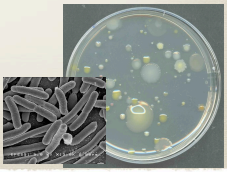
METAGENOMICS

My Lab

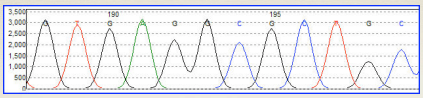
Traditional Genomics

Environmental Sample

Isolate Organism and Culture



Once cultured -- Sequence



Whole Genome Shotgun Sequencing Method

Genomic DNA

Sequence Each Fragment with Shotgun Approach

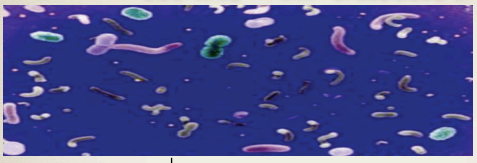
Align Contiguous Sequences

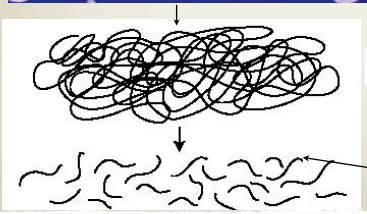
Generate Finished Sequence

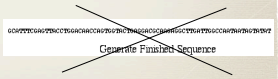
Only one possible genome in

So, only one possible out

Metagenomics - No longer "need" to culture





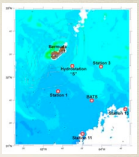


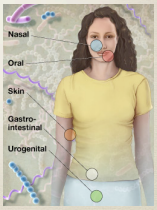
Each is different sequence

Why do we want to know the sample composition?

What's this good for? Types of Environmental Samples

- From the ocean: (e.g. Sargasso Sea very salty and thus nutrient deprived)
  - Venter et al. (2004) sampled it and found diverse prokaryotic life despite doubts.
- From the human body: (e.g. our gastrointestinal tract)
- From extinct, ancient animals (e.g. separate a Mammoth's DNA from the other sample's 40% of bugs and plants)





Trillions are depending on you (mostly - you're depending on them): The Human Microbiome

- You are 100 trillion cells (yours have a nucleus and therefore, you are eukaryotic)
- You contain 1000-2000 trillion Microbes (10/20-to-1) -- (they have no nucleus -- they are prokaryotic)
- Microbes make up 1-2% of your body mass
- Every gram of your intestines have ~100 billion bacteria
- Microbes in our intestines help produce certain essential vitamins such as vitamin B12.
- Most microbes deny disease-causing microbes the foothold they need to colonise our bodies and do damage

Implications that can revolutionize technology

Steven Chu, Secretary of Energy

- Can't use plants as fuel because don't know how to break it down
- Termites' guts can digest cellulose efficiently and "fighteningly quickly"

The new anti-cancer drug -- Tunicates?

A tunicate from a Thai coral reef: a potential source of new anticancer compounds

Chavanich, S.; Koeysin, P.; Viyakarn, V.; Piyatiratitivorakul, S.; Menasveta, P.; Suwanborirux, K.; Poovachiranon, S.

Coral Reefs, Volume 24, Issue 4, pp.621-621

Scientists have shown how a microbe that lives inside sea squirts could be used to biosynthesize a chemical compound that may help fight cancer.

(In collaboration with the Drexel Allegheny Institute)

What makes a healthy, fertile soil?

(Mary Ann Bruns at Penn State University is interested)

C. Pelagibacter -- most compact genome on earth

- Can't culture - only metagenomic reads
- Estimated 20 billion billion billion Pelagibacter microbes
- No junk
- The fewest genes "life" can get-away-with and still be efficient

(In collaboration with Oregon State University)

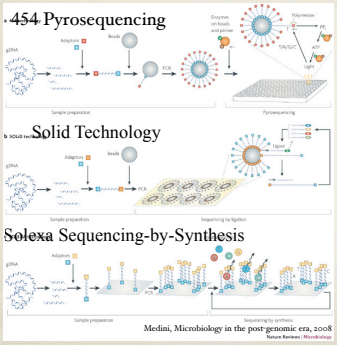
Microbes in Ants and Fruit Flies

- Dr. Russell's research
- Co-evolution of arthropod gut microbes with different arthropod lineages
- 3:30 Talk by Noah Whiteman in Stratton 113 on 'ecology and evolution of host-symbiont interactions'

