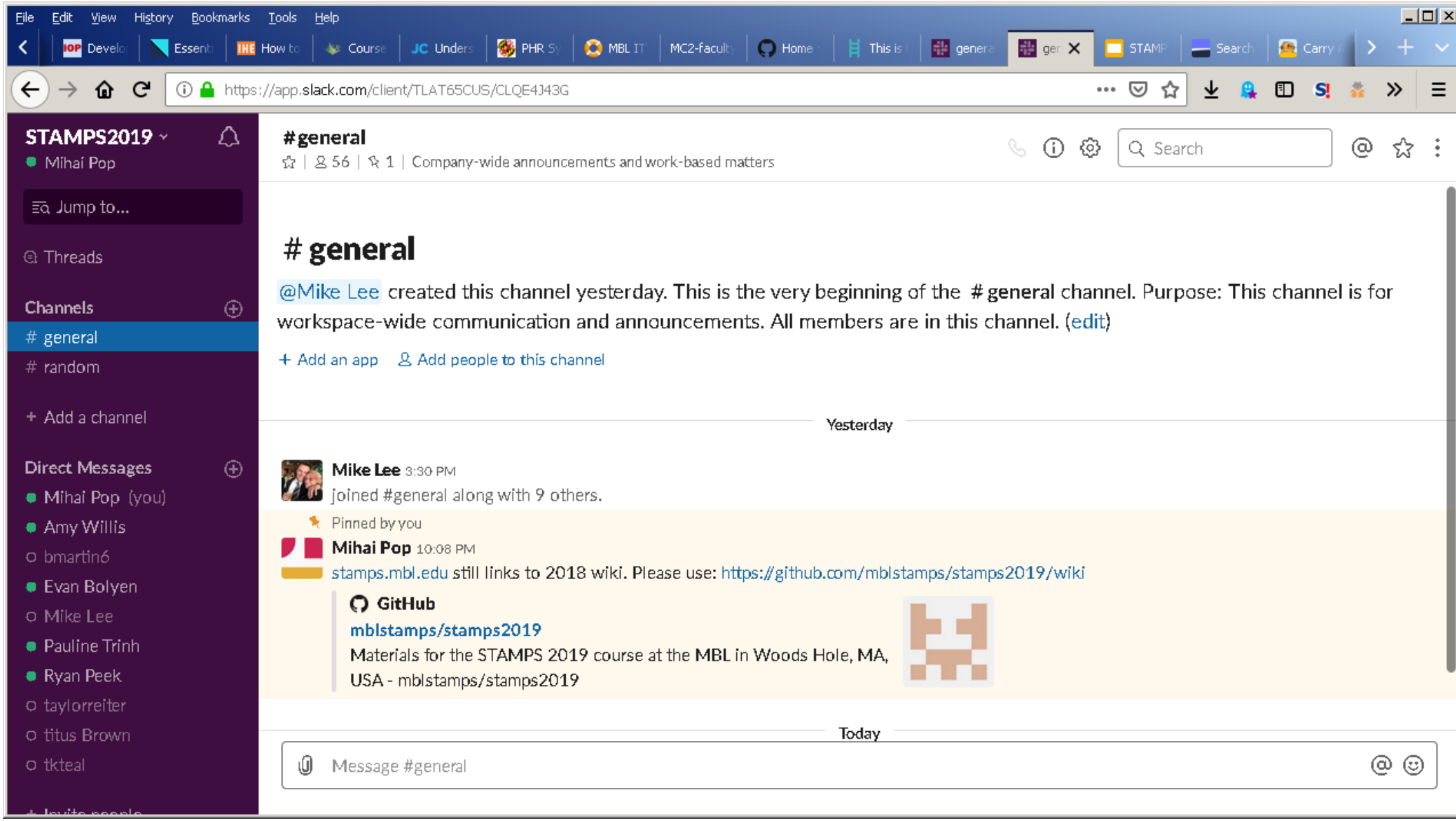
Metagenomics & STAMPS Overview

Mihai Pop
University of Maryland
College Park


File Edit View History Bookmarks Tools Help
https://app.slack.com/client/TLAT65CUS/CLQE4J43G
STAMPS2019
Mihai Pop
Jump to...
Threads
Channels
general
random
+ Add a channel
Direct Messages
Mihai Pop (you)
Amy Willis
bmartin6
Evan Bolyen
Mike Lee
Pauline Trinh
Ryan Peek
taylorreiter
titus Brown
tkteal
#general
56 | 1 | Company-wide announcements and work-based matters
Search
general
@Mike Lee created this channel yesterday. This is the very beginning of the # general channel. Purpose: This channel is for workspace-wide communication and announcements. All members are in this channel. (edit)
+ Add an app
Add people to this channel
Yesterday
Mike Lee 3:30 PM
joined #general along with 9 others.
Pinned by you
Mihai Pop 10:08 PM
stamps.mbl.edu still links to 2018 wiki. Please use: https://github.com/mblstamps/stamps2019/wiki
GitHub
mblstamps/stamps2019
Materials for the STAMPS 2019 course at the MBL in Woods Hole, MA, USA - mblstamps/stamps2019
Today
Message #general

About STAMPS

- Learning goals
- Course format
- Code of Conduct
- Course materials
- Communication channels
- Introductions

Learning goals

Strategies and Techniques for Analyzing Microbial Population Structure

- Foundational skills to work with metagenomic data
