Metagenomics

Smithsonian Institution - Comparative Genomics Course 10/6/2011

Jay Evans and Scott Cornman Bee Research Lab, USDA-ARS Beltsville, MD

- FIRST PRINCIPLES
- DATA COLLECTION
- OCEANS, TERMITES, MAMMALS
- BEE STORIES
- RETURN TO FUNCTION
- ANALYSIS

- FIRST PRINCIPLES
- DATA COLLECTION
- OCEANS, TERMITES, MAMMALS
- BEE STORIES
- RETURN TO FUNCTION
- ANALYSIS

Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

Jo Handelsman[1], Michelle R Rondon[1], Sean F Brady[2], Jon Clardy[2] and Robert M Goodman[1]

Cultured soil microorganisms have provided a rich source of natural-product chemistry. Because only a tiny fraction of soil microbes from soil are readily cultured, soil might be the greatest untapped resource for novel chemistry. The concept of cloning the metagenome to access the collective genomes and the biosynthetic machinery of soil microflora is explored here.

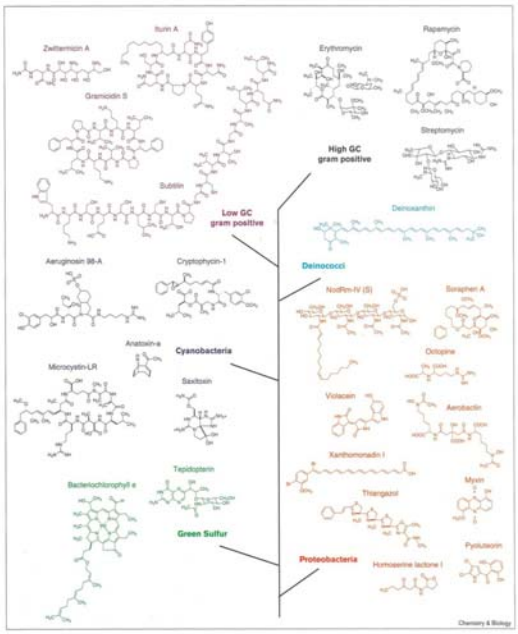
Figure 1

Morphological diversity typical of microorganisms cultured from soil on a broad spectrum medium, tryptic soy agar.

"Tapping into this source should be a great, joint adventure for biologists and chemists"

Chemistry and Biology, Oct 1998

Figure 2

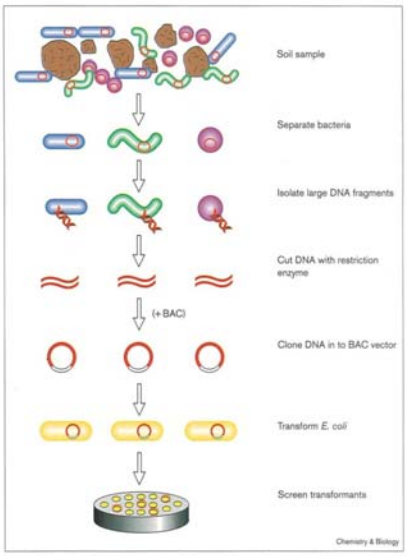


Examples of the chemical diversity in some of the major phyla of bacteria.

Chemistry and Biology, Oct 1998

Figure 3

Cloning the metagenome is our process for isolating new pathways for the synthesis of bioactive molecules from noncultured soil microorganisms. DNA is extracted directly from soil, using gentle methods to preserve high-molecular-weight DNA. The DNA is cut using a restriction enzyme and cloned into a bacterial artificial chromosome (BAC), a vector which can carry large fragments of DNA in *E. coli*. The BAC clones are then screened for biological activity and for the production of novel natural products.



Chemistry and Biology, Oct 1998

"The excitement surrounding this new field lies in the vast diversity of unknown soil microflora and the chemical richness that they are thought to contain"

"The methodology has been made possible by advances in molecular biology and eukaryotic genomics, which have laid the groundwork for **cloning and functional analysis of the collective genomes** of soil microflora, which we term the metagenome of the soil"

Daughter of metagenomics….

Plus combinatorial chemistry, HTP proteomics, other advances in databases and thinking…

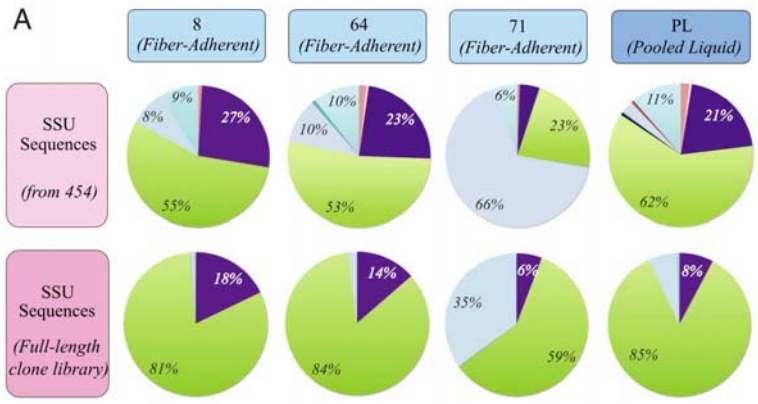
- FIRST PRINCIPLES
- DATA COLLECTION
- OCEANS, TERMITES, MAMMALS
- BEE STORIES
- RETURN TO FUNCTION
- ANALYSIS

Issues in design & analysis

Focus on deep sequencing

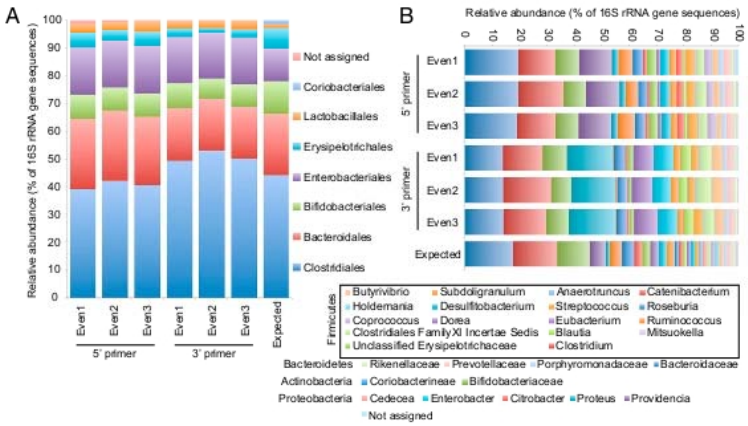
- Does it work?
- Metadata
- Assembly vs. mapping
- Measuring diversity (rDNA) vs. function (enzymes)
  - OTU-based analysis
  - Gene-based analysis

NextGen sequencing has less bias than cloning



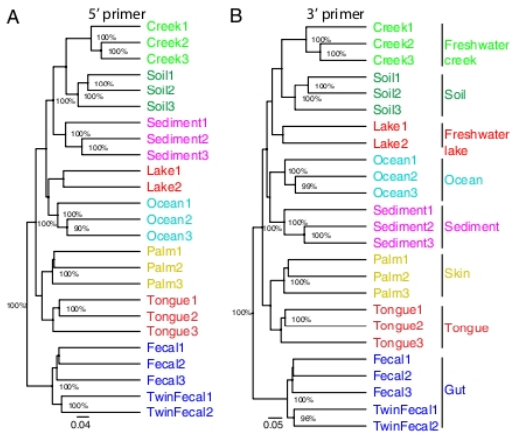
Brulc et al. PNAS 2008

Replication of community diversity estimates with Illumina short reads



Caporaso et al. PNAS 2010

Replication of community diversity estimates with Illumina short reads



Caporaso et al. PNAS 2010

Sequencing platform comparison

(per minimum unit: plate/lane/chip)

454: ~400 bp length, several hundred thousand reads

Illumina (GA/HiSeq): 120 bp, 30 million/150 million reads

ABI Solid: 75 bp max, total output ≥ Illumina

-highest accuracy, fewer software options

Ion Torrent: 200 bp, 4-8 million reads

-fast

Sample preparation

Filtration

Extraction biases exist (use a consistent method)

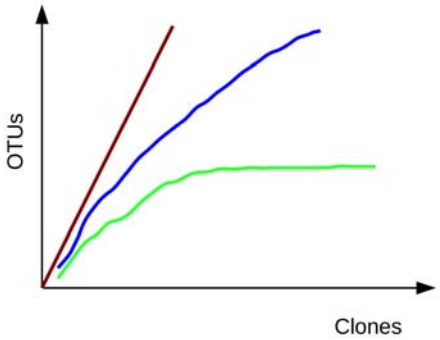
Amplicons vs. whole genome

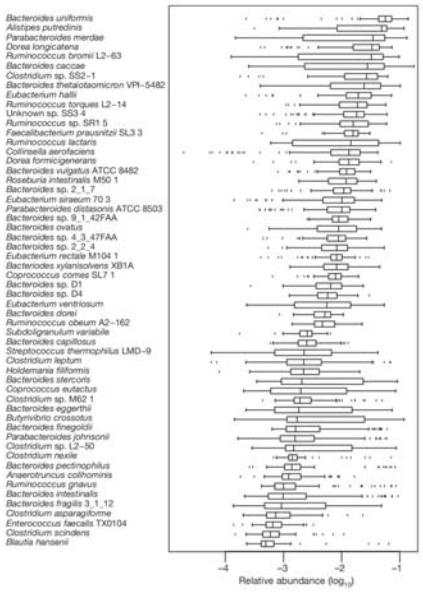
- limitations of universal primers

Normalization of RNA (transcriptomics)

Library preparation for NGS can be difficult, talk with your sequencing center

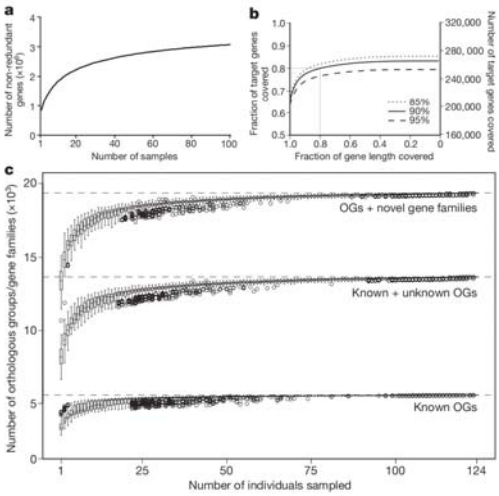
Rarefaction curves estimate completeness





Qin et al., *Nature*, 2010

Rarefaction curves estimate completeness



Qin et al., *Nature,* 2010

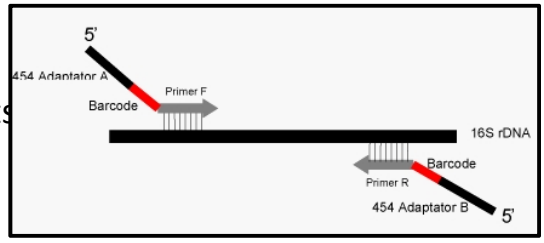
Multiplexing / Barcoding

Independent sequencing libraries run together

- identified by distinct adaptor primers
- divides basic sequencing run into multiple subsets

Applications:

- Biological replicates (key for statistical comparisons)
- Time series
- Multiple treatments



Collection of metadata

Table 1. Sampling Locations and Environmental Data

ID	Sample Location	Country	Date, mm/dd/yy	Time	Location	Sample Depth, m	Water Depth, m	T (°C)[a]	S[b] (ppt)	Size Fraction (µm)	Habitat Type	Chl *a* Sample Month (Annual ± SE) mg/m^{-3}	Good Sequences
G500a	Sargasso Stations 13 and 11	Bermuda (UK)	02/26/03	3:00	31°32′6″ n; 63°35′42″ w	5.0	>4,200	20.0 20.5	36.6	0.1–0.8	Open ocean	0.17 (0.0.9 ± 0.02)	644,551
				10:10	31°10′30″ n; 64°19′27″ w				36.7				
G500b	Sargasso Stations 13 and 11	Bermuda (UK)	02/26/03	3:35	31°32′10″ n; 63°35′70″ w	5.0	>4,200	20.0 20.5	36.6	0.22 0.8	Open ocean	0.17 (0.0.9 ± 0.02)	317,180
				10:43	31°10′50n; 64°19′27″ w				36.7				
G500c	Sargasso Stations 3	Bermuda (UK)	02/25/03	13:00	32°09′30″ n; 64°00′36″ w	5.0	>4,200	19.8	36.7	0.22–0.8	Open ocean	0.17 (0.0.9 ± 0.02)	368,835
G500d	Sargasso Stations 13	Bermuda (UK)	02/25/03	17:00	31°32′6″ n; 63°35′42″ w	5.0	>4,200	20.0	36.6	0.22–0.8	Open ocean	0.17 (0.0.9 ± 0.02)	332,240
G501a	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10′00″ n 64°30′00″ w	5.0	>4,200	22.9	36.7	0.8–3.0	Open ocean	0.10 (0.10 ± 0.01)	142,652
G501b	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10′00″ n; 64°30′00″ w	5.0	>4,201	22.9	36.7	0.8 3.0	Open ocean	0.10 (0.10 ± 0.01)	90,005
G501c	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10′00″ n; 64°30′00″ w	5.0	>4,202	22.9	36.7	0.1 0.8	Open ocean	0.1 (0.1 ± 0.01)	92,851
G502	Gulf of Maine	USA	08/21/03	6:32	42°30′11″ n; 67°14′24″ w	1.0	106	18.7	29.7	0.1–0.8	Coastal	1.4 (1.17 ± 0.19)	121,590
G503	Browns Bank, Gulf of Maine	Canada	08/21/03	11:50	42°51′10″ n; 66°13′2″ w	1.0	119	11.7	29.9	0.1–0.8	Coastal	1.4 (1.12 ± 0.19)	61,605
G504	Outside Halifax, Nova Scotia	Canada	08/22/03	5:25	44°8′14″ n; 63°48′40″ w	2.0	142	17.4	28.1	0.1–0.8	Coastal	0.4 (0.78 ± 0.17)	52,959
G505	Bedford Basin, Nova Scotia	Canada	08/22/03	16:21	44°41′25″ n; 63°38′14″ w	1.0	64	15.0	30.2	0.1–0.8	Embayment	6 (6.76 ± 0.90)	61,131
G506	Bay of Fundy, Nova Scotia	Canada	08/23/03	10:47	45°6′47″ n; 64°55′48″ w	1.0	11	11.2		0.1–0.8	Estuary	2.8 (1.87 ± 0.18)	59,679
G507	Northern Gulf of Maine	Canada	08/25/03	8:25	43°37′56″ n; 66°50′50″ w	1.0	139	17.9	31.7[c]	0.1–0.8	Coastal	1.4 (1.12 ± 0.19)	50,980
G508	Newport Harbor, RI	USA	11/16/03	16:45	41°29′9″ n; 71°21′4″ w	1.0	12	9.4	20.5[c]	0.1–0.8	Coastal	2.3 (1.59 ± 0.17)	139,655
G509	Block Island, NY	USA	11/17/03	10:30	41°5′28″ n; 71°35′8″ w	1.0	32	11.0	31.0[c]	0.1–0.8	Coastal	4.0 (2.72 ± 0.24)	79,303
G510	Cape May, NJ	USA	11/18/03	9:40	38°56′24″ n; 74°41′6″ w	1.0	10	12.0	31.0[c]	0.1–0.8	Coastal	2.0 (2.75 ± 0.33)	78,304
G511	Delaware Bay, NJ	USA	11/18/03	11:30	39°25′4″ n; 75°30′15″ w	1.0	8	11.0		0.1–0.8	Estuary	4.8 (5.73 ± 1.02)	174,435
G512	Chesapeake Bay, MD	USA	12/18/03	11:42	38°56′49″ n; 76°25′2″ w	1.0	25	3.7	3.47[c]	0.1–0.8	Estuary	21.0 (15.0 ± 1.01)	126,152
G513	Off Nags Head, NC	USA	12/19/03	6:38	36°0′14″ n; 75°23′41″ w	1.0	20	9.3		0.1–0.8	Coastal	3.0 (2.24 ± 0.25)	158,044
G514	South of Charleston, SC	USA	12/20/03	17:12	32°30′25″ n; 79°15′50″ w	1.0	31	18.6		0.1–0.8	Coastal	1.70 (1.92 ± 0.25)	128,885
G515	Off Key West, FL	USA	01/08/04	6:25	24°29′18″ n; 83°4′12″ w	2.0	47	25.3	36.0	0.1–0.8	Coastal	0.2 (0.27 ± 0.09)	127,362
G516	Gulf of Mexico	USA	01/08/04	14:15	24°10′29″ n; 84°20′40″ w	2.0	3,333	26.4	35.8	0.1 0.8	Coastal sea	0.16 (0.11 ± 0.01)	177,177
G517	Yucatan Channel	Mexico	01/09/04	13:47	20°31′21″ n; 85°24′49″ w	2.0	4,513	27.0	35.8	0.1–0.8	Open ocean	0.13 (0.09 ± 0.01)	257,581
G518	Roseriо Bank	Honduras	01/10/04	8:12	18°2′12″ n; 83°47′5″ w	2.0	4,470	27.4	35.4	0.1–0.8	Open ocean	0.19 (0.09 ± 0.01)	162,435
G519	Northeast of Colón	Panama	01/13/04	9:03	10°43′59″ n; 80°15′16″ w	2.0	3,196	27.7	35.4	0.1 0.8	Coastal	0.23 (0.15 ± 0.02)	138,325
G520	Lake Gatun	Panama	01/15/04	10:24	9°9′52″ n; 79°50′10″ w	2.0	4	28.5	0.06	0.1–0.8	Fresh water		296,350
G521	Gulf of Panama	Panama	01/19/04	16:40	8°7′45″ n; 79°41′28″ w	2.0	70	27.6	30.7	0.1 0.8	Coastal	0.50 (0.73 ± 0.22)	131,798
G522	250 miles from Panama City	Panama	01/20/04	16:39	6°29′34″ n; 82°54′14″ w	2.0	2,431	29.3	32.3	0.1–0.8	Open ocean	0.33 (0.28 ± 0.02)	121,662
G523	30 miles from Cocos Island	Costa Rica	01/21/04	15:00	5°38′24″ n; 86°33′55″ w	2.0	1,139	28.7	32.6	0.1–0.8	Open ocean	0.07 (0.19 ± 0.02)	155,051
G525	Dirty Rock, Cocos Island	Costa Rica	01/26/04	10:51	5°55′10″ n; 87°5′16″ w	1.1	30	28.3	31.4	0.8 3.0	Fringing reef	0.11 (0.19 ± 0.01)	120,671
G526	134 miles NL of Galapagos	Ecuador	02/01/04	16:16	1°15′51″ n; 90°17′42″ w	2.0	2,376	27.8	32.6	0.1–0.8	Open ocean	0.22 (0.28 ± 0.02)	102,708
G527	Devil's Crown, Floreana	Ecuador	02/04/04	11:41	1°12′58″ s; 90°25′22″ w	2.0	2.3	25.5	34.0	0.1 0.8	Coastal	0.40 (0.38 ± 0.03)	222,080
G528	Coastal Floreana	Ecuador	02/04/04	15:47	1°13′11″ s; 90°16′11″ w	2.0	156	25.0[c]		0.1–0.8	Coastal	0.35 (0.35 ± 0.02)	189,032
G529	North James Bay, Santiago	Ecuador	02/08/04	18:03	0°12′0″ s; 90°50′7″ w	2.0	12	26.0	34.5	0.1–0.8	Coastal	0.10 (0.39 ± 0.03)	113,529
G530	Warm seep, Roca Redonda	Ecuador	02/09/04	11:52	0°16′20″ n; 91°58′0″ w	19.0	19	26.9		0.1–0.8	Warm seep		359,152
G531	Upwelling, Fernandina	Ecuador	02/10/04	14:43	0°19′4″ s; 91°29′6″ w	12.0	19	18.6		0.1–0.8	Coastal upwelling	0.35 (0.39 ± 0.03)	430,481
G532	Mangrove, Isabella	Ecuador	02/11/04	11:30	0°35′38″ s; 91°4′10″ w	0.3	0.67	25.4		0.1–0.8	Mangrove		148,018
G533	Punta Cormorant Lagoon, Floreana	Ecuador	02/19/04	13:35	1°13′42″ s; 90°25′45″ w	0.2	0.33	37.6	46[c]	0.1–0.8	Hypersaline		692,255