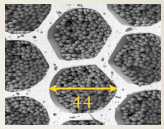
Precursory Steps to Analysis

- Obtain Environmental Sample (Do not taint!)
- Sequence the Sample / DNA Microarrays (detect gene or gene expression)/ Metaproteomics (mass spectroscopy to identify proteins in sample)
- Sequencing: Traditionally (standard for 30 years), Sanger (Chain-Termination Method used)
  - ~850 bp read (usually good because unique and have less "chunks" to sequence)
  - Needs an amplification step (cloning) -- SLOW

New Sequencing Technologies:
Short and Fast



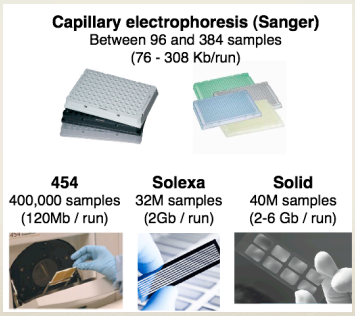
100-250 bp reads



25-35 bp reads

25-40 bp reads

High Throughput Sequencing



Comparing metrics and performance of next-generation DNA sequencers

	Platform		
	Roche(454)	Illumina	SOLiD
Sequencing chemistry	Pyrosequencing	Polymerase-based sequencing-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired ends/separation	Yes/3 kb	yes/200 bp	Yes/3 kb
Mb/run	100 Mb	1300 Mb	3000 Mb
Time/run (paired ends)	7 h	4 days	5 days
Read length	250 bp	32-40 bp	35 bp
Cost per run (total direct[a])	$8439	$8950	$17 447
Cost per Mb	$84.39	$5.97	$5.81

[a] Total direct costs include the reagents and consumables, the labor, instrument amortization cost and the disc storage

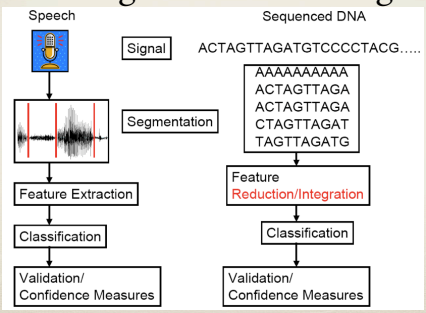
E. R. Mardis, "The impact of next-generation sequencing technology on genetics," Elsevier Trends in Genetics, 2008.

Sequencers generate fragments (sampling underlying distribution)

Problem: If billions of organisms with 100's of trillions of nucleotides (and we can sample a billion bp on a good day) -- what is Nyquist for this type of problem?

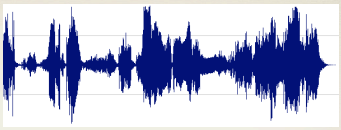
First step: Reads need to be classified

Parallel of Speech Processing to Metagenomic Processing



Signal Processing for Recognition

* Speech Recognition (Gene and Function)
* **Speaker Identification** (Taxa ID)



y= s+n (signal + noise)
noise is assumed to be Gaussian

Who's voice?
(Which taxa?)

Speaker Identification Problems and their parallel - I

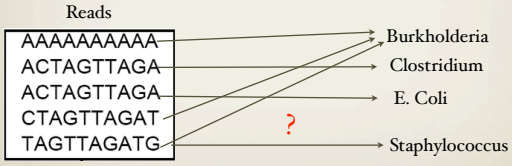
- How do you account for noise in the signal?
  - Microphone noise, environmental "noises"
  - Our equivalent -- Sequencing error, "background DNA" (currently everything is of interest)
  - Sequence-to-Taxa mapping

http://cnx.org/content/m14200/latest/

Noise in Taxa Recognition

y, DNA reads, are assumed noiseless (not true assumption)

Noise is in the mapping from sequence to Taxa



Speaker Identification Problems and their parallel - II

- How do you account for the different volumes of speakers
  - Imbalanced amounts of Taxa
  - Not the nebulous "Normalization of Metagenomic data"
- How do you characterize a voice?
  - Spectral Formant Features are "Fingerprint" of Voice
  - Genomic features, Genes that characterize a Taxa

Speaker Identification Problems and their parallel - II

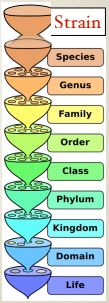
- How does the system handle varying speeds of inputs
  - Time-Frequency trade-off in Speech
  - Same with metagenomic reads -- shorter the read, the "less resolution" we have
- How can you account for imitating speech patterns?
  - Mimicking and Interference
  - Close strains and species

Classified as what?

- Taxa
- Genes
- Function

2007: 1.8 million species known to Science

- A species is often defined as a group of organisms capable of interbreeding and producing fertile offspring.
- Does not apply to **asexual single-cells**
- Horizontal gene transfer violates the getting-genes-only from parents assumption (Vertical)



Traditional Taxa Classification - 16S

Highly Conserved 16S RNA sequences

- 1200 bp average length
- 3' end of 16S RNA bind to mRNA to start translation

1. Universally distributed, allowing the comparison of phylogenetic (tree-of-life) relationships (present in all 3 kingdoms)
2. Core of information genes which are only weakly affected by horizontal gene transfer
3. Functionally highly constrained mosaics of sequence stretches ranging from conserved to more variable (most organisms will die with too high of mutation)

11 Gram Negative Bacteria 16S regions:
http://www.jlindquist.com/sequence.html

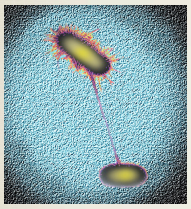
Taxonomy Standards Based on 16S

- Intra-species 16S variation -- 3%
  - New: Can sometimes be as low as 1% and as high as 5%
- Intra-genus 16S variation -- not well studied but can be 6-9%

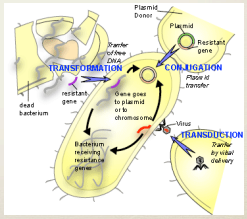
16S tells part of the story

- ✳ 1200bp out of millions of bases (other parts of genome may contain clues)
- ✳ Other parts may be laterally transferred

Extremely Mobile Elements: Horizontal Gene Transfer



- Like your neighbors drug-resilience?
- Get a few plasmids from them

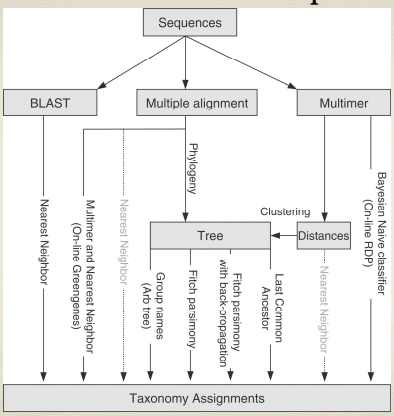


Anomalies

Type of relationship	First strain	Second strain	No. of bp differences	Reason for this example
Same genotype but different phenotypes	*Mycobacterium tuberculosis* ATCC 27294	*M. bovis* ATCC 19210 or *M. africanum* ATCC 25420	0	Although given names that appear to be species, the three strains are now designated subspecies and show small or no genetic difference, but there is a definite phenotypic and pathological difference. See reference 43 for other examples within the mycobacteria. This rare situation is a limitation to the use of sequencing as gold standard.
Similar genotype but different phenotypes	*E. coli* ATCC 11775	*Shigella dysenteriae* ATCC 13313	3	Two high-level pathogens are genotypically close enough to be considered the same species but have kept clinically and historically important separate names.
	Streptococcus pneumoniae ATCC 33400	*S. mitis* ATCC 49456	3	A high-level pathogen and a commensal are genotypically very similar.
	Streptococcus bovis	*S. equinus*	2	Very close genogroups which are difficult to distinguish phenotypically in the clinical laboratory. However, their reported difference in lactose reaction and different niches may justify the different names.
Similar phenotypes but different genotype	*Nocardia asteroides* ATCC 19247	*N. farcinica* ATCC 3318	13	Separate species which are relatively difficult to distinguish phenotypically are easy to distinguish by sequence.
	S. bovis ATCC 33317T	*S. bovis* ATCC 43143	13	Separate genogroups which are difficult to distinguish phenotypically are easy to distinguish by genotype (11).
Too distant to be the same species	*Enterobacter (Pantoea. agglomerans* (hg1)	*E. (Pantoea. agglomerans* (hg2)	27	Strains that were originally thought to be biogroups within the same species are genotypically distant enough to be considered separate genera.
Too distant to be the same genus	*Clostridium tetani* ATCC 19406	*Clostridium tonocuum* ATCC 14501	About 104	Although these two organisms have been given the same genus name, the large difference (20%) means that at least one has been taxonomically misplaced. Species in the same genera should not differ by more than about 5 to 6%.
Too close to have different names	*Mycobacterium duvalii*	*"M. valentiae"*	0	Sometimes names get into the literature without full justification.
Too close to be three different genera	*Enterobacter cloacae* ATCC 13047	*Leclercia (Enterobacter. adecarboxylata* ATCC 23216	1-2	These three organisms, which have been placed in different genera, are genotypically close enough to be considered the same species.
	E. cloacae ATCC 13047	*Citrobacter werkmanii*	6	
Subspecies	*Streptococcus dysgalactiae* subsp. *dysgalactiae*	*Streptococcus dysgalactiae* subsp. *equisimilis*	14/500 and 161,500	The difference between these two "subspecies" is greater than that between the genera above.
	Streptococcus equi subsp. *equi*	*Streptococcus equi* subsp. *zooepidemicus*	1/500 and 1/1,500	The only difference between the two subspecies was at bp 204.
	Staphylococcus cohnii subsp. *cohnii*	*Staphylococcus cohnii* subsp. *urealyticum*	1/500 and 3/1,500	The difference of 0.33% was the same calculated using either the 500-bp or the 1,500-bp sequence.

