity sampled to date, which, with the exception of a few habitats such as the human gut, remains severely undersampled. Further, estimation of the degree of coverage of a microbial community by a metagenomic data set is prohibitively time-consuming for large data sets, and coverage values may not be directly comparable between data sets obtained with different sequencing technologies. Here, we extend Nonpareil, a database-independent tool for the estimation of coverage in metagenomic data sets, to a high-performance computing implementation that scales up to hundreds of cores and includes, in addition, a *k*-mer-based estimation as sensitive as the original alignment-based version but about three hundred times as fast. Further, we propose a metric of sequence diversity (N_d) derived directly from Nonpareil curves that correlates well with alpha diversity assessed by traditional metrics. We use this metric in different experiments demonstrating the correlation with the Shannon index estimated on 16S

Nonpareil-3

20,000,000
Metagenomic Reads

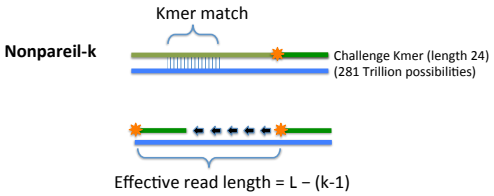


100,000,000
Metagenomic Reads



10,000
Challenge Reads

Nonpareil-3
Measure relative abundance



- Each challenge read represents 0.01% of the metagenome.
- By measuring the abundance of each challenge read, we can:
  - Determine the coverage curve
  - Estimate the fraction with high, moderate and low coverage.

Multi-Gene Phylogenetic Analysis

- Uses additional marker genes
  - Universal: transcription – translation – replication
  - Choose for no horizontal gene transfer
  - Unfortunately, few genes meet these criteria
    - 100 – 130 genes commonly used
- Two main methods
  - Concatenate multiple gene sequences
  - Supertrees (combine multiple trees)

RESOURCE

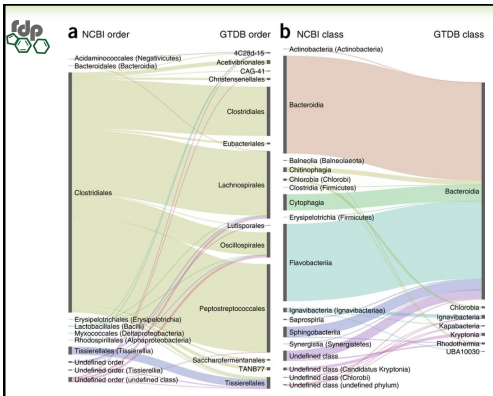
nature biotechnology

A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz

Taxonomy is an organizing principle of biology and is ideally based on evolutionary relationships among organisms. Development of a robust bacterial taxonomy has been hindered by an inability to obtain most bacteria in pure culture and, to a lesser extent, by the historical use of phenotypes to guide classification. Culture-independent sequencing technologies have matured sufficiently that a comprehensive genome-based taxonomy is now possible. We used a concatenated protein phylogeny as the basis for a bacterial taxonomy that conservatively removes polyphyletic groups and normalizes taxonomic ranks on the basis of relative evolutionary divergence. Under this approach, 58% of the 94,759 genomes comprising the Genome Taxonomy Database had changes to their existing taxonomy. This result includes the description of 99 phyla, including six major monophyletic units from the subdivision of the Proteobacteria, and amalgamation of the Candidate Phyla Radiation into a single phylum. Our taxonomy should enable improved classification of uncultured bacteria and provide a sound basis for ecological and evolutionary studies.

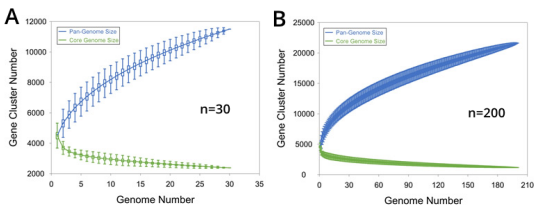
Nature Biotechnology 36:996–1004 (2018)



Nature Biotechnology 36:996–1004 (2018)

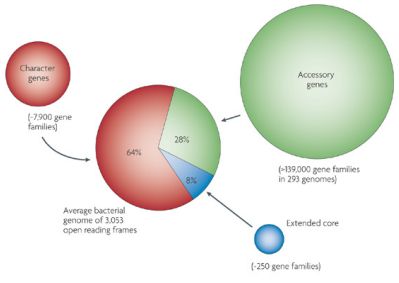
Introduction to the Pangenome

Gene Content of a Species



Of Terms in Biology: The Pan-Genome
by Christoph Weigel

In *'Small Things Considered'*
June 12, 2014 – schaechter.asmblog.org

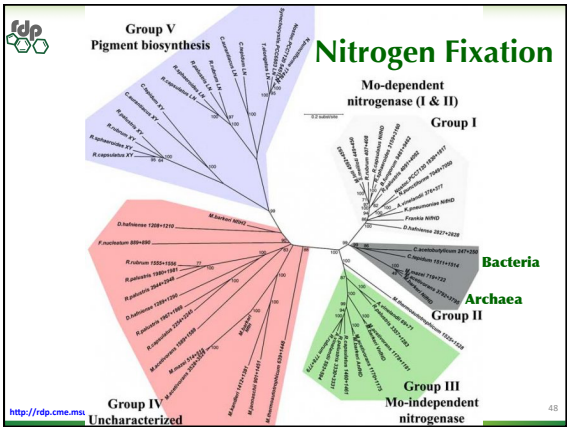


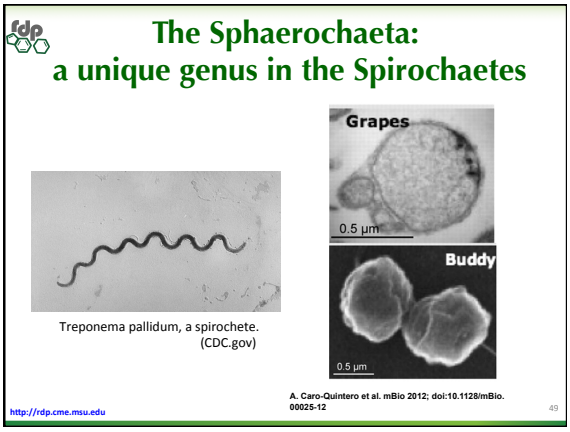
Horizontal gene transfer occurs more readily between closely related organisms

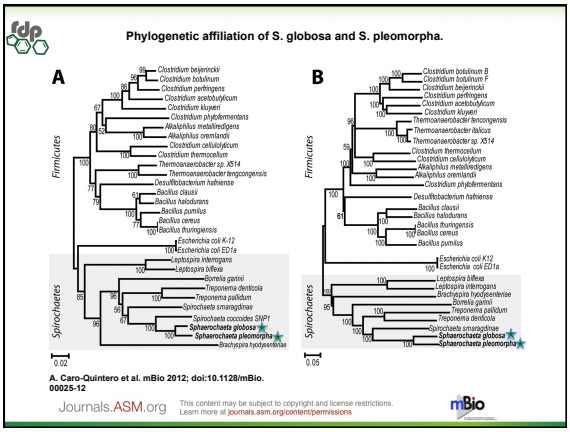
Nature Reviews | Microbiology

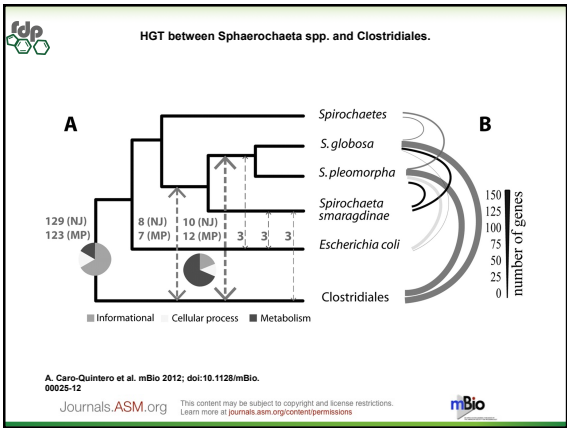
Horizontal Gene Transfer

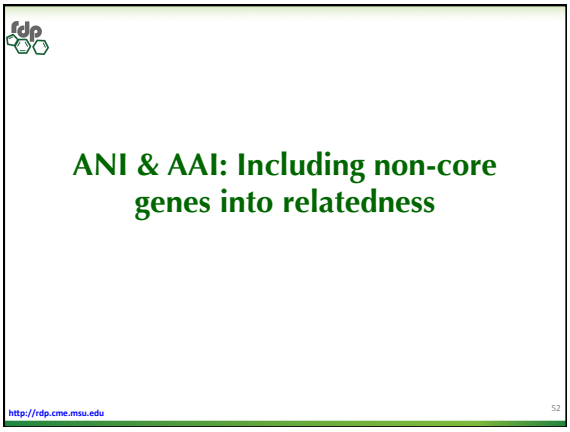
- Many important character traits have evolutionary history that is not coherent with core phylogeny.
- Pathogenicity, drug resistance are prime examples.
- Also more basic metabolic traits, e.g. substrate utilization, sulfate reduction, nitrogen fixation and many others have been horizontally transferred multiple times during evolution.

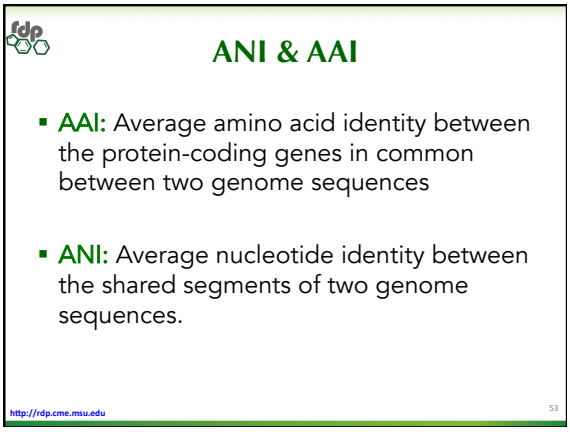

Nitrogen Fixation
Group V
Pigment biosynthesis
Mo-dependent
nitrogenase (I & II)
Group I
Bacteria
Archaea
Group II
Group III
Mo-independent
nitrogenase
Group IV
Uncharacterized
http://rdp.cme.msu
48


The Sphaerochaeta:
a unique genus in the Spirochaetes
Treponema pallidum, a spirochete.
(CDC.gov)
Grapes
0.5 μm
Buddy
0.5 μm
A. Caro-Quintero et al. mBio 2012; doi:10.1128/mBio.
00025-12
http://rdp.cme.msu.edu
49


Phylogenetic affiliation of S. globosa and S. pleomorpha.
A
B
Firmicutes
Spirochaetes
A. Caro-Quintero et al. mBio 2012; doi:10.1128/mBio.
00025-12
Journals.ASM.org
This content may be subject to copyright and license restrictions.
Learn more at journals.asm.org/content/permissions
mBio


HGT between Sphaerochaeta spp. and Clostridiales.
A
B
Spirochaetes
S.globosa
S.pleomorpha
Spirochaeta
smaragdinae
Escherichia coli
Clostridiales
129 (NJ)
123 (MP)
8 (NJ)
7 (MP)
10 (NJ)
12 (MP)
3
3
3
150
125
100
75
50
25
0
number of genes
Informational
Cellular process
Metabolism
A. Caro-Quintero et al. mBio 2012; doi:10.1128/mBio.
00025-12
Journals.ASM.org
This content may be subject to copyright and license restrictions.
Learn more at journals.asm.org/content/permissions
mBio


ANI & AAI: Including non-core
genes into relatedness
http://rdp.cme.msu.edu
52


ANI & AAI
AAI: Average amino acid identity between the protein-coding genes in common between two genome sequences
ANI: Average nucleotide identity between the shared segments of two genome sequences.
http://rdp.cme.msu.edu
53

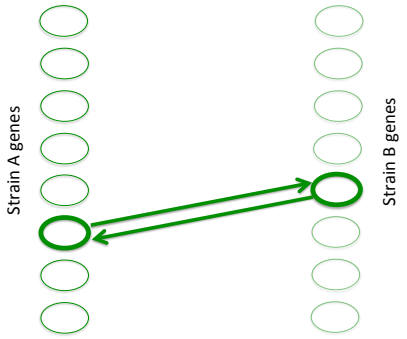
Average Identity Methodology

- Determine all orthologous pairs
- Measure the percent identity for each pair
- Take the average

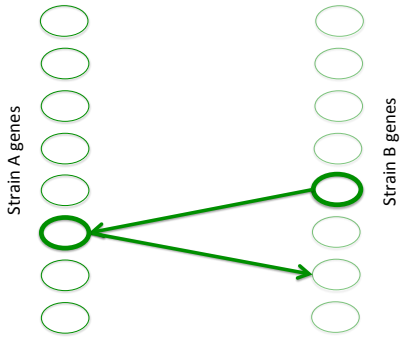
Need to find comparable genes for Average Identity methods

- Homologous:
  - The existence of shared ancestry between a pair of genes.
- Orthologous:
  - Inherited by two organisms from the same ancestral sequence. (Usually same function.)
- Paralogous:
  - Originally created by a duplication event within a single genome. (May have different functions.)

Reciprocal Best Matches - Likely Orthologs



Best matches not reciprocal - Potential Paralogs?



ANI & AAI: Pros and Cons

- Advantages of Average Identity:
  - Takes into account all related data
  - Can be used to classify organisms into existing or new clades
- Disadvantages of Average Identity:
  - Can not be used directly for phylogenetic analysis
  - Collapses multidimensional into a single distance measure

MiGA
Microbial Genomes Atlas

LM Rodriguez-R [@_lrr_] GT, S Gunturu MSU, C Luo GT,
JM Tiedje MSU, J Cole MSU, KT Konstantinidis GT

Georgia Tech
MICHIGAN STATE UNIVERSITY

W282–W288 Nucleic Acids Research, 2018, Vol. 46, Web Server issue Published online 14 June 2018
doi: 10.1093/nar/gky467

The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of *Archaea* and *Bacteria* at the whole genome level

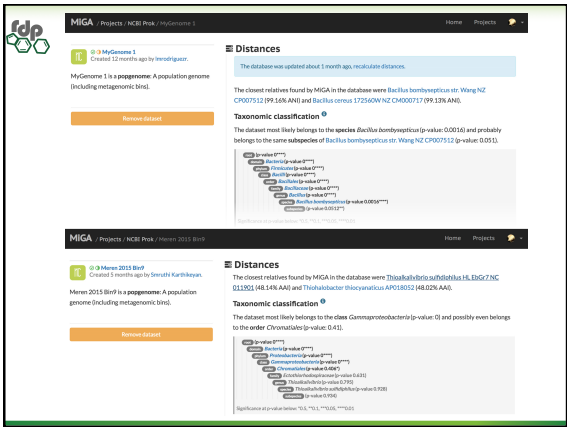
Luis M. Rodriguez-R[1], Santosh Gunturu[2], William T. Harvey[1], Ramon Rosselló-Mora[3], James M. Tiedje[2,4], James R. Cole[2] and Konstantinos T. Konstantinidis[1,5,*]

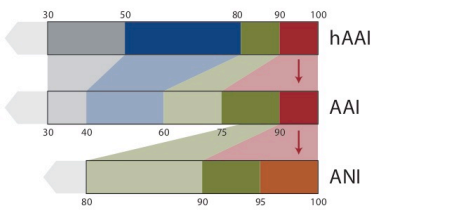
ABSTRACT

The small subunit ribosomal RNA gene (16S rRNA) has been successfully used to catalogue and study the diversity of prokaryotic species and communities but it offers limited resolution at the species and finer levels, and cannot represent the whole-genome diversity and fluidity. To overcome these limitations, we introduced the Microbial Genomes Atlas (MiGA), a webserver that allows the classification of an unknown query genomic sequence, complete or partial, against all taxonomically classified taxa with available genome sequences, as well as comparisons to other related genomes including uncultivated ones, based on the genome-aggregate Average Nucleotide and Amino Acid Identity (ANI/AAI) concepts. MiGA integrates best practices in sequence quality trimming and assembly and allows input to be raw reads or assemblies from isolate genomes, single-cell sequences, and metagenome-assembled genomes (MAGs). Further, MiGA can take as input hundreds of closely related genomes of the same or