Submission of metadata

Specialized Structured Comments

1. **MIGS/MIMS/MIENS**

Minimum information checklists have been developed by the Genomic Standards Consortium (GSC) as a means of reporting core descriptive information about the environment from which an organism(s) was collected. Core descriptors include information about the origins of the nucleic acid sequence (genome), its environment (eg, latitude and longitude, date and time of sampling, habitat) and sequence processing (sequencing and assembly methods).

Three different metadata lists have been developed to describe genomic, metagenomic, and environmental sequences:

- MIGS - Minimum Information About a Genome Sequence
- MIMS - Minimum Information About a Metagenome Sequence
- MIENS - Minimum Information About an Environmental Sequence

The tag-value pairs that are included for each submission type can be validated for compliance with the GSC recommended list. The recommended lists of core descriptors that should be included for each of these sequence types can be found here.

Validation tools within Sequin and tbl2asn will report if structured comments include all of the GSC recommended compliant core descriptors. Submissions that include of all the compliant tags will have a Keyword included within the GenBank flatfile:

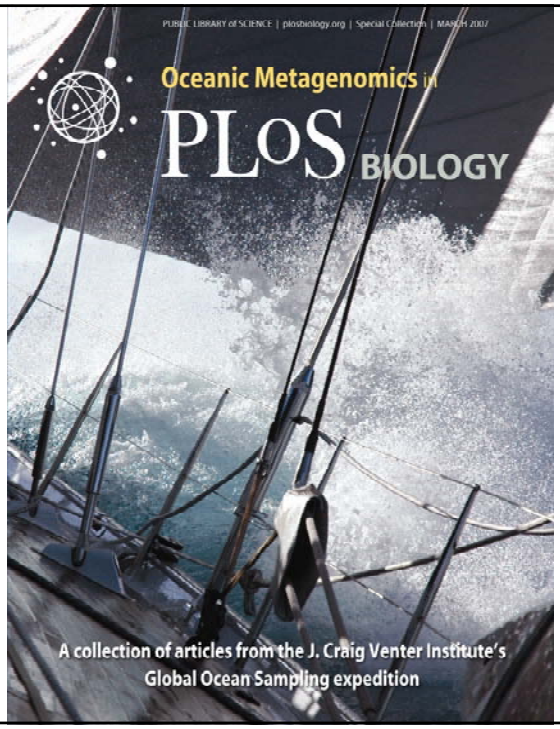
KEYWORD GSC:MIGS:2.1

Structured comments that are not compliant based on the GSC guidelines can still be included within GenBank submissions - they just will not include the keyword.

In order for this validation to occur, you will need to include within the first column in your table a tag that defines the prefix and suffix for the start and end tags within the structured comment, for example:

StructuredCommentPrefix MIGS-Data

- FIRST PRINCIPLES
- DATA COLLECTION
- OCEANS, TERMITES, MAMMALS
- BEE STORIES
- RETURN TO FUNCTION
- ANALYSIS



Various authors, 2007

doi:10.1371/journal.pbio.0050074.g001

Figure 1. Two of a Kind?
The young Charles Darwin (left) and Craig Venter (right). (Photo: J. Craig Venter Institute)

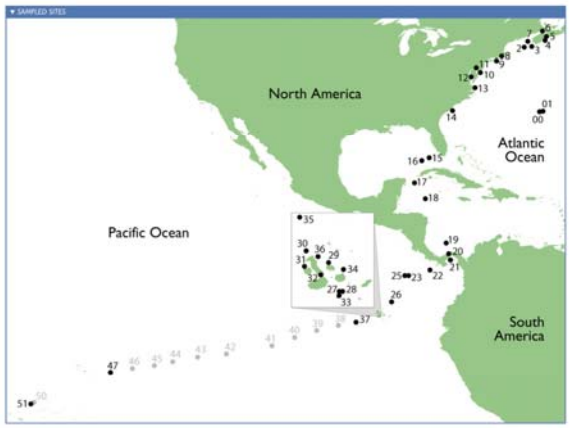


Figure 1. Sampling Sites
Microbial populations were sampled from locations in the order shown. Samples were collected at approximately 200 miles (320 km) intervals along the eastern North American coast through the Gulf of Mexico into the equatorial Pacific. Samples 00 and 01 identify sets of sites sampled as part of the Sargasso Sea pilot study [19]. Samples 27 through 36 were sampled off the Galapagos Islands (see inset). Sites shown in gray were not analyzed as part of this study.
doi:10.1371/journal.pbio.0050077.g001

(Rusch et al., Plos Biol 2007)

Table 1. Some Major Methods for Studying Individual Microbes Found in the Environment

Method	Summary	Comments
Microscopy	Microbial phenotypes can be studied by making them more visible. In conjunction with other methods, such as staining, microscopy can also be used to count taxa and make inferences about biological processes.	The appearance of microbes is not a reliable indicator of what type of microbe one is looking at.
Culturing	Single cells of a particular microbial type are grown in isolation from other organisms. This can be done in liquid or solid growth media.	This is the best way to learn about the biology of a particular organism. However, many microbes are uncultured (i.e., have never been grown in the lab in isolation from other organisms) and may be unculturable (i.e., may not be able to grow without other organisms).
rRNA-PCR	The key aspects of this method are the following: (a) all cell-based organisms possess the same rRNA genes (albeit with different underlying sequences); (b) PCR is used to make billions of copies of basically each and every rRNA gene present in a sample; this amplifies the rRNA signal relative to the noise of thousands of other genes present in each organism's DNA; (c) sequencing and phylogenetic analysis places rRNA genes on the rRNA tree of life; the position on the tree is used to infer what type of organism (a.k.a. phylotype) the gene came from; and (d) the numbers of each microbe type are estimated from the number of times the same rRNA gene is seen.	This method revolutionized microbiology in the 1980s by allowing the types and numbers of microbes present in a sample to be rapidly characterized. However, there are some biases in the process that make it not perfect for all aspects of typing and counting.
Shotgun genome sequencing of cultured species	The DNA from an organism is isolated and broken into small fragments, and then portions of these fragments are sequenced, usually with the aid of sequencing machines. The fragments are then assembled into larger pieces by looking for overlaps in the sequence each possesses. The complete genome can be determined by filling in gaps between the larger pieces.	This has now been applied to over 1,000 microbes, as well as some multicellular species, and has provided a much deeper understanding of the biology and evolution of life. One limitation is that each genome sequence is usually a snapshot of one or a few individuals.
Metagenomics	DNA is directly isolated from an environmental sample and then sequenced. One approach to doing this is to select particular pieces of interest (e.g., those containing interesting rRNA genes) and sequence them. An alternative is ESS, which is shotgun genome sequencing as described above, but applied to an environmental sample with multiple organisms, rather than to a single cultured organism.	This method allows one to sample the genomes of microbes without culturing them. It can be used both for typing and counting taxa and for making predictions of their biological functions.

doi:10.1371/journal.pbio.0050082.t001

(Rusch et al., Plos Biol 2007)

Table 2. Methods of Binning

Method	Description	Comments
Genome assembly	Identify regions of overlap between different fragments from the same organism to build larger contiguous pieces (contigs).	Getting deep enough sampling for this to work is very expensive except for low diversity systems or for very abundant taxa.
Reference genome alignment	Identify ESS fragments or contigs that are very similar to already assembled sections of the genome of single microbial types.	(a) One of the most effective ways to sort through ESS data, if the reference genome is very closely related to an organism in the sample; (b) the reason why more reference genomes are needed; (c) does not handle regions present in uncultured organisms but not in the reference.
Phylogenetic analysis	Build evolutionary trees of genes encoded by ESS fragments or contigs. Assign fragments or contigs to taxonomic groups based on nearest neighbor(s) in trees.	(a) Very powerful, but level of resolution depends on whether fragments encode useful phylogenetic markers and on how well sampled the database is for the neighbor analysis; (b) would work much better if more genomes were available from across the tree of life.
Word frequency and nucleotide composition analysis	Measure word frequency and composition of each fragment. Group by clustering algorithms or principal component analysis.	(a) Has the potential to work because organisms sometimes have "signatures" of word frequencies that are found throughout the genome and are different between species; (b) very challenging for small fragments.
Population genetics	Build alignments of fragments or contigs with similarity to each other (but not as much as needed for assembly). Examine haplotype structure, predicted effective population size, and synonymous and non synonymous substitution patterns.	May be most useful as a way of subdividing bins created by other methods.

Note that some methods can be applied to ESS fragments or to bins identified by other methods.
doi:10.1371/journal.pbio.0050082.t002

(Rusch et al., Plos Biol 2007)

Table 1. The Complete Dataset Consisted of Sequences from NCBI-nr, ENS, TGI-EST, PG, and GOS, for a Total of 28,610,944 Sequences

Dataset	Source	Number of Amino Acid Sequences	Mean Sequence Length	Brief Description
NCBI-nr	NCBI	2,317,995	339	Consists of protein sequences submitted to SWISS-PROT, PDB, PIR, and PRF, and also predicted proteins from both finished and unfinished genomes in GenBank, EMBL, and DDBJ.
PG ORFs	NCBI	3,049,695	160	ORFs identified from 222 prokaryotic genome projects. Organisms are listed in Protocol S1.
TGI-EST ORFs	TIGR Gene Index	5,458,820	119	ORFs identified from 72 datasets in which each dataset consists of EST assemblies. Organisms are listed in Protocol S1.
ENS	Ensembl	361,668	466	Sequences from 12 species, including human, mouse, rat, chimp, zebrafish, fruit fly, mosquito, honey bee, dog, two species of puffer fish, chicken, and worm.
GOS ORFs	J. Craig Venter Institute	17,422,766	134	ORFs identified from an assembly of 7.7 million reads. These reads include both the reads from the *Sorcerer II* GOS Expedition and the reads from the earlier Sargasso Sea study. Also included are 36,318 ORFs identified from an assembly of sequences collected from the viral size (< 0.1 μm) fraction of one sample.

doi:10.1371/journal.pbio.0050016.t001

(Rusch et al., Plos Biol 2007)

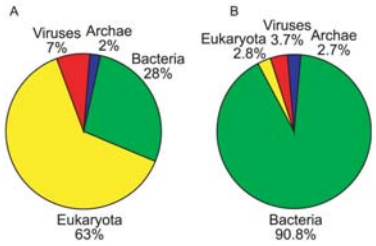


Figure 1. Proportion of Sequences for Each Kingdom

(A) The combined set of NCBI-nr, PG, TGI-EST, and ENS has 3,167,979 sequences. The eukaryotes account for the largest portion and is more than twice the bacterial fraction.

(B) Predicted kingdom proportion of sequences in GOS. Out of the 5,654,638 GOS sequences, 5,058,757 are assigned kingdoms using a BLAST-based scheme. The bacterial kingdom forms by far the largest fraction in the GOS set.

doi:10.1371/journal.pbio.0050016.g001

Table 7. Taxonomic Makeup of GOS Samples Based on 16S Data from Shotgun Sequencing

Phylum or Class	Fraction[a]
Alpha *Proteobacteria*	0.32
Unclassified *Proteobacteria*	0.155
Gamma *Proteobacteria*	0.132
Bacteroidetes	0.13
Cyanobacteria	0.079
Firmicutes	0.075
Actinobacteria	0.046
Marine Group A	0.022
Beta *Proteobacteria*	0.017
OP11	0.008
Unclassified *Bacteria*	0.008
Delta *Proteobacteria*	0.005
Planctomycetes	0.002
Epsilon *Proteobacteria*	0.001

(Rusch et al., Plos Biol 2007)

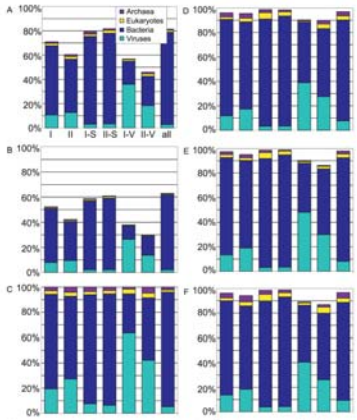


Figure 16. GOS-Only Clusters Are Enriched for Sequences of Viral Origin Independently of the Kingdom Assignment Method Employed

For each panel, clusters are as in Figure 4. For (A–C), a kingdom is assigned to each neighboring ORF within each cluster set; the percentage of all neighboring ORFs with a given kingdom assignment is plotted. For (D–F), a kingdom is assigned to each cluster if more than 50% of all that cluster's neighbors with a kingdom assignment share the same assignment; the percentage of clusters in each set with a given assignment is plotted. In (A) and (D), a kingdom is assigned to a neighboring ORF by a majority vote of the top four BLAST matches to a protein in NCBI-nr (Materials and Methods). In (B) and (E), a kingdom is assigned if all eight highest-scoring BLAST matches agree in kingdom. In (C) and (F), all ORFs on a scaffold are assigned the same kingdom by voting among all ORFs with BLAST matches to NCBI-nr on that scaffold (Materials and Methods). In all graphs, only clusters with at least one assignable neighbor are considered. When compared to the size-matched controls, in all cases the GOS-only clusters show enrichment for viral sequences.

doi:10.1371/journal.pbio.0050016.g016

(Rusch et al., Plos Biol 2007)

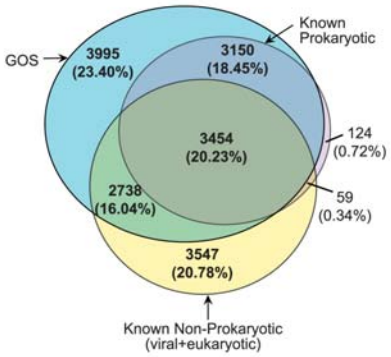


Figure 3. Venn Diagram Showing Breakdown of the 17,067 Medium and Large Clusters by Three Categories—GOS, Known Prokaryotic, and Known Nonprokaryotic

doi:10.1371/journal.pbio.0050016.g003

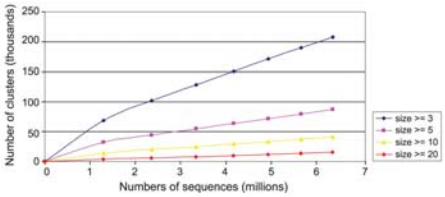
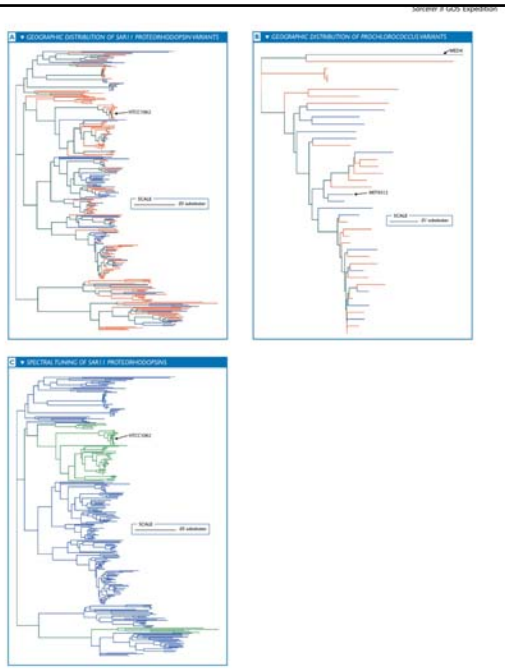


