

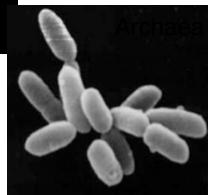
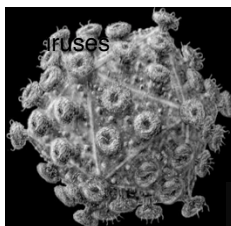
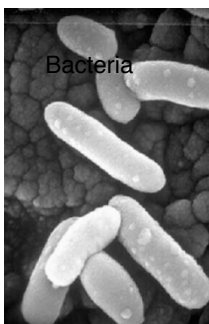
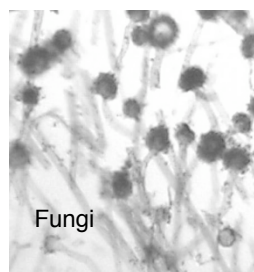


Microbes and Microbiome

Julie Segre, PhD
Senior Investigator,
National Human
Genome Research
Institute, NIH



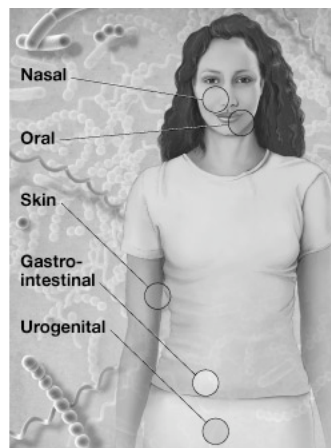
Why the Human Microbiome?



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

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)



 NIH Roadmap FOR MEDICAL RESEARCH

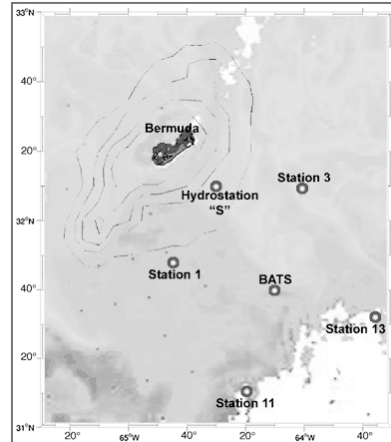
3

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.

4

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

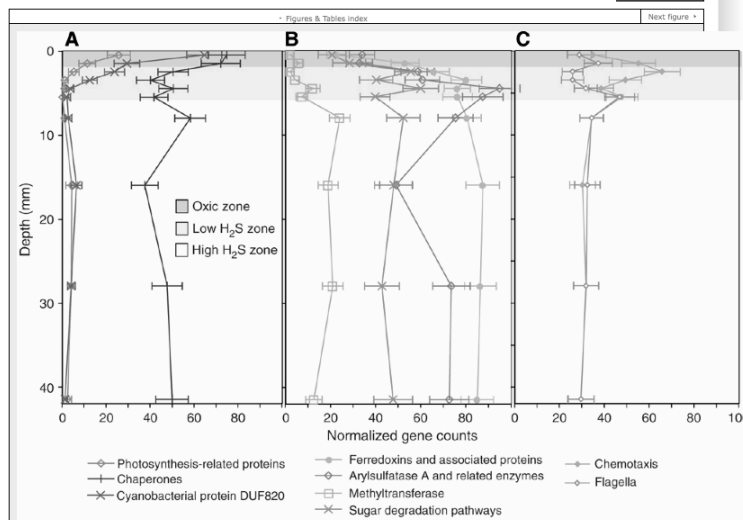
Environmental Genome Shotgun Sequencing of the Sargasso Sea

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 Neelson,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. Smith¹

5

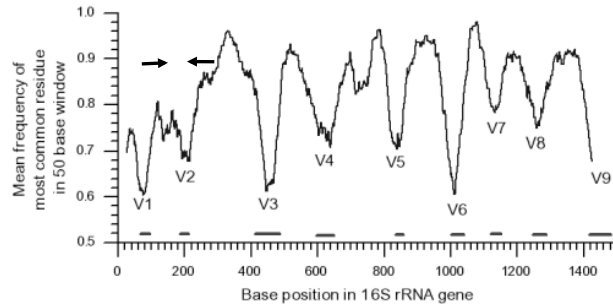
HyperSaline mat diversity Guerrero Negro, MX

FROM: **Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat**
 Victor Krumholz, Jeroen Raaij, J Kirk Harris, John R Soer, Jeffrey J Walker, Natalia Ivanova, Christian von Marburg, Brad M Boboud, Norman R Pace, Peer Bork & Philip Hugenholtz
 doi:10.1038/nmb.2008.35



6

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)

9

Calculating Bacterial Load

Human DNA	300 pg		Bacterial DNA 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/cm²
- Scrape yields 50,000 bacteria/cm²
- Biopsy yields 1,000,000 bacteria/cm²

Grice et al, Genome Research 2008

10

How to study microbial diversity

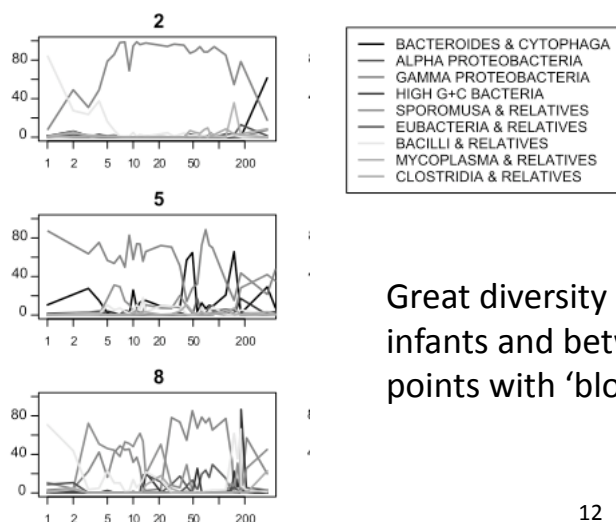
- Fingerprinting: cheapest, but very limited (Anderson and Cairney, *Envir Microbiol* 2004)
- PhyloChip or GeoChip: like microarray,
will be powerful to assess changes in diversity (when predominate species enumerated) but like all Chips will never find UNIQUE species (Wilson *Appl Environ Microbiol* 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.

11

PhyloChip to examine intestinal microbiota in first year of life

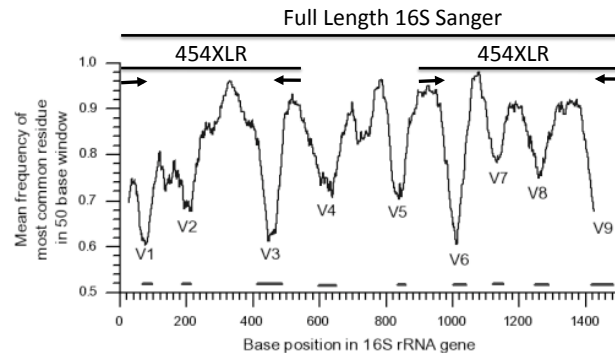
Palmer, Relman, Brown 2007 *PLOS Bio*



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

12

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 ELSE'S IF DIFFERENT PRIMER OR AMPLIFICATION CONDITIONS WERE USED

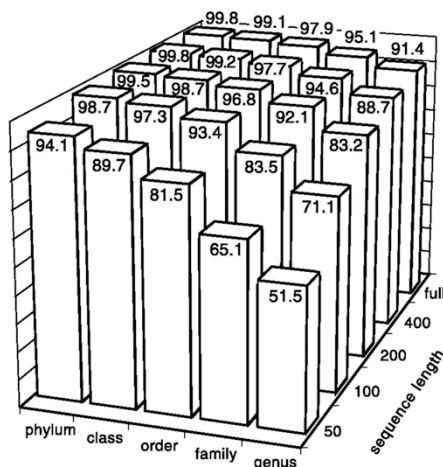
13

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, 10-50,000 reads/sample
- Solexa/Illumina: 100 bp tags still too small to identify bacterial genera (but great for whole genome bacterial sequencing)

14

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. Cole¹
 Also see: Liu, DeSantis, Andersen and Knight, NAR 2008

DEFINITION *Staphylococcus aureus* strain 16S ribosomal RNA gene,

```

1 ggcgcgcgtt atcgtgtggc gcggtcctaa tacatgcaag tcgagcgaac ggacgagaag
61 cttctctgat gttagcggcg gacgggtgag taacacgtgg ataacctacc tataagactg
121 ggataacttc ggaaaccgg agctaatacc ggataatatt ttgaaccgca tgggtcaaaa
181 gtgaaagacg gtccttgtgt cacttataga tggatccgcg ctgcattagc tagttggtaa
241 ggtaacggct taccaaggca acgatgcata gccgacctga gagggtgatc ggccacactg
301 gaactgagac acggtccaga ctctacggg aggcagcagt agggaatcct ccgcaatggg
361 cgaaagcctg acggagcaac ccgcgtgagt gatgaaggtc ttcggatcgt aaaactctgt
421 tattagggaa gaacatatgt gtaagtaact gtgcacatct tgacggtact aatcagaag
481 ccacggctaa ctacgtgcca gcagccgagg taatacgtag gtggcaagcg ttatccggaa
541 ttattgggag taaagcgcgc gtaggcgggt ttttaagtct gatgtgaaag cccacggctc
601 aaccgtggag ggtcattgga aactgaaaaa cttgagtga gaagaggaaa gtggaattcc
661 atgtgtagcg gtgaaatcgc cagagatatg gaggaacacc agtggcgaag gcgactctcg
721 gtctgtaact gacgctgatg tgcgaaagcg tggggatcaa acaggattag ataccctggt
781 agtccacgcc gtaaacgatg agtgctaagt gttagggggt ttccgcccct tagtgctgca
841 gctaacgcat taagcactcc gcctggggag tacgaccgca aggttgaaac tcaaaggaa
901 tgacggggac ccgcacaagc ggtggagcat gtggtttaat tcgaagcaac gccaaacctt
961 accaaatcct gacatccttt gacaactcta gagatagagc cttccccttc gggggacaaa
1021 gtgacagggt gtgcatggtt gtcgtcagct cgtgctgtga gatgttgggt taagtcccgc
1081 aacgagcgca acccttaagc ttagttgcca tcattaagtt gggcactcta agttgactgc
1141 cggtgacaaa ccggaggaag gtggggatga cgtcaaatca tcatgcccct tatgatttgg
1201 gctacacacg tgtacaatg gacaatacaa agggcagcga aaccgagagg tcaacaaatc
1261 ccataaagtt gttctcagtt cggattgtag tctgcaactc gactacatga agctggaatc
1321 gctagtaatc gtatagcagc atgctacggg gaatacgttc ccgggtcttg tacacaccgc
1381 ccgtcacacc acgagagttt gtaacacccc aagccggtgg agtaaccttt tagggctagc
1441 cggaaggtgg g
    
```


How to identify a bacterial sequence and align sequences?

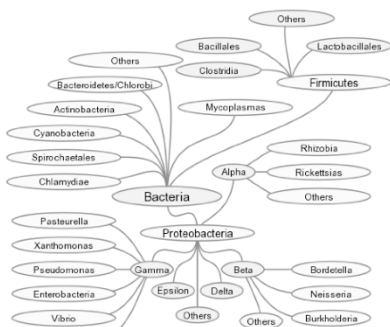


BLAST with bacterial genomes ([text table](#))

Enter your query sequence as Accession/GI or FASTA:

Select type of query and database or BLAST-program
 Query: Database: Both: Blast-program: MegaBlast:

You may change BLAST options
 Expect: Filter: Descriptions: Alignments:



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 Bayesian classifier based on Bergey's taxonomy. (Note: other taxonomies such as Euzebey and NCBI exist).
- Tools: RDP classifier, Seqmatch, Probematch

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Aug. 2007, p. 5261-5267
 0099-2240/07/508.00+0 doi:10.1128/AEM.00062-07
 Copyright © 2007, American Society for Microbiology. All Rights Reserved.

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*}



RIBOSOMAL DATABASE PROJECT

BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE BUILDER | PYRO | TAXOMATIC | SEQCART | ASSIGNGEN

RDP Release 10, Update 18 :: Jan 25, 2010 :: 1,358,426 16S rRNAs

The Ribosomal Database Project (RDP) provides ribosome related data and services to the scientific community, including online data analysis and aligned and annotated Bacterial and Archaeal small-subunit 16S rRNA sequences.

Cite RDP's NAR article [↗](#)



News

18

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.

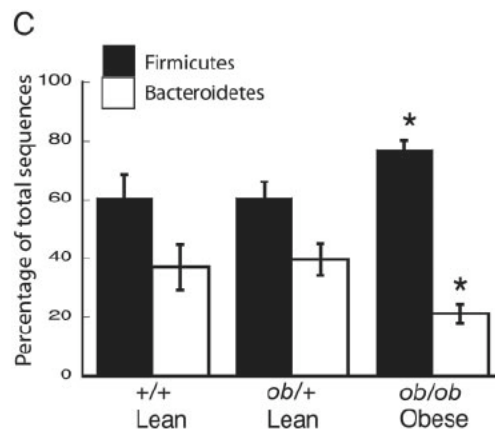
Formats for Common Programs:

- 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

Gordon: lean versus obese mice



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

Ley, ... Gordon PNAS 2005

Also true in humans

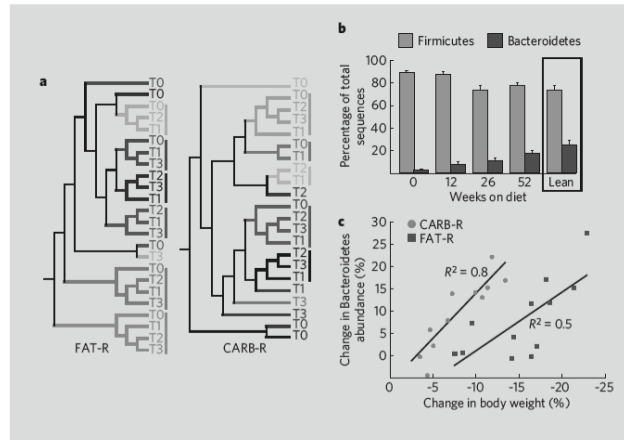


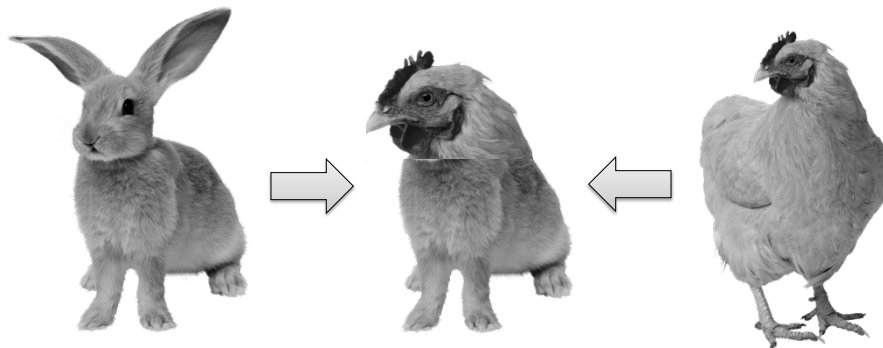
Figure 1 | Correlation between body-weight loss and gut microbial ecology. **a**, Clustering of 16S ribosomal RNA gene sequence libraries of faecal microbiota for each person (in different colours) and time point in diet therapy (T0, baseline; T1, 12 weeks; T2, 26 weeks; T3, 52 weeks) in the two diet-treatment 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

21

Chimeras: PCR generated (template switching)

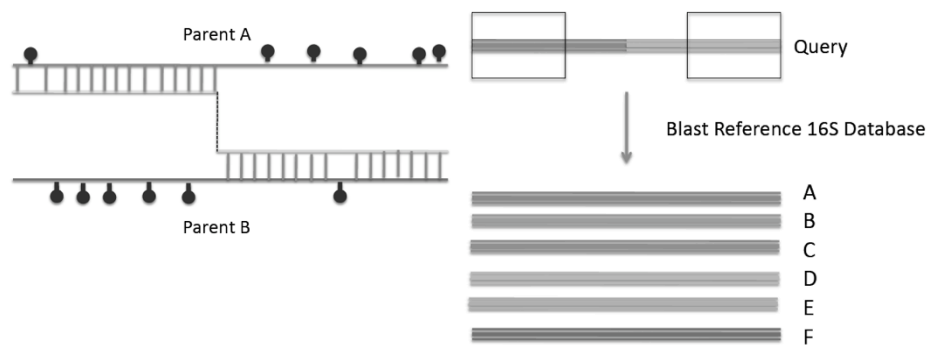


Evaluate Accuracy:

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

22

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



23

Pintail Chimera Detection Program

Given a dubious query sequence and a trusted reference sequence with computed average divergence, there is an expected divergence profile. Pintail identifies sequence anomalies. Requires checking sequences individually. Primarily used at RDP. Reimplemented by Broad Institute as WigeoN.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2005, p. 7724-7736
0099-2240/05/\$08.00+0 doi:10.1128/AEM.71.12.7724-7736.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Vol. 71, No. 12

At Least 1 in 20 16S rRNA Sequence Records Currently Held
in Public Repositories Is Estimated To Contain
Substantial Anomalies

Kevin E. Ashelford,^{1*} Nadia A. Chuzhanova,³ John C. Fry,¹ Antonia J. Jones,²
and Andrew J. Weightman¹

24

Bellerophon chimera checker <http://greengenes.lbl.gov/>

...weighted the likelihood of a sequence being chimeric according to the similarity of the parent sequences. The more distantly related the parent sequences were to each other relative, the greater the likelihood that the inferred chimera was real.

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 14 2004, pages 2317-2319
doi:10.1093/bioinformatics/bth226



Bellerophon: a program to detect chimeric sequences in multiple sequence alignments

Thomas Huber^{1,*}, Geoffrey Faulkner¹ and Philip Hugenholtz²

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, July 2006, p. 5069-5072
0099-2240/06/\$08.00+0 doi:10.1128/AEM.03006-05
Copyright © 2006, American Society for Microbiology. All Rights Reserved.

Vol. 72, No. 7

Greengenes, a Chimera-Checked 16S rRNA Gene Database
and Workbench Compatible with ARB

T. Z. DeSantis,¹ P. Hugenholtz,² N. Larsen,³ M. Rojas,⁴ E. L. Brodie,¹ K. Keller,⁵
T. Huber,⁶ D. Dalevi,⁷ P. Hu,¹ and G. L. Andersen^{1*}

25

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



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. Hugenholtz², K. Keller^{5,4}, E. L. Brodie¹, N. Larsen³, Y. M. Piceno¹, R. Phan^{1,4} and G. L. Andersen^{1,4,*}

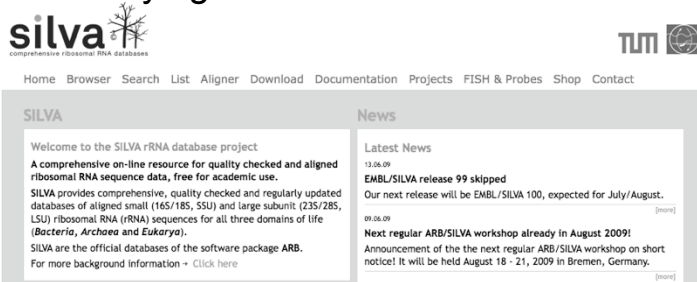
NAST-iEr

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.

29

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.

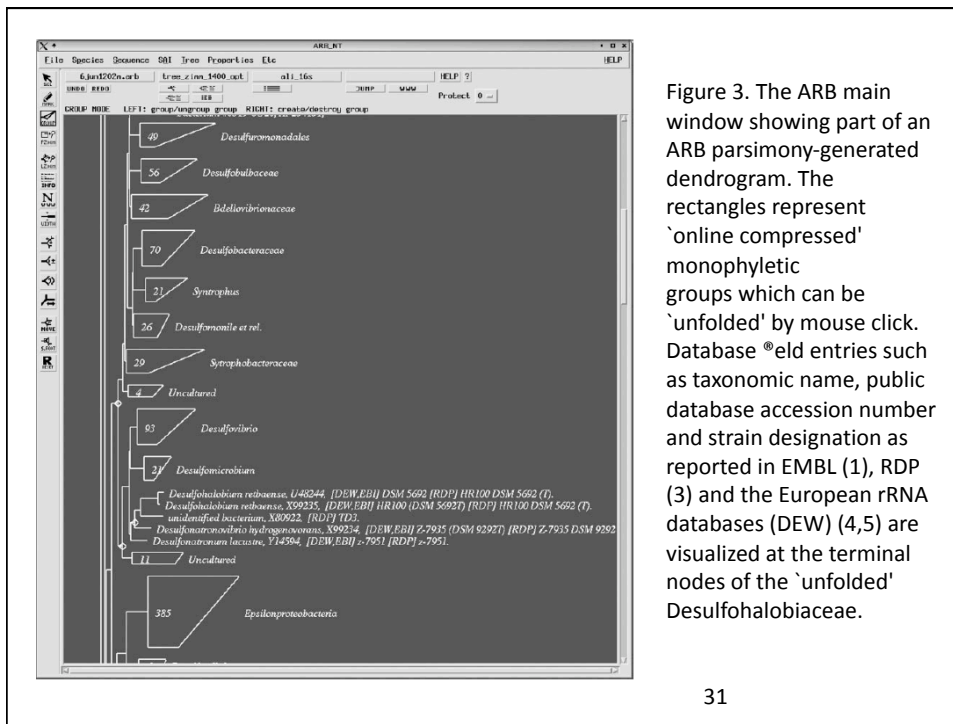
[Nuc. Acids Res. 2007; Vol. 35, No. 21, p. 7188-7196](#)

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

ARB: a software environment for sequence data

Wolfgang Ludwig¹, Oliver Strunk, Ralf Westram, Lothar Richter, Harald Meier¹, Yadhukumar, Arno Buchner, Tina Lai, Susanne Steppi, Gangolf Jobb¹, Wolfram Förster¹, Igor Brettske, Stefan Gerber, Anton W. Ginhart¹, Oliver Gross, Silke Grumann¹, Stefan Hermann¹, Ralf Jost¹, Andreas König¹, Thomas Liss¹, Ralph Lüßmann¹, Michael May¹, Björn Nonhoff¹, Boris Reichel¹, Robert Strehlow¹, Alexandros Stamatakis¹, Norbert Stuckmann¹, Alexander Vilbig¹, Michael Lenke¹, Thomas Ludwig², Arndt Bode¹ and Karl-Heinz Schleifer

30



31

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

mothur

Download	Wiki	Forum	facebook
----------	------	-------	----------

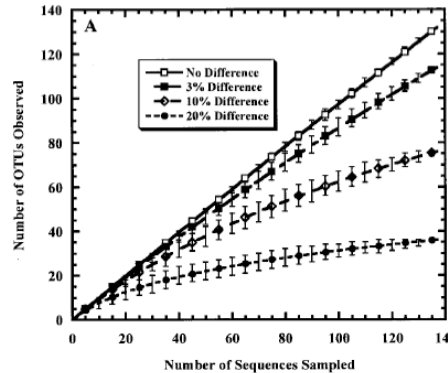
Welcome to the website for the mothur project, initiated by [Dr. Patrick Schloss](#) and his software development team in the Department of Microbiology & Immunology at [The University of Michigan](#). This project seeks to develop a single piece of open-source, expandable software to fill the bioinformatics needs of the microbial ecology community. In February 2009 we released the first version of mothur, which had accelerated versions of the popular DOTUR and SONS programs. Since then we have added the functionality of a number of other popular tools including s-libshuff, TreeClimber (i.e. the parsimony test), UniFrac, distance calculation, visualization tools, a NAST-based aligner, and many other features. If you would

32

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
 Copyright © 2005, American Society for Microbiology. All Rights Reserved.

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

Patrick D. Schloss and Jo Handelsman*

33

Comparing Bacterial Diversity: Community Membership & Structure

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

Community Membership

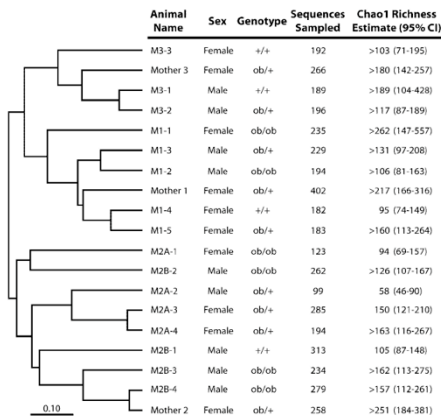
(Categories of fruit in common)
 $= 2/5 = 0.4$

Community Structure

(Pieces of fruit in common)
 $= \sim 0.9$

34

Community Membership: Pups are most like their mothers



APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 2006, p. 6773-6779
 0099-2240/06/\$08.00+0 doi:10.1128/AEM.00474-06
 Copyright © 2006, American Society for Microbiology. All Rights Reserved.

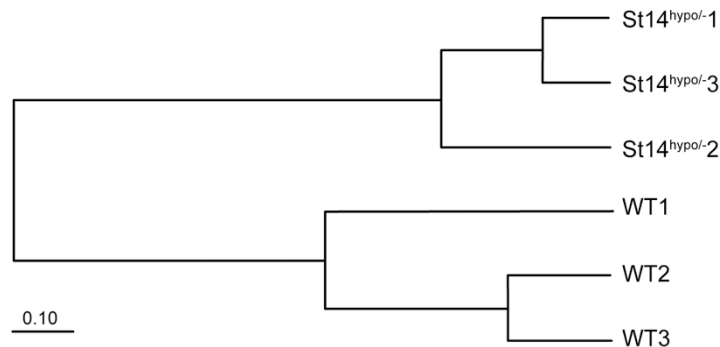
Vol. 72, No. 10

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

Patrick D. Schloss† and Jo Handelsman*

35

Community Structure: Pups cluster according to genotype



Scharschmidt et al. JID 2009

36

UniFrac: Unique Fraction Metric

- Measures fraction of branch length in a tree that is unique to a community
- Weighted or unweighted for abundance
- Can be used with multivariate statistical methods (UPGMA and PCA) for visualization
- Calculate parsimonious changes to obtain p value

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2005, p. 8228–8235
0099-2240/05/\$08.00+0 doi:10.1128/AEM.71.12.8228–8235.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Vol. 71, No. 12

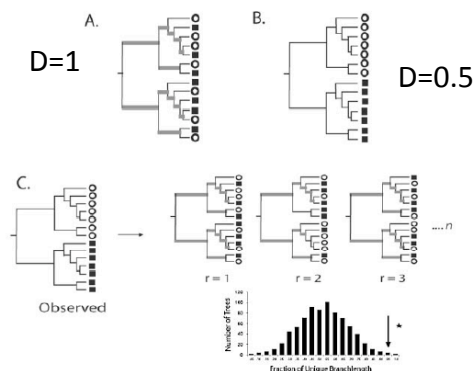
UniFrac: a New Phylogenetic Method for Comparing Microbial Communities

Catherine Lozupone¹ and Rob Knight^{2*}

37

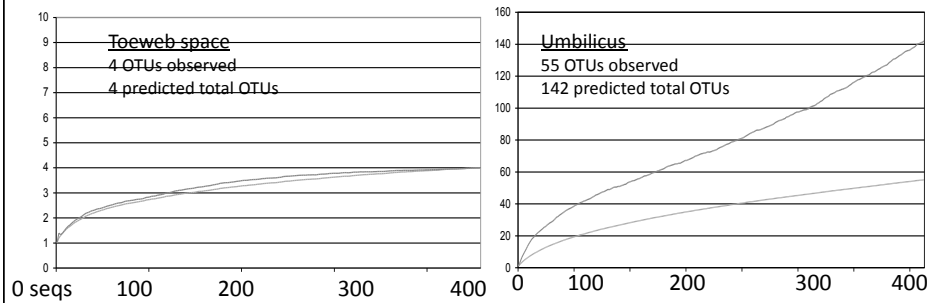
UniFrac allows you to:

1. Determine if the environments in the input phylogenetic tree have significantly different microbial communities.
2. Determine if community differences are concentrated within particular lineages of the phylogenetic tree.



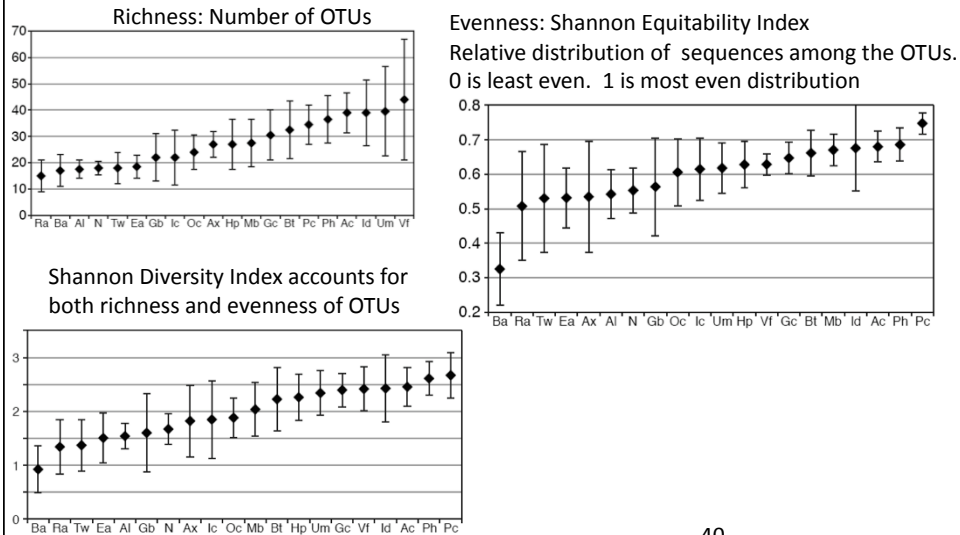
38

How much diversity is there in the population? Have you sequenced enough to capture the diversity? Chao1 rarefaction curves



39

Richness, evenness, diversity: Shannon and Simpson diversity



40

If you are using 454 sequences,
consider VAMPS to form OTUs
<http://vamps.mbl.edu/>

Revised: October 9th, 2009.

VAMPS
The Visualization and Analysis
of Microbial Population Structures

VAMPS is an integrated collection of tools for researchers to visualize and analyze data for microbial population structures and distributions. For more information on the VAMPS project, visit our **VAMPS Overview page**.

There are two essential elements to VAMPS:

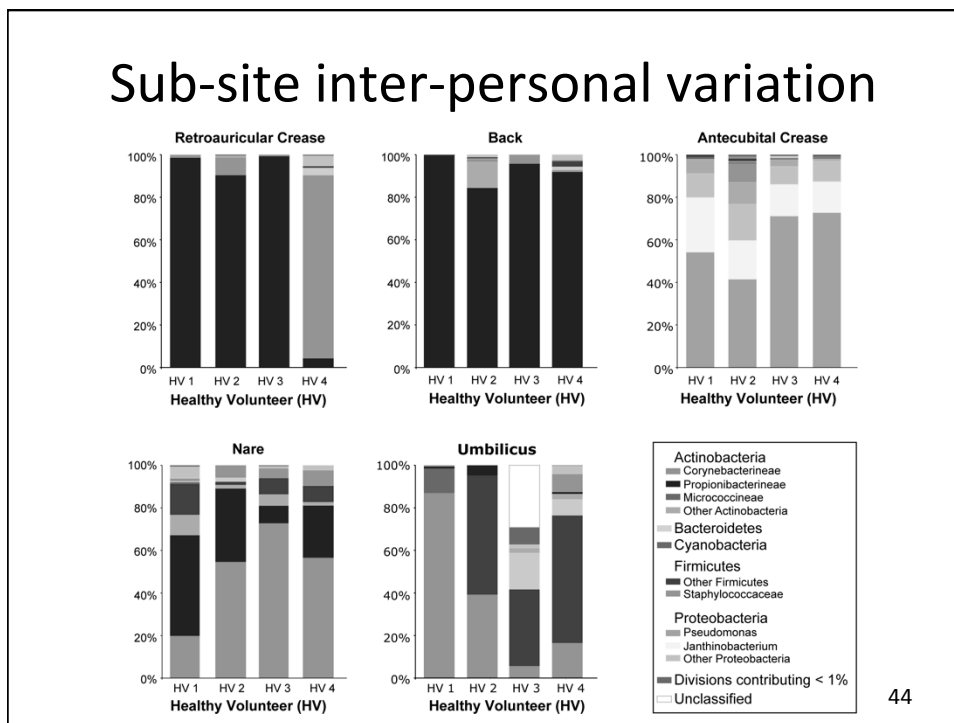
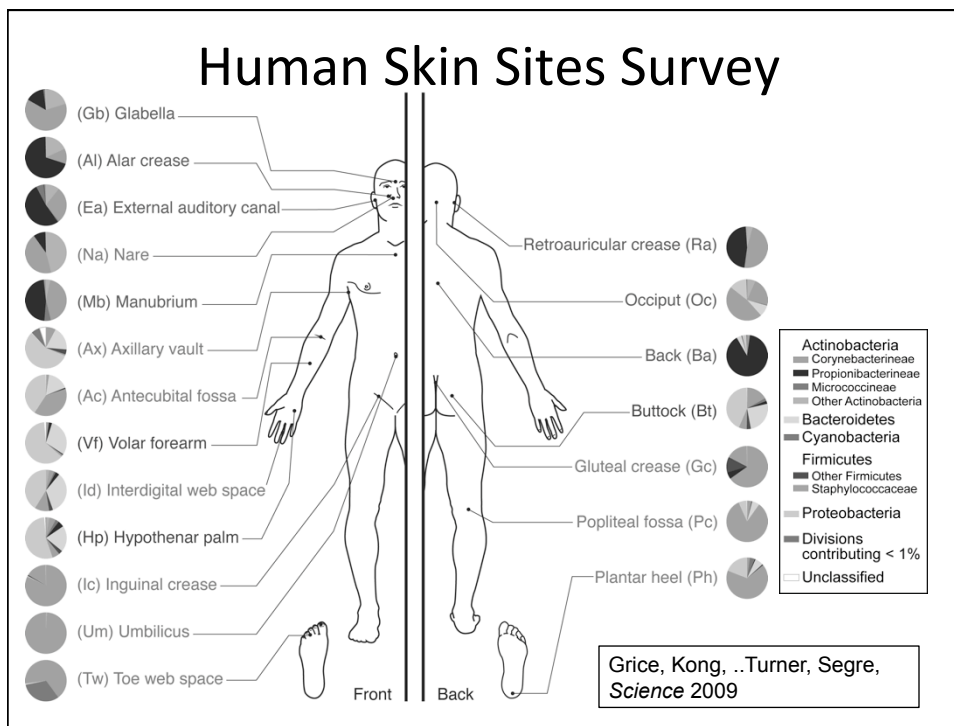
- **Visualization and Analysis** - *Community Visualization* including heat maps and comparative pie charts, as well as diversity estimates, rarefaction curves and spreadsheet-style output provide researchers with analytical tools for assessing individual microbial populations, based on either taxonomic assignments or independently-derived operational taxonomic units (OTUs).
- **Data Ramp** - Researchers who want to use the VAMPS tools with their own data can enter their sequence or taxonomy data to the VAMPS website and merge it with the existing shared datasets for individual or comparative analyses. Researchers will be given a user name and password, and their data will be visible only to registered users of their choice.

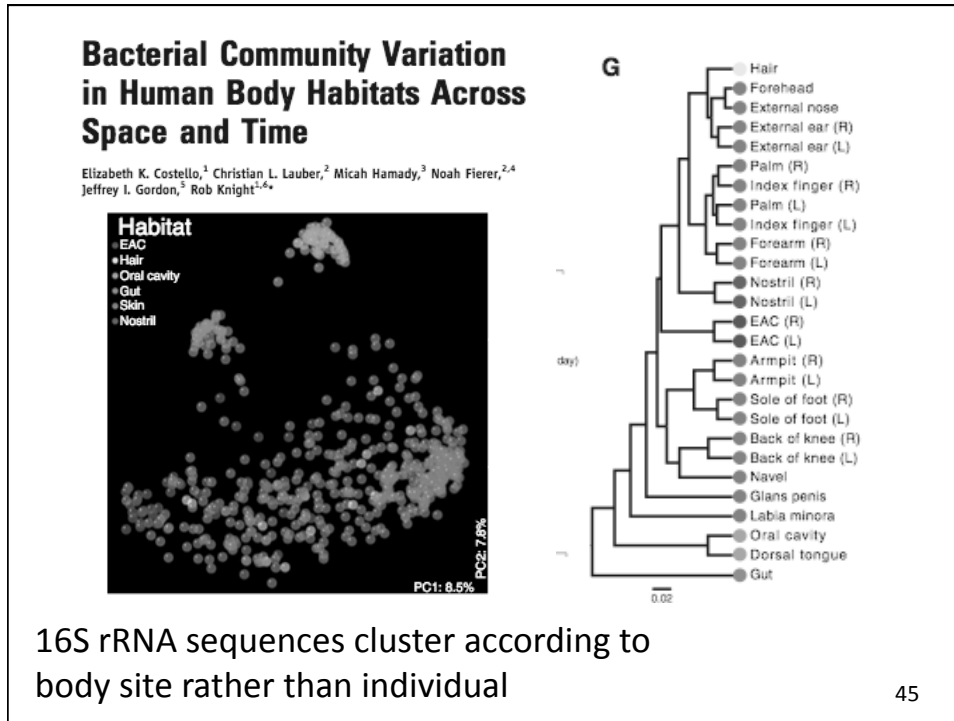
41

Fungal Diversity

- Similar strategy can be used to classify the 18S rRNA or the intervening sequence (ITS) of fungi

42





Microbial community profiling for human microbiome projects: Tools, techniques, and challenges

Micah Hamady and Rob Knight

Genome Res. 2009 19: 1141-1152 originally published online April 21, 2009
 Access the most recent version at doi:10.1101/gr.085464.108

INSIGHT FEATURE NATURE Vol 449 | 18 October 2007 | doi:10.1038/nature06244

The Human Microbiome Project

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight & Jeffrey I. Gordon

A strategy to understand the microbial components of the human genetic and metabolic landscape and how they contribute to normal physiology and predisposition to disease.

The NIH Human Microbiome Project

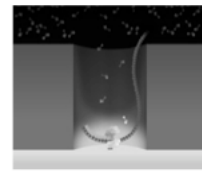
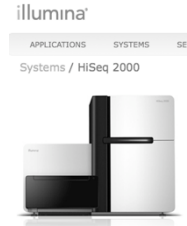
The NIH HMP Working Group, Jane Peterson, Susan Garges, et al.

Genome Res. 2009 19: 2317-2323 originally published online October 9, 2009
 Access the most recent version at doi:10.1101/gr.096651.109

46

Topic 2: Sequencing Bacterial Genomes

- Roche/454 generates 1, 250,000 reads of ~400+ bp (5 Gbp).
- Illumina/Solexa generate shorter reads (75 bp) but generate more sequence data per run for cheaper price/base pair.



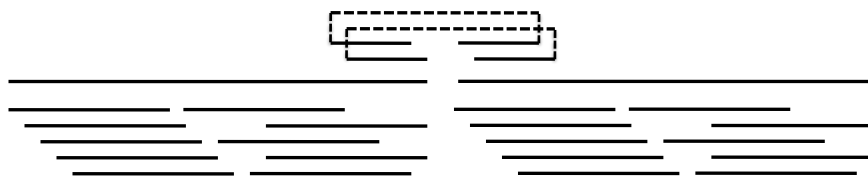
Roche/454-XLR <i>Pyrosequencing</i>	Illumina/Solexa GA2 <i>Sequencing by synthesis</i>
•Emulsion PCR •400-bp read (avg)	•Bridge PCR •75-bp read



* Manufacturer specifications from Holt and Jones, *Genome Research* 18:839-46 (2008)

47

Paired end reads (8 kb inserts) scaffold contigs



Unidirectional reads form contigs

48

Assemblers (*de novo*)

- Phrap
- Newbler (454)
- Velvet
- ALL-PATHS, SSAKE, VCAKE, SHARCGS, Edena, AMOS
- CAP3/PCAP



49

Newbler (gsAssembler)

- Works in base-space and flow-space
- Overlap-Layout-Consensus method
- Homopolymer correction
- 1. Identify pairwise read overlaps
- 2. Build graph
 - 1. Nodes are contiguous alignments
 - 2. Edges connect nodes with branch points representing repeat boundaries
- 3. Detangle
- 4. Build consensus alignment

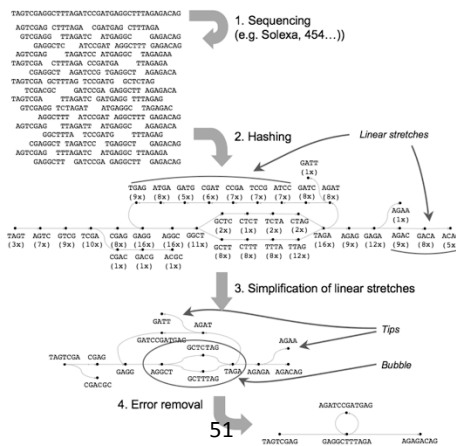
50

Velvet (Zerbino and Birney, 2008)

- Works in base-space and color-space
- Good for small genomes
- Agnostic of read length

- Construct k-mer hash
- Build De Bruijn graph
- Simplify graph
- Resolve

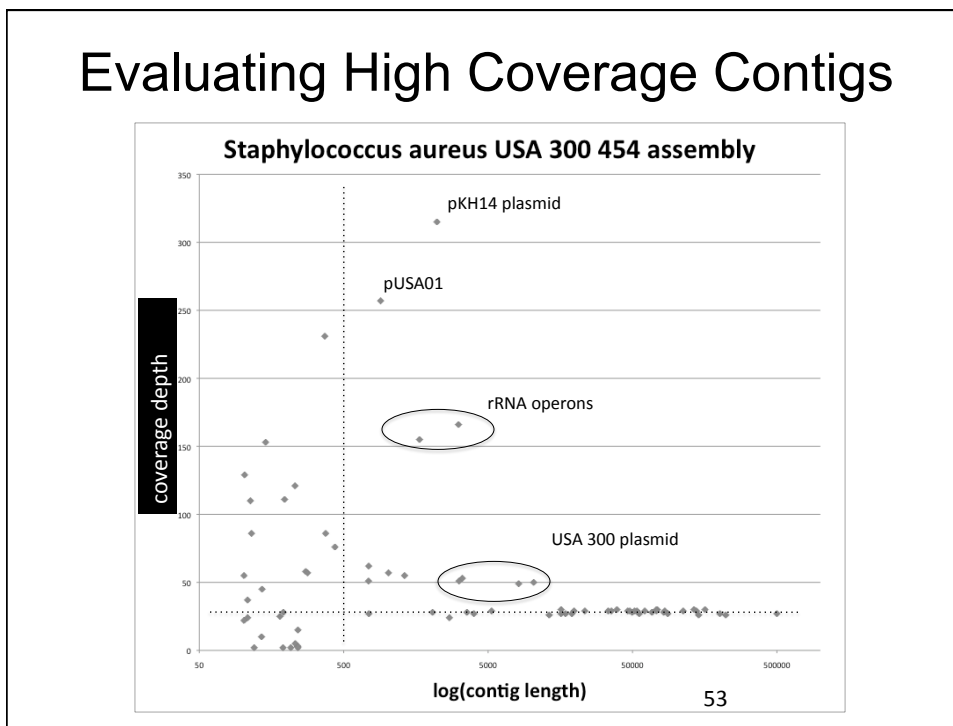
- Tips
- Bubbles



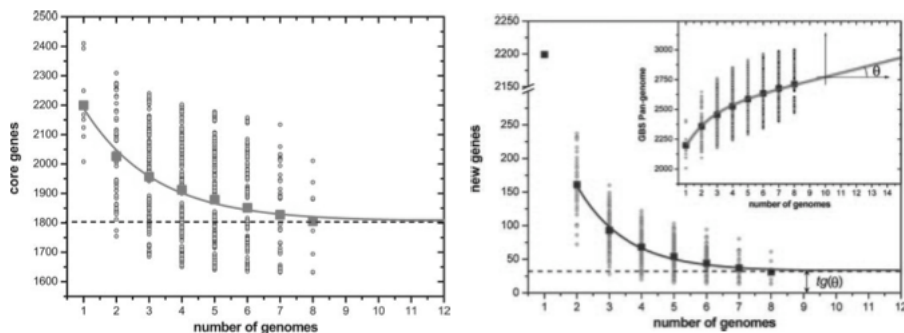
Evaluating Assemblies

- Coverage is a measure of how deeply a region has been sequenced
- The Lander-Waterman model predicts 8-10 fold coverage is needed to minimize the number of contigs for a 1 Mbp genome
- The N50 size is the point at which 50% of bases are in contigs this size or greater

Evaluating High Coverage Contigs



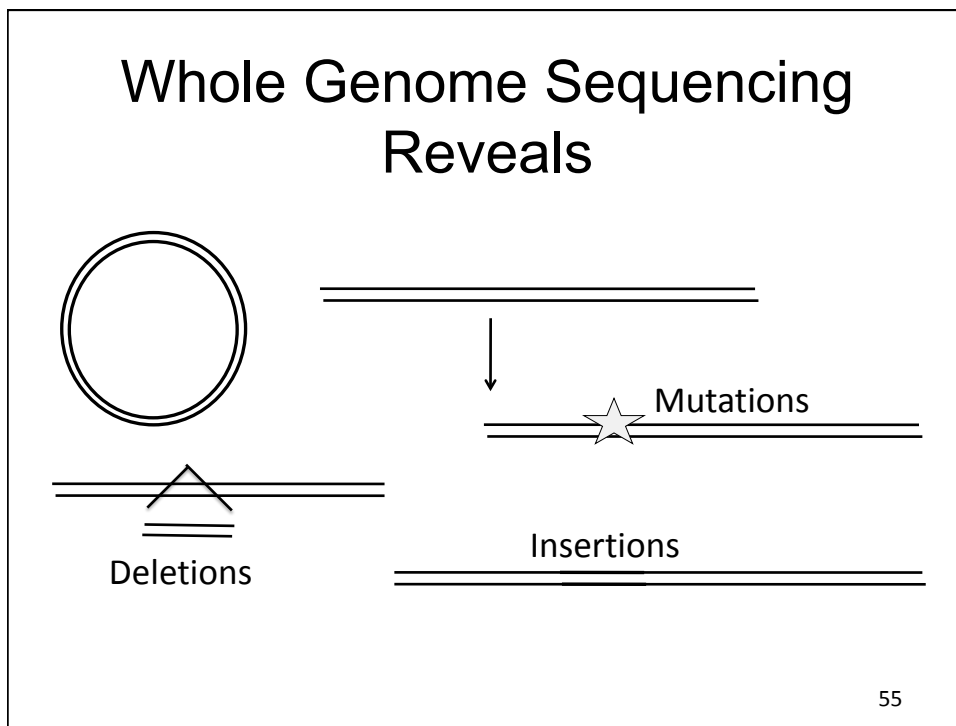
Is there a reference genome? Is it a fixed genome? Bacteria exchange information with horizontal gene transfer



Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

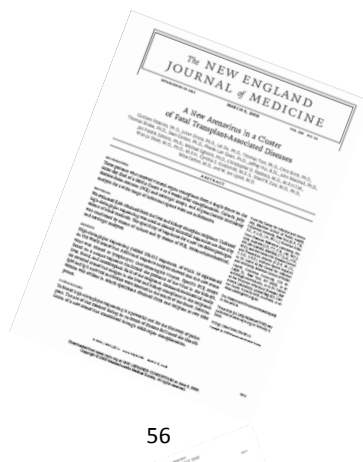
Hervé Tettelin^{1,2}, Vega Masignani^{1,2}, Michael J. Cieslewicz^{1,2,3,4}, Claudio Donati¹, Duccio Medini¹, Naomi L. Ward¹, Samuel V. Angiuoli¹, Jonathan Crabtree¹, Amanda L. Jones¹, A. Scott Durkin¹, Robert T. DeBoy¹, Tanja M. Davidsen¹, Marimón Mora¹, Maria Scarselli¹, Immaculada Margarit y Ros¹, Jeremy D. Peterson¹, Christopher R. Hauser¹, Jaideep P. Sundaram¹, William C. Nelson¹, Ramana Madupati¹, Lauren M. Brinkac¹, Robert J. Dodson¹, Mary J. Rosovitz¹, Steven A. Sullivan¹, Sean C. Daugherty¹, Daniel H. Haft¹, Jeremy Selengut¹, Michelle L. Gwinn¹, Liwei Zhou¹, Nikhat Zafar¹, Hoda Khouri¹, Diana Radune¹, George Dimitrov¹, Kisha Watkins¹, Kevin J. B. O'Connor¹, Shannon Smith¹, Teresa R. Utterback¹, Owen White¹, Craig E. Rubens¹, Guido Grandi¹, Lawrence C. Madoff¹, Dennis L. Kasper¹, John L. Telford¹, Michael R. Wessels^{1,2}, Rino Rappuoli^{1,2}, and Claire M. Fraser^{1,2,3,4,5,6}

54



TOPIC 3. Identifying Novel Virus: Transplant Associated Arenavirus (also SARS, Merkel cell carcinoma)

Resequencing the human genome to identify viral associated disease is getting EASIER and CHEAPER. Once you find them once, finding them again is PCR-based. Very cheap and easy!