- How to work effectively and reproducibly with data
- Familiarity and practice with particular bioinformatics tools
- Ability to learn about and evaluate other tools and understand caveats
- Perspective and confidence to apply these skills in your own work
- Empower you to ask and answer the questions you have of your own data
- Building a framework and a network

Course format

- Hands-on, active learning
- Materials available during and after the course
- Mix of lectures, tutorials and practice
- Ask questions!
- Feedback and formative assessment
- Learn from each other as well as instructors and TAs
- Learn by teaching

Communication channels

- https://stamps.mbl.edu
- https://github.com/mblstamps/stamps2019/wiki
- Slack (announcements, links and conversations)
  - https://stamps2019.slack.com
- Twitter #stamps2019

- Green and red stickies

Code of Conduct

STAMPS Code of Conduct

All STAMPS attendees are expected to agree with the following code of conduct. We will enforce this code as needed. We expect cooperation from all attendees to help ensuring a safe environment for everybody. MBL also has a formal Code of Conduct.

The Quick Version

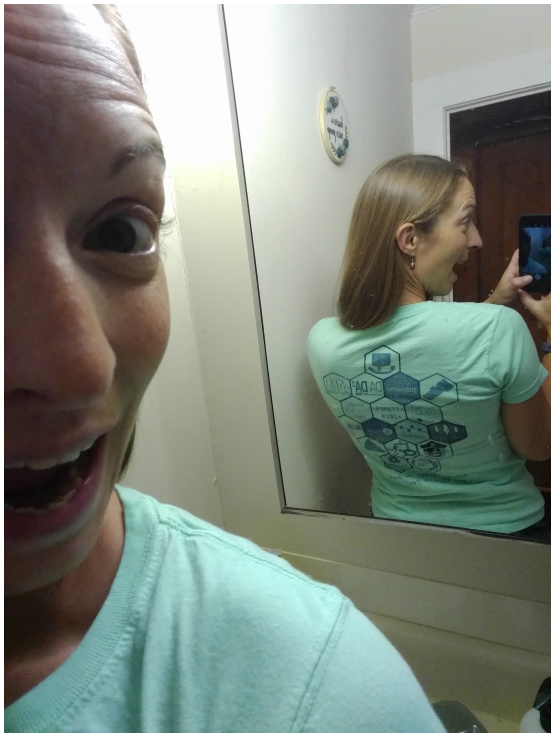
STAMPS events are neither a dating scene nor an intellectual contest.

STAMPS is dedicated to providing a harassment-free experience for everyone, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, or religