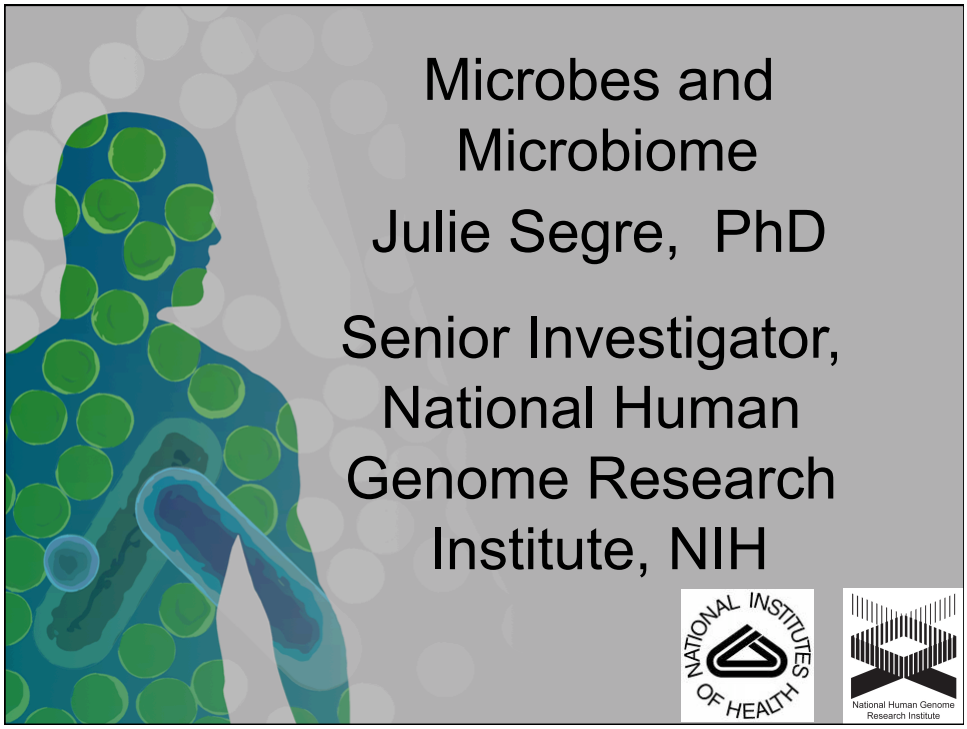
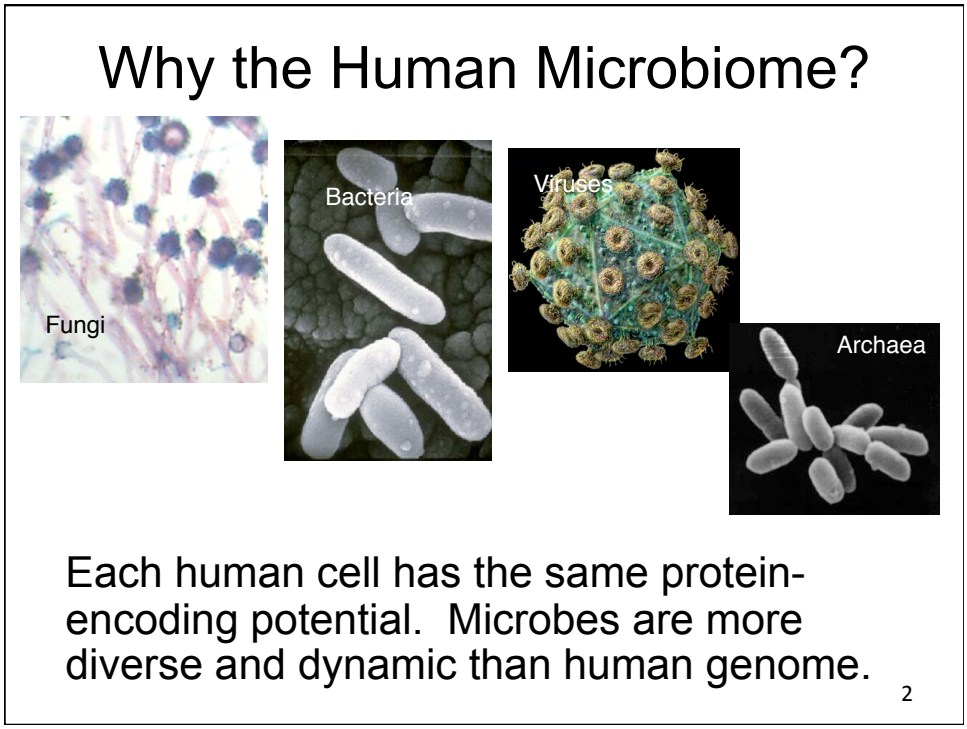
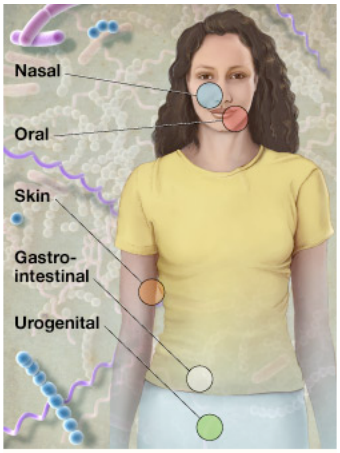

Microbes and Microbiome
Julie Segre, PhD
Senior Investigator, National Human Genome Research Institute, NIH
NATIONAL INSTITUTES OF HEALTH
National Human Genome Research Institute


Why the Human Microbiome?
Fungi
Bacteria
Viruses
Archaea
Each human cell has the same protein-encoding potential. Microbes are more diverse and dynamic than human genome.
2

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

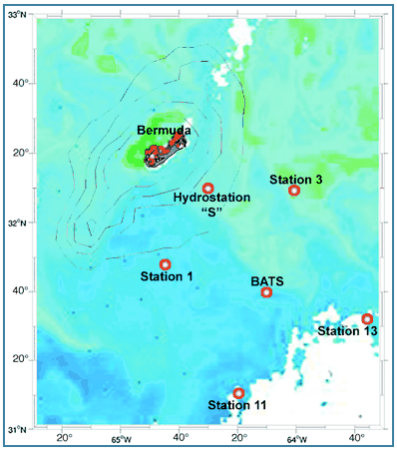
3

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

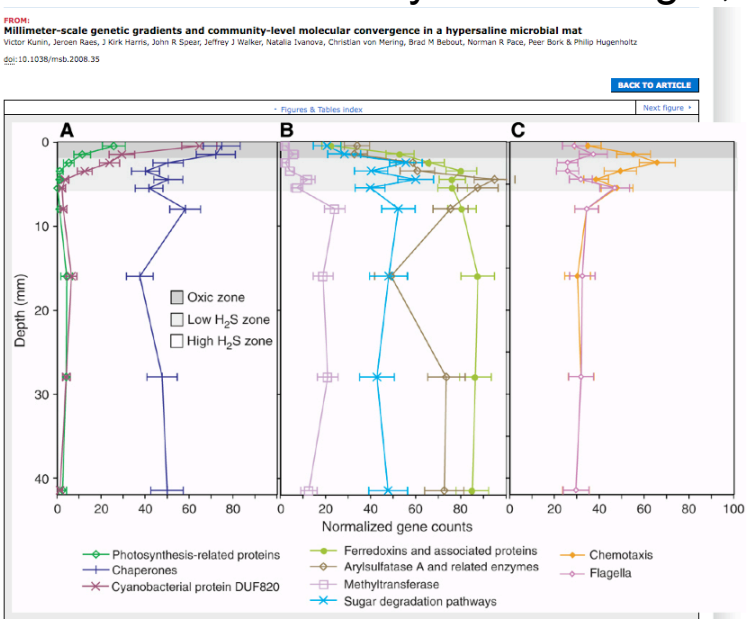
RESEARCH ARTICLES

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 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]

5

HyperSaline mat diversity Guerro Negro, MX



6

And human-environment diversity: shower heads across USA

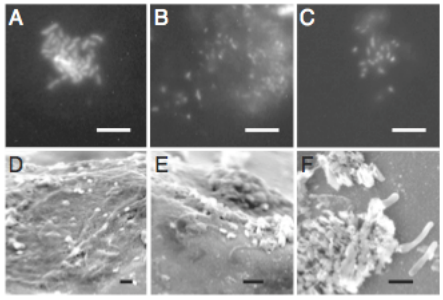
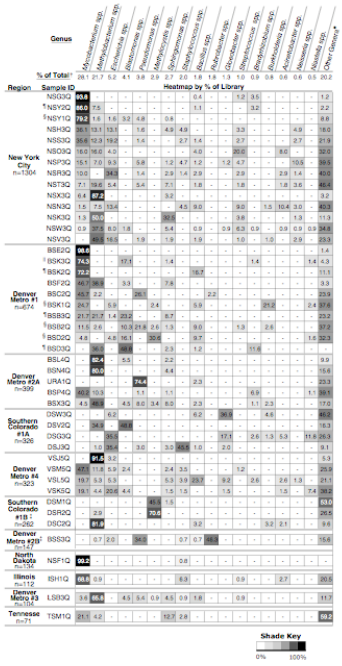


Fig. 1. Fluorescence and SEM images of showerhead biofilm. (*A–C*) Epifluorescence microscopy of biofilm samples stained with DAPI; scale bars, 10 μm. (*D–F*) SEM micrographs of increasing magnification of in situ showerhead

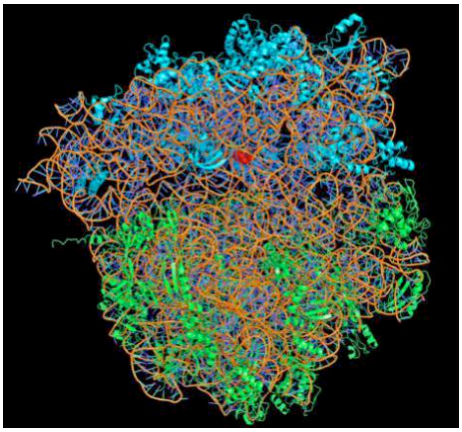


Opportunistic pathogens enriched in showerhead biofilms

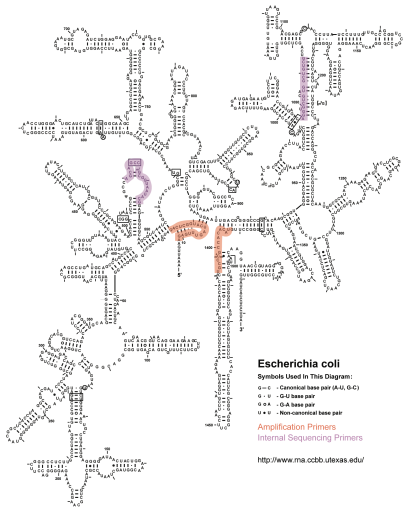
Leah M. Feazel[a], Laura K. Baumgartner[a], Kristen L. Peterson[a], Daniel N. Frank[a], J. Kirk Harris[b], and Norman R. Pace[a,1]

7

TOPIC 1. Bacterial Diversity: 16S rRNA gene



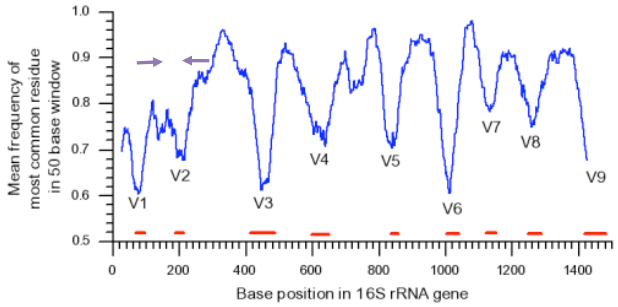
Orange= rRNA ;
Blue = small subunit proteins
Green = large subunit proteins



Secondary Structure: 16S small subunit ribosomal RNA

8

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

Grice et al, Genome Research 2008

10

How to study microbial diversity

- Fingerprinting: cheapest, but very limited (Anderson and Cairney, Envir Microbiol 2004)
- PhyloChip orGeoChip: 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)

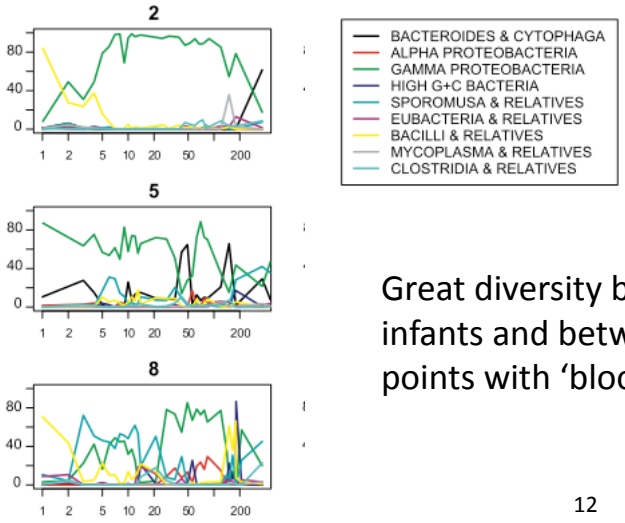
- Sequencing: taxonomic classification and function, dynamic range and compare multiple complex samples.

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

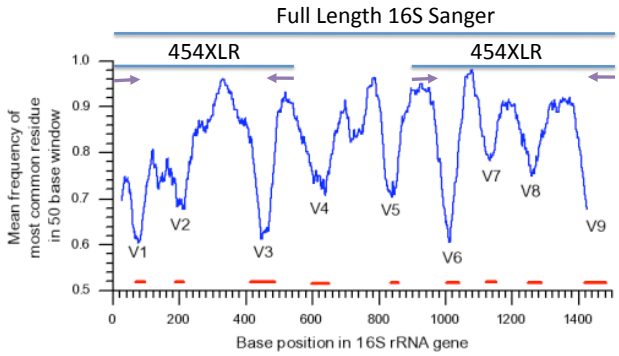
Palmer, Relman, Brown 2007 PLOS Bio



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 ELSES IF DIFFERENT PRIMER OR AMPLIFICATION CONDITIONS WERE USED

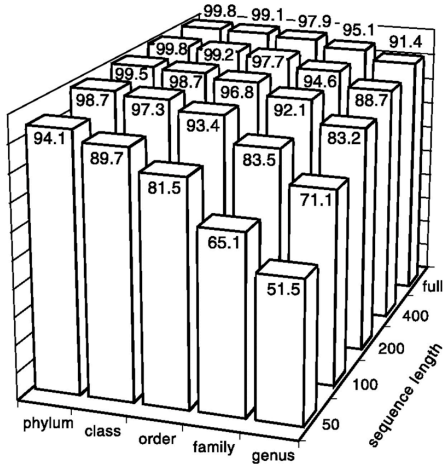
13

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.



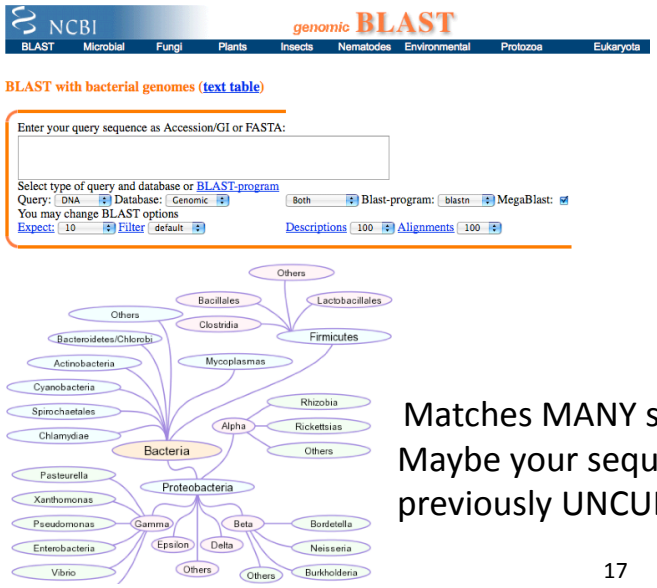
Applied and Environmental Microbiology, August 2007, p. 5261-5267, Vol. 73, No. 16
Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy · Qiong Wang,[1] George M. Garrity,[1,2] James M. Tiedje,[1,2] and James R. Cole[1]
Also see: Liu, DeSantis, Andersen and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

15

```
DEFINITION  Staphylococcus aureus strain 16S ribosomal RNA gene,
        1 ggcgcgcgtt atcgctggcg gcgtgcctaa tacatgcaag tcgagcgaac ggacgagaag
       61 cttctctgat gttagcggcg gacgggtgag taacacgtgg ataacctacc tataagactg
      121 ggataacttc gggaaaccgg agctaatacc ggataatatt ttgaaccgca tggttcaaaa
      181 gtgaaagacg gtcttgctgt cacttataga tggatccgcg ctgcattagc tagttggtaa
      241 ggtaacggct taccaaggca acgatgcata gccgacctga gagggtgatc ggccacactg
      301 gaactgagac acggtccaga ctcctacggg aggcagcagt agggaatctt ccgcaatggg
      361 cgaaagcctg acggagcaac ccgcgtgagt gatgaaggtc ttcggatcgt aaaactctgt
      421 tattagggaa gaacatatgt gtaagtaact gtgcacatct tgacggtact aatcagaaag
      481 ccacggctaa ctacgtgcca gcagccgcgg taatacgtag gtggcaagcg ttatccggaa
      541 ttattgggcg taaagcgcgc gtaggcggtt ttttaagtct gatgtgaaag cccacggctc
      601 aaccgtggag ggtcattgga aactggaaaa cttgagtgca gaagaggaaa gtggaattcc
      661 atgtgtagcg gtgaaatgcg cagagatatg gaggaacacc agtggcgaag gcgacttctg
      721 gtctgtaact gacgctgatg tgcgaaagcg tggggatcaa acaggattag ataccctggt
      781 agtccacgcc gtaaacgatg agtgctaagt gttaggggtt ttccgcccct tagtgctgca
      841 gctaacgcat taagcactcc gcctggggag tacgaccgca aggttgaaac tcaaaggaat
      901 tgacggggac ccgcacaagc ggtggagcat gtggtttaat tcgaagcaac gcgaaacctt
      961 accaaatctt gacatccttt gacaactcta gagatagagc cttccccttc gggggacaaa
     1021 gtgacaggtg gtgcatggtt gtcgtcagct cgtgtcgtga gatgttgggt taagtcccgc
     1081 aacgagcgca acccttaagc ttagttgcca tcattaagtt gggcactcta agttgactgc
     1141 cggtgacaaa ccggaggaag gtggggatga cgtcaaatca tcatgccccc tatgatttgg
     1201 gctacacacg tgctacaatg gacaatacaa agggcagcga aaccgcgagg tcaacaaatc
     1261 ccataaagtt gttctcagtt cggattgtag tctgcaactc gactacatga agctggaatc
     1321 gctagtaatc gtagatcagc atgctacggt gaatacgttc ccgggtcttg tacacaccgc
     1381 ccgtcacacc acgagagttt gtaacacccg aagccggtgg agtaaccttt tagggctagc
     1441 cggaaggtgg g
```

16

How to identify a bacterial sequence and align sequences?



Matches MANY sequences. Maybe your sequence is previously UNCULTURED?

17

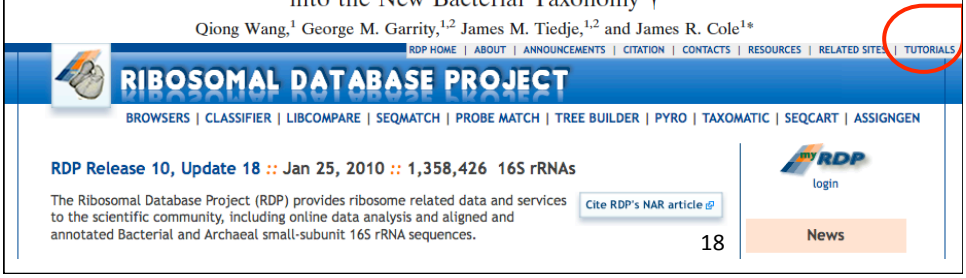
RDP Database http://rdp.cme.msu.edu/

- 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

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]*



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:

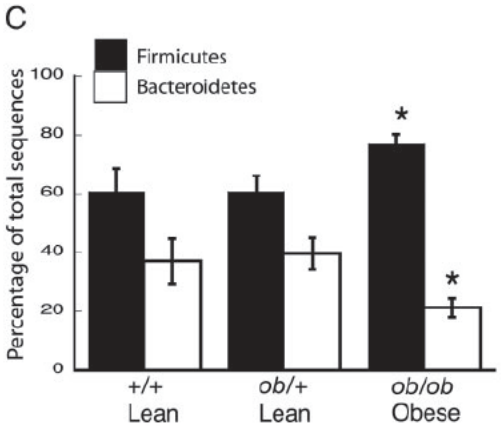
- **SPADE Formatter** - make a SPADE compatible input format.
- **R Formatter** - make a R compatible input format.
- **EstimateS Formatter** - make an EstimateS compatible input format. Can also be used with PAST.
- **Mothur: Column Distance Matrix** - create a column distance matrix compatible with 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

19

Gordon: lean versus obese mice



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

Ley, ... Gordon PNAS 2005

20

Also true in humans

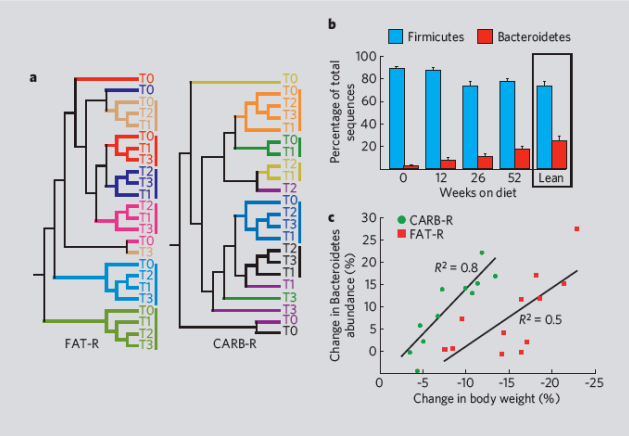


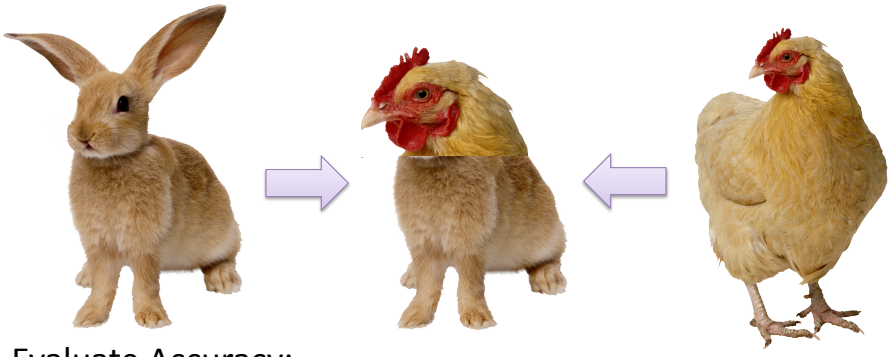
Figure 1 | Correlation between body-weight loss and gut microbial ecology. **a,** Clustering of 16S ribosomal RNA gene sequence libraries of faecal microbiota for each person (in different colours) and time point in diet therapy (T0, baseline; T1, 12 weeks; T2, 26 weeks; T3, 52 weeks) in the two diet-treatment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



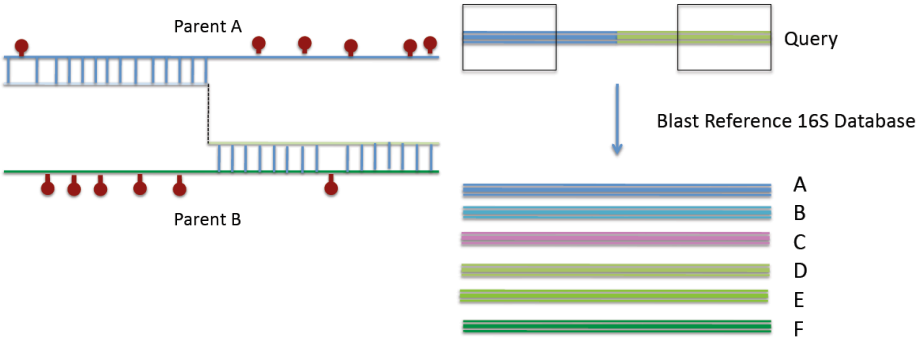
Evaluate Accuracy:

– 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

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. Weightman[1]

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 Faulkner[1] and Philip Hugenholtz[2]*

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, July 2006, p. 5069–5072 Vol. 72, No. 7
0099-2240/06/$08.00+0 doi:10.1128/AEM.03006-05

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

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

27

How to align sequences?

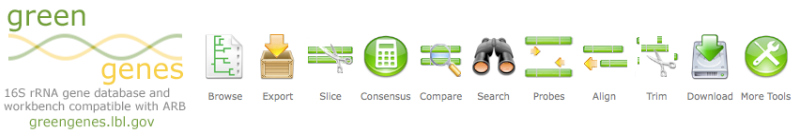
```
Query  225   ATTAGCTAGTTGGTAAGGTAACGGCT---TACCAAGGC-A-ACG-ATGCATAGCC-GACC  277
             |||||||||| |||  ||||||||||   | || |||| |  || || | ||||| |  |
Sbjct  212   ATTAGCTAGTAGGTGGGGTAACGGCTCCATCCCTAGGCGAGCCGAATCCTTAGCCTGGTC  271

Query  278   TGAGAGG-GTGATCGGCCACACTGGAACTGAG-ACACGGTCCAGACTCCTACGGGAGGCA  335
             |||||||  ||| | |||||||||| |||||| ||||||||||||||||||||||||||
Sbjct  272   TGAGAGGAATGACCAGCCACACTGGGACTGAGAACACGGTCCAGACTCCTACGGGAGGCA  331
```

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

28

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

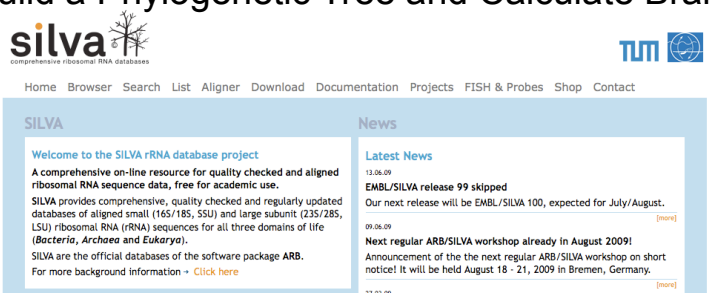
The NAST-iEr alignment utility (download) aligns a single raw nucleotide sequence against one or more NAST formatted sequences.

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.