Clarridge. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases, CMR, 2004.

Current 16S Techniques



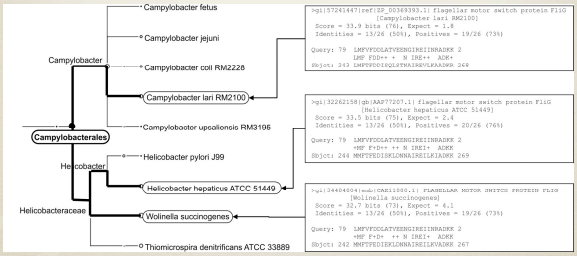
Liu et al. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers, Nucleic Acids Research, 2008

Multimer vs. BLAST (Composition vs. Homology)

- A Multimer is a string of length N (or what we refer to as Nmer)
  - A= 1mer, AT = 2mer, ATG = 3mer, ATCG = 4mer, etc.
- Multimer approaches "count Nmer frequencies". These features characterize the full gene.
- Classifies according to Probability of finding Multimers in sequence. The more a sequence diverges, the less likely this is to happen.
- BLAST uses dynamic programming -- maximize score between two sequences. (and find sequence similarity)
- Arguably more robust to a high-level of mutations

Least Common Ancestor

- Assigns each read to the lowest common ancestor (LCA) of the set of taxa that it hit in the comparison



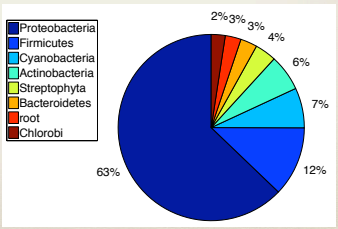
Techniques for full-genomes

Features	Classifier	Published Method
Homology-based	Nearest-Neighbor	BLAST [34]
	Nearest-Neighbor & Last Common Ancestor	MEGAN [54]
Composition-based	Naïve Bayesian	Sandberg et al. [55]
		RDP classifier (16S sequences only) [56]
		Rosen et al. [57]
	Support Vector Machines	PhyloPythia [58]

Processing: Identify the Content from Reads

- Identify the Kingdom, Phyla, Class, Genera, and Species distribution of a sample
- Example in Matlab (2008a) Metagenomics Demo:

Analyzes first 100 reads of the Sargasso Sea (each around 850 bp)



Tree: BLAST + Least Common Ancestor

MEGAN (MEtaGenome ANalyzer): a graphical tool that uses BLAST reports to assess content of a sample

- Ambiguous hits for SHORT fragments

assigns each read to the lowest common ancestor (LCA) of the set of taxa that it hit in the comparison

Top-percent filter is used to retain only those hits for a given read r whose scores lie within a given percentage of the highest score involving r.