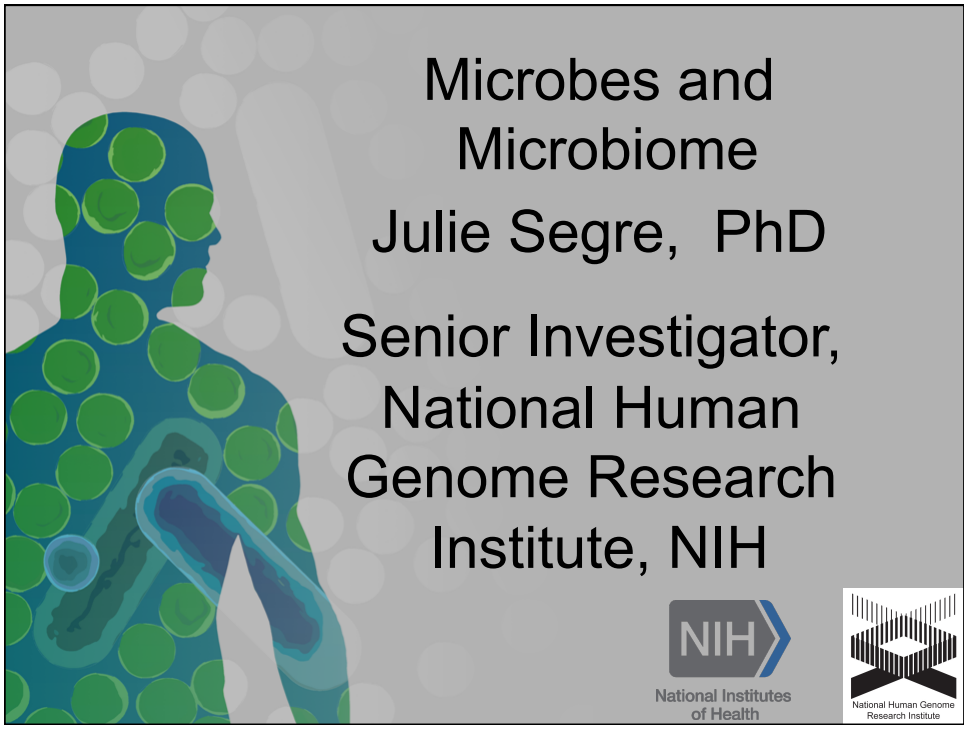


JOHNS HOPKINS MEDICINE
CONTINUING MEDICAL EDUCATION

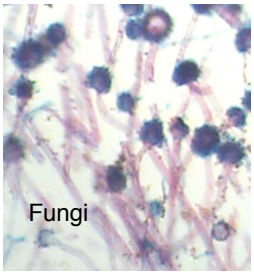
Current Topics in Genome Analysis 2014

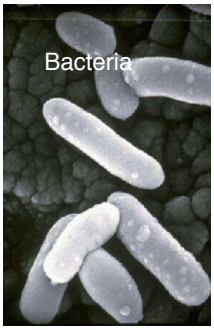
Julia Segre

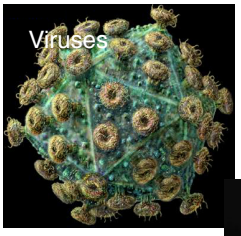
No Relevant Financial Relationships with Commercial Interests

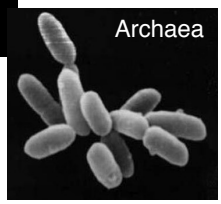
2

Why the Human Microbiome?



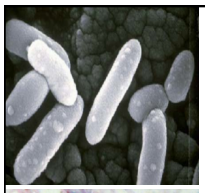


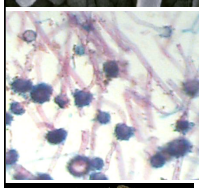


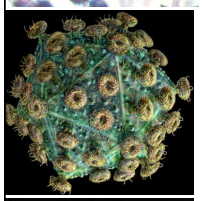


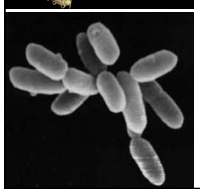
Each human cell has the same protein-encoding potential. Microbes are more diverse and dynamic than human genome.

3



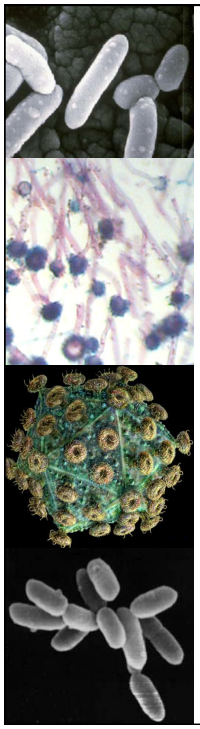






Human Microbiome

- Humans are hosts to many microbes (bacteria, fungi, viruses)
- Microbiome is totality of microbial community DNA
- Microbial cells outnumber human cells
- Many unknown functions of microbes
- Many microbes are often considered pathogenic
  - Mycobacterium tuberculosis
  - *Staphylococcus aureus*



Not all microbes are bad:
Beneficial microbes perform functions essential for human health

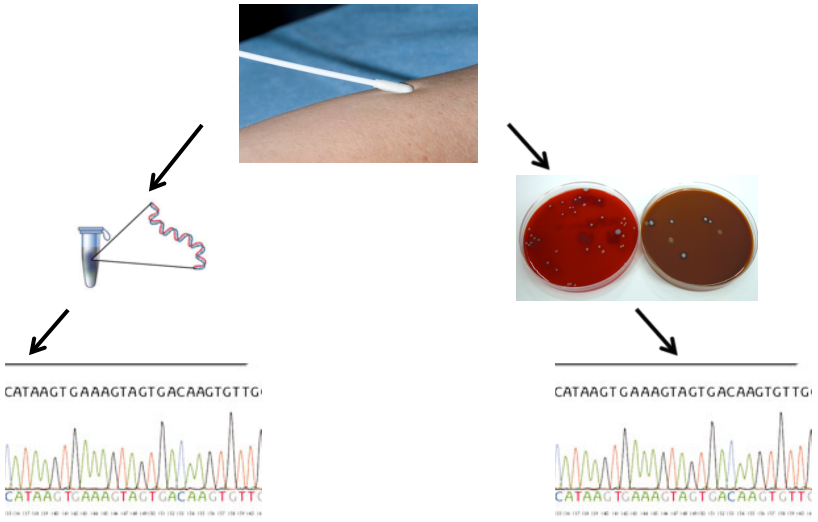
- Vitamin synthesis
- Digestion
- Education and activation of immune system
- Inhibition of skin colonization by pathogens

Many microbial-host and microbial-microbial interactions remain unknown

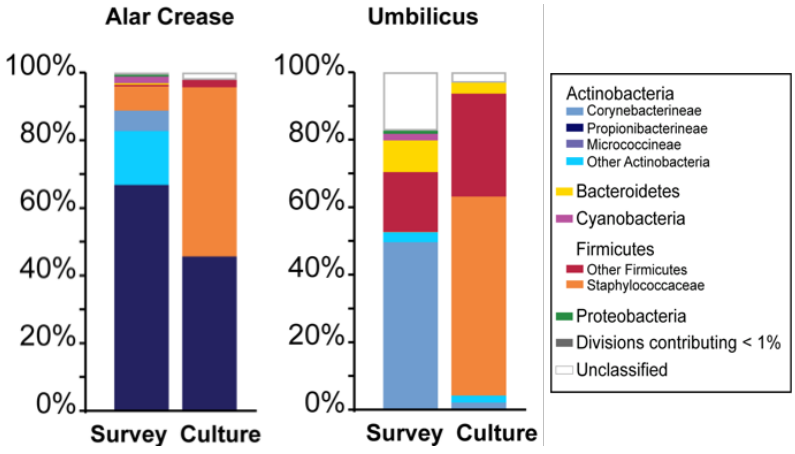
Elucidating the diversity of the human microbiome

- Traditional approaches rely on isolating bacteria in pure culture
- The majority of bacterial species do not grow in culture = “the great plate count anomaly”
- Culturing favors lab weeds--not necessarily the most dominant or influential species
- Excludes microbes that rely on community interactions

Direct sequencing vs. culture-based methods



Direct sequencing vs. culture data

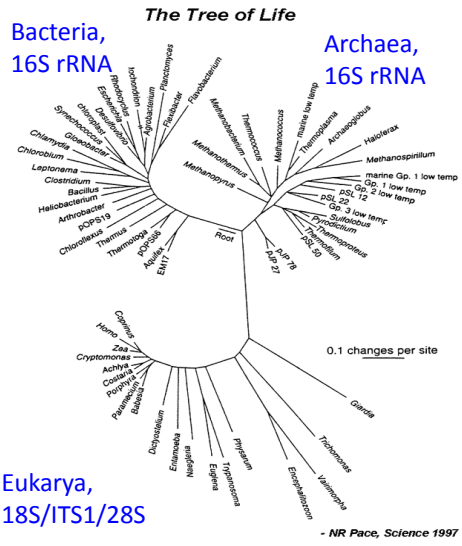


Topics for today's talk

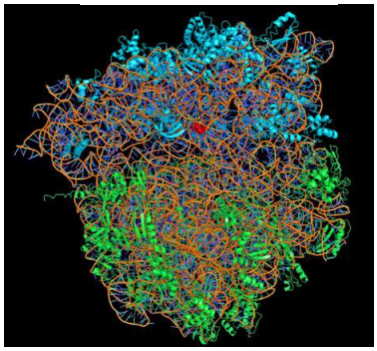
1. Bacterial diversity studies: 16S rRNA
2. Fungal diversity studies: ITS1
3. Bacterial genomes: Shotgun sequencing
4. Metagenomics
5. Where is the technology going?

9

Core marker genes are specific to taxonomic clades.



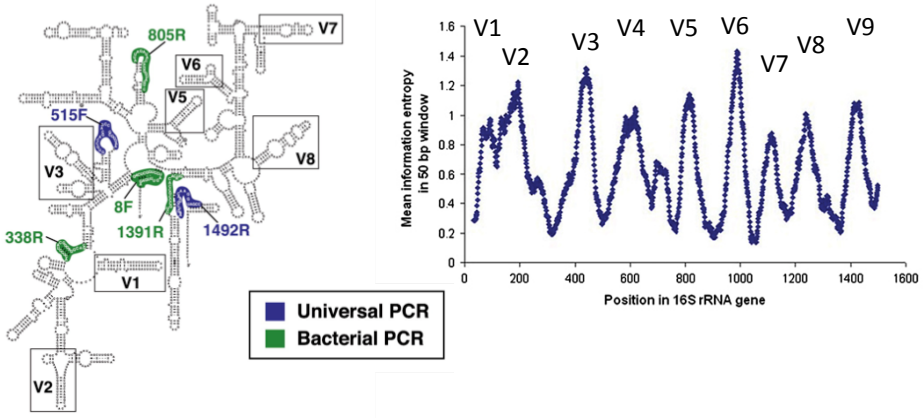
Bacterial Ribosome



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

TOPIC 1.
Bacterial Diversity: 16S rRNA gene

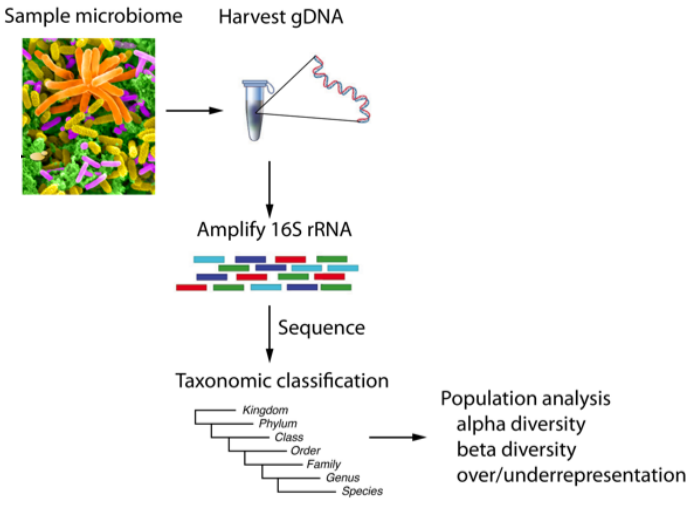
Universal and variable regions of 16S rRNA used for PCR amplification & classification



Peterson et al., Cell Host & Microbe (2008)

Siqueria et al., Journal of Oral Microbiology (2012)

Basic workflow



Kong, Trends in Mol Medicine, 2011

Important Issues to Consider Before Initiating Experiment

1. Study Design. Define the question as precisely as possible; e.g. 'I want to compare wild-type with knock-out mice.' →Are these mice littermates? Because there is a lot of variation between individuals, cages and facilities. What controls do you need?
2. What sequencing platform will you use?
3. What region of the 16S rRNA gene will you amplify?
4. How many reads do you need per sample?
5. What are hidden technical issues? CHIMERAS
6. What analysis tool will you use?
7. How will you display your data?
8. How will you compare your results with other published studies?
9. What information will yield a testable hypothesis?

13

Calculating Bacterial Load: qPCR with primers in conserved region of 16S rRNA gene

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. Must also consider how to normalize sample. /cm^2 or /g stool?

- 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
Castillo M…Gasa J…2006

14

DNA Sequencing to assess bacterial diversity

Illumina Mi-Seq (2 x 300 bp paired-end reads)

- 2 runs/week on one instrument.
- Costs $2K, which is $4/sample if you multiplex 500 samples.
- Scale is the issue. Need to dual-index bar-code primers for multiplexing since platform generates >10 million reads per lane. Assume 10,000 reads is more than enough per sample, you can multiplex 500+ samples together in one lane.
- Short reads, but can link paired reads.

Primer: 8F_____ _____505R primer

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

Fadrosh DW…Ravel J Microbiome 2014;
Kozich JJ….Schloss PD Appl Environ Microbiol 2013;
Caporaso JG…Knight R ISME J 2012

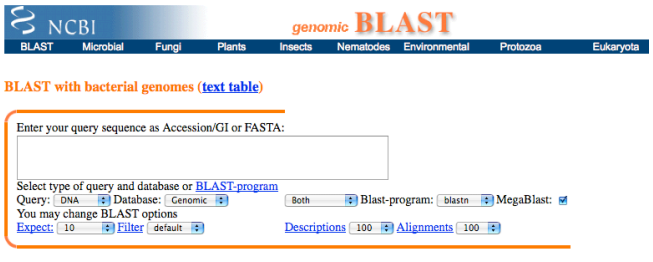
Other means of sequence data acquisition

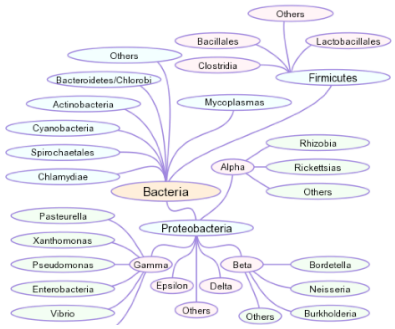
- 454 pyrosequencing (~500bp)
  - Limited to known taxa, but can get species-level designations
  - More expensive than Illumina.
  - Roche is no longer supporting this sequencing platform.

- Phylochip (16S rRNA microarray)
  - Limited to known taxa, but can get species-level designations
  - More expensive.
  - will never find unique or novel species

- Hi-Seq Illumina (2 x 100 bp paired-end reads)
  - Production sequencing. High output mode (TruSeq v3 chemistry) runs for 10 days and produces 4 billion clusters.

How to identify a bacterial sequence and align sequences?





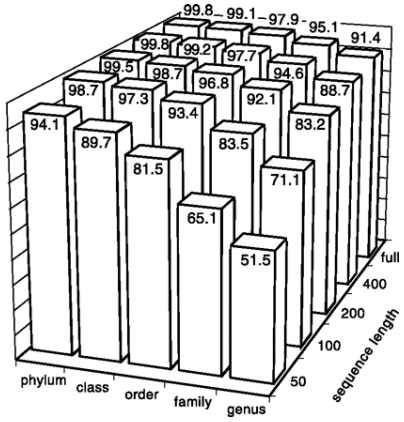
Matches MANY sequences. And many of them are not type strains, but UNCULTURED from 16S rRNA sequencing study.