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 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

30

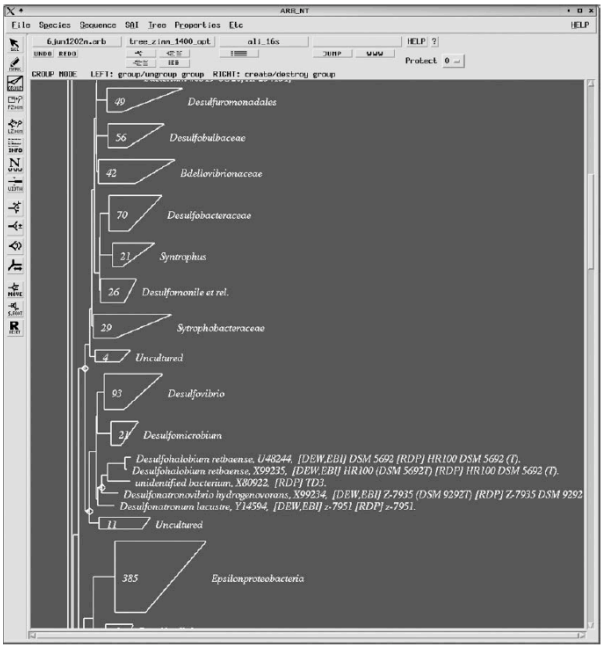
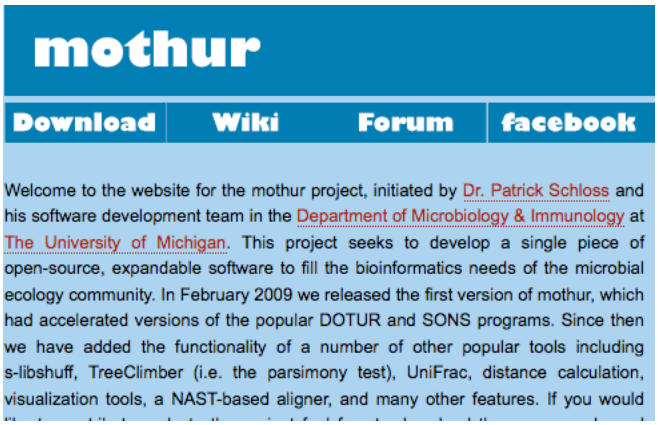


Figure 3. The ARB main window showing part of an ARB parsimony-generated dendrogram. The rectangles represent `online compressed' monophyletic groups which can be `unfolded' by mouse click. Database ®eld entries such as taxonomic name, public database accession number and strain designation as reported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

31

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

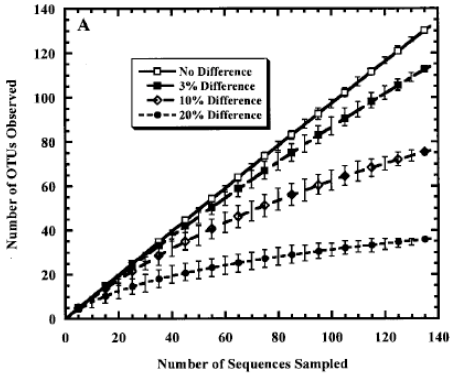


32

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

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

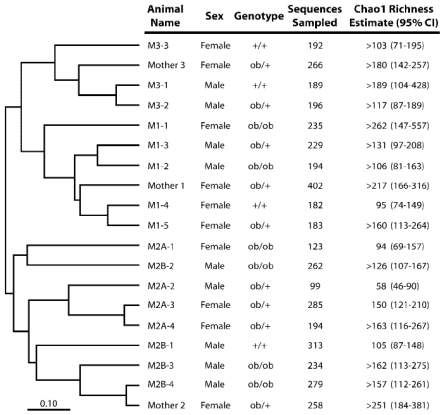
	Grp A	Grp B
	60	50
	34	50
	2	0
	2	0
	2	0

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

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

34

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
0099-2240/06/$08.00+0 doi:10.1128/AEM.00474-06

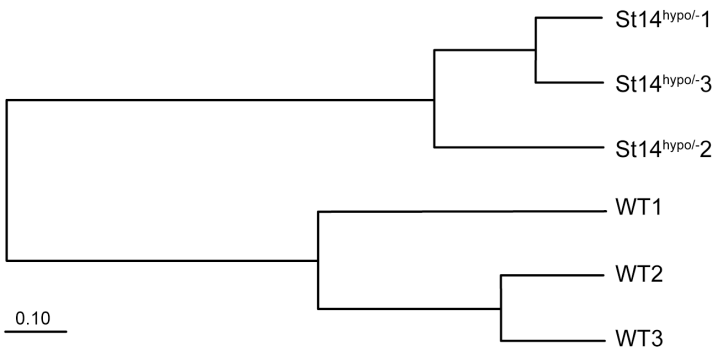
Vol. 72, No. 10

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

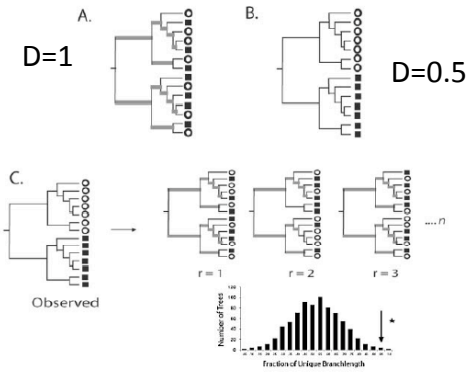
APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2005, p. 8228–8235 Vol. 71, No. 12
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.12.8228–8235.2005

UniFrac: a New Phylogenetic Method for Comparing Microbial Communities

Catherine Lozupone[1] and Rob Knight[2]*

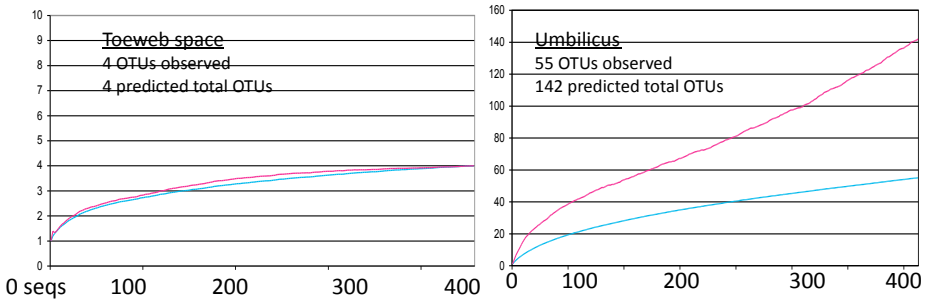
37

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

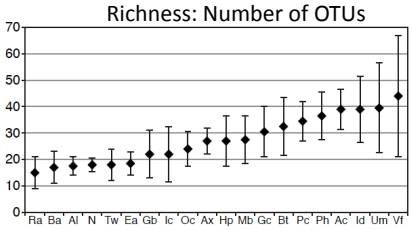
How much diversity is there in the population? Have you sequenced enough to capture the diversity? Chao1 rarefaction curves



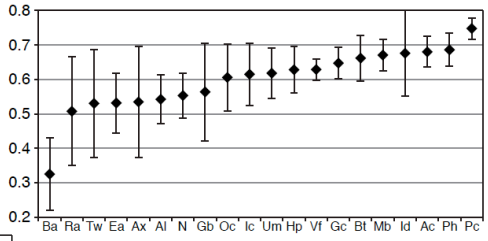
39

Richness, evenness, diversity: Shannon and Simpson diversity

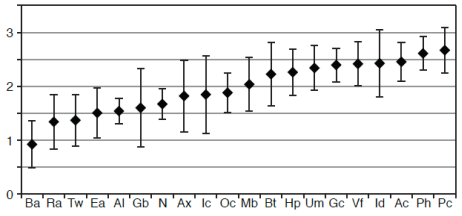
Richness: Number of OTUs



Evenness: Shannon Equitability Index
Relative distribution of sequences among the OTUs.
0 is least even. 1 is most even distribution

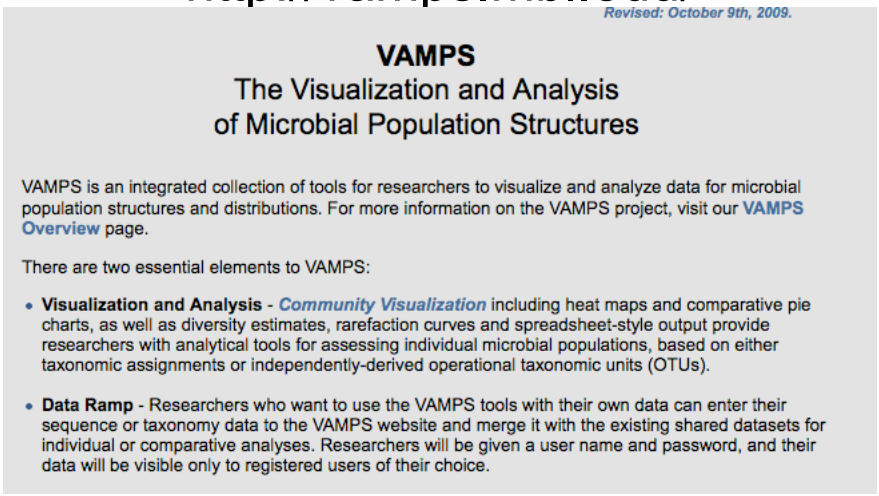


Shannon Diversity Index accounts for both richness and evenness of OTUs



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:

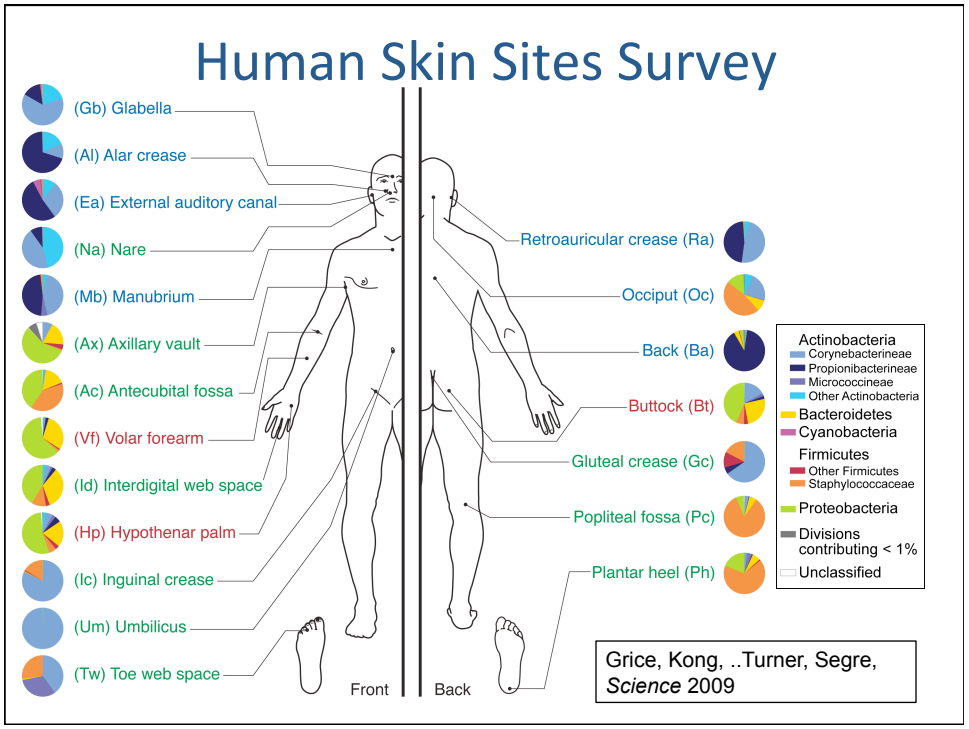
- **Visualization and Analysis** - ***Community Visualization*** including heat maps and comparative pie charts, as well as diversity estimates, rarefaction curves and spreadsheet-style output provide researchers with analytical tools for assessing individual microbial populations, based on either taxonomic 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.

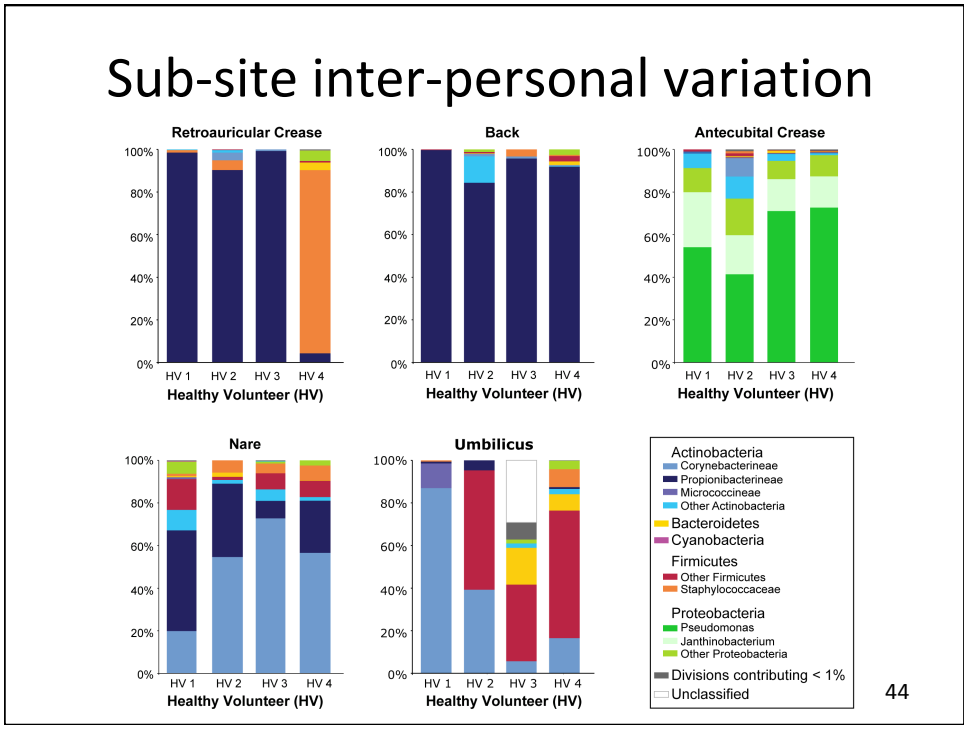
41

Fungal Diversity

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

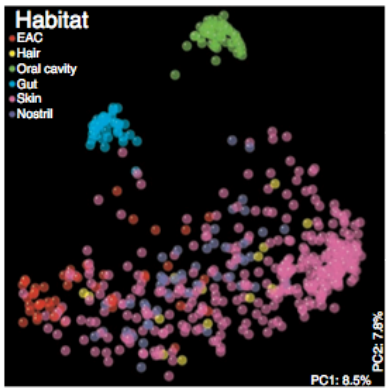
42

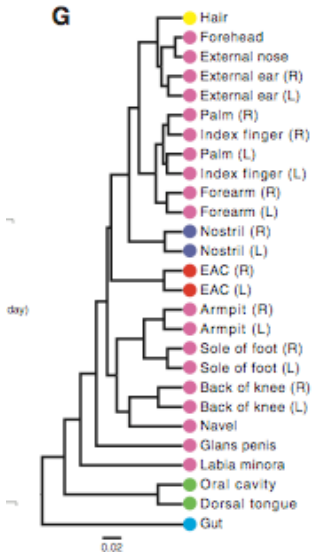

Human Skin Sites Survey
(Gb) Glabella
(Al) Alar crease
(Ea) External auditory canal
(Na) Nare
(Mb) Manubrium
(Ax) Axillary vault
(Ac) Antecubital fossa
(Vf) Volar forearm
(Id) Interdigital web space
(Hp) Hypothenar palm
(Ic) Inguinal crease
(Um) Umbilicus
(Tw) Toe web space
Retroauricular crease (Ra)
Occiput (Oc)
Back (Ba)
Buttock (Bt)
Gluteal crease (Gc)
Popliteal fossa (Pc)
Plantar heel (Ph)
Front
Back
Actinobacteria
Corynebacterineae
Propionibacterineae
Micrococcineae
Other Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Other Firmicutes
Staphylococcaceae
Proteobacteria
Divisions contributing < 1%
Unclassified
Grice, Kong, ..Turner, Segre, *Science* 2009


Sub-site inter-personal variation
Retroauricular Crease
Back
Antecubital Crease
Nare
Umbilicus
100%
80%
60%
40%
20%
0%
HV 1
HV 2
HV 3
HV 4
Healthy Volunteer (HV)
Actinobacteria
Corynebacterineae
Propionibacterineae
Micrococcineae
Other Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Other Firmicutes
Staphylococcaceae
Proteobacteria
Pseudomonas
Janthinobacterium
Other Proteobacteria
Divisions contributing < 1%
Unclassified
44

Bacterial Community Variation in Human Body Habitats Across Space and Time

Elizabeth K. Costello,[1] Christian L. Lauber,[2] Micah Hamady,[3] Noah Fierer,[2,4] Jeffrey I. Gordon,[5] Rob Knight[3,4]*





16S rRNA sequences cluster according to body site rather than individual

45

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

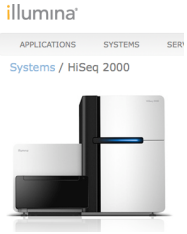
The NIH HMP Working Group, Jane Peterson, Susan Garges, et al.

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

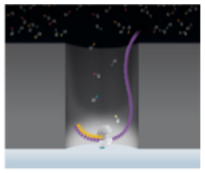
46

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.



PACIFIC BIOSCIENCES℠



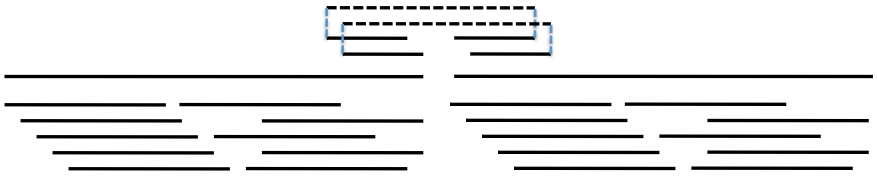
ion torrent

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

47

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

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

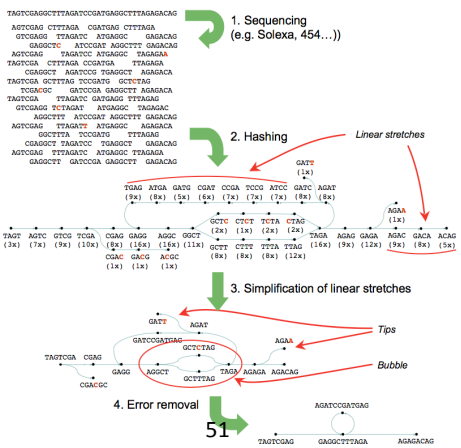
1. Identify pairwise read overlaps
2. Build graph
   1. Nodes are contiguous alignments
   2. Edges connect nodes with branch points representing repeat boundaries
3. Detangle
4. Build consensus alignment

50

Velvet (Zerbino and Birney, 2008)

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

1. Construct k-mer hash
2. Build De Bruijn graph
3. Simplify graph
4. Resolve
   1. Tips
   2. Bubbles



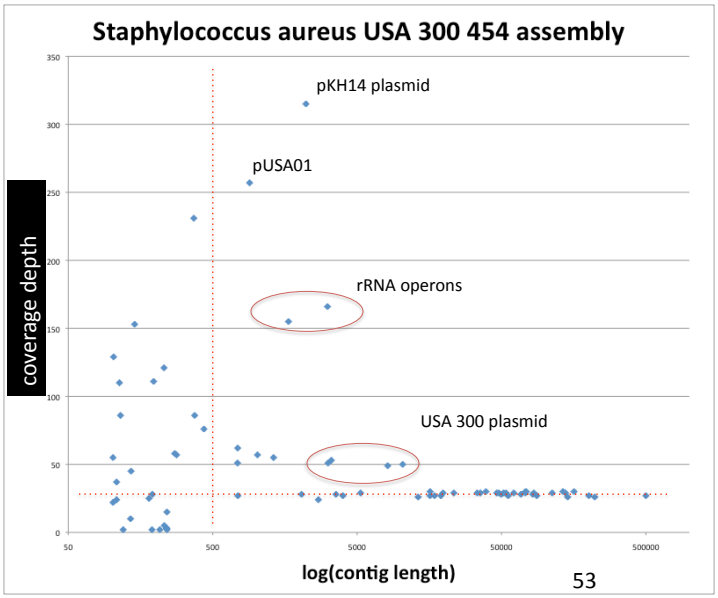
51

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 minimze 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

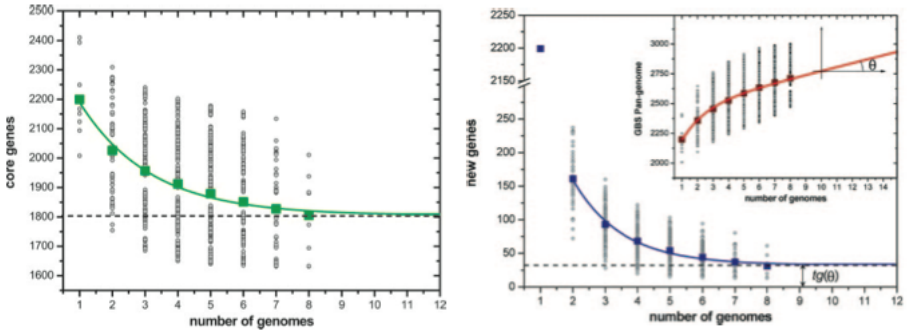
52

Evaluating High Coverage Contigs



53

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

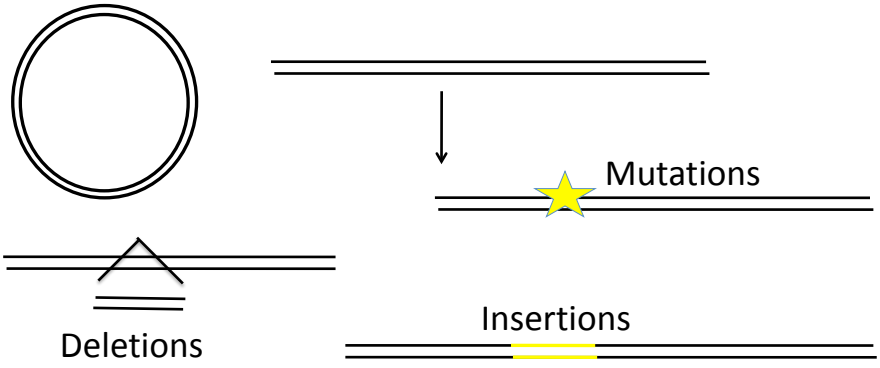


Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin[a,b], Vega Masignani[b,c], Michael J. Cieslewicz[b,d,e], Claudio Donati[c], Duccio Medini[c], Naomi L. Ward[a,f], Samuel V. Angiuoli[a], Jonathan Crabtree[a], Amanda L. Jones[a], A. Scott Durkin[a], Robert T. DeBoy[a], Tanja M. Davidsen[a], Marirosa Mora[c], Maria Scarselli[c], Immaculada Margarit y Ros[c], Jeremy D. Peterson[a], Christopher R. Hauser[a], Jaideep P. Sundaram[a], William C. Nelson[a], Ramana Madupu[a], Lauren M. Brinkac[a], Robert J. Dodson[a], Mary J. Rosovitz[a], Steven A. Sullivan[a], Sean C. Daugherty[a], Daniel H. Haft[a], Jeremy Selengut[a], Michelle L. Gwinn[a], Liwei Zhou[a], Nikhat Zafar[a], Hoda Khouri[a], Diana Radune[a], George Dimitrov[a], Kisha Watkins[a], Kevin J. B. O'Connor[a], Shannon Smith[a], Teresa R. Utterback[a], Owen White[a], Craig E. Rubens[g], Guido Grandi[c], Lawrence C. Madoff[h,i], Dennis L. Kasper[e,j], John L. Telford[c], Michael R. Wessels[d,e], Rino Rappuoli[c,k,l], and Claire M. Fraser[a,b,k,m]

54

Whole Genome Sequencing Reveals



55

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!

56

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

Table 1. Characteristics of the Organ-Transplant Recipients.

Recipient No.	Age	Diagnosis	Organ Transplanted	Clinical Course	Interval between Transplantation and Death
	yr				*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

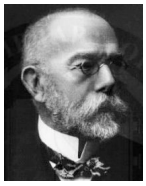
Table 2. Nucleotide and Amino Acid Homologies of the New Arenavirus to Other Arenaviruses.*

Gene	Accession No.	LCMV Strain	Homology	
			Amino Acid	Nucleotide
			percent	
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

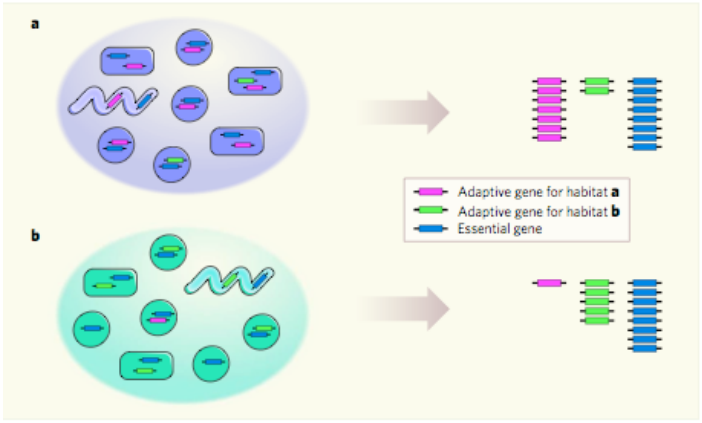
Metagenomics

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. Turnbaugh[1], Micah Hamady[3], Tanya Yatsunenko[1], Brandi L. Cantarel[5], Alexis Duncan[2], Ruth E. Ley[1], Mitchell L. Sogin[6], William J. Jones[7], Bruce A. Roe[8], Jason P. Affourtit[9], Michael Egholm[9], Bernard Henrissat[5], Andrew C. Heath[2], Rob Knight[4] & Jeffrey I. Gordon[1]

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[7], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

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

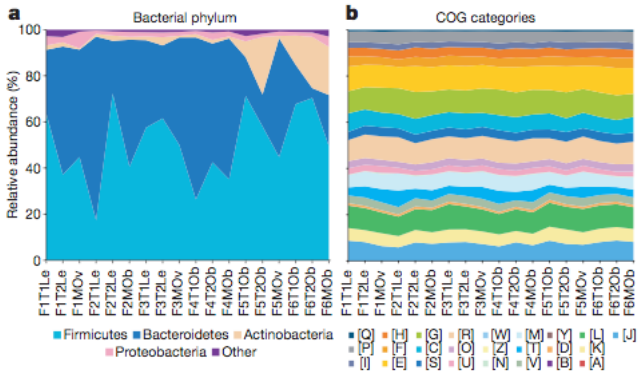


Figure 3 | Comparison of taxonomic and functional variations in the human gut microbiome. a, Relative abundance of major phyla across 18 faecal microbiomes from monozygotic twins and their mothers, based on BLASTX

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