Figure 2. Rate of Discovery of Clusters as (Nonredundant) Sequences Are Added

The *x*-axis denotes the number of sequences (in millions) and the *y*-axis denotes the number of clusters (in thousands). Seven datasets with increasing numbers of (nonredundant) sequences are chosen as described in the text. The blue curve shows the number of core sets of size ≥ 3 for the seven datasets. Curves for core set sizes ≥ 5, ≥ 10, and ≥ 20 are also shown. Linear regression gives slopes 0.027 ($R^2 = 0.999$), 0.011 ($R^2 = 0.999$), 0.0053 ($R^2 = 0.999$), and 0.0024 ($R^2 = 0.996$) for size ≥ 3, size ≥ 5, size ≥ 10, and size ≥ 20, respectively.

doi:10.1371/journal.pbio.0050016.g002

(Rusch et al., Plos Biol



(Rusch et al., Plos Biol 2007)

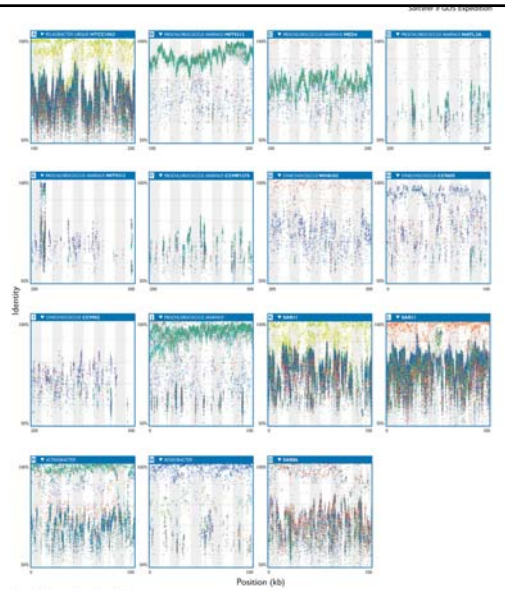
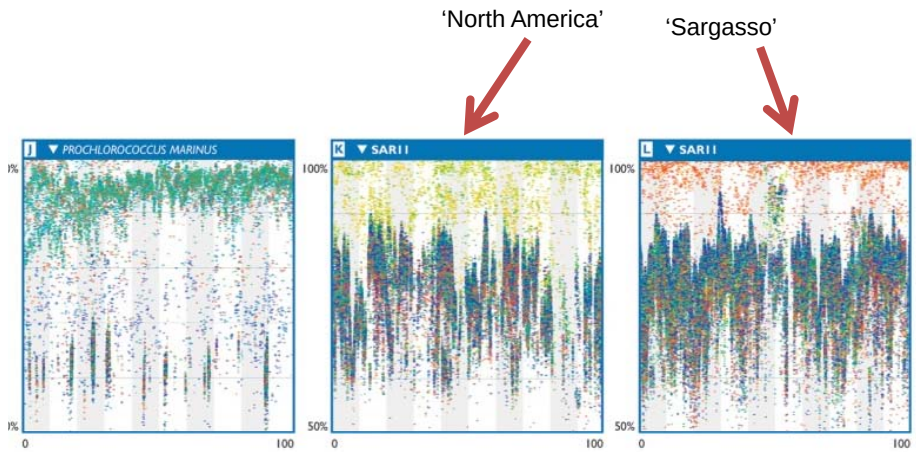


Figure 2. Fragment Recruitment Plots

The horizontal axis of each panel corresponds to a 100-kb segment of genomic sequence from the indicated reference microbial genome. The vertical axis indicates the sequence identity of an alignment between a GOS sequence and the reference genomic sequence. The identity ranges from 100% (top) to 50% (bottom). Individual GOS sequencing reads were colored to reflect the sample from which they were isolated. Geographically nearby samples have similar colors (see Poster S1 for key). Each organism shows a distinct pattern of recruitment reflecting its origin and relationship to the environmental data collected during the course of this study.

(A) *P. ubique* HTCC1062 recruits the greatest density of GOS sequences of any genome examined to date. The GOS sequences show geographic

(Rusch et al., Plos Biol 2007)



(*Pelagibacter ubique* strains)

(Rusch et al., Plos Biol 2007)

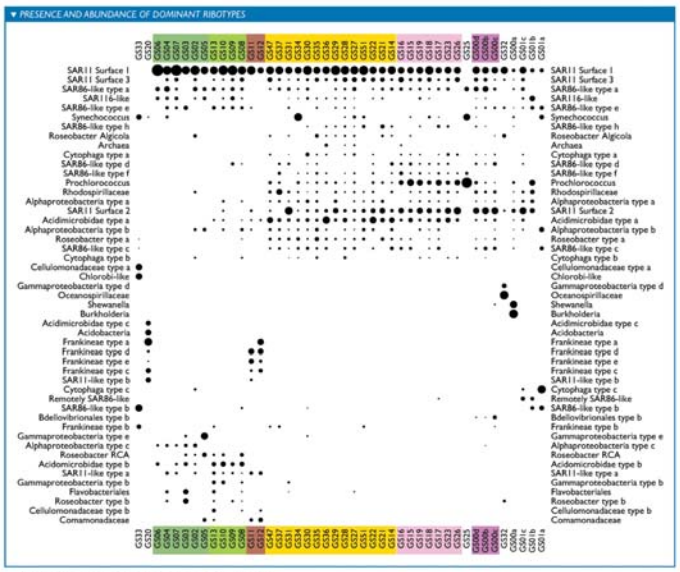
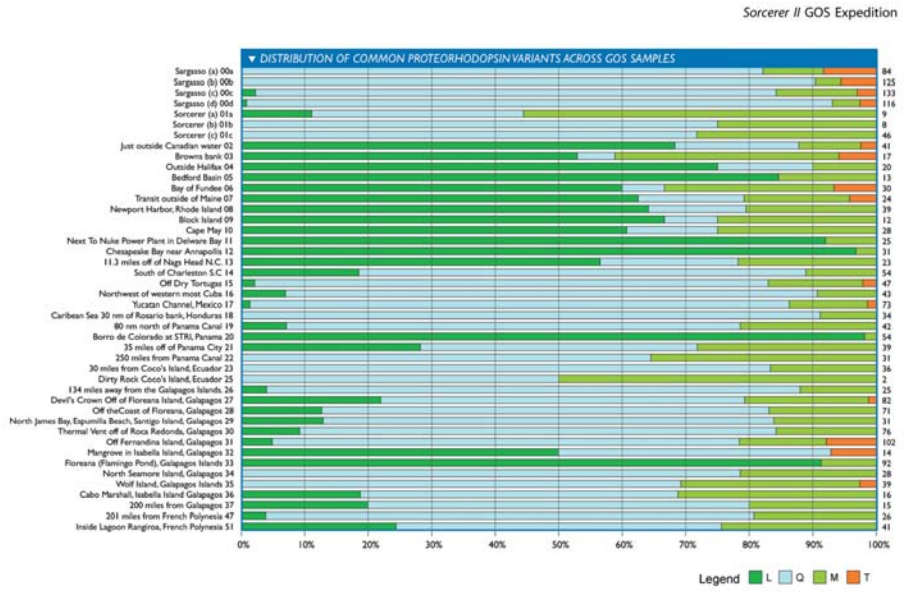


Figure 9. Presence and Abundance of Dominant Ribotypes

The relative abundance of various ribotypes (rows) in each filter (columns) is represented by the area of the corresponding spot (if any). The listed ribotypes each satisfied the following criteria in at least one filter: the ribotype was among the five most abundant ribotypes detected in the shotgun data, and was represented by at least three sequencing reads. Relative abundance is based on the total number of 16S sequences in a given filter. Order and grouping of filters is based on the clustering of genomic similarity shown in Figure 11. Ribotype order was determined based on similarity of sample distribution. A marked contrast between temperate and tropical groups is visible. Estuarine samples GS11 and GS12 contained a mix of ribotypes seen in freshwater and temperate marine samples, while samples from nonmarine habitats or larger filter sizes were pronounced outliers. The presence of large amounts of *Burkholderia* and *Shewanella* in one Sargasso Sea sample (GS00a) makes this sample look much less like other Sargasso and tropical marine samples than it otherwise would. Note that 16S is not a measure of cell abundance since 16S genes can be multicopy.
doi:10.1371/journal.pbio.0050077.g009

(Rusch et al., Plos Biol 2007)



(Rusch et al., Plos Biol 2007)

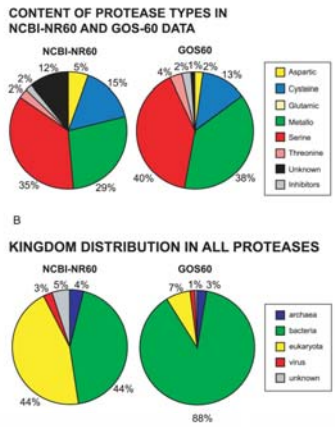


Figure 17. Content of Protease Types in NCBI-nr and GOS, and Kingdom Distribution of All Proteases

Due to the highly redundant nature of some NCBI-nr protease groups, nonredundant sets for both NCBI-nr and GOS are computed; these nonredundant sets are referred to as NCBI-nr60 and GOS60.
doi:10.1371/journal.pbio.0050016.g017

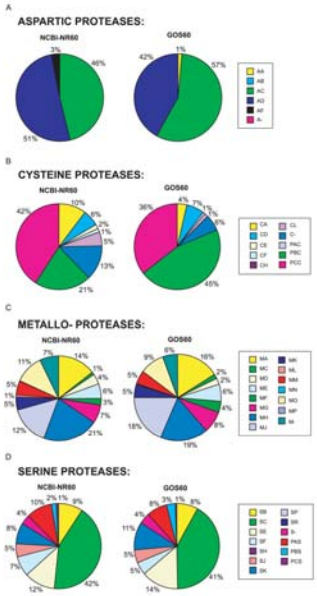


Figure 18. Content of Bacterial Protease Clans
doi:10.1371/journal.pbio.0050016.g018

(Warnecke et al., *Nature* 2007)

LETTERS

Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite

Falk Warnecke[1]*, Peter Luginbühl[2]*, Natalia Ivanova[1], Majid Ghassemian[2], Toby H. Richardson[2]†, Justin T. Stege[2], Michelle Cayouette[2], Alice C. McHardy[3]†, Gordana Djordjevic[2], Nahla Aboushadi[2], Rotem Sorek[1], Susannah G. Tringe[1], Mircea Podar[4], Hector Garcia Martin[1], Victor Kunin[1], Daniel Dalevi[1], Julita Madejska[1], Edward Kirton[1], Darren Platt[1], Ernest Szeto[1], Asaf Salamov[1], Kerrie Barry[1], Natalia Mikhailova[1], Nikos C. Kyrpides[1], Eric G. Matson[5], Elizabeth A. Ottesen[6], Xinning Zhang[3], Myriam Hernández[7], Catalina Murillo[7], Luis G. Acosta[7], Isidore Rigoutsos[3], Giselle Tamayo[7], Brian D. Green[2], Cathy Chang[2], Edward M. Rubin[1], Eric J. Mathur[2]†, Dan E. Robertson[2], Philip Hugenholtz[1] & Jared R. Leadbetter[5]*

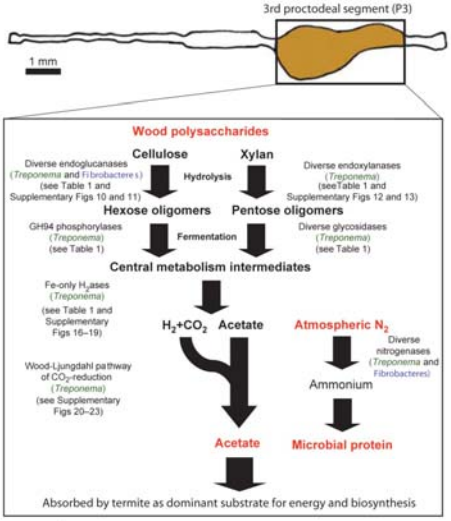


Figure 2 | Model of nutritional symbiosis-relevant metabolism by *Nasutitermes* P3 luminal bacteria. Wood is triturated by the insect's mandibles into small particles and predigested by poorly studied upstream processes before transit into the P3 compartment. The P3 lumen is dominated by diverse species of *Treponema* (Spirochaetes) and Fibrobacteres. There was no evidence for methanogenesis or lignin degradation in the metagenomic data set.

(Warnecke et al., *Nature* 2007)

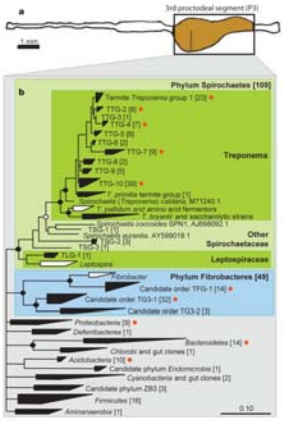
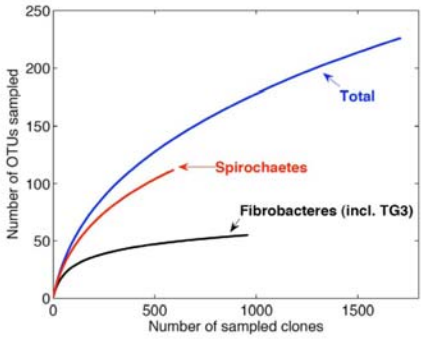


Figure 1 | Source and composition of the bacterial community analysed. **a,** Diagram of the gut tract extracted from the *Nasutitermes* host insect. The anterior of the P3 paunch was incised as indicated (dotted line) with a syringe needle. The luminal contents from about 165 specimens were collected and pooled for nucleic acid and protein analyses. **b,** Phylogenetic diversity of the P3 luminal microbiota. From a PCR-based inventory and from the metagenome libraries, 1,703 almost full-length and 41 full-to-partial 16S rRNA gene sequences were used in a maximum-likelihood analysis (RAxML). The number of distinct P3 luminal community phylotypes contained within each grouping is given in brackets (a total of 216 operational taxonomical units sharing at least 99% sequence identity); black shading denotes that at least one of these phylotypes was represented within the PCR inventory or metagenome libraries (red dots). White shading denotes reference groups having no representation in this or other public inventories derived from termite gut. The phylogram was constructed from 1,289 unambiguously aligned and filtered nucleotide positions. Branching pattern confidence is visualized as follows: no circle, more than 50%; open circle, more than 70%; filled circle, more than 90%. Scale bar indicates 10% estimated sequence divergence. See Supplementary Tables 1, 2 and 11 and Supplementary Figs 2–9 for accession numbers and detailed phylogenetic analyses.



(Warnecke et al., *Nature* 2007)

The ISME Journal (2011) 5, 1133–1142

www.nature.com/ismej

ORIGINAL ARTICLE

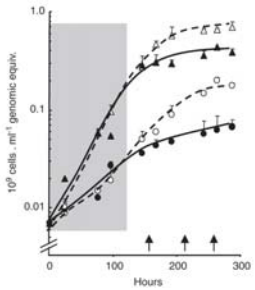
RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture

Adam Z Rosenthal[1], Eric G Matson[1], Avigdor Eldar[2] and Jared R Leadbetter[1]
[1]*Ronald and Maxine Linde Center for Global Environmental Science, California Institute of Technology, Mailcode 138-78, Pasadena, CA, USA and* [2]*Howard Hughes Medical Institute and Division of Biology and Department of Applied Physics, California Institute of Technology, Pasadena, CA, USA*

Table 1 A hypothetical short sequence data set and experimental RNA-sequencing data preferentially aligned to the cognate genome

	Total DB size	*Total hits (% rRNA)*	*Non-rRNA hits*	*Unique loci*[a]	*Genes with a hit (%)*
Hypothetical					
Exact	3 855 671	6518 (45%)	2923	2923	109 (2.8%)
1 miss	3 855 671	14 086 (27%)	10 283	10 283	290 (7.6%)
Actual[b]					
Before mask	10 943 994	1 936 998 (99.9%)	1525	575	161 (4.1%)
After mask[c]	542 600	646 (0%)	646	340	139 (3.6%)
Actual[d]					
Before mask	15 151 014	2 400 846 (99.9%)	1280	409	194 (4.8%)
After mask[c]	427 444	380 (0%)	380	204	151 (3.8%)

Abbreviation: rRNA, ribosomal RNA.
[a]Unique loci refer to the number of distinct non-ribosomal sequences of *T. primitia* that had at least one hit.
An *in-silico*-generated data set of all possible 37 base-pair sequences in the genome of *T. azotonutricium* (hypothetical) and the RNA-seq data from a sample of *T. azotonutricium* (denoted as 'Actual[b]' in the table) were mapped to the genome of *T. primitia*. Database size describes the number of short sequences in the data sets. The total hits column displays the number of short sequences that mapped to the *T. primitia* genome, and the percentage of hits that align to ribosomal 16S or 23S.
[c]After mask sequences are those remaining after the most similar sequences between the two genomes were removed from sampling.
[d]Refers to the RNA-Seq data from a sample of *T. primitia*, when mapped to the genome of *T. azotonutricium*.