http://rdp.cme.msu.edu

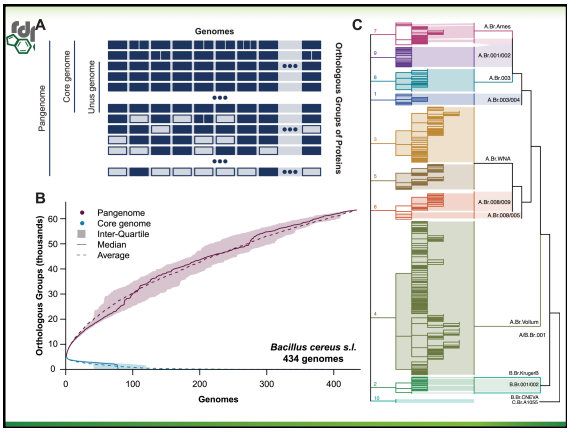
60



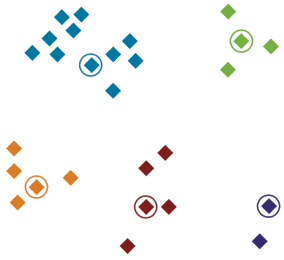


Hierarchical approach to genome classification

10



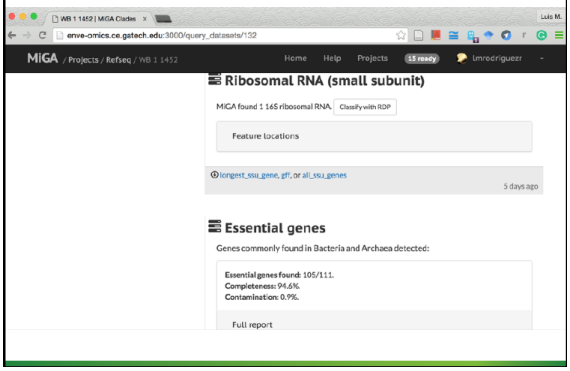
PAM + Silhouette

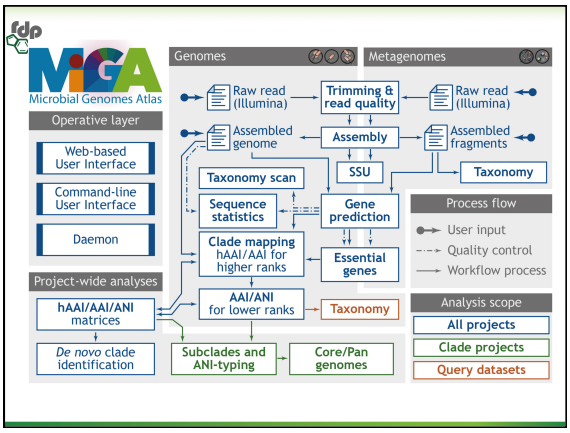


Medoid clustering

64

Many other functions in MiGA





Taxonomy in the Genomics Era

- A set of core genes have evolved together with little horizontal transfer and produce more reliable phylogenetic inferences than rRNA alone.
- Incorporating these data can improve our current taxonomy, especially for the more basal ranks.
- Methods such as Average Identity are better able to discriminate at the species and subspecies levels.
- A strictly hierarchical taxonomy can hide important (dis)similarities and in the future taxonomists may explore new structures that accommodate these extra information.

http://rdp.cme.msu.edu 67

Acknowledgements

http://rdp.cme.msu.edu http://fungene.cme.msu.edu
http://microbial-genomes.org

MSU Collaborators:
C. Titus Brown (UCD)
Sakti Pramanik
Yanni Sun
Jim Tiedje

GaTech Collaborators:
Kostas Konstantinidis
Miguel Rodrigues R.
Alex Luo

RDP Bioinformatics Group:
Benli Chai
Alex Fall
Jordan Fish (Google)
Mariah Gilman (Google)
Santosh Gunturu
Jiarong Guo
Donna McGarrell
Michael Okeefe (MSU)
Leo Tift
Qiong Wang (DuPont/Pioneer)
Ziye Xing (UCLA)
Tae-Kwon Lee (U Vienna)
All Past RDP Group members

All Current and Past Collaborators

Support from DOE, NSF, NIH, USDA, NIEHS