17

Alignment & Classification

- Reference-dependent
  - Ribosomal Database Project (RDP), SILVA, Greengenes
- But what about species? Amplify the appropriate region of 16S rRNA gene (V1-3 for Staphylococcus[1]; or Lactobacillus[2]) and use custom database.
- Sequences with no reference? Not so many of those, might have to consider other explanations

[1]Conlan, PLoS One 2012; [2]Ravel PNAS 2011



Wang et al., Applied and Environmental Microbiology (2007)

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

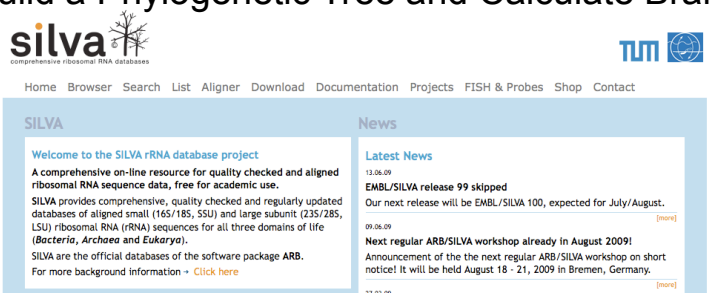
Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy▽†

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

19

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

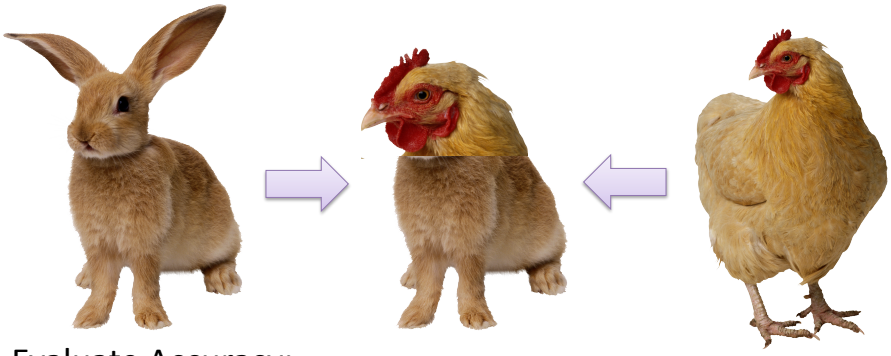
Nucleic Acids Research, 2004, Vol. 32, No. 4 **1363–1371**
DOI: 10.1093/nar/gkh293

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

20

Chimeras: PCR generated (template switching)

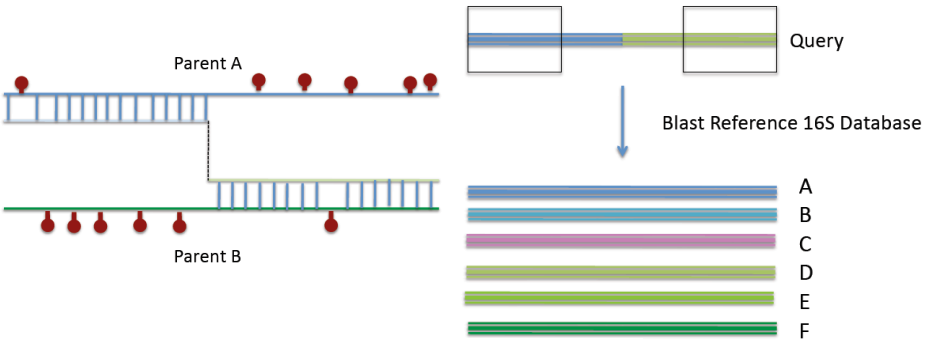


Evaluate Accuracy:
– True Positives (TP): artificial chimeras flagged
– False Positives (FP): reference (non-chimera) flagged

21

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

ChimeraSlayer Detection Program
http://microbiomeutil.sourceforge.net

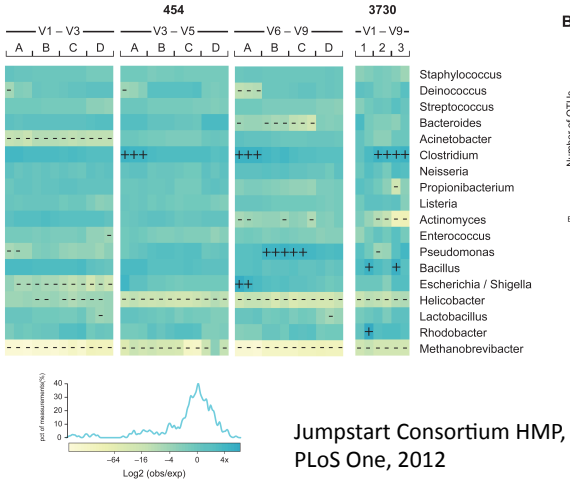


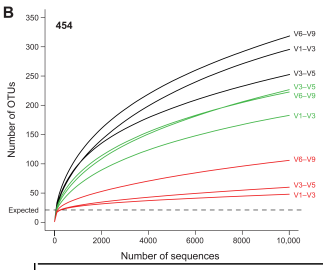
Microbiome Utilities Portal of the Broad Institute

BROAD INSTITUTE

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

Do not underestimate primer bias or chimeras! Mock community contains 20 bacteria. Amplified and sequenced with various methods



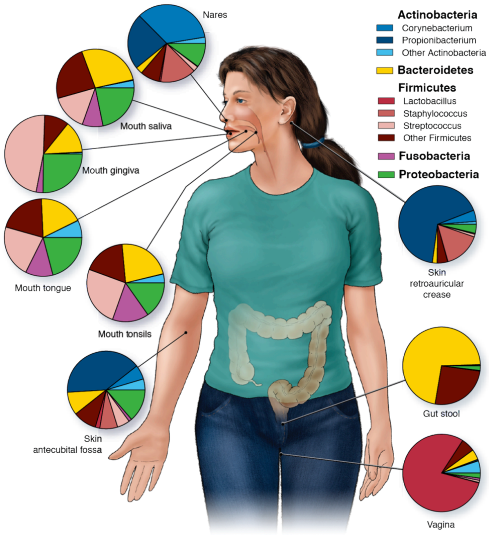


Samples	% Observed Chimera content			
	ABI3730	454 FLX Titanium		
	V1–V9	V1–V3	V3–V5	V6–V9
MC	5.99±3.07	14.26±10.34	14.75±9.45	13.49±8.52
gut	7.71±6.46	22.90±8.56	16.03±2.86	17.76±3.76
oral	7.22±6.35	20.55±11.73	10.98±4.01	9.10±5.02
skin	3.49±5.77	11.15±1.36	7.51±2.49	5.73±1.69
vaginal	6.31±6.64	12.60±6.70	6.62±3.51	3.00±1.65

*Values are averages ± STDEV calculated from multiple replicates of MC, and from replicates of multiple clinical samples originating from different body sites.
doi:10.1371/journal.pone.0039315.t001

Jumpstart Consortium HMP, PLoS One, 2012

NIH Common Fund: Human Microbiome Project

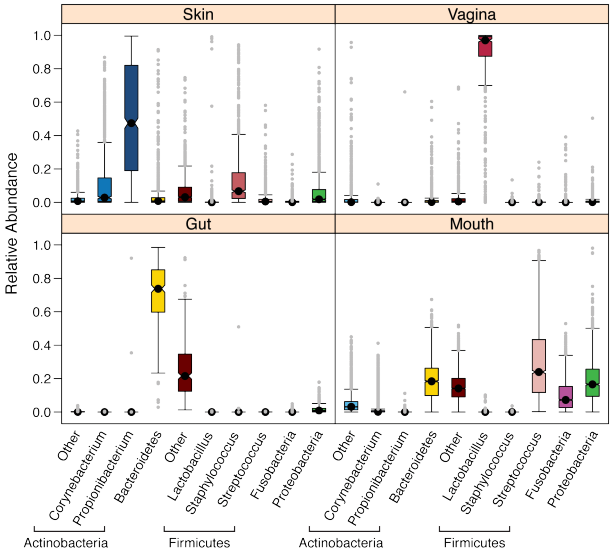


Longitudinally assess microbial diversity of 250 healthy subjects at 5 major body sites

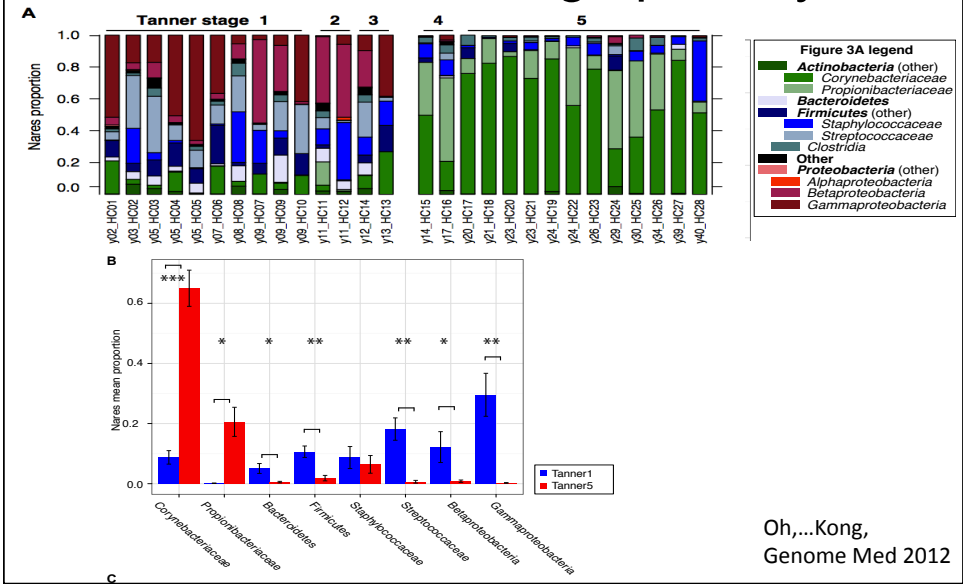
http://commonfund.nih.gov/hmp/

Grice and Segre, Annual Review of Human Genetics and Genomics, 2012

Signature Taxa for each major body site, large variation between subjects



Skin microbiome changes with the transition through puberty

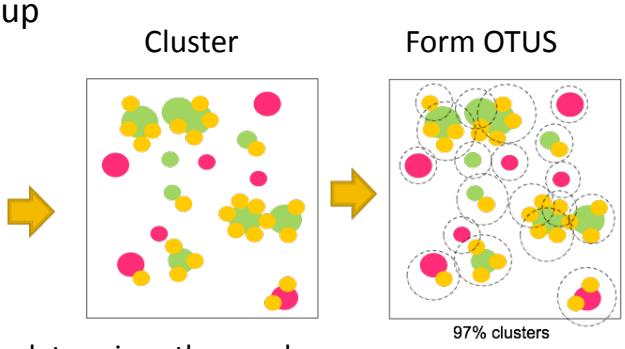


Oh,…Kong,
Genome Med 2012

Ecological Measures, 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

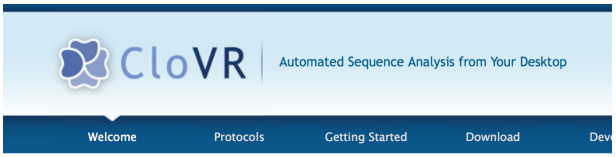
Calculate all pairwise distances between sequences



% 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.

Adapted from Robert Edgar

Pipeline tools



Comparing Bacterial Diversity: Community Membership & Structure

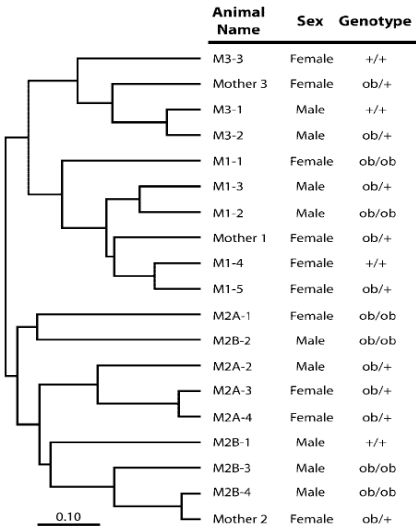
	Grp A	Grp B
	60	50
	34	50
	2	0
	2	0
	2	0

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

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

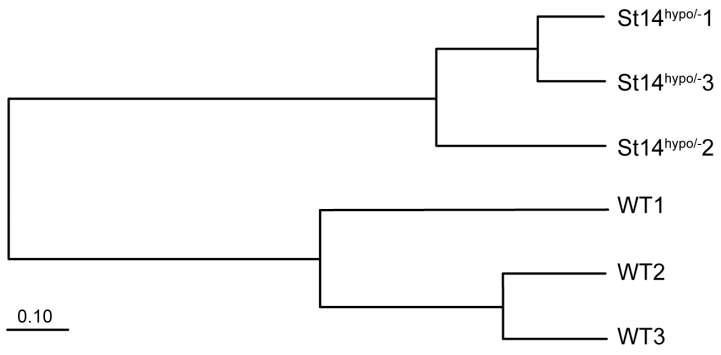
29

Community Membership:
Pups are most like their mothers



30

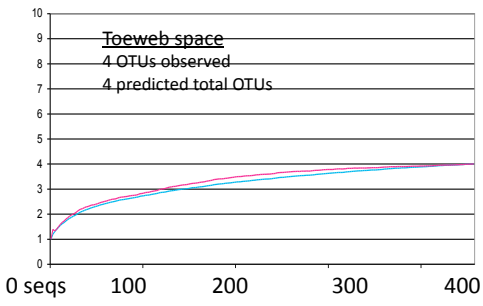
Community Structure: Pups cluster according to genotype

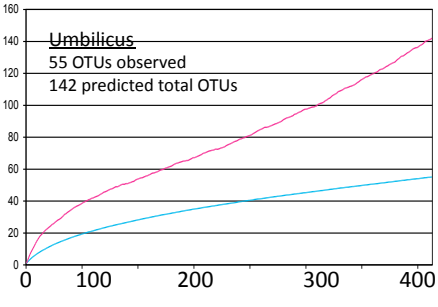


Scharschmidt et al. JID 2009

31

How many reads do you need? Ballpark 1,000 sequences for first pass analysis. With high throughput sequencing, no longer as relevant. How much diversity is there in the population? Have you sequenced enough to capture the diversity? Chao1 rarefaction curves

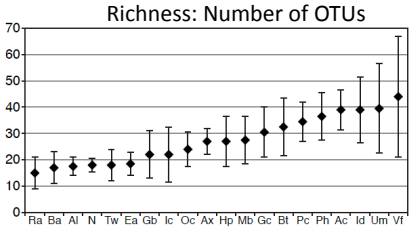




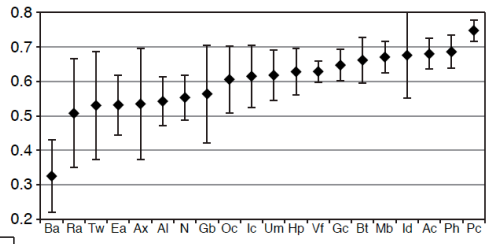
32

Richness, evenness, diversity:
Shannon and Simpson diversity

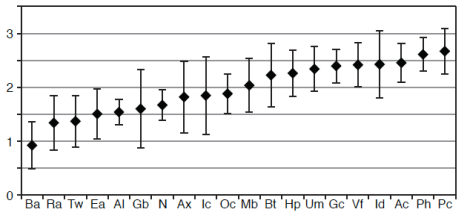
Richness: Number of OTUs



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



33

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

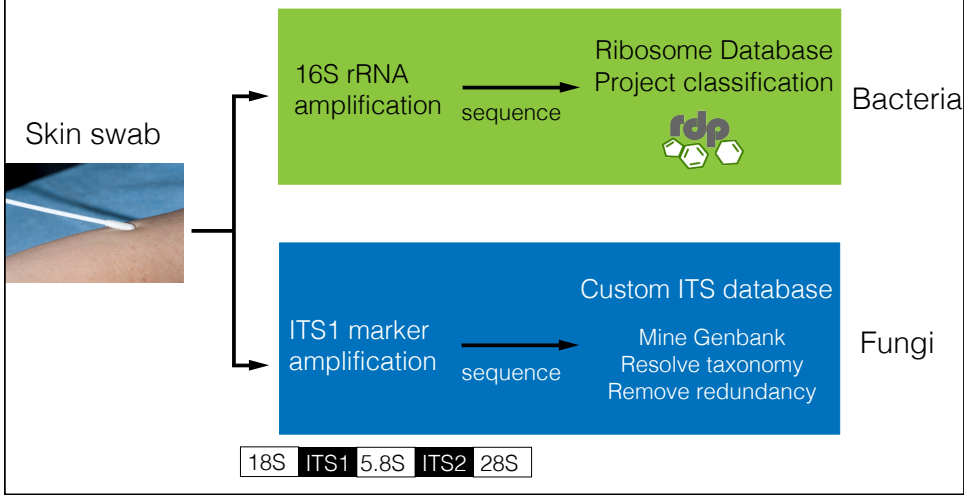
STUDY DESIGNS

Experimental and analytical tools for studying the human microbiome

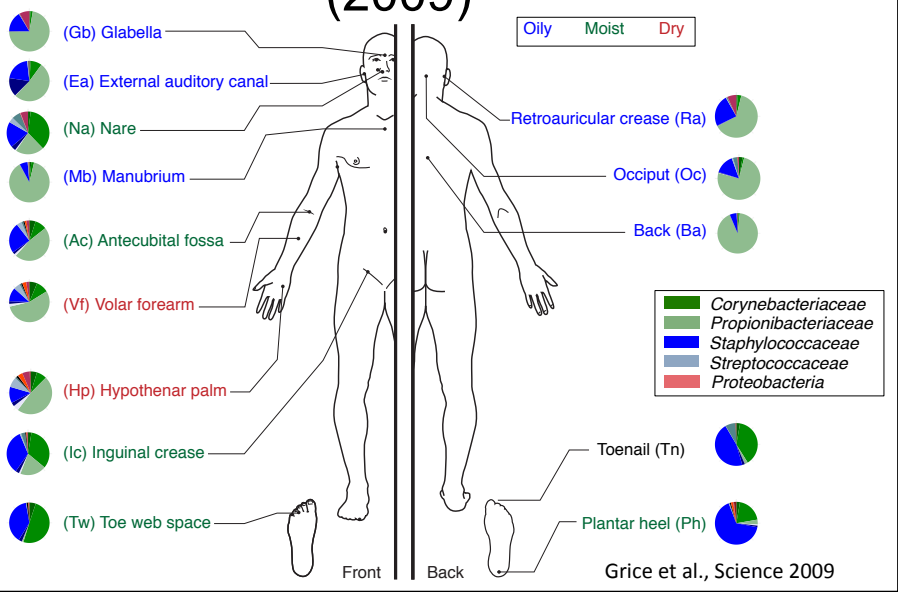
Justin Kuczynski[1], Christian L. Lauber[2], William A. Walters[1], Laura Wegener Parfrey[3], José C. Clemente[3], Dirk Gevers[4] and Rob Knight[3,5]

34

TOPIC 2: Fungal Diversity

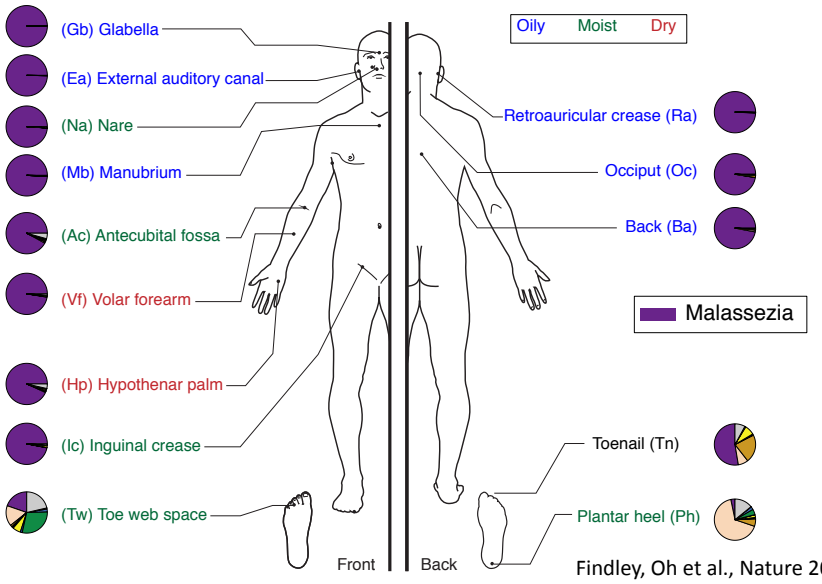


Healthy skin bacterial survey (2009)



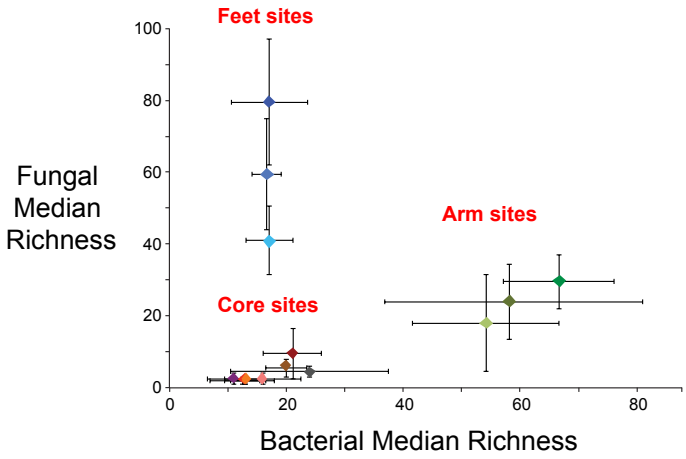
Grice et al., Science 2009

Healthy skin fungal survey (2013)



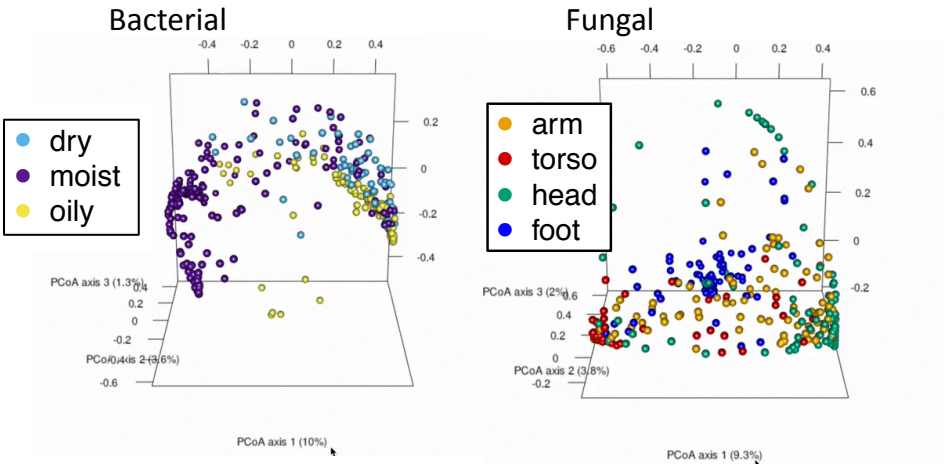
Findley, Oh et al., Nature 2013

Core sites are low diversity; divergent at peripheral sites



Findley, Oh et al., Nature 2013

Different forces shape bacterial and fungal communities



Findley, Oh et al., Nature 2013

TOPIC 3. BACTERIAL GENOME

1. What is study objective? E.g. Determine if two hospital isolates are clonal? Or Determine what genes are encoded by diverse set of Staphylococcus epidermidis?
2. What reference genomes exist for phylogenetic comparison?
3. What sequencing platform will you use?
4. What depth of sequencing do you need for assembly?
5. What assembly tool will you use? What alignment tool will you use?
6. How will you display your data?
7. How will you compare your results with other published studies?
8. What information will yield a testable hypothesis?

TOPIC 3. BACTERIAL GENOME

How to Assemble a Bacterial Genome: Gram-negative is ~6,000,000 base pair

Shotgun sequence 2x300 bp fragments on Illumina MiSeq at 30-fold redundancy.

Overlapping reads form large DNA contigs with N50 of ~100 kb.

Or very low coverage (3-5X) just to define species and strain

Assemblers (*de novo*)

- mira
- Velvet
- SPAdes
- MaSuRCA
- SOAPdenovo2
- Newbler (454)
- ALL-PATHS, DISCOVAR

Hunt *et al. Genome Biology* 2014, **15**:R42
http://genomebiology.com/2014/15/3/R42

Genome Biology

RESEARCH Open Access

A comprehensive evaluation of assembly scaffolding tools

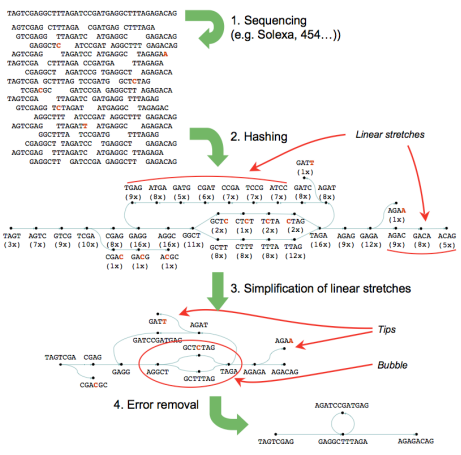
Martin Hunt[1*], Chris Newbold[2,1], Matthew Berriman[1] and Thomas D Otto[1]

42

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

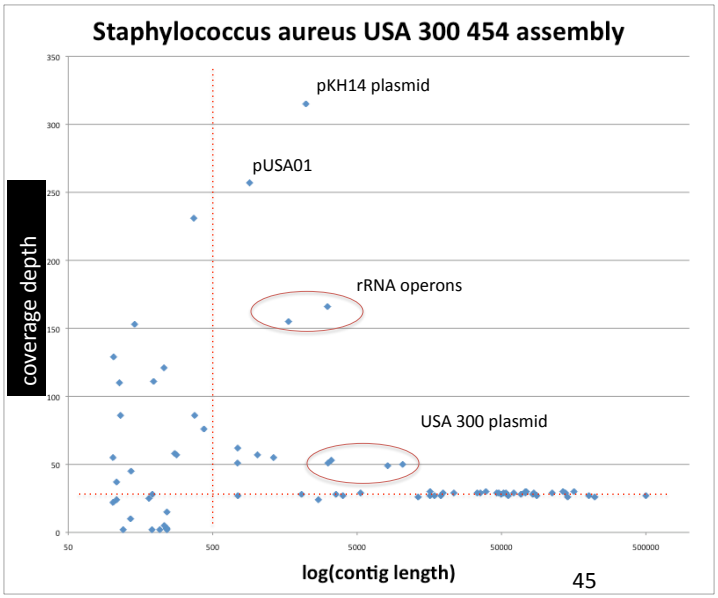


Evaluating Assemblies

- 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

44

Evaluating High Coverage Contigs



45

Genome Aligners: Compare sequences to identify sequence nucleotide variants, Insertion/Deletions

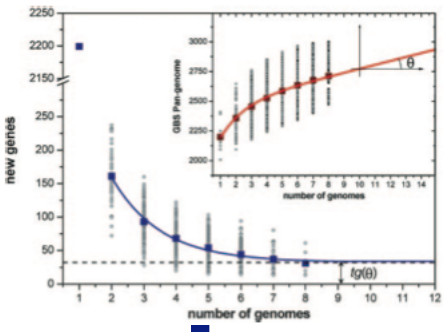
1. MumMER
2. MUGSY
3. MAUVE

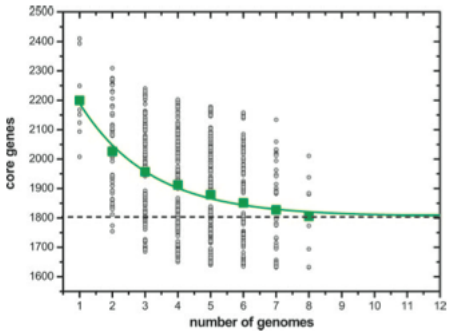
Genome Annotation: Predicting and naming genes encoding proteins

1. PGAAP (NCBI)
2. IMG (JGI)
3. Glimmer, GeneMark

46

Is there a reference genome? Is the 'pan-genome' open or closed?





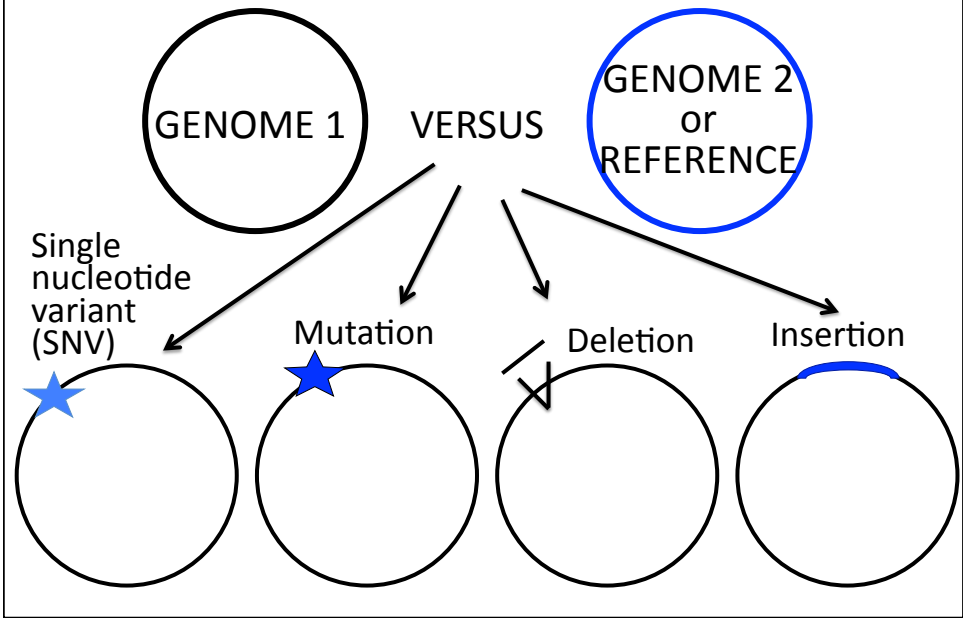
PNAS

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

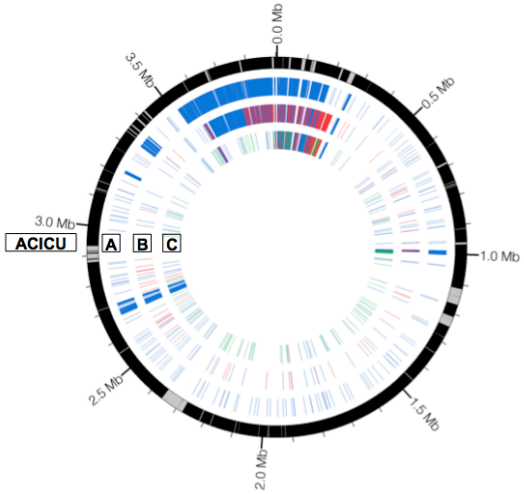
Hervé Tettelin[a,b], Vega Masignani[b,c], Michael J. Cieslewicz[b,d,e], Claudio Donati[c], Duccio Medini[c], Naomi L. Ward[a,f], Samuel V. Angiuoli[g], Jonathan Crabtree[a], Amanda L. Jones[g], A. Scott Durkin[a], Robert T. DeBoy[a], Tanja M. Davidsen[a], Marirosa Mora[c], Maria Scarselli[c], Immaculada 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[h], Shannon Smith[i], Teresa R. Utterback[i], Owen White[a], Craig E. Rubens[d], Guido Grandi[c], Lawrence C. Madoff[e,j], 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]

47

Whole Genome Sequence Comparison



Polyclonal outbreak of multi-drug resistant *Acinetobacter baumannii* (A,B, C). Clusters of variants = recombination



SNP found in A

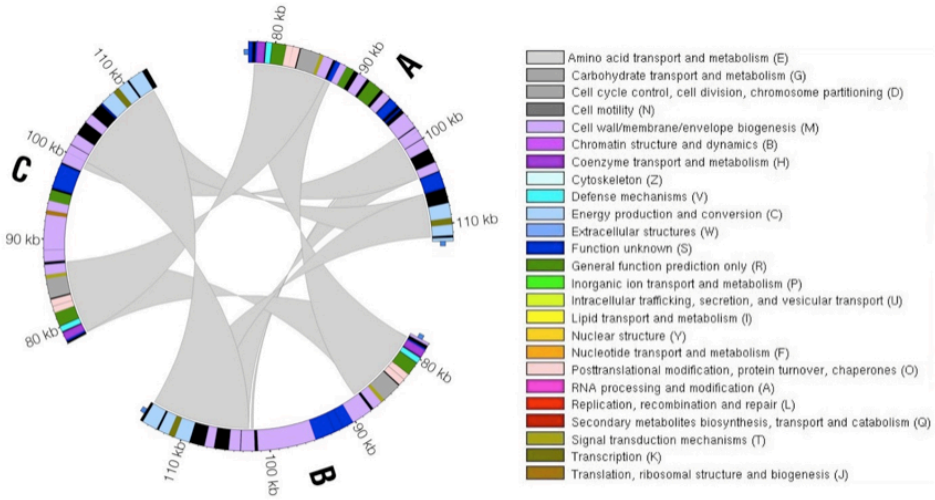
SNP found in B, but not A,

SNP found in C, but not A or B

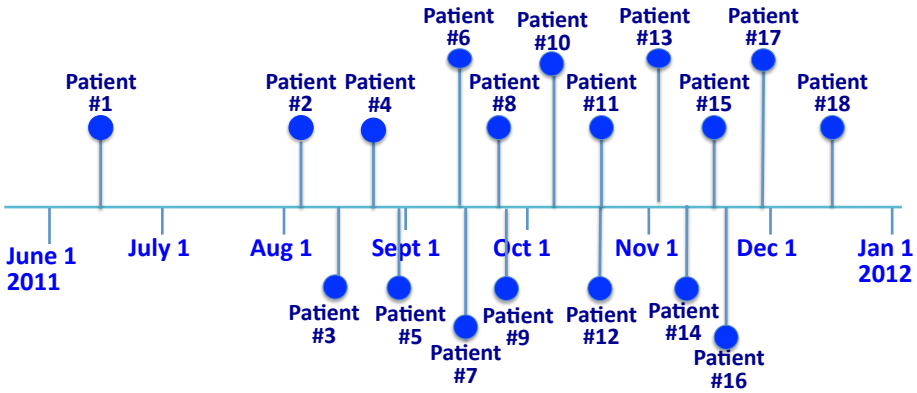
Circos: an information aesthetic for comparative genomics. Marra MA. Genome Res. 2009 Sep; 19(9):1639-45