	T_d 1	T_d 2	Yield
▲ *T. primitia*	32.2	179	0.43
△ *T. prim co-culture*	30.9	78.5	0.69
● *T. azotonutricium*	64	158.8	0.2
○ *T. azoto co-culture*	49	59.5	0.07

Figure 1 Synergistic growth effects in co-culture of termite-gut treponemes. Quantitative PCR measurements are given for *T. primitia* and *T. azotonutricium* abundance in either pure culture (solid lines) or co-culture (dashed lines). The growth phase in which bacteria in both pure cultures and co-cultures grow at roughly equal rates is marked by a shaded box. RNA-sequencing samples were taken at approximately 160 hours (co-culture), 210 h (*T. primitia*) and 260 h (*T. azotonutricium*). These collection times are in the later growth phase and are marked by arrows on the *X* axis. Samples were run in triplicates, and the s.d. is denoted by error bars. Growth rates and yield of each sample are summarized below the curve. Yields are in number of genomic equivalents per ml. Doubling times were calculated from data points either within the first 117 h of growth (Td1) or within the productive stage of growth that proceeds (Td2).

(Rosenthal et al, ISME, 2011)

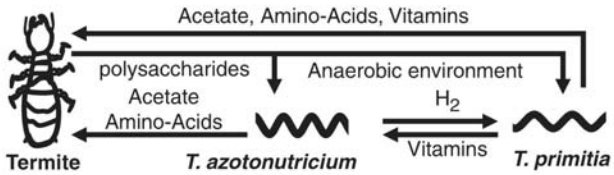


Figure 4 Schematic representation of the symbiosis between *T. primitia* and *T. azotonutricium* and the termite host.

1138

Table 2 Major groups of transcriptionally up- and downregulated genes in co-culture

Process	*Function*	*Up in co-culture*	*Down in co-culture*
Treponema primitia			
Metabolism	Hydrogen and C1 metabolism	5	7
Vitamins and cofactors	B_{12} and corrinoid related	2	23
	Tryptophan/phenylalanine/tyrosine biosynthesis	10	
Amino acids	Methionine synthesis and transport	3	
	Isoleucine/leucine/valine transport	2	
Treponema azotonutricium			
Vitamins and cofactors	B_{12} and corrinoid-related		8
	Biotin transport, regulation, metabolism		3
	Vitamin B_6 precursor synthesis		1
	Enzymes requiring B6 for activity		3
Amino acids	Isoleucine/leucine/valine biosynthesis	4	
	Serine[a]		3
	Cysteine[a]		2

Regulated genes and gene clusters of *T. primitia* and *T. azotonutricium* are listed by major cellular pathways. Included are genes with clear annotation and fold change that is above background (see Materials and methods), and which are discussed in the manuscript. The values in the up and down columns describe the number of genes associated with a specific process.
[a]Some genes involved in the serine and cysteine biosynthesis pathways require vitamin B_6, and appear in both categories.

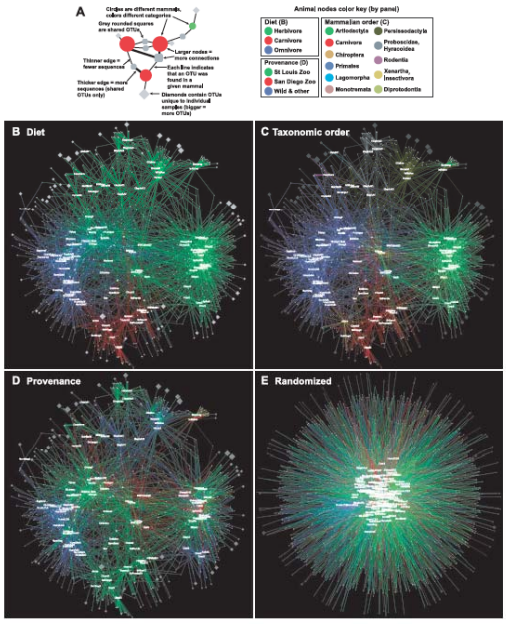
(Rosenthal et al, ISME, 2011)

Evolution of Mammals and Their Gut Microbes

Ruth E. Ley,[1] Micah Hamady,[2] Catherine Lozupone,[1,3] Peter J. Turnbaugh,[1] Rob Roy Ramey,[4] J. Stephen Bircher,[5] Michael L. Schlegel,[6] Tammy A. Tucker,[6] Mark D. Schrenzel,[6] Rob Knight,[3] Jeffrey I. Gordon[1]*

Mammals are metagenomic in that they are composed of not only their own gene complements but also those of all of their associated microbes. To understand the coevolution of the mammals and their indigenous microbial communities, we conducted a network-based analysis of bacterial 16S ribosomal RNA gene sequences from the fecal microbiota of humans and 59 other mammalian species living in two zoos and in the wild. The results indicate that host die and phylogeny both influence bacterial diversity, which increases from carnivory to omnivory to herbivory; that bacterial communities codiversified with their hosts; and that the gut microbiota of humans living a modern life-style is typical of omnivorous primates.

SCIENCE VOL 320 20 JUNE 2008



(Ley et al., Science, 2008)

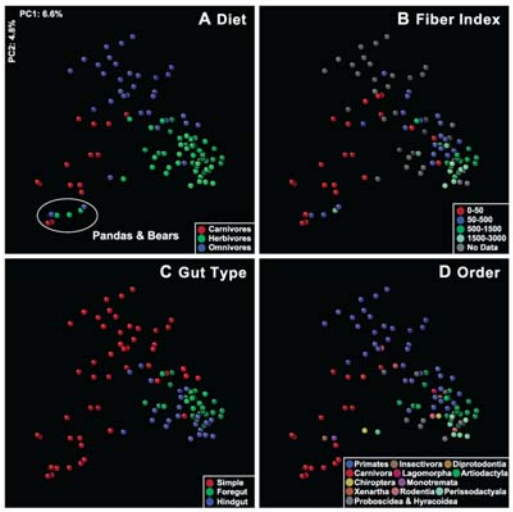
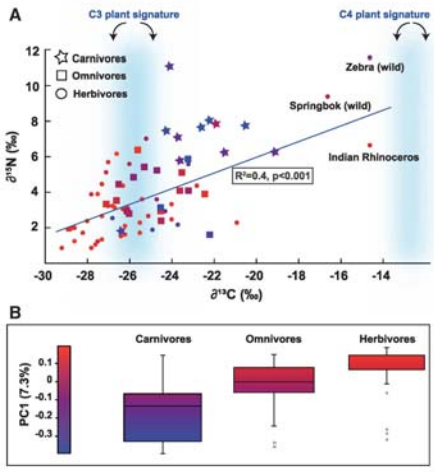


Fig. 2. Mammalian fecal bacterial communities clustered using principal coordinates analysis (PCoA) of the UniFrac metric matrix. PC1 and PC2 are plotted on *x* and *y* axes. Each circle corresponds to a fecal sample colored according to (**A**) diet, (**B**), diet fiber index, (**C**), gut morphology/physiology, and (**D**) host taxonomic order. The same data (samples) are shown in each panel. The percentage of the variation explained by the plotted principal coordinates is indicated on the axes.

(Ley et al., Science, 2008)

Fig. 3. Markers of trophic level mapped onto the variance in fecal microbial community diversity. (**A**) Stable isotope values for C and N plotted for each fecal sample, presented according to diet group. Symbols are colored according to their PC1 value; PC1 is the first principal coordinate of the PCoA of the unweighted UniFrac metric. $\delta^{13}C$ ranges for C3 and C4 plants [per mil (‰)] are highlighted in blue. R^2 is for $\delta^{13}C$ versus $\delta^{15}N$. (**B**) Box plots are shown for the three diet groups (central line is the mean; box outline equals 1 SD; the bar denotes 2 SD; circles are outliers). The majority of fecal $\delta^{13}C$ values are intermediate between the average for C4 plants (–12.5%) and C3 plants (–26.7%).



(Ley et al., Science, 2008)

ARTICLES

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[2], 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[1,3], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

Vol 464|4 March 2010| doi:10.1038/nature08821

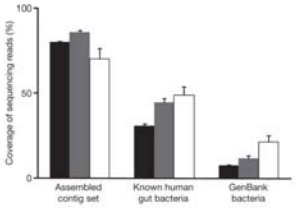


Figure 1 | Coverage of human gut microbiome. The three human microbial sequencing read sets—Illumina GA reads generated from 124 individuals in this study (black; $n = 124$), Roche/454 reads from 18 human twins and their mothers (grey; $n = 18$) and Sanger reads from 13 Japanese individuals (white; $n = 13$)—were aligned to each of the reference sequence sets. Mean values ± s.e.m. are plotted.

NATURE|Vol 464|4 March 2010

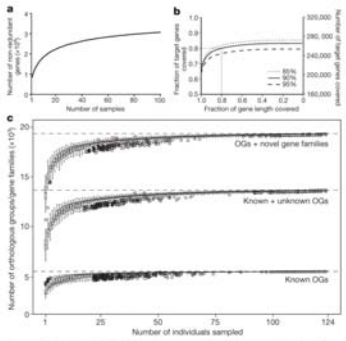
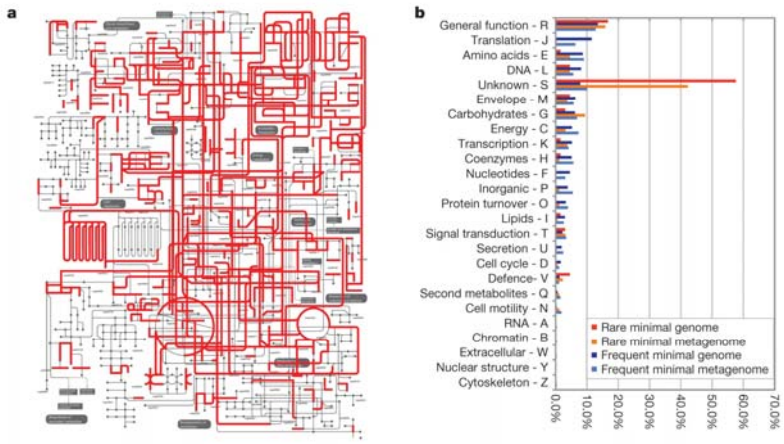


Figure 2 | Predicted ORFs in the human gut microbiome. **a**, Number of unique genes as a function of the extent of sequencing. The gene accumulation curve corresponds to the S_{obs} (Mao Tau) values (number of observed genes), calculated using EstimateS[27] (version 8.2.0) on randomly chosen 100 samples (due to memory limitation). **b**, Coverage of genes from 89 frequent gut microbial species (Supplementary Table 12). **c**, Number of functions captured by number of samples investigated, based on known (well characterized) orthologous groups (OGs; bottom), known plus unknown orthologous groups (including, for example, putative, predicted, conserved hypothetical functions; middle) and orthologous groups plus novel gene families (>20 proteins) recovered from the metagenome (top). Boxes denote the interquartile range (IQR) between the first and third quartiles (25th and 75th percentiles, respectively) and the line inside denotes the median. Whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote outliers beyond the whiskers.

Qin et al., *Nature,* 2010



Qin et al., *Nature,* 2010

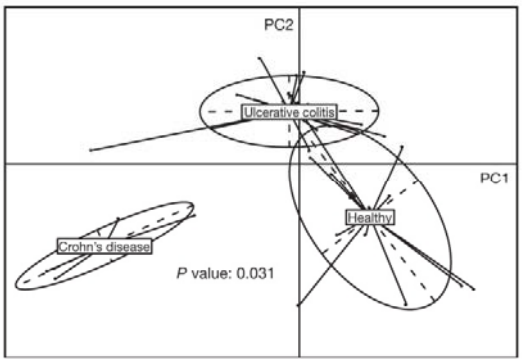


Figure 4 | Bacterial species abundance differentiates IBD patients and healthy individuals. Principal component analysis with health status as instrumental variables, based on the abundance of 155 species with ≥1% genome coverage by the Illumina reads in at least 1 individual of the cohort, was carried out with 14 healthy individuals and 25 IBD patients (21 ulcerative colitis and 4 Crohn's disease) from Spain (Supplementary Table 1). Two first components (PC1 and PC2) were plotted and represented 7.3% of whole inertia. Individuals (represented by points) were clustered and centre of gravity computed for each class; *P*-value of the link between health status and species abundance was assessed using a Monte-Carlo test (999 replicates).

Qin et al., *Nature,* 2010

MICROBIAL ECOLOGY

Human gut microbes associated with obesity

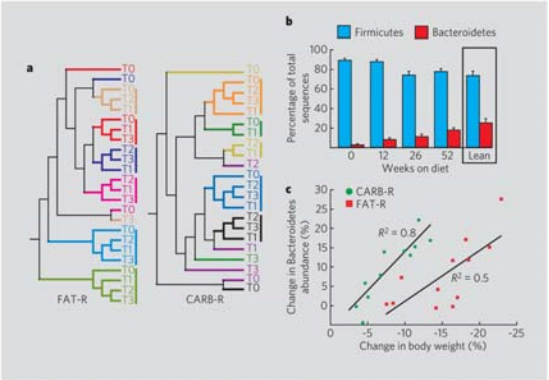


Figure 1 | Correlation between body-weight loss and gut microbial ecology. a, Clustering of 16S ribosomal RNA gene sequence libraries of faecal microbiota for each person (in different colours) and time point in diet therapy (T0, baseline; T1, 12 weeks; T2, 26 weeks; T3, 52 weeks) in the two diet-treatment 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-subject controls include four stool samples from two people taken 1 year apart, plus three other stool samples[6]. Mean values ± s.e. are plotted. **c,** Change in relative abundance of Bacteroidetes in subjects with weight loss above a threshold of 2% weight loss for the CARB-R diet and 6% for the FAT-R diet.

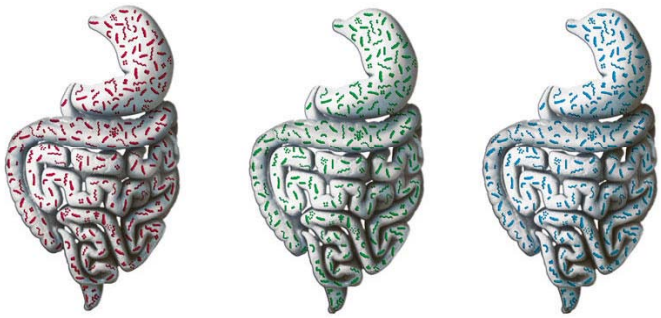
Ley et al., NATURE|Vol 444|21/28 December 2006 BRIEF COMMUNICATIONS

ARTICLE

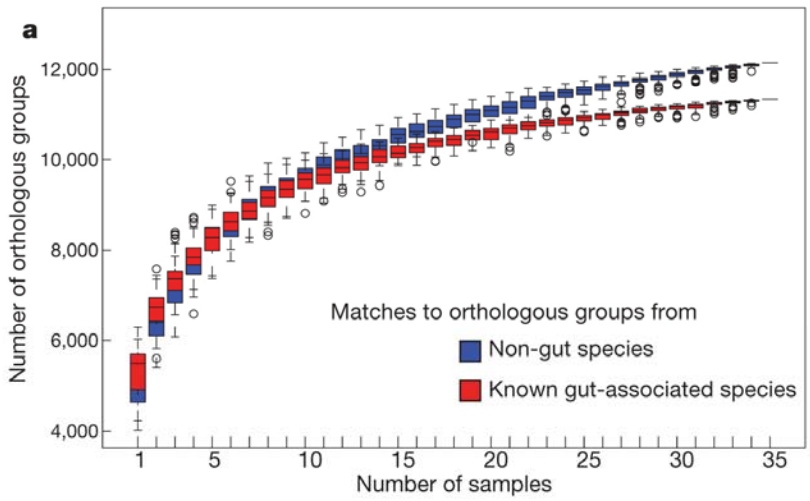
doi:10.1038/nature09944

Enterotypes of the human gut microbiome

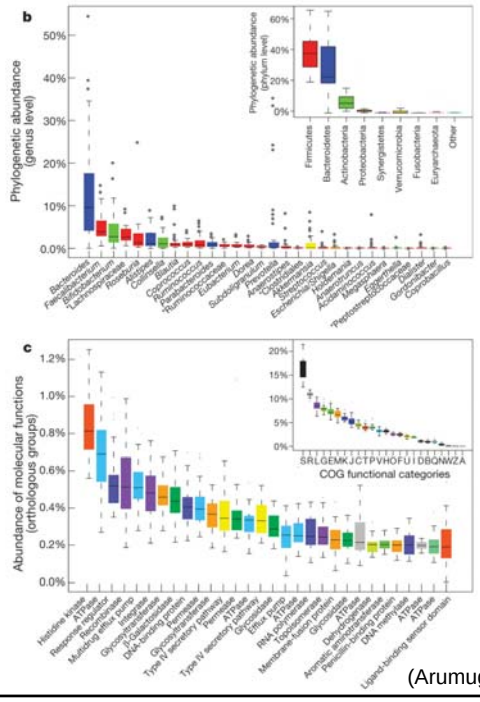
Manimozhiyan Arumugam[1*], Jeroen Raes[1,2*], Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[10,19], Edgardo Ugarte[3], Erwin G. Zoetendal[15], Jun Wang[17,20], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24], Søren Brunak[8], Joel Doré[7], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[7] & Peer Bork[1,25]



