Microbes and Microbiome

Julie Segre, PhD

Senior Investigator,
National Human
Genome Research
Institute, NIH

JOHNS HOPKINS
MEDICINE
CONTINUING MEDICAL EDUCATION

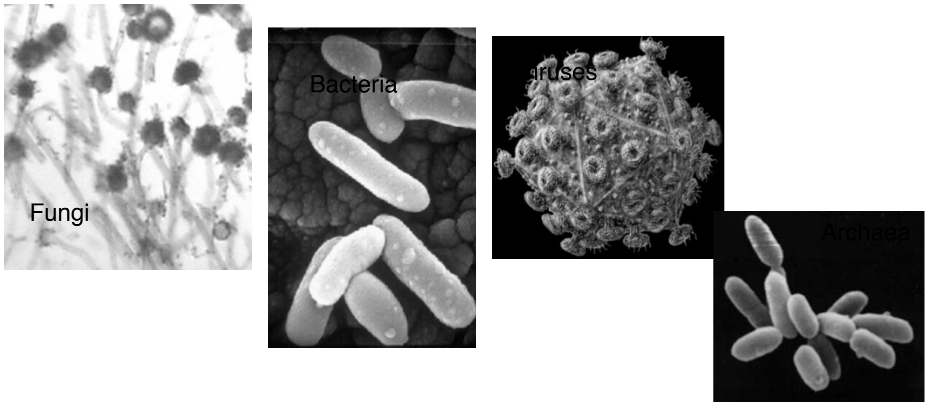
Current Topics in Genome Analysis 2012

Julia Segre

No Relevant Financial Relationships with Commercial Interests

2

Why the Human Microbiome?

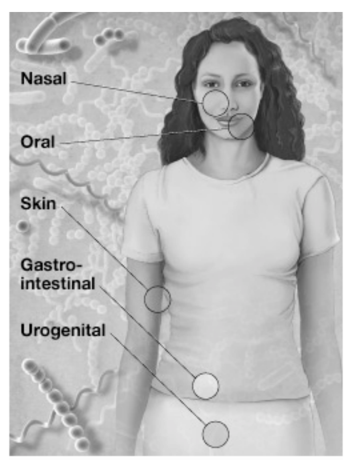


Each human cell has the same protein-encoding potential. Microbes are more diverse and dynamic than human genome.

3

Human Microbiome Project (HMP) Goals: Baseline to empower future clinical studies

Assess microbial diversity of 250 healthy individuals at 5 sites (gut, nasal, oral, vaginal and skin)



HMP

NIH Roadmap FOR MEDICAL RESEARCH

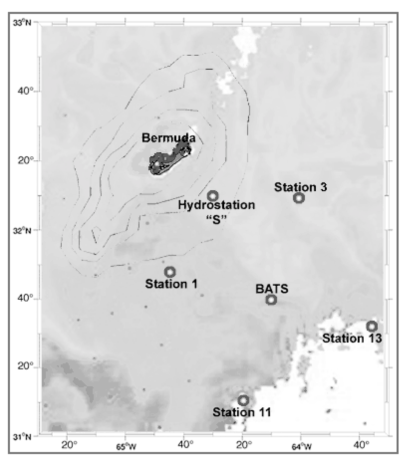
4

HMP Research Goals

- Sequence bacterial reference genomes
- Metagenomics, the analysis of the combined coding potential of a mixed population.
- Correlation of changes in microbial communities with disease states.
- Explore ethical, legal and social implications of this new field of research.

5

Microbial Diversity Studied in the Environment



Originally published in *Science* Express on 4 March 2004
Science 2 April 2004:
Vol. 304. no. 5667, pp. 66 – 74
DOI: 10.1126/science.1093857