*ACICU is a previously sequenced strain of the EC II lineage from a 2005 outbreak in Rome

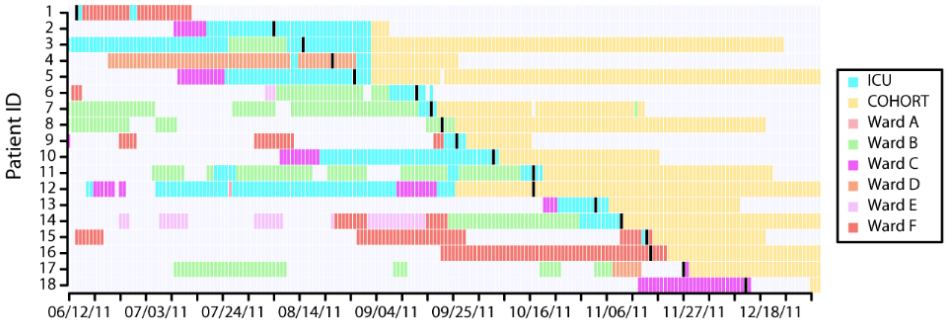
Alternate O-antigen biosynthetic clusters



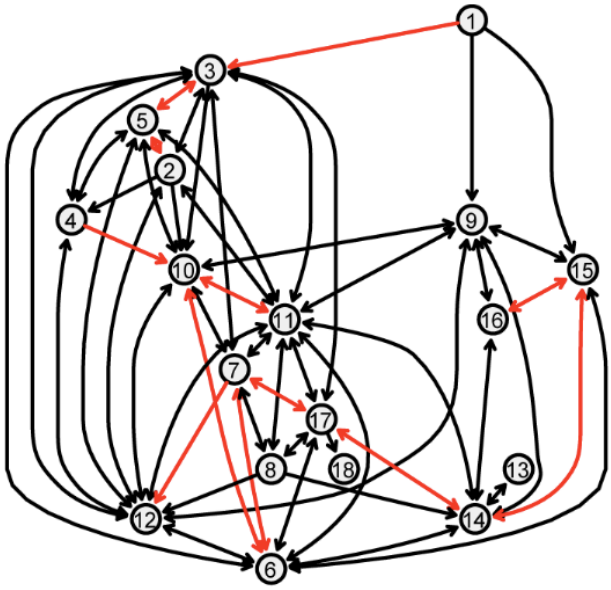
2011 outbreak at NIH of carbapenem-resistant *Klebsiella pneumoniae*:



Reconstructing transmission

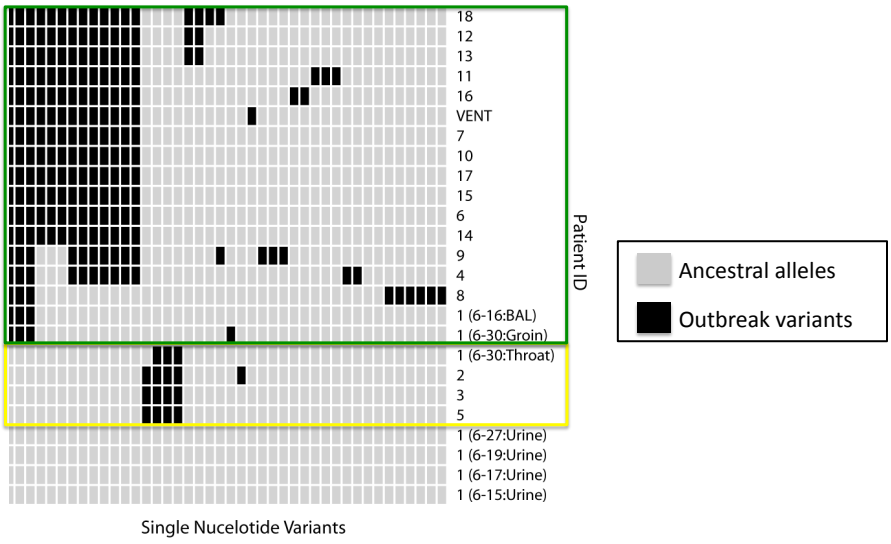


Reconstructing transmission

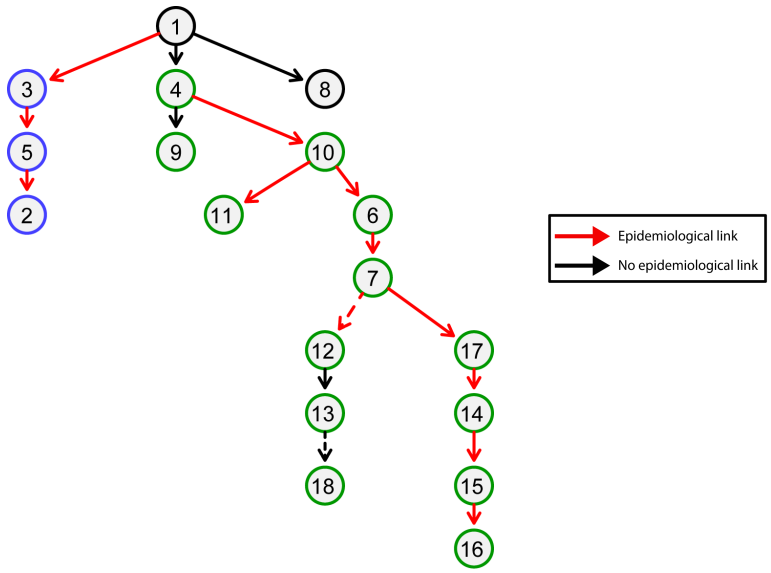


Genetic heterogeneity among isolates taken from patient 1

Evidence for transmissions originating from distinct sites on patient 1



Reconstructing transmission



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

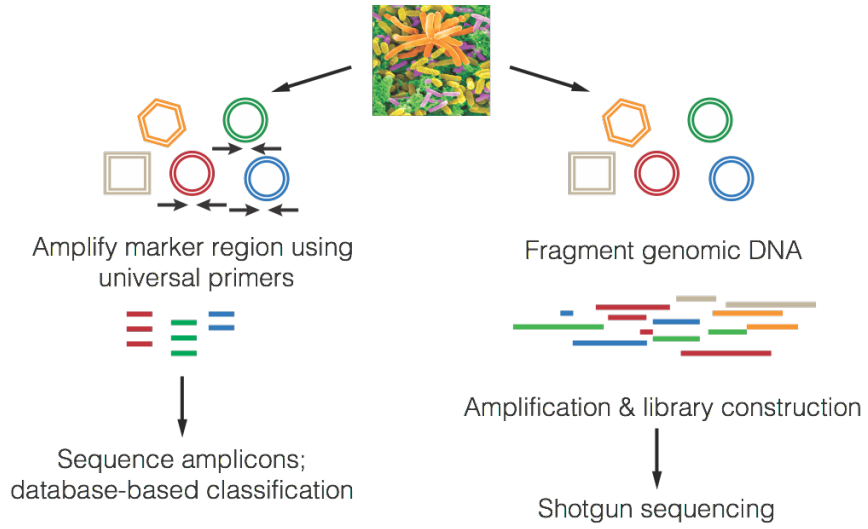
Ten years after the term metagenomics was coined, the approach continues to gather momentum. This culture-independent, molecular way of analysing environmental samples of cohabiting microbial populations has opened up fresh perspectives on microbiology.