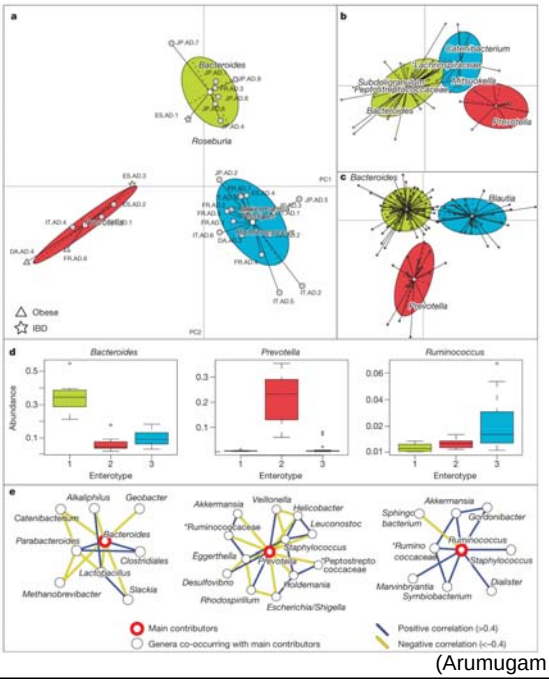
(Arumugam et al., *Nature*, 2011)



(Arumugam et al., *Nature*, 2011)



(Arumugam et al., *Nature*, 2011)



(Arumugam et al., *Nature*, 2011)

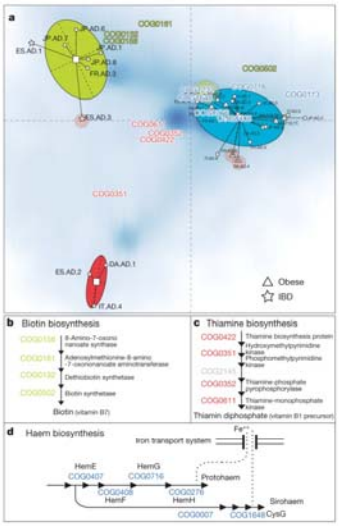


Figure 3 | Functional differences between enterotypes. a, Between-class analysis (see Fig. 2) of orthologous group abundances showing only minor disagreements with enterotypes (unfilled circles indicate the differing samples). The blue cloud represents the local density estimated from the coordinates of orthologous groups; positions of selected orthologous groups are highlighted. **b,** Four enzymes in the biotin biosynthesis pathway (COG0132, COG0156, COG0161 and COG0502) are overrepresented in enterotype 1. **c,** Four enzymes in the thiamine biosynthesis pathway (COG0422, COG0351, COG0352 and COG0611) are overrepresented in enterotype 2. **d,** Six enzymes in the haem biosynthesis pathway (COG0007, COG0276, COG407, COG0408, COG0716 and COG1648) are overrepresented in enterotype 3.

(Arumugam et al., *Nature*, 2011)

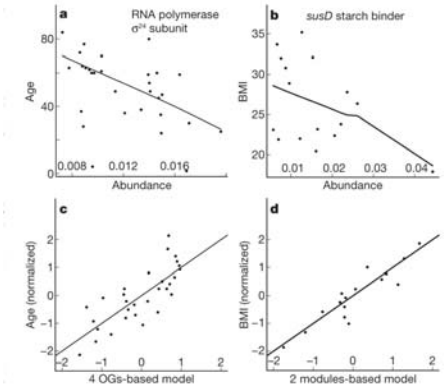


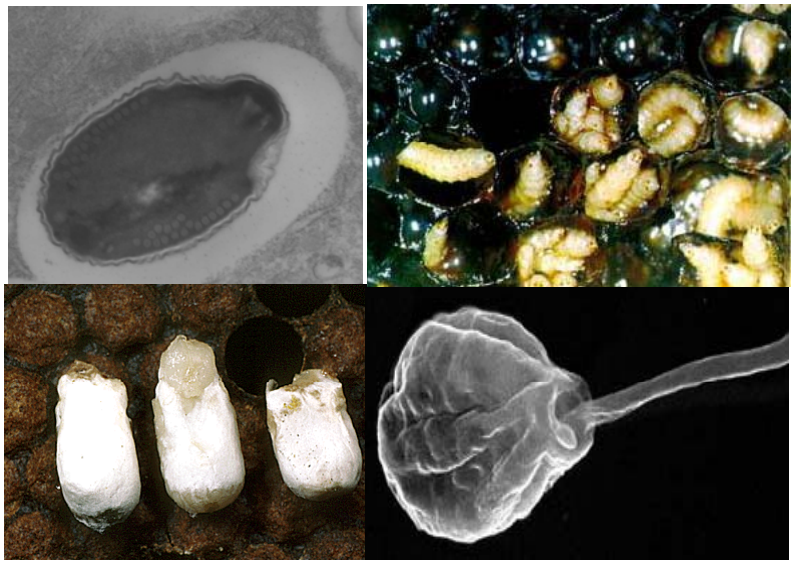
Figure 4 | Correlations with host properties. a, Pairwise correlation of RNA polymerase facultative σ^{24} subunit (COG1595) with age ($P = 0.03$, rho $= -0.59$). **b,** Pairwise correlation of SusD, a family of proteins that bind glycan molecules before they are transported into the cell, and BMI ($P = 0.27$, rho $= -0.29$, weak correlation). **c,** Multiple orthologous groups (OGs) (COG0085, COG0086, COG0438 and COG0739; see Supplementary Table 18) significantly correlating with age when combined into a linear model (see Supplementary Methods section 13 and ref. 40 for details; $P = 2.75 \times 10^{-5}$, adjusted $R^2 = 0.57$). **d,** Two modules, ATPase complex and ectoine biosynthesis (M00051), significantly correlating with BMI when combined into a linear model ($P = 6.786 \times 10^{-6}$, adjusted $R^2 = 0.82$).

(Arumugam et al., *Nature*, 2011)

- ❖ **FIRST PRINCIPLES**
- ❖ **DATA COLLECTION**
- ❖ **OCEANS, TERMITES, MAMMALS**
- ❖ **BEE STORIES**
- ❖ **RETURN TO FUNCTION**
- ❖ **ANALYSIS**

HONEY BEE PARASITES

AND MORE HONEY BEE PARASITES



The New York Times
nytimes.com

PRINTER-FRIENDLY FORMAT SPONSORED BY CHI

February 27, 2007

Honeybees Vanish, Leaving Keepers in Peril

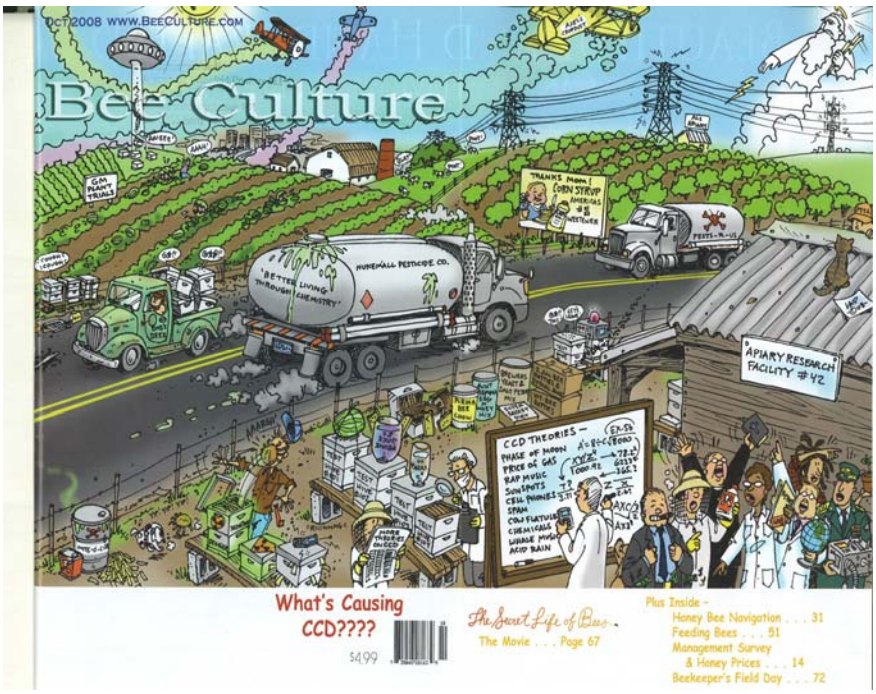
By ALEXEI BARRIONUEVO

VISALIA, Calif., Feb. 23 — David Bradshaw has endured countless stings during his life as a beekeeper, but he got the shock of his career when he opened his boxes last month and found half of his 100 million bees missing.

In 24 states throughout the country, beekeepers have gone through similar shocks as their bees have been disappearing inexplicably at an alarming rate, threatening not only their livelihoods but also the production of numerous crops, including California almonds, one of the nation's

CCD TRAITS

'RAPID' WORKER LOSS
NO DEAD BODIES
EARLY SPRING
PATCHY IN SPACE/TIME




HORIZONTAL TRANSMISSION + STRESS?

A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder

Diana L. Cox-Foster,[1] Sean Conlan,[2] Edward C. Holmes,[3,4] Gustavo Palacios,[2] Jay D. Evans,[5] Nancy A. Moran,[6] Phenix-Lan Quan,[2] Thomas Briese,[2] Mady Hornig,[2] David M. Geiser,[7] Vince Martinson,[8] Dennis vanEngelsdorp,[1,9] Abby L. Kalkstein,[1] Andrew Drysdale,[2] Jeffrey Hui,[2] Junhui Zhai,[2] Liwang Cui,[1] Stephen K. Hutchison,[10] Jan Fredrik Simons,[10] Michael Egholm,[10] Jeffery S. Pettis,[5] W. Ian Lipkin[2]*

In colony collapse disorder (CCD), honey bee colonies inexplicably lose their workers. CCD has resulted in a loss of 50 to 90% of colonies in beekeeping operations across the United States. The observation that irradiated combs from affected colonies can be repopulated with naive bees suggests that infection may contribute to CCD. We used an unbiased metagenomic approach to survey microflora in CCD hives, normal hives, and imported royal jelly. Candidate pathogens were screened for significance of association with CCD by the examination of samples collected from several sites over a period of 3 years. One organism, Israeli acute paralysis virus of bees, was strongly correlated with CCD.

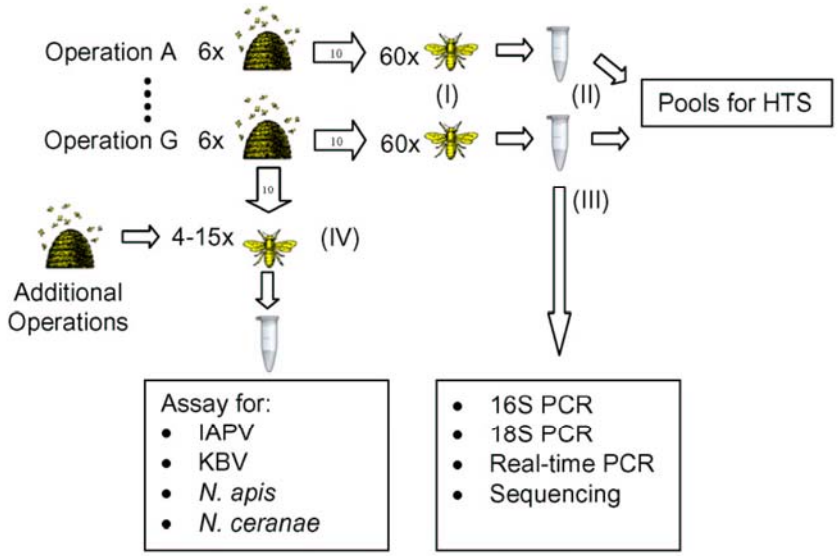


Table 1. Closest sequenced relatives identified through BLAST analysis of the high-throughput sequence data.

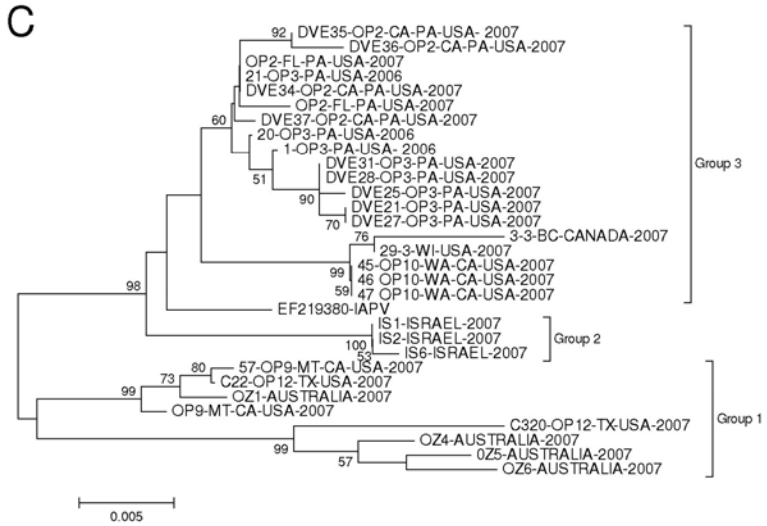
Kingdom	Taxon (rank)	Organism
Bacteria	Firmicutes (phylum)	*Lactobacillus* sp.*†
		Uncultured Firmicutes†
Bacteria	Actinobacteria (class)	*Bifidobacterium* sp.*
Bacteria	Alphaproteobacteria (class)	*Bartonella* sp.*†
		Gluconacetobacter sp.*†
Bacteria	Betaproteobacteria (class)	*Simonsiella* sp.*†
Bacteria	Gammaproteobacteria (class)	Two uncultured species*†
Fungus	Entomophthorales (order)	*Pandora delphacis*
Fungus	Mucorales (order)	*Mucor* spp.
Fungus/microsporidian	Nosematidae (family)	*Nosema ceranae*
Fungus/microsporidian	Nosematidae (family)	*Nosema apis*
Eukaryota	Trypanosomatidae (family)	*Leishmania/Leptomonas* sp.
Metazoan	Varroidae (family)	*Varroa destructor*
Virus	(Unclassified)	CBPV‡
Virus	*Iflavirus* (genus)	SBV
Virus	*Iflavirus* (genus)	DWV‡
Virus	Dicistroviridae (family)	BQCV
Virus	Dicistroviridae (family)	KBV‡
Virus	Dicistroviridae (family)	ABPV
Virus	Dicistroviridae (family)	IAPV of bees‡

*Found by Jeyaprakash *et al.* (*10*). †Found by Babendreier *et al.* (*9*). ‡Indicates viruses not yet classified by the International Committee on the Taxonomy of Viruses but that exhibit the key features of the indicated taxon.

IDENTIFICATION OF CANDIDATES TO PURSUE

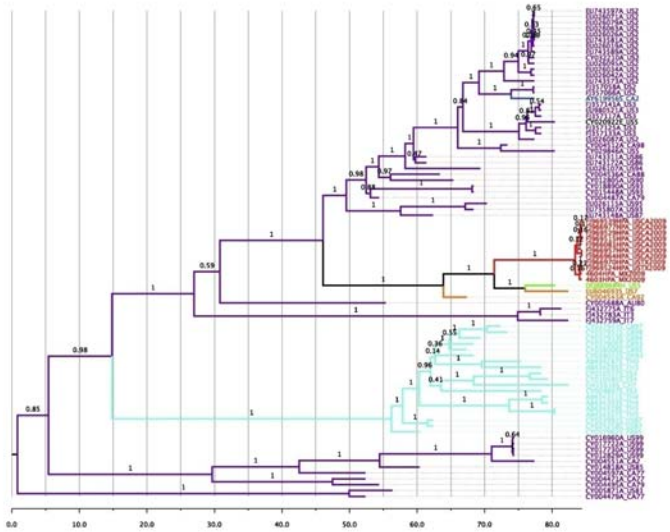
Agent	Number of positive samples *n* (% positive of samples tested)			Positive Predictive Value (%)	Sensitivity (%)	Specificity (%)
	CCD ($n = 30$)	non-CCD ($n = 21$)	Total ($n = 51$)			
IAPV	25 (83.3%)	1 (4.8%)	26 (51.0%)	96.1	83.3	95.2
KBV	30 (100%)	16 (76.2%)	46 (90.2%)	65.2	100	23.8
N. apis	27 (90%)	10 (47.6%)	37 (72.5%)	73.0	90.0	52.4
N. ceranae	30 (100%)	17 (80.9%)	47 (92.1%)	63.8	100	19.0
All 4 agents	23 (76.7%)	0 (0%)	23 (45.0%)	100	76.7	100

ISRAELI ACUTE PARALYSIS VIRUS, IGR PHYLOGENY



Palacios et al., 2008

H1N1 INFLUENZA PHYLOGEOGRAPHY, 2009



1) Identification of Candidates to Pursue

Agent	Number of positive samples *n* (% positive of samples tested)			Positive Predictive Value (%)	Sensitivity (%)	Specificity (%)
	CCD ($n = 30$)	non-CCD ($n = 21$)	Total ($n = 51$)			
IAPV	25 (83.3%)	1 (4.8%)	26 (51.0%)	96.1	83.3	95.2
KBV	30 (100%)	16 (76.2%)	46 (90.2%)	65.2	100	23.8
N. apis	27 (90%)	10 (47.6%)	37 (72.5%)	73.0	90.0	52.4
N. ceranae	30 (100%)	17 (80.9%)	47 (92.1%)	63.8	100	19.0
All 4 agents	23 (76.7%)	0 (0%)	23 (45.0%)	100	76.7	100

(Cox-Foster et al., *Science, 2007*)

REPORTS

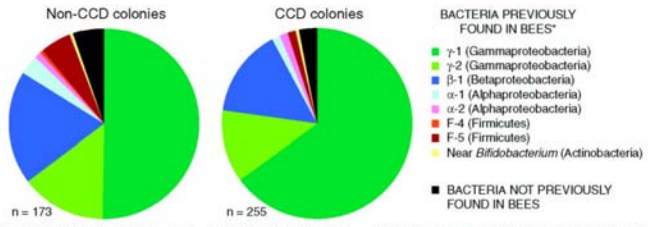
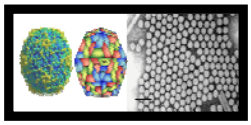


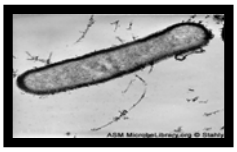
Fig. 1. Summary of bacterial groups from *A. mellifera* derived from colonies categorized as non-CCD and CCD. For both categories, the top BLAST hit for over 96% of sequences from 16S rRNA clones was a sequence obtained in previous studies on bacterial associates of *A. mellifera*. Asterisk indicates that the bacteria were categorized according to the cluster designations of Babendreier *et al.* (*9*) [except for the Bifidobacterium-like sequence of Jeyaprakash *et al.* (*10*)]. *n*, number of sequences. GenBank accession numbers corresponding to the categories are: γ-1 (AY370191-AY370192 and DQ837602-DQ837609), γ-2 (DQ837610-DQ837611), β-1 (AY370189-AY370190 and DQ837616-DQ837621), α-1 (AY370185-AY370187 and DQ837622-DQ837624), α-2 (AY370188 and DQ837625-DQ837626), F-4 (DQ837632-DQ837633), F-5 (AY370183 and DQ837634-DQ837637), near Bifidobacterium (AY370184).