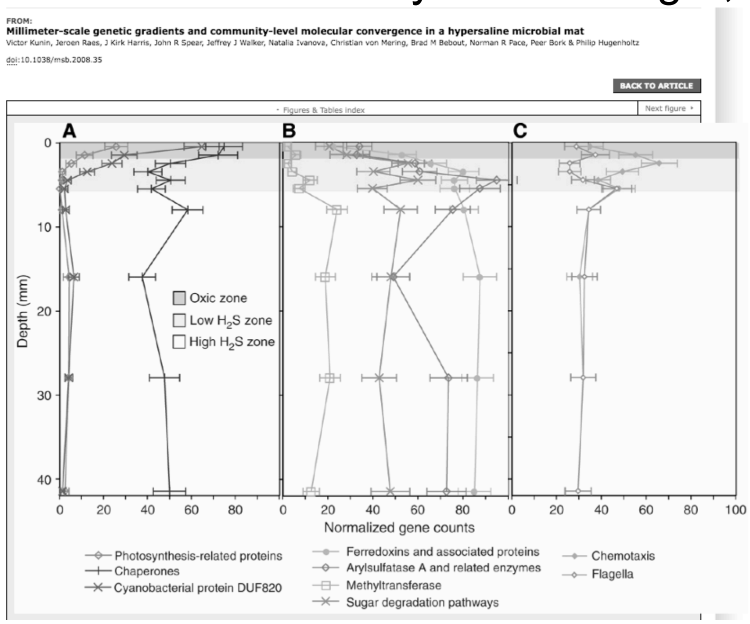
RESEARCH ARTICLES

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1*] Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

6

HyperSaline mat diversity Guerro Negro, MX



7

And human-environment diversity: shower heads across USA

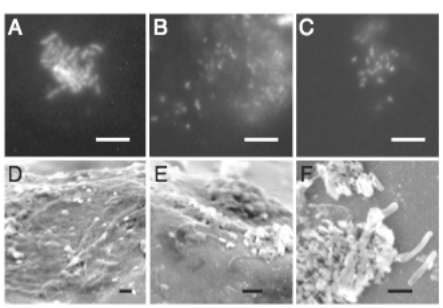
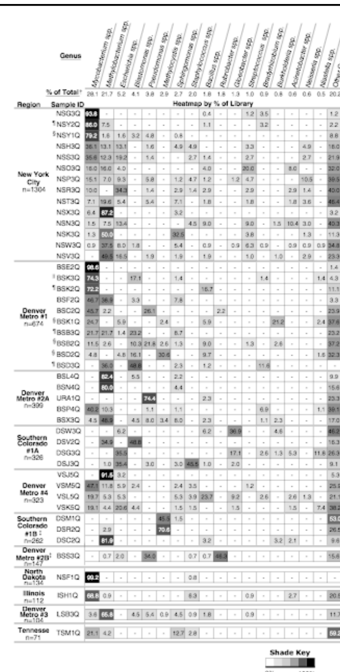


Fig. 1. Fluorescence and SEM images of showerhead biofilm. (*A–C*) Epifluorescence microscopy of biofilm samples stained with DAPI; scale bars, 10 μm. (*D–F*) SEM micrographs of increasing magnification of in situ showerhead

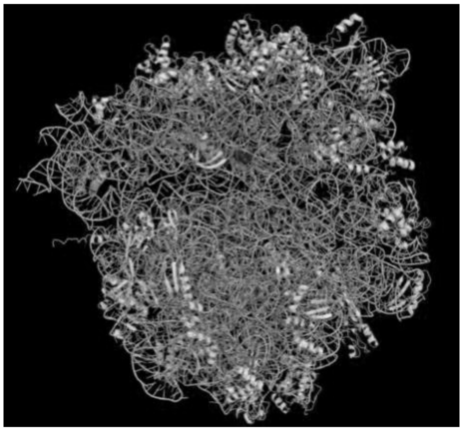


Opportunistic pathogens enriched in showerhead biofilms

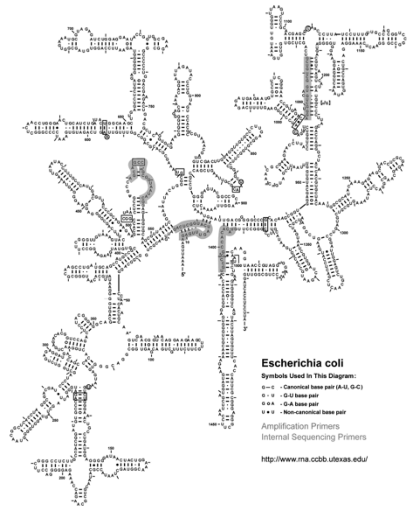
Leah M. Feazel[a], Laura K. Baumgartner[a], Kristen L. Peterson[a], Daniel N. Frank[a], J. Kirk Harris[b], and Norman R. Pace[a,1]

8

TOPIC 1.
Bacterial Diversity: 16S rRNA gene



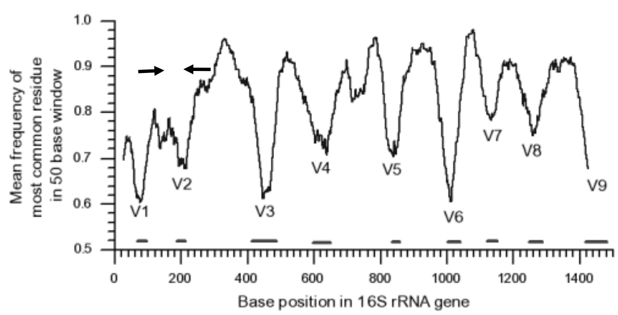
Orange= rRNA ;
Blue = small subunit proteins
Green = large subunit proteins



Secondary Structure: 16S small subunit ribosomal RNA

9

Bacterial Load:
qPCR wth primers in conserved regions



16S gene was amplified using forward primer 63F (5-GCAGGCCTAACACATGCAAGTC-3) and reverse primer 355R (5-CTGCTGCCTCCCGTAGGAGT-3) to yield a 292-bp PCR product. (Castillo M…Gasa J…2006)

10

Calculating Bacterial Load

Human DNA	Bacterial DNA 300 pg		30 pg		3 pg	
	Ct	copy #	Ct	copy #	Ct	copy #
0 g	17.85	54924.50	20.92	6951.93	24.24	743.61
0.3 ng	17.78	57575.00	20.93	6905.28	24.42	658.74

C_t of qPCR of bacterial DNA to calculate relative bacterial counts of each sampling method. The function used to calculate copy number is as follows: $C_t = -3.42x + 34.06$; $R^2 = 0.99$; where C_t = threshold cycle and x = log copy number.

- Swab yields 10,000 bacteria/cm^2
- Scrape yields 50,000 bacteria/cm^2
- Biopsy yields 1,000,000 bacteria/cm^2

Grice et al, Genome Research 2008

11

How to study microbial diversity

- Fingerprinting: cheapest, but very limited (Anderson and Cairney, Envir Microbiol 2004)
- PhyloChip or GeoChip: like microarray,

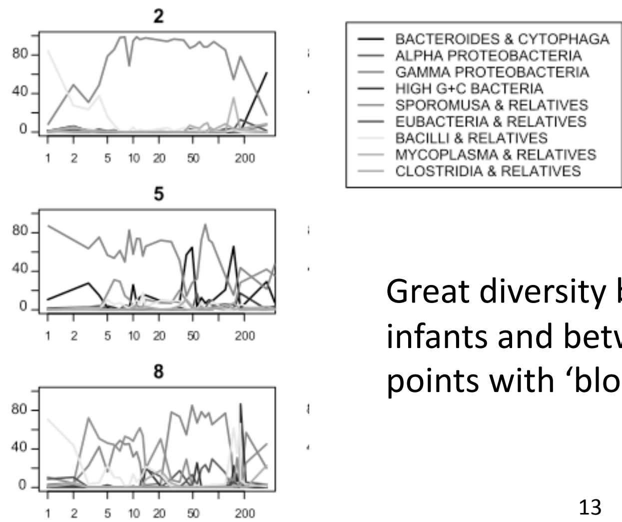
  will be powerful to assess changes in diversity (when predominate species enumerated) but like all Chips will never find UNIQUE species (Wilson Appl Environ Microbiol 2002 and He ISME J 2007)
- Sequencing: taxonomic classification and function, dynamic range and compare multiple complex samples.

For a SMALL study, SEQUENCE is limiting;
For a LARGE study, BIOINFORMATICS is limiting.

12

PhyloChip to examine intestinal microbiota in first year of life

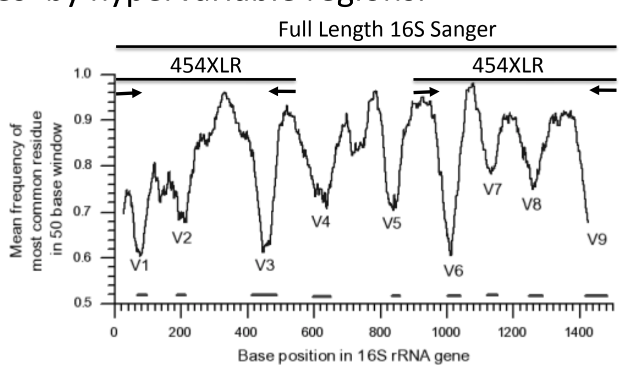
Palmer, Relman, Brown 2007 PLOS Bio



Great diversity between infants and between time points with 'blooms'

13

16S Bacterial rRNA gene conserved, variable and hypervariable regions. Primers put into conserved regions, phylogeny determined by variable regions, 'species' by hypervariable regions.



PRIMERS SIGNIFICANTLY DETERMINE MICROBIAL DIVERSITY RECOVERED. CAN NOT A PRIORI COMPARE YOUR DATASET TO SOMEONE ELSES IF DIFFERENT PRIMER OR AMPLIFICATION CONDITIONS WERE USED

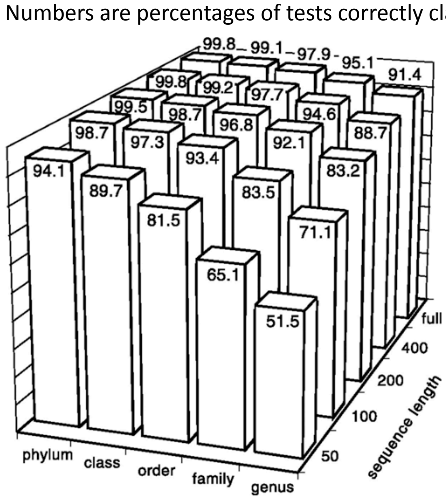
14

How many reads do you need? Depends on site diversity (slide 34,35) and taxonomic aim of study

- Sanger: Full-length 1.6 kb gives you a match to a cultured isolate, 384 sequences/sample
- 454/Roche: 400 bp V1-V3 or V6-V9 region, allows you to assign to genera, 3,000 reads/ sample
- Illumina: 100 bp tags (2x150 bp on MiSeq) identify bacterial genera, not species (and great for whole genome bacterial sequencing)

15

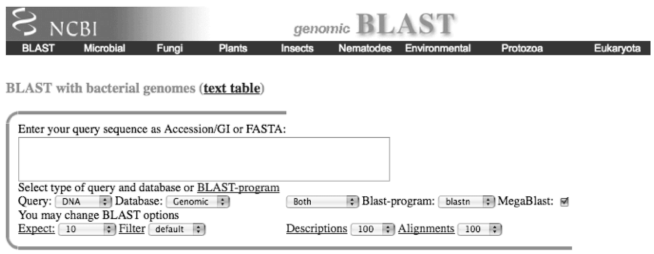
FIG. 1. Overall classification accuracy by query size (exhaustive leave-one-out testing using the Bergey corpus). Numbers are percentages of tests correctly classified.

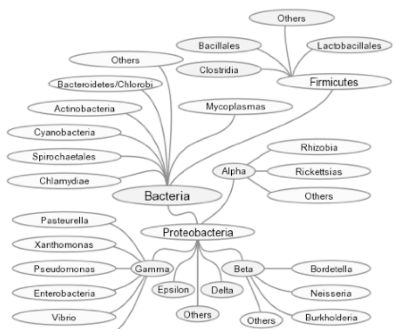


Applied and Environmental Microbiology, August 2007, p. 5261-5267, Vol. 73, No. 16
Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy , Qiong Wang,[1] George M. Garrity,[1,2] James M. Tiedje,[1,2] and James R. Cole[1]
Also see: Liu, DeSantis, Andersen and Knight, NAR 2008

16

How to identify a bacterial sequence and align sequences?





Matches MANY sequences.
Maybe your sequence is previously UNCULTURED?

17

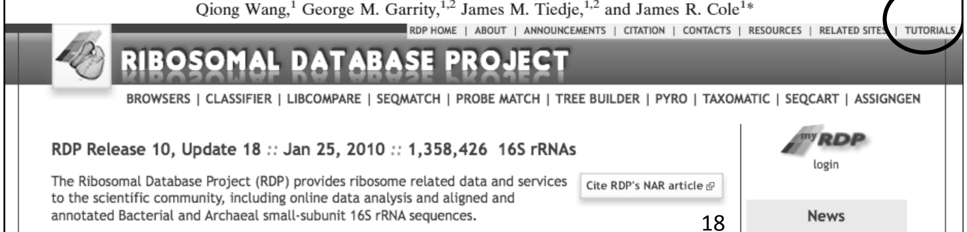
RDP Database http://rdp.cme.msu.edu/

- RDP 10.18 consists of 920,643 aligned and annotated 16S rRNA sequences. Naïve Baysian classifier based on Bergey's taxonomy. (Note: other taxonomies such as Euzeby and NCBI exist).
- Tools: RDP classifier, Seqmatch, Probematch

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Aug. 2007, p. 5261–5267 Vol. 73, No. 16
0099-2240/07/$08.00+0 doi:10.1128/AEM.00062-07
Copyright © 2007, American Society for Microbiology. All Rights Reserved.

Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy▽†

Qiong Wang,[1] George M. Garrity,[1,2] James M. Tiedje,[1,2] and James R. Cole[1]*



RDP Pyrosequencing Pipeline

About the RDP's Pyrosequencing Pipeline

The Ribosomal Database Project's Pyrosequencing Pipeline aims to simplify the processing of large 16s rRNA sequence libraries obtained through pyrosequencing. This site processes and converts the data to formats suitable for common ecological and statistical packages such as SPADE, EstimateS, and R.

Data Processing Steps:

- **Pipeline Initial Process** - sort and trim the raw reads, filter low quality sequences.
- **Aligner** - align sequences using the fast, secondary-structure aware Infernal aligner.
- **Complete Linkage Clustering** - cluster sequences by the complete-linkage clustering method.

Formats for Common Programs:

- **SPADE Formatter** - make a SPADE compatible input format.
- **R Formatter** - make a R compatible input format.
- **EstimateS Formatter** - make an EstimateS compatible input format. Can also be used with PAST.
- **Mothur: Column Distance Matrix** - create a column distance matrix compatible with Mothur.
- **Mothur: Phylip Distance Matrix** - create a matrix and sample group file compatible with Mothur's LIBSHUFF function.

Analysis Tools:

- **Shannon & Chao1 Index** - calculate Shannon Index & Chao1 estimator from a single sample file.
- **Rarefaction** - calculate Rarefaction from a single sample file.
- **RDP Classifier** - assign 16S rRNA sequences to our taxonomical hierarchy.
- **RDP LibCompare** - compare two sequence libraries using the RDP Classifier

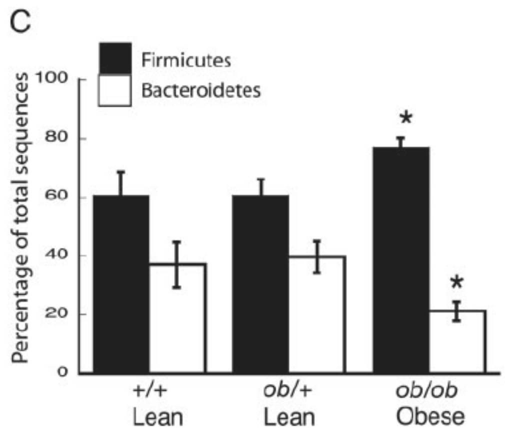
19

Host Sequence Contamination

- Important when dealing with human-derived samples
- Ethically, projects should attempt to filter human subject sequences before submission to public databases
- This is actually harder than it sounds

20

Gordon: lean versus obese mice



Obesity (in mice) correlates with an increase in Firmicutes/ Bacteriodetes ratio

Ley, ... Gordon PNAS 2005

21

Also true in humans

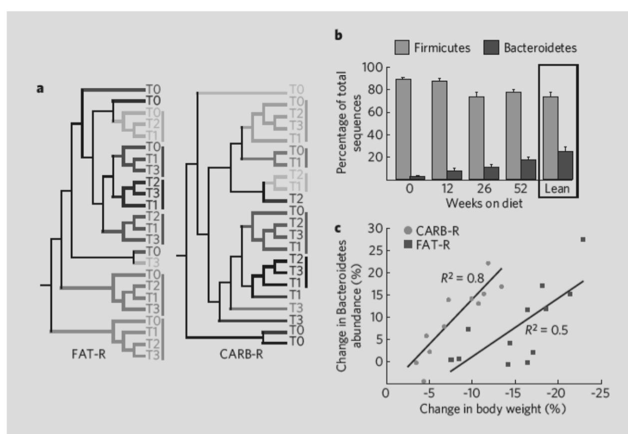


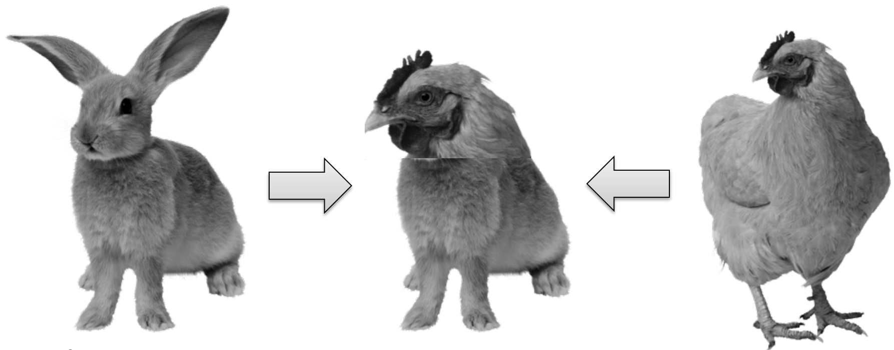
Figure 1 | Correlation between body-weight loss and gut microbial ecology. a, Clustering of 16S ribosomal RNA gene sequence libraries of faecal microbiota for each person (in different colours) and time point in diet therapy (T0, baseline; T1, 12 weeks; T2, 26 weeks; T3, 52 weeks) in the two diet-treatment groups (fat restricted, FAT-R; carbohydrate restricted, CARB-R), based on UniFrac analysis of the 18,348-sequence phylogenetic tree. **b,** Relative abundance of Bacteroidetes and Firmicutes. For each time point, values from all available samples were averaged (*n* was 11 or 12 per time point). Lean-

NATURE|Vol 444|21/28 December 2006

Ruth E. Ley, Peter J. Turnbaugh, Samuel Klein, Jeffrey I. Gordon

22

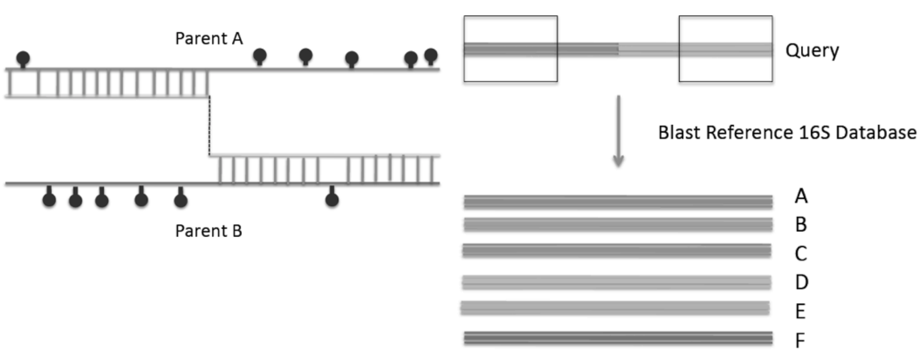
Chimeras: PCR generated (template switching)



Evaluate Accuracy:
– True Positives (TP): artificial chimeras flagged
– False Positives (FP): reference (non-chimera) flagged

23

How Do Chimeras Occur? Incomplete extension of PCR, Template Switching at Conserved Regions



24

ChimeraSlayer Detection Program

http://microbiomeutil.sourceforge.net

Compatible with near-full length Sanger sequences and shorter 454-FLX sequences (~500 bp).
Given a candidate chimera query sequence, candidate parental sequences of a chimera are identified by a homology search. The ends of the query sequence are searched separately to identify candidate parental sequences. ... Those candidate parents identified by this alignment fitting procedure are tested in all pairwise combinations as potential parents of the putative chimeric query sequence using a modified Bellerophon-like algorithm.

Microbiome Utilities Portal of the Broad Institute

Genome Res. 2011 Mar;21(3):494-504.

25

How to align sequences?

```
Query  225   ATTAGCTAGTTGGTAAGGTAACGGCT---TACCAAGGC-A-ACG-ATGCATAGCC-GACC  277
             |||||||||| |||  ||||||||||   | || |||| |  || || | ||||| |  |
Sbjct  212   ATTAGCTAGTAGGTGGGGTAACGGCTCCATCCCTAGGCGAGCCGAATCCTTAGCCTGGTC  271

Query  278   TGAGAGG-GTGATCGGCCACACTGGAACTGAG-ACACGGTCCAGACTCCTACGGAGGCA  335
             |||||||  ||| | |||||||||| |||||| ||||||||||||||||||||||||||
Sbjct  272   TGAGAGGAATGACCAGCCACACTGGGACTGAGAACACGGTCCAGACTCCTACGGAGGCA  331
```

WANT TO USE A PROGRAM THAT TAKES 16S STRUCTURE INTO CONSIDERATION. GAPS ARE MORE LIKELY IN LOOPS THAN STEMS

26

NAST and NASTier

fixed-width character alignment format

W394–W399 Nucleic Acids Research, 2006, Vol. 34, Web Server issue
doi:10.1093/nar/gkl244

NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes

T. Z. DeSantis[1,4,*], P. Hugenholtz[2], K. Keller[5,4], E. L. Brodie[1], N. Larsen[3], Y. M. Piceno[1], R. Phan[1,4] and G. L. Andersen[1,4,*]

NAST-iEr

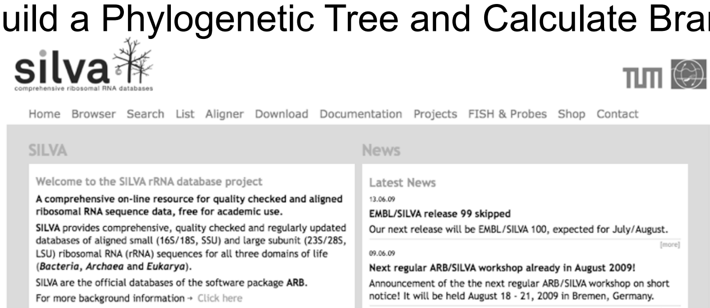
The NAST-iEr alignment utility (download) aligns a single raw nucleotide sequence against one or more NAST formatted sequences.

The alignment algorithm involves global dynamic programming alignment to a fixed template sequences without any end-gap penalty similar in principle to Pearson's align0 program with a fixed template sequence containing arbitrary gap positions.

27

Silva Database (ARB): http://www.arb-silva.de/

Build a Phylogenetic Tree and Calculate Branch Length



Pruesse, E., C. Quast, K. Knittel, B. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner.
SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.
Nuc. Acids Res. 2007; Vol. 35, No. 21, p. 7188-7196

Nucleic Acids Research, 2004, Vol. 32, No. 4 ***1363–1371***
DOI: 10.1093/nar/gkh293

ARB: a software environment for sequence data

Wolfgang Ludwig*, Oliver Strunk, Ralf Westram, Lothar Richter, Harald Meier[1], Yadhukumar, Arno Buchner, Tina Lai, Susanne Steppi, Gangolf Jobb[1], Wolfram Förster[1], Igor Brettske, Stefan Gerber, Anton W. Ginhart[1], Oliver Gross, Silke Grumann[1], Stefan Hermann[1], Ralf Jost[1], Andreas König[1], Thomas Liss[1], Ralph Lüßmann[1], Michael May[1], Björn Nonhoff[1], Boris Reichel[1], Robert Strehlow[1], Alexandros Stamatakis[1], Norbert Stuckmann[1], Alexander Vilbig[1], Michael Lenke[1], Thomas Ludwig[2], Arndt Bode[1] and Karl-Heinz Schleifer

28

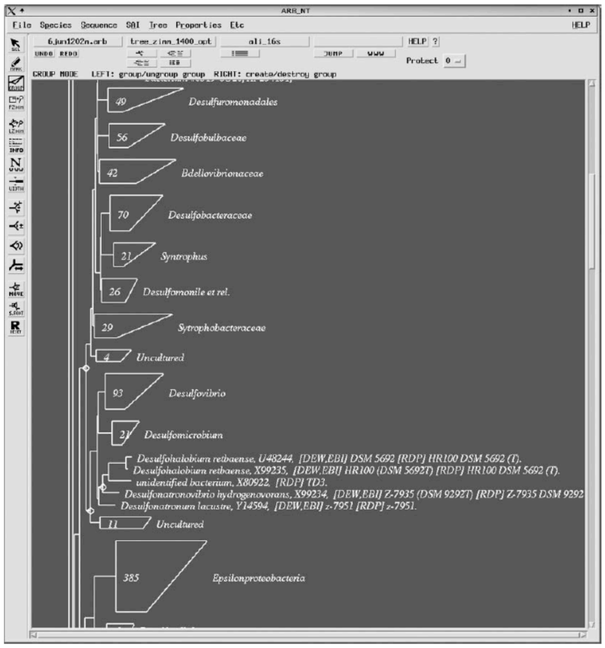


Figure 3. The ARB main window showing part of an ARB parsimony-generated dendrogram. The rectangles represent `online compressed' monophyletic groups which can be `unfolded' by mouse click. Database ®eld entries such as taxonomic name, public database accession number and strain designation as reported in EMBL (1), RDP (3) and the European rRNA databases (DEW) (4,5) are visualized at the terminal nodes of the `unfolded' Desulfohalobiaceae.