57

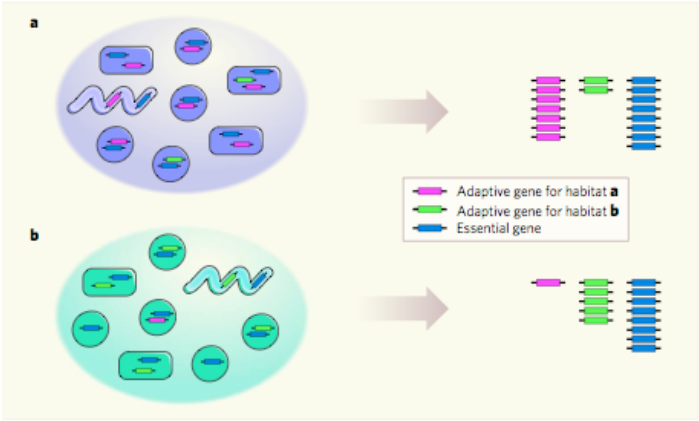
Goals of whole genome shotgun metagenomic analysis

1. Want to know who's there & abundance
2. Want to know what they do (function)
   - Want to know what genes are present
   - Can we identify pathways
3. Can we recover genomes

Whole genome shotgun sequencing vs. amplicon



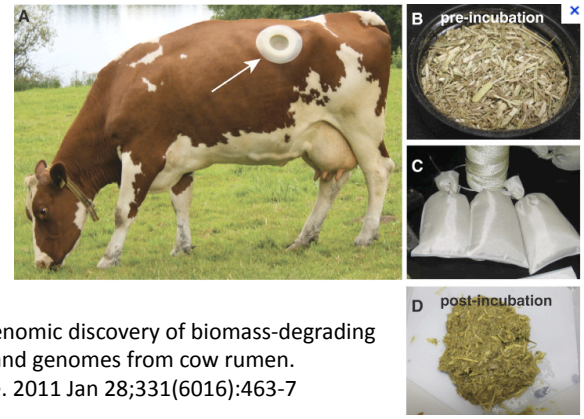
Metagenomics: types of bacteria similar between 2 populations, but pink genes enriched in top population



60

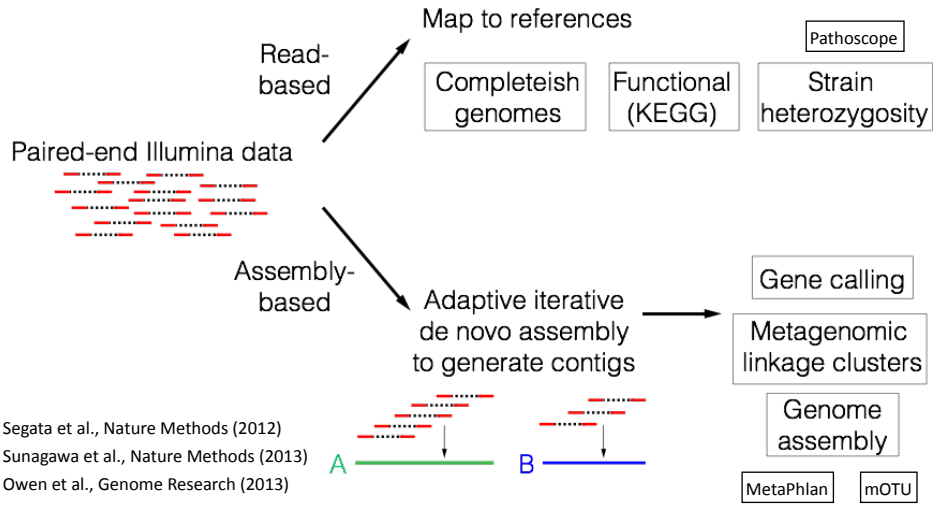
Using metagenomic sequencing to find new metabolic enzymes

Nature. 2007 Nov 22;450(7169):560-5. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.



Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. 2011 Jan 28;331(6016):463-7

Metagenomics: computational infancy to develop pipelines for analysis



Segata et al., Nature Methods (2012)
Sunagawa et al., Nature Methods (2013)
Owen et al., Genome Research (2013)

Looking for function

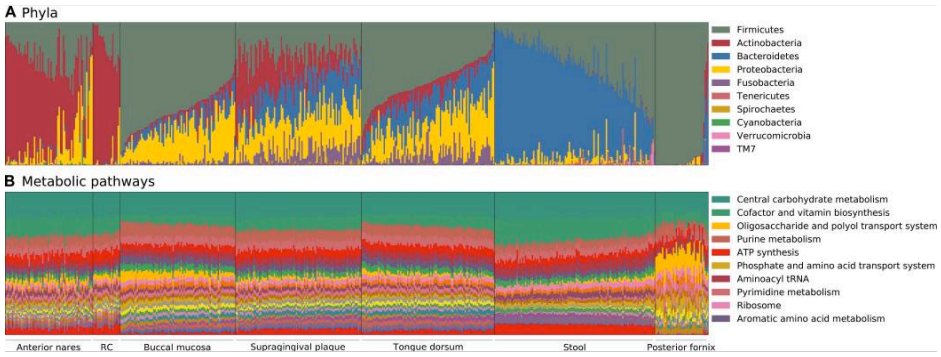
- Leverage functional databases like

- Generally, use blastx-like programs to map reads to these databases

HUMANN (The HMP Unified Metabolic Analysis Network): post-processes BLASTx hits into pathways

Abubucker et al., PLoS Computational Biology (2012)

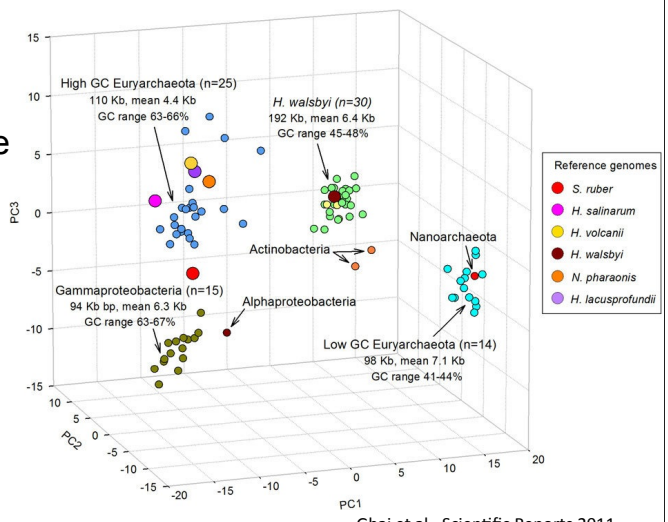
Example output



Human Microbiome Project
Nature. 2012

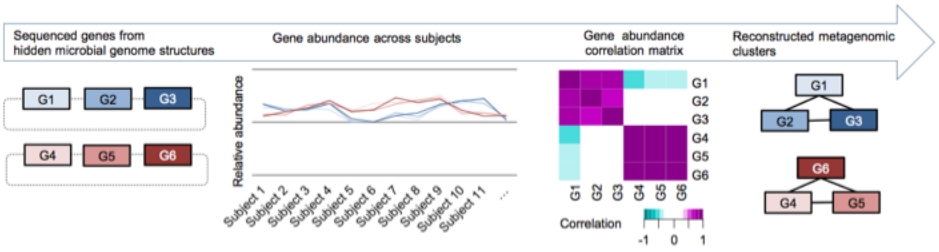
Calling genomes

- Binning methods
  - GC content/ tetranucleotide frequencies



Ghai et al., Scientific Reports 2011

Metagenomic linkage groups to aggregate contigs



Karlsson et al., Nature (2013)

Human DNA Admixture

- 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

68

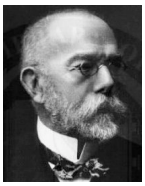
Topic 5: Where is sequencing technology going?

Now: Illumina MiSeq generates 2x300 bp paired end for amplicon and whole-genome sequencing. Costs ~$100K

Future: ? (REFERENCE GENOMES for hospital pathogens is my #1 priority; CLINICAL REPORTS from genomic sequence data is also my #1 priority.

Roche/454-XLR	Illumina MiSeq, HiSeq	Pacific Biosciences
•Emulsion PCR •400-bp read (avg)	•Bridge PCR •300 or 100 bp read, paired end	LONG reads, accurate full genome assemblies with end-to-end coverage of chromosome and plasmids

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.

70