(Cox-Foster et al., *Science, 2007*)

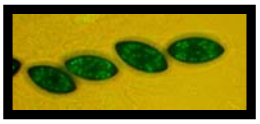


Viruses
ABPV
BQCV
CBPV
DWV
IAPV
KBV
SBV
VDV

Beenomics 2011



American foulbrood bacterium



Nosema ceranae
Nosema apis

Chalkbrood fungus

***Varroa* mites**

9/9/11

624 V. G. MARTINSON *ET AL.*

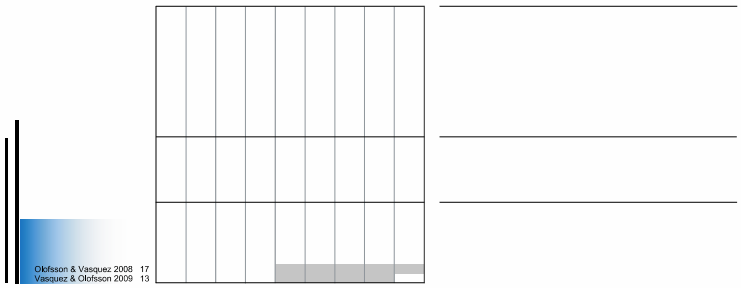
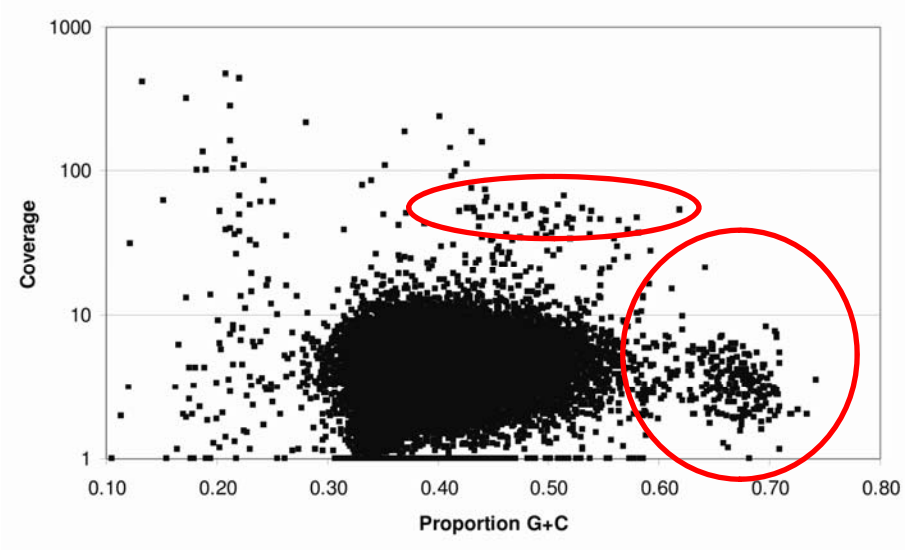


Fig. 2 Frequency of bacterial types in the microbiotae of 21 specimens (three wasps and 16 bees) represented as a heatmap. (a) OTUs correspond to one of the distinctive *A. mellifera* bacterial phylotypes. Column '*n*' denotes number of bacterial 16S rRNA sequences for each specimen. Lower section displays the presence or absence of each bacterial type in previously published studies. (b) OTUs recovered that did not have a top BLASTn hit to the *A. mellifera* bacterial types.

(Martinson et al., *Molecular Ecology,* 2011)

Varroa draft assembly 1.0

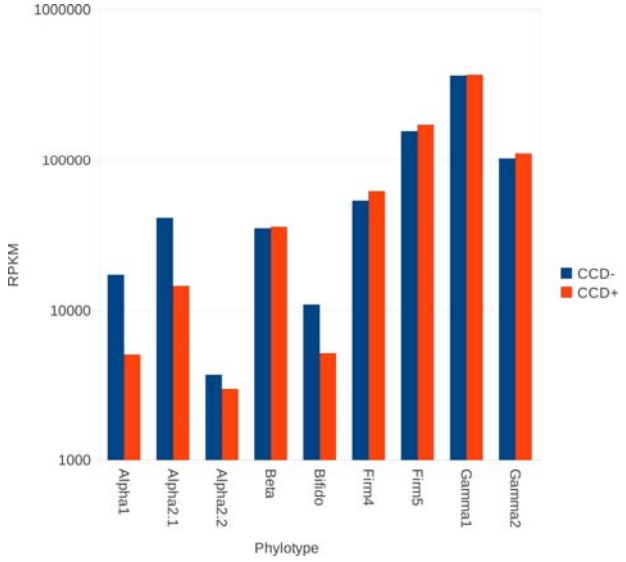


ILLUMINA CCD +/- survey

Healthy (n=63) and collapsed (n=61) colonies sampled in 2007 from eastern and western U.S.

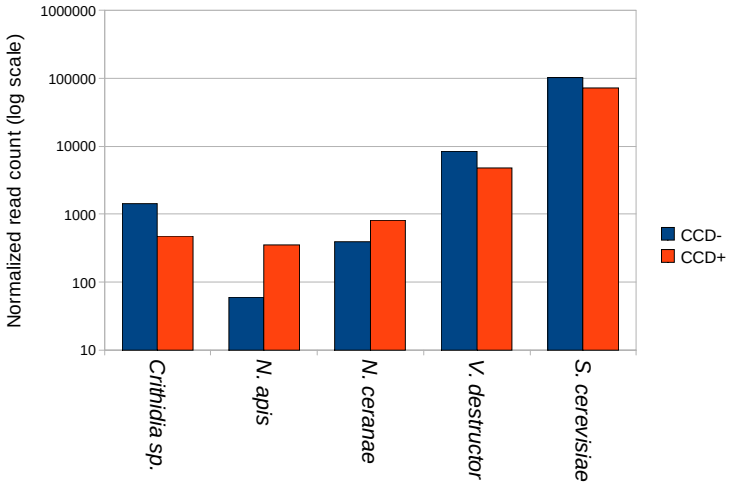
- 8 workers each, random-primed cDNA libraries
- Illumina RNA single-end (CCD-) and paired-end sequencing (CCD+), one lane each
- >19 million reads for CCD-, >20 million paired reads for CCD+
- Assembly with Velvet, mapping with Bowtie

Major honey bee gut bacteria

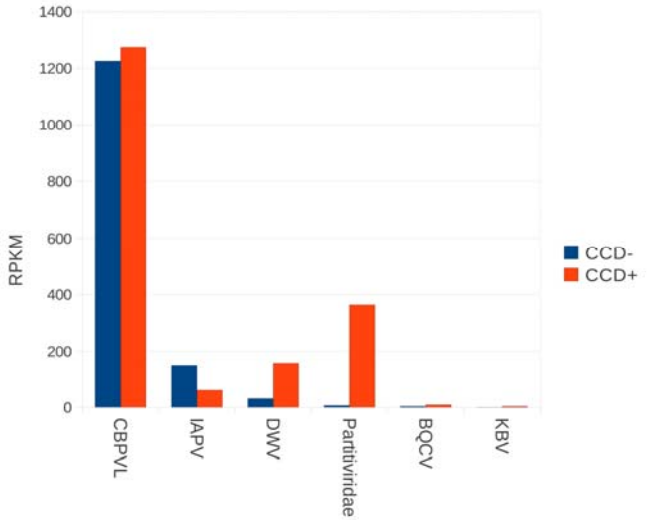


OTUs from Martinson et al. 2011 Molecular Ecology 20:619

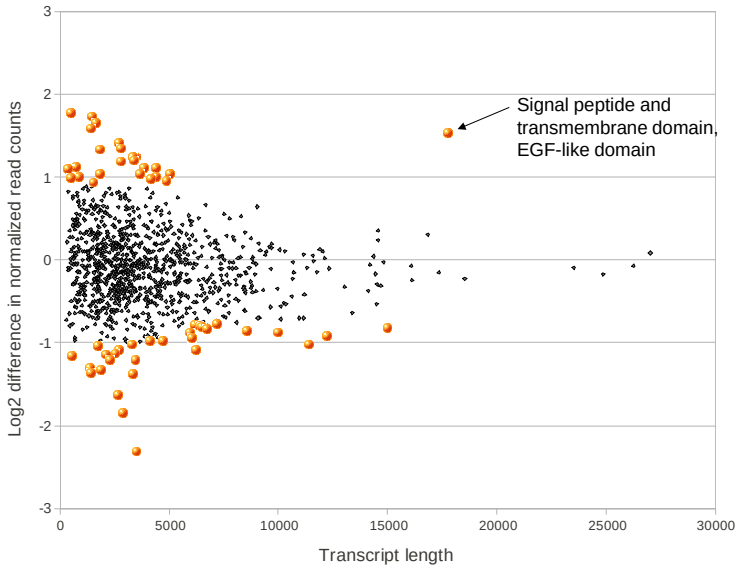
Eukaryotic associates



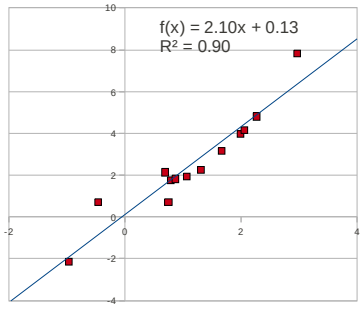
RNA viruses



Gene expression difference in CCD+



Gene	Log2Diff	Homology	Class
GB11923	1.8	HSP20-like	Chaperone
GB10737	1.9	Apidermin2	Cuticle
GB10674	2.1	Apidermin3	Cuticle
GB30202	2.3	Apidermin1	Cuticle
GB19612	1.5	Paxillin	Focal adhesion-associated adaptor
GB17331	1.8	Peritrophic matrix	Peritrophic matrix
GB16446	-1.5	RBM5	RNA binding protein
GB13519	-2.1	RAP55	mRNA silencing/degradation



Gene	Log2Diff	Homology	Class
GB11923	1.8	HSP20-like	Chaperone
GB10737	1.9	Apidermin2	Cuticle
GB10674	2.1	Apidermin3	Cuticle
GB30202	2.3	Apidermin1	Cuticle
GB19612	1.5	Paxillin	Focal adhesion-associated adaptor
GB17331	1.8	Peritrophic matrix	Peritrophic matrix
GB16446	-1.5	RBM5	RNA binding protein
GB13519	-2.1	RAP55	mRNA silencing/degradation

Evidence of age structure (older bees missing in CCD), consistent with CCD phenotype?

Gene	Log2Diff	Homology	Class
GB11923	1.8	HSP20-like	Chaperone
GB10737	1.9	Apidermin2	Cuticle
GB10674	2.1	Apidermin3	Cuticle
GB30202	2.3	Apidermin1	Cuticle
GB19612	1.5	Paxillin	Focal adhesion-associated adaptor
GB17331	1.8	Peritrophic matrix	Peritrophic matrix
GB16446	-1.5	RBM5	RNA binding protein
GB13519	-2.1	RAP55	mRNA silencing/degradation

Related to higher viral loads in CCD+? RBM5 promotes apoptosis, RAP55 silences mRNA within cytoplasmic P bodies. Both functions are potential responses to viral stress.

Conclusions

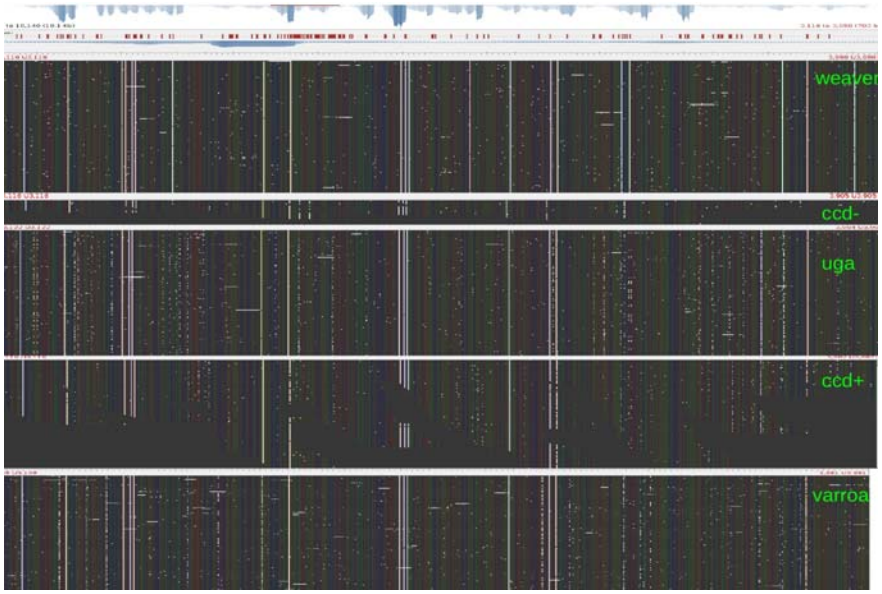
CCD survey:

- Bacterial signature
- Known viruses increased, novel viruses found
- *Nosema* increased
- No strong immune/detox signal

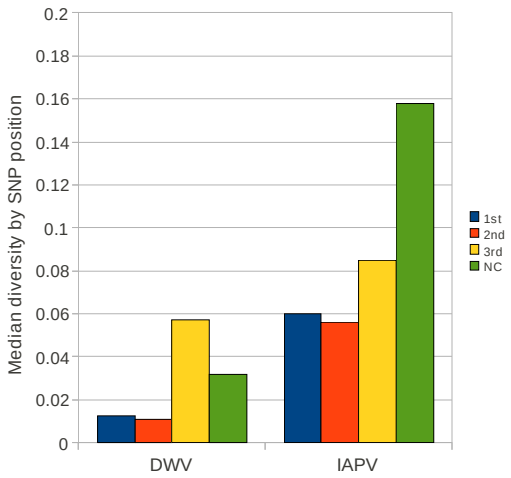
Viral polymorphism

Quantify genetic variation among viral strains of the same species

- Virulence genotypes
- Recombination, chimeras
- Population structure
- Host-specific genotypes



Lower noncoding polymorphism in DWV



Deformed wing virus

	Median per-site Fst				
bp					
hfb	0.002				
varroa	0.001	0.002			
ugasick	0.001	0.001	0.002		
weaver-	0.001	0.001	0.001	0.001	
ccd+	0.012	0.008	0.013	0.010	0.011

Deformed wing virus

	Median per-site Fst				
bp					
hfb	0.002				
varroa	0.001	0.002			
ugasick	0.001	0.001	0.002		
weaver-	0.001	0.001	0.001	0.001	
ccd+	0.012	0.008	0.013	0.010	0.011

Israel acute paralysis virus

	Median per-site Fst	
bp		
ccd-	0.83542	
ccd+	0.17019	0.3333

- FIRST PRINCIPLES
- DATA COLLECTION
- OCEANS, TERMITES, MAMMALS
- BEE STORIES
- RETURN TO FUNCTION
- ANALYSIS

Social evolution in multispecies biofilms

Sara Mitri[a,b,1], João B. Xavier[c], and Kevin R. Foster[a,b,1]

[a]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom; [b]Oxford Centre for Integrative Systems Biology, Oxford University, Oxford OX1 3QU, United Kingdom; and [c]Program in Computational Biology, Memorial Sloan-Kettering Cancer Center, New York, NY 10065

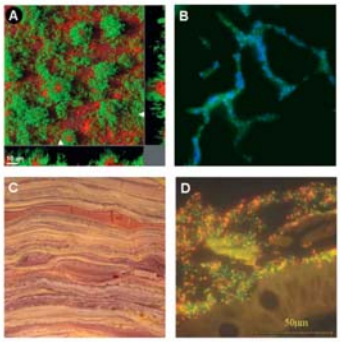


Fig. 1. Microbial diversity: examples of natural microbial communities. (*A*) A two-species bacterial biofilm cultivated in the laboratory in which one strain evolves to increase its exploitation of the other. Adapted by permission from Macmillian Publishers Ltd: *Nature* (78), copyright 2007. (*B*) A two-strain bacterial aggregate detected on a bean leaf surface (magnification 500×) [*Appl Environ Microbiol* (2005) 71(9):5484–5493, 10.1128/AEM.71.9.5484–5493.2005. Reproduced with permission from the American Society for Microbiology] (79). (*C*) Stromatolite fossil that is ~2 billion y old. Modern stromatolites consist of multilayered sheets of microorganisms, and are a good example of very diverse, yet spatially structured microbial communities (copyright Merv Feick, http://www.Indiana9Fossils.com). (*D*) The detection of two of the species present in a bacterial biofilm covering the intestinal mucosae of a self-limiting colitis patient, imaged using triple-color fluorescence in situ hybridization [*J Clin Microbiol* (2005) 43(7):3380–3389, 10.1128/JCM.43.7.3380–3389.2005. Reproduced with permission from the American Society for Microbiology] (80).

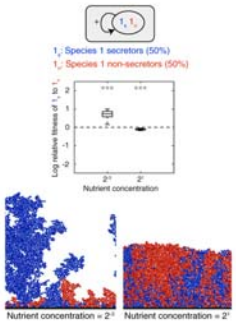


Fig. 2. Secretors and nonsecretors of a single species. Equal proportions of two strains of the same species 1 are inoculated and left to grow to a fixed total biomass. Strain 1_s secretes a product that benefits both strains. 1_n does not secrete the product. Product secretion incurs a cost of 30% of the cells' growth rate. Boxplots show log relative fitness (*Materials and Methods*) of secreting to nonsecreting cells [$log(w(1_s):w(1_n))$] in 40 replicates with high and low nutrient concentrations. The dashed line shows the level at which the two phenotypes are equally fit. Asterisks indicate the significance of the difference between secretor and nonsecretor fitness, ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$; ns, not significant. Below each boxplot is an image generated using the simulation from one of the 40 simulations that was closest to the median in the boxplot. It is shown that secretors can outcompete nonsecretors when the two phenotypes are well segregated, whereas they are at a disadvantage under conditions leading to high mixing.

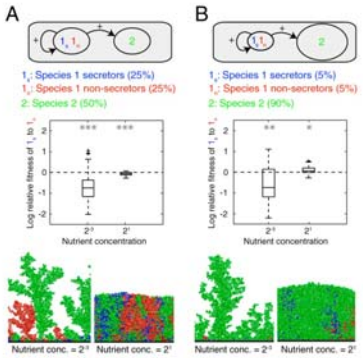


Fig. 3. Ecological competition with a second species. A second species is added to the competition between secretors and nonsecretors (Fig. 2). This second species is intended to also approximate the effects of a mixture of many species (Box 1). Species 1 is equally divided into secretor and nonsecretor strains, whereas species 2 represents either 50% (*A*) or 90% of the cells inoculated (*B*). All cells are then left to grow to a fixed total biomass. Strain 1_s secretes a product that benefits both strains of its own species, as well as species 2. 1_n and species 2 do not secrete any products. Product secretion incurs a cost of 30% of the cells' growth rate. See Fig. 2 legend for explanations on data representation. It is shown that when cells are highly segregated, secretor cells lose their advantage (compared with Fig. 2, *Bottom Left*), independently of the two proportions of species 2. At high levels of mixing, however, secretors can outcompete nonsecretors when there is a high proportion of species 2 cells. The image (*B*, *Bottom Right*) shows the social insulation effect discussed in the text.

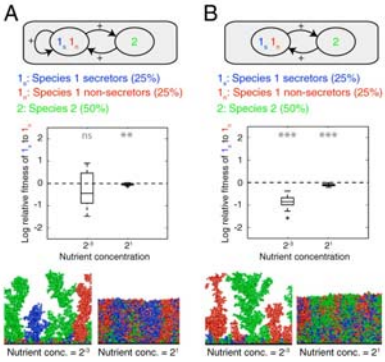
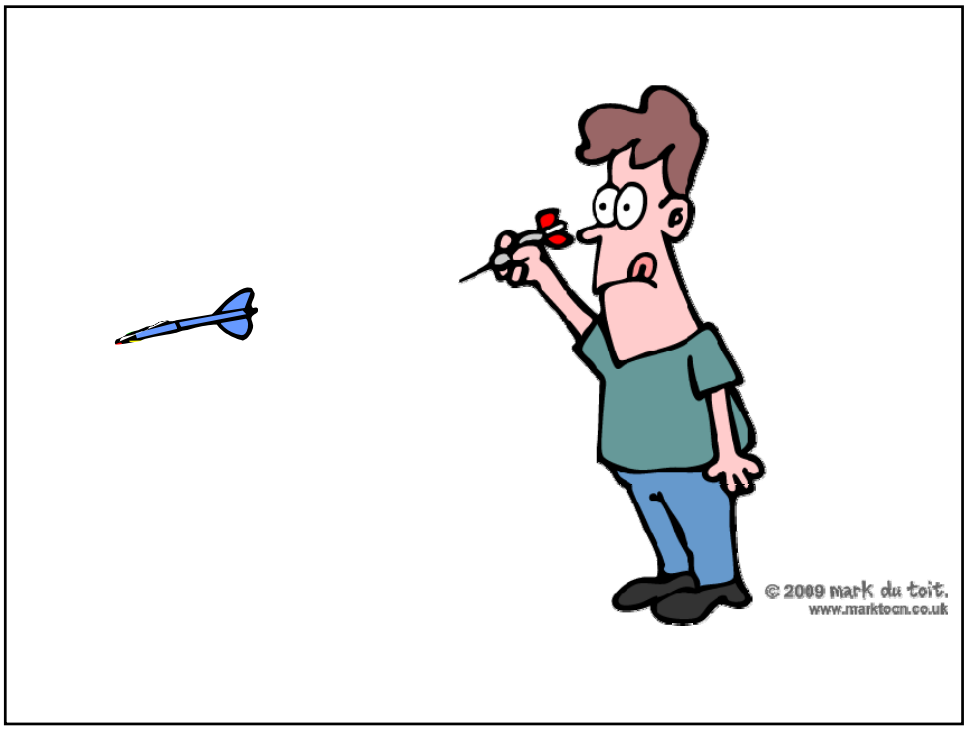
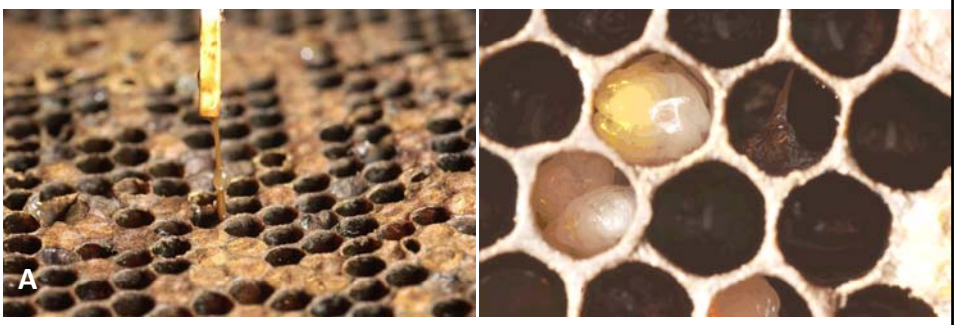


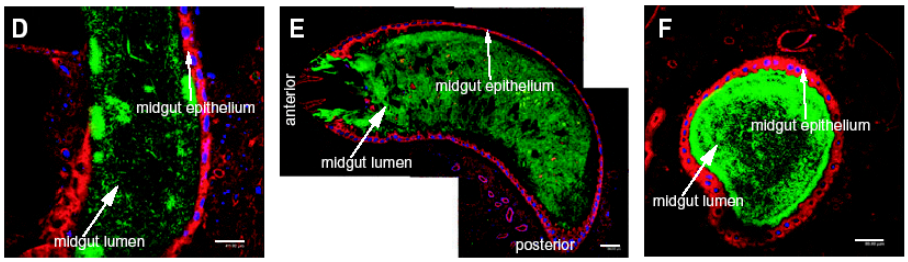
Fig. 4. Multispecies mutualism. Species 2 now secretes a product that is beneficial to species 1, resulting in a mutualism between the two species. Species 1 is equally divided into secretor and nonsecretor strains, whereas species 1 and 2 are inoculated in equal proportions and left to grow to a fixed total biomass. Strain 1_s secretes a product that either benefits both strains of its own species, as well as species 2 (*A*), or species 2 only (*B*). Product secretion by 1_s incurs a cost of 30% of the cells' growth rate. In turn, species 2 secretes a cost-free product that benefits species 1. 1_n does not secrete any products. See Fig. 2 legend for explanations on data representation. It is shown that secretor cells do not have a clear advantage over nonsecretors in any one of the four conditions considered here. This result is because mixing is important for the benefits of the two secreting strains to be shared, but is detrimental because it allows nonsecretors to grow faster than secretors, thereby undermining the mutualistic interaction.


© 2009 mark du toit.
www.marktoon.co.uk

Paenibacillus larvae (American foulbrood disease)



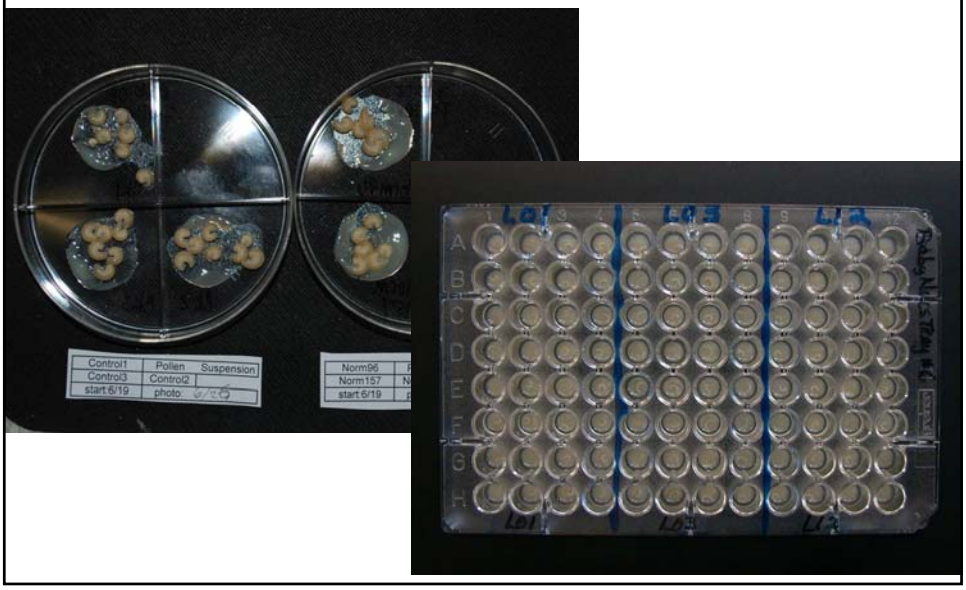
Stephen Pernal, AgCanada



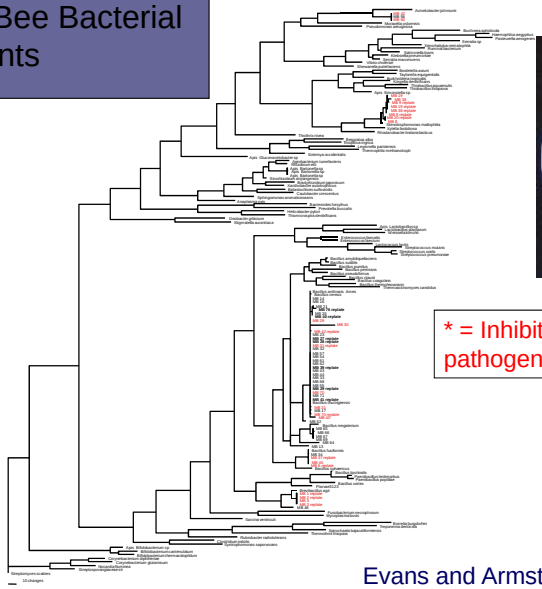
2 days PI 4 days PI 4 days PI

YUE ET AL. ENV MICROB. 2008

Challenge Experiments

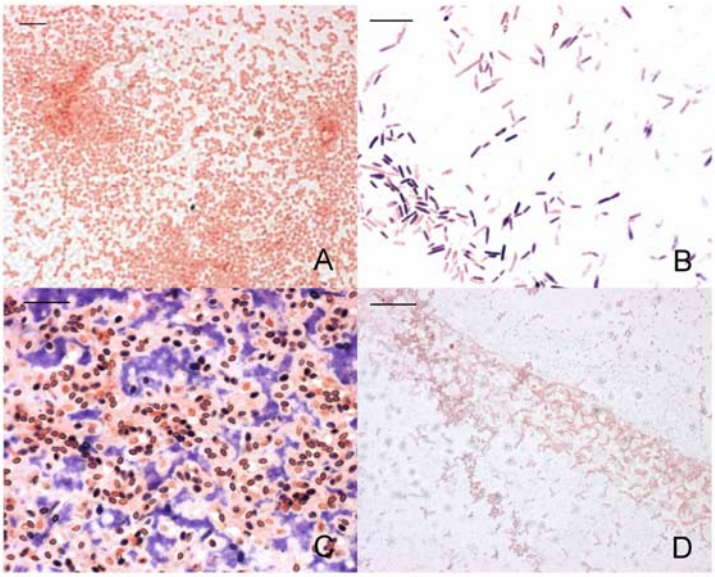


16S Placement of Honey Bee Bacterial Symbionts



* = Inhibited the bee pathogen, *P. larvae*

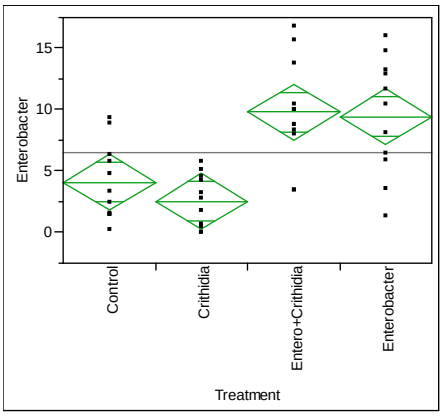
Evans and Armstrong, *JAR* (2005)
BMC Ecology, 6:4 (2006)

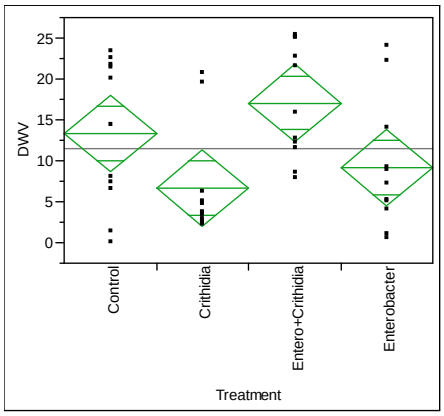


Evans and Armstrong, BMC Ecology, 2006

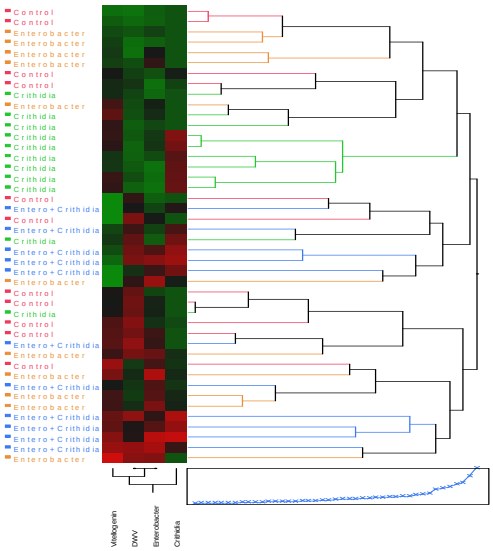
Bacteria x Pathogen Interactions

Controlled Inoculations, Bacteria Plus Trypanosome

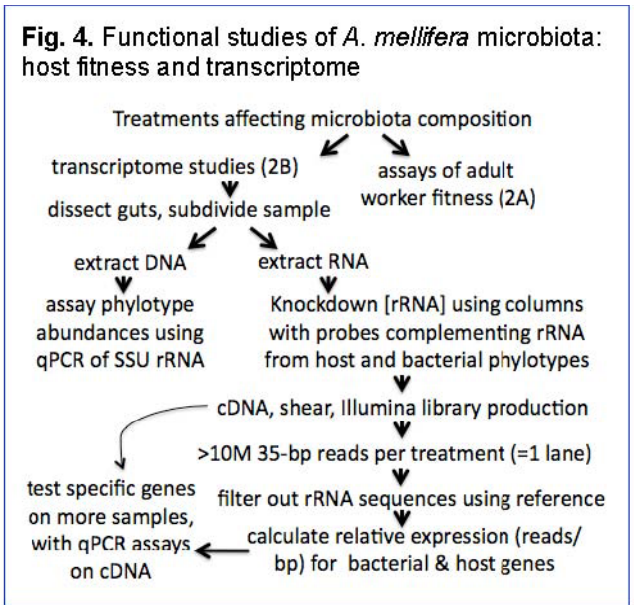




Controlled Inoculations, Bacteria Plus Trypanosome

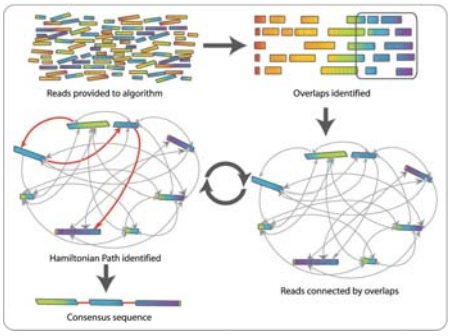


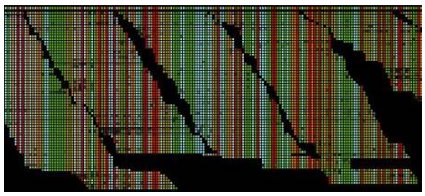
Moran/Evans/Winfree 2011-2015



- FIRST PRINCIPLES
- DATA COLLECTION
- OCEANS, TERMITES, MAMMALS
- BEE STORIES
- RETURN TO FUNCTION
- ANALYSIS

Assembly vs. mapping





Assembly: every read aligned to each other and then resolved to optimal 'path'. Computationally intensive. Outputs contigs.

Mapping: reads aligned one at a time to an existing reference. Computationally easy. Outputs an alignment file.

OTUs

Operational, *a priori* way to describe the number of taxonomic groups

Usually based on rDNA sequence (16S)

Clustering at arbitrary %ID (97-98% typical for bacteria, 95% for virus). Different programs give different clusters.

Comparison of OTUs

Phylogenetic tree

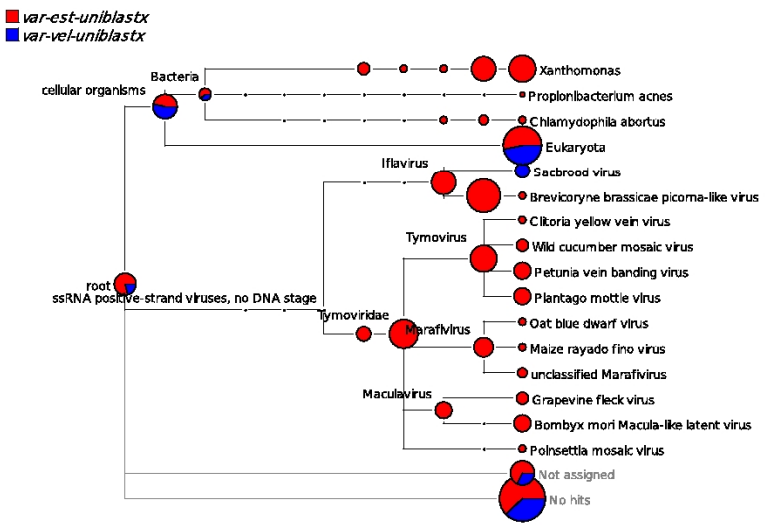
- MEGAN

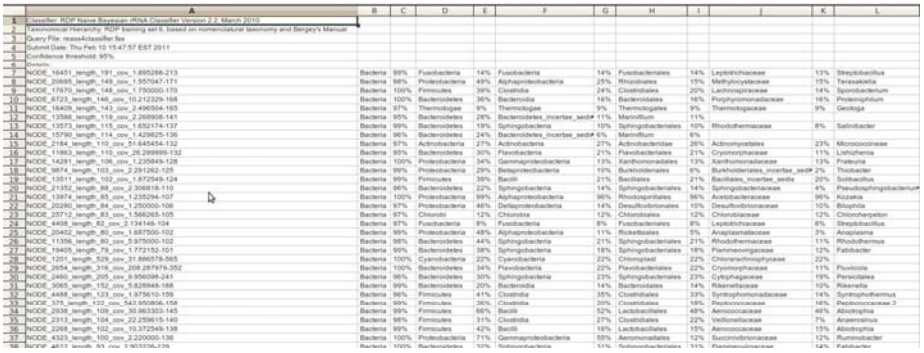
Ribosomal classifier, align to reference database

- RDP Classifier

Nucleotide composition model

- PhyloPythia, PhymmBL





	A	B	C	D	E	F	G	H	I	J	K	L
1	Classifier: RDP Naive Bayesian rRNA Classifier Version 2.2, March 2010											
2	Taxonomical Hierarchy: RDP training set 6, based on nomenclatural taxonomy and Bergey's Manual											
3	Query File: mass4classifier.fas											
4	Submit Date: Thu Feb 10 15:47:57 EST 2011											
5	Confidence threshold: 95%											
6	Domain											
7	NODE_16451_length_191_cov_1.895286-213	Bacteria	99%	Fusobacteria	14%	Fusobacteria	14%	Fusobacteriales	14%	Leptotrichiaceae	13%	Streptobacillus
8	NODE_20895_length_149_cov_1.557047-171	Bacteria	98%	Proteobacteria	49%	Alphaproteobacteria	25%	Rhizobiales	15%	Methylocystaceae	15%	Terasakiella
9	NODE_17670_length_148_cov_1.750000-170	Bacteria	100%	Firmicutes	39%	Clostridia	24%	Clostridiales	20%	Lachnospiraceae	14%	Sporobacterium
10	NODE_6723_length_146_cov_10.212329-168	Bacteria	100%	Bacteroidetes	36%	Bacteroidia	16%	Bacteroidales	16%	Porphyromonadaceae	16%	Proteiniphilum
11	NODE_16409_length_143_cov_2.496504-165	Bacteria	97%	Thermologae	9%	Thermologae	9%	Thermologales	9%	Thermologaceae	9%	Geologa
12	NODE_13586_length_119_cov_2.268908-141	Bacteria	95%	Bacteroidetes	28%	Bacteroidetes_incertae_sedis	11%	Marinifilum	11%			
13	NODE_13573_length_115_cov_1.852174-137	Bacteria	99%	Bacteroidetes	19%	Sphingobacteria	10%	Sphingobacteriales	10%	Rhodothermaceae	8%	Salinibacter
14	NODE_15790_length_114_cov_1.429825-136	Bacteria	96%	Bacteroidetes	24%	Bacteroidetes_incertae_sedis	6%	Marinifilum	6%			
15	NODE_2184_length_110_cov_51.645454-132	Bacteria	97%	Actinobacteria	27%	Actinobacteria	27%	Actinobacteridae	26%	Actinomycetales	23%	Micrococcineae
16	NODE_11863_length_110_cov_26.299999-132	Bacteria	95%	Bacteroidetes	30%	Flavobacteria	21%	Flavobacteriales	21%	Cryomorphaceae	11%	Lishizhenia
17	NODE_14281_length_106_cov_1.235849-128	Bacteria	100%	Proteobacteria	34%	Gammaproteobacteria	13%	Xanthomonadales	13%	Xanthomonadaceae	13%	Frateuria
18	NODE_9874_length_103_cov_2.291262-125	Bacteria	99%	Proteobacteria	29%	Betaproteobacteria	10%	Burkholderiales	6%	Burkholderiales_incertae_sedis	2%	Thiobacter
19	NODE_13511_length_102_cov_1.872549-124	Bacteria	99%	Firmicutes	39%	Bacilli	21%	Bacillales	21%	Bacillales_incertae_sedis	20%	Solibacillus
20	NODE_21352_length_88_cov_2.308618-110	Bacteria	96%	Bacteroidetes	22%	Sphingobacteria	14%	Sphingobacteriales	14%	Sphingobacteriaceae	4%	Pseudosphingobacterium
21	NODE_13974_length_85_cov_1.235294-107	Bacteria	100%	Proteobacteria	99%	Alphaproteobacteria	96%	Rhodospirillales	96%	Acetobacteraceae	96%	Kozakia
22	NODE_20280_length_84_cov_1.250000-106	Bacteria	97%	Proteobacteria	46%	Deltaproteobacteria	14%	Desulfovibrionales	10%	Desulfovibrionaceae	10%	Bilophila
23	NODE_20712_length_83_cov_1.566265-105	Bacteria	97%	Chlorobi	12%	Chlorobia	12%	Chlorobiales	12%	Chlorobiaceae	12%	Chlorochrogatum
24	NODE_4408_length_82_cov_2.134146-104	Bacteria	97%	Fusobacteria	8%	Fusobacteria	8%	Fusobacteriales	8%	Leptotrichiaceae	8%	Streptobacillus
25	NODE_20402_length_80_cov_1.687500-102	Bacteria	99%	Proteobacteria	48%	Alphaproteobacteria	11%	Rickettsiales	5%	Anaplasmataceae	3%	Anaplasma
26	NODE_11356_length_80_cov_5.975000-102	Bacteria	98%	Bacteroidetes	44%	Sphingobacteria	21%	Sphingobacteriales	21%	Rhodothermaceae	11%	Rhodothermus
27	NODE_19405_length_79_cov_1.772152-101	Bacteria	99%	Bacteroidetes	38%	Sphingobacteria	18%	Sphingobacteriales	18%	Flammeovirgaceae	12%	Fabibacter
28	NODE_1201_length_529_cov_31.886578-565	Bacteria	100%	Cyanobacteria	22%	Cyanobacteria	22%	Chloroplast	22%	Chlorarachniophyceae	22%	
29	NODE_2654_length_318_cov_208.287979-352	Bacteria	100%	Bacteroidetes	34%	Flavobacteria	22%	Flavobacteriales	22%	Cryomorphaceae	11%	Fluviicola
30	NODE_2460_length_205_cov_9.956098-241	Bacteria	96%	Bacteroidetes	30%	Sphingobacteria	23%	Sphingobacteriales	23%	Cytophagaceae	19%	Persicitalea
31	NODE_3065_length_152_cov_5.828948-188	Bacteria	99%	Bacteroidetes	20%	Bacteroidia	14%	Bacteroidales	14%	Rikenellaceae	10%	Rikenella
32	NODE_4488_length_123_cov_1.875610-159	Bacteria	96%	Firmicutes	41%	Clostridia	35%	Clostridiales	33%	Syntrophomonadaceae	14%	Syntrophothermus
33	NODE_375_length_122_cov_542.950806-158	Bacteria	99%	Firmicutes	26%	Clostridia	20%	Clostridiales	18%	Peptococcaceae	16%	Peptococcaceae 2
34	NODE_2938_length_109_cov_30.963303-145	Bacteria	99%	Firmicutes	66%	Bacilli	52%	Lactobacillales	48%	Aerococcaceae	46%	Abiotrophia
35	NODE_2313_length_104_cov_22.259615-140	Bacteria	98%	Firmicutes	31%	Clostridia	27%	Clostridiales	22%	Veillonellaceae	7%	Anaerovorax
36	NODE_2268_length_102_cov_10.372549-138	Bacteria	99%	Firmicutes	42%	Bacilli	16%	Lactobacillales	15%	Aerococcaceae	15%	Abiotrophia
37	NODE_4323_length_102_cov_2.220000-138	Bacteria	100%	Proteobacteria	71%	Gammaproteobacteria	55%	Aeromonadales	12%	Succinivibrionaceae	12%	Ruminobacter
38	NODE_4637_length_93_cov_2.903226-126	Bacteria	100%	Bacteroidetes	33%	Sphingobacteria	31%	Sphingobacteriales	31%	Flammeovirgaceae	14%	Fabibacter

Species richness

Alpha: the diversity of species at one site/habitat

Beta: how distinct different sites/habitats are from each other

Typically use some kind of index that considers how much of each species is present, rather than just total number of species

Testing difference between samples

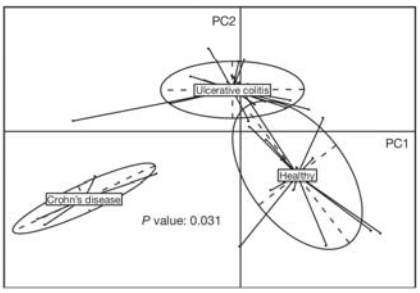
Need a metric/index

- Overlap/relative abundance of OTUs

Clustering/PCA/NMDS

Some programs:

- EstimateS
- Unifrac



Identification of genes

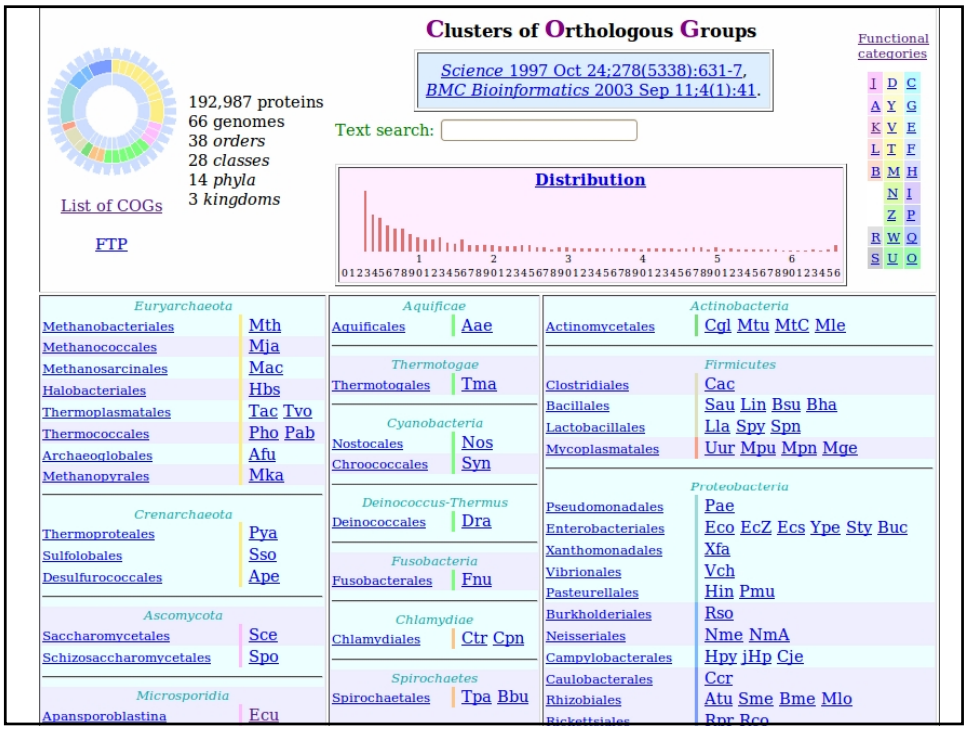
BlastX of contigs or reads

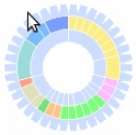
ORF identification

- any general purpose sequence analysis package

Domain identification (Pfam etc.)

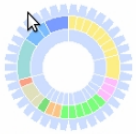
Orthologous groups





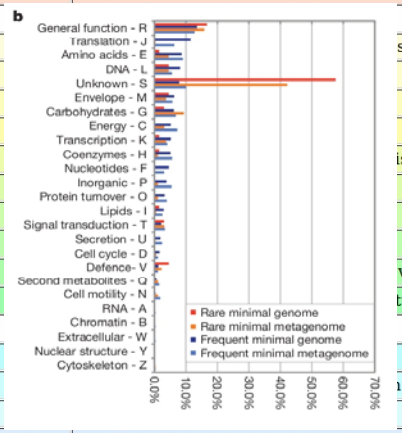
Functional categories

Code	COGs	Domains	Description
			Information storage and processing
J	245	10,572	Translation, ribosomal structure and biogenesis
A	25	137	RNA processing and modification
K	231	11,271	Transcription
L	238	10,338	Replication, recombination and repair
B	19	228	Chromatin structure and dynamics
			Cellular processes and signaling
D	72	1,678	Cell cycle control, cell division, chromosome partitioning
Y	-	-	Nuclear structure
V	46	2,380	Defense mechanisms
T	152	7,683	Signal transduction mechanisms
M	188	7,858	Cell wall/membrane/envelope biogenesis
N	96	2,747	Cell motility
Z	12	128	Cytoskeleton
W	1	25	Extracellular structures
U	159	3,743	Intracellular trafficking, secretion, and vesicular transport
O	203	6,206	Posttranslational modification, protein turnover, chaperones
			Metabolism
C	258	9,830	Energy production and conversion
G	230	10,816	Carbohydrate transport and metabolism
E	270	14,939	Amino acid transport and metabolism
F	95	3,922	Nucleotide transport and metabolism
H	179	6,582	Coenzyme transport and metabolism
I	94	5,201	Lipid transport and metabolism



Functional categories

Code	COGs	Domains	Description
			Information storage and processing
J	245	10,572	Translation, ribosomal structure and biogenesis
A	25	137	RNA processing and modification
K	231	11,271	Transcription
L	238	10,338	Replication, recombination and repair
B	19	228	Chromatin structure and dynamics
D	72		some partitioning
Y	-		
V	46		
T	152		
M	188		is
N	96		
Z	12		
W	1		
U	159		vesicular transport
O	203		turnover, chaperones
C	258		
G	230		
E	270		
F	95	3,922	Nucleotide transport and metabolism
H	179	6,582	Coenzyme transport and metabolism
I	94	5,201	Lipid transport and metabolism



Gene ontology terms

Domain/ontology mapping programs

Pfam/Interpro scan

Blast2GO

KEGG Mapper

eggNOG

Read mapping to a reference

- Bowtie, BWA, Stampy, Novoalign
- Competitive mapping
  - Matches uniquely?
- Parameters for allowing match (% mismatch, indels)
- Measure abudance of taxa, expression of genes independent of assembled contigs

Questions??