Three organ-transplant recipients died with a month of the transplant

Table 1. Characteristics of the Organ-Transplant Recipients.

Recipient No.	Age yr	Diagnosis	Organ Transplanted	Clinical Course	Interval between Transplantation and Death days
1	63	End-stage renal failure due to polycystic kidney disease	Kidney	Fever, sepsis, encephalopathy, acute tubular necrosis, graft rejection, radiographic evidence of chest infiltrates	36
2	64	Decompensated cirrhosis and hepatocellular cancer due to hepatitis C infection	Liver	Fever, confusion, encephalopathy with myoclonus, chest infiltrates	30
3	44	End-stage renal failure due to polycystic kidney disease	Kidney	Fever, graft rejection, intraabdominal hematomas and effusion, transplant nephrectomy, encephalopathic illness	29

57

The needle(s) in the haystack...

103,632 reads from 454 FLX lane
 (length= 45-337 nt, mean=162.)
 94,043 reads after filtering



BLASTN largely uninformative



BLASTX analysis identified 14 fragments that were consistent with Old World arenaviruses (12 S-segment and 2 L-segment).



PCR using primers based on the pyrosequencing reads and consensus information from sequenced Arenaviruses

Table 2. Nucleotide and Amino Acid Homologies of the New Arenavirus to Other Arenaviruses.*

Gene	Accession No.	LCMV Strain	Homology	
			Amino Acid	Nucleotide
GPC	AB261990	M2	94	86
NP	AB261990	M2	97	87
L	DQ286932	Marseille 12	82	79
Z	DQ286932	Marseille 12	79	72

* LCMV denotes lymphocytic choriomeningitis virus.

58

Sequencing is just the start... Koch's postulates



- The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy animals.
- The microorganism must be isolated from a diseased organism and grown in pure culture.
- The cultured microorganism should cause disease when introduced into a healthy organism.
- The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.

59

TOPIC 4. METAGENOMICS: DNA sequence from multiple organisms

Fungal, Bacterial, Viral, Archaeal DNA all together
(with human DNA).

Very Complex mixture and very complex computationally.

Vol 455|25 September 2008

nature

MICROBIOLOGY

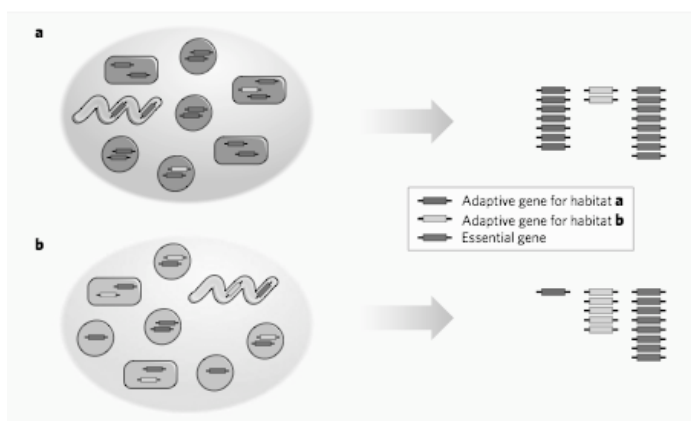
Metagenomics

Philip Hugenholtz and Gene W. Tyson

Ten years after the term metagenomics was coined, the approach continues to gather momentum. This culture-independent, molecular way of analysing environmental samples of cohabiting microbial populations has opened up fresh perspectives on microbiology.

60

Metagenomics: types of bacteria similar between 2 populations, but pink genes enriched in top population



61

A core gut microbiome in obese and lean twins

Peter J. Turnbaugh¹, Micah Hamady³, Tanya Yatsunenok¹, Brandi L. Cantarel⁵, Alexis Duncan², Ruth E. Ley¹, Mitchell L. Sogin⁶, William J. Jones⁷, Bruce A. Roe⁸, Jason P. Affourtit⁹, Michael Egholm⁹, Bernard Henrissat⁵, Andrew C. Heath², Rob Knight⁴ & Jeffrey I. Gordon¹

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin^{1*}, Ruiqiang Li^{1*}, Jeroen Raes^{2,3}, Manimozhiyan Arumugam⁷, Kristoffer Solvsten Burgdorf⁴, Chaysavanh Manichanh⁵, Trine Nielsen⁴, Nicolas Pons⁶, Florence Levenez², Takuji Yamada¹, Daniel R. Mende⁷, Junhua Li^{1,7}, Junming Xu¹, Shaochuan Li¹, Dongfang Li^{1,8}, Jianjun Cao¹, Bo Wang¹, Huiqing Liang¹, Huisong Zheng¹, Yinlong Xie^{1,7}, Julien Tap⁹, Patricia Lepage⁶, Marcelo Bertalan⁹, Jean-Michel Batto⁹, Torben Hansen¹, Denis Le Paslier¹⁰, Allan Linneberg¹¹, H. Bjørn Nielsen⁹, Eric Pelletier¹⁰, Pierre Renault⁶, Thomas Sicheritz-Ponten⁹, Keith Turner¹², Hongmei Zhu¹, Chang Yu¹, Shengting Li¹, Min Jian¹, Yan Zhou¹, Yingrui Li¹, Xiuqing Zhang¹, Songgang Li¹, Nan Qin¹, Huanming Yang¹, Jian Wang¹, Søren Brunak⁹, Joel Doré⁶, Francisco Guarner¹, Karsten Kristiansen¹³, Oluf Pedersen^{4,14}, Julian Parkhill¹², Jean Weissenbach¹⁰, MetaHIT Consortium¹, Peer Bork², S. Dusko Ehrlich⁶ & Jun Wang^{1,15}

The human body contains about ten times as many microbes as human cells, and most of them live in the gut. The new study, published today in *Nature*¹, shows that, between them, those microbes contain 3.3 million genes, dwarfing the human genome's 23,000. The authors also find that the bacterial species in one person's gut are not as different from those of others as had been expected.

Tools do not yet exist to catalogue and comprehend metagenomic complexity

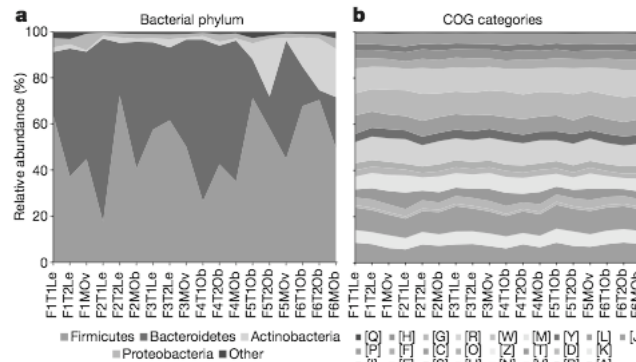


Figure 3 | Comparison of taxonomic and functional variations in the human gut microbiome. a, Relative abundance of major phyla across 18 faecal microbiomes from monozygotic twins and their mothers, based on BLASTX

63

Why do we sterilize our exterior (skin) and populate our interior (gut)?



Is this promoting health or disease? Is this neutral to our health? To our environment?

64