29

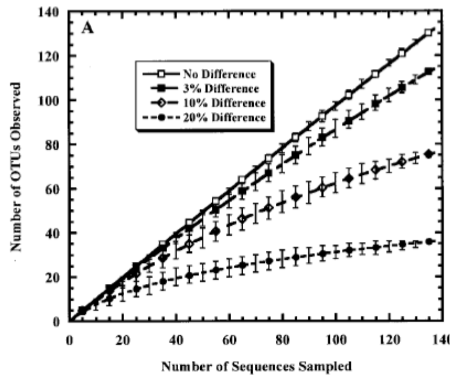
Defining Taxonomic Groups by sequence similarity: DOTUR, SONS and MOTHUR http://www.mothur.org

30

OTU: Operational Taxonomic Unit

Cluster Sequences Based on Furthest Joining Method; i.e. Every sequence is at most X% different from every other sequence in the group

% identity within group determines the number of OTUs produced. This should be done on the TOTAL dataset. Most experiments classify at the 97% or 99% identity.



Applied and Environmental Microbiology, Mar. 2005, p. 1501–1506
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.3.1501–1506.2005

Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness

Patrick D. Schloss and Jo Handelsman*

31

Comparing Bacterial Diversity: Community Membership & Structure

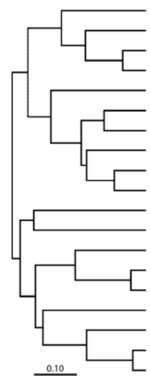
Grp A	Grp B
60	50
34	50
2	0
2	0
2	0

Community Membership
(Categories of fruit in common)
= 2/5= 0.4

Community Structure
(Pieces of fruit in common)
= ~ 0.9

32

Community Membership:
Pups are most like their mothers



Animal Name	Sex	Genotype	Sequences Sampled	Chao1 Richness Estimate (95% CI)
M3-3	Female	+/+	192	>103 (71-195)
Mother 3	Female	ob/+	266	>180 (142-257)
M3-1	Male	+/+	189	>189 (104-428)
M3-2	Male	ob/+	196	>117 (87-189)
M1-1	Female	ob/ob	235	>262 (147-557)
M1-3	Male	ob/+	229	>131 (97-208)
M1-2	Male	ob/ob	194	>106 (81-163)
Mother 1	Female	ob/+	402	>217 (166-316)
M1-4	Female	+/+	182	95 (74-149)
M1-5	Female	ob/+	183	>160 (113-264)
M2A-1	Female	ob/ob	123	94 (69-157)
M2B-2	Male	ob/ob	262	>126 (107-167)
M2A-2	Male	ob/+	99	58 (46-90)
M2A-3	Female	ob/+	285	150 (121-210)
M2A-4	Female	ob/+	194	>163 (116-267)
M2B-1	Male	+/+	313	105 (87-148)
M2B-3	Male	ob/ob	234	>162 (113-275)
M2B-4	Male	ob/ob	279	>157 (112-261)
Mother 2	Female	ob/+	258	>251 (184-381)

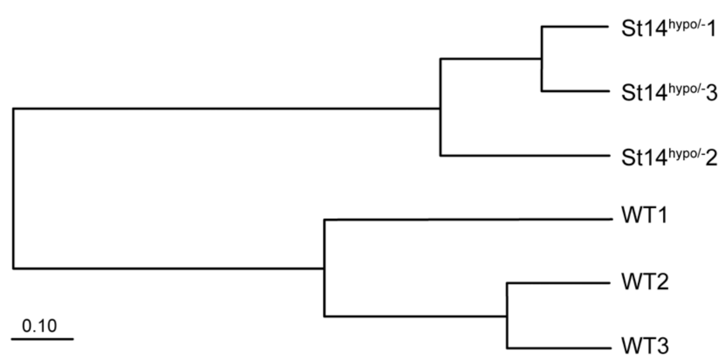
APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 2006, p. 6773–6779 Vol. 72, No. 10
0099-2240/06/$08.00+0 doi:10.1128/AEM.00474-06

Introducing SONS, a Tool for Operational Taxonomic Unit-Based Comparisons of Microbial Community Memberships and Structures

Patrick D. Schloss† and Jo Handelsman*

33

Community Structure:
Pups cluster according to genotype



Scharschmidt et al. JID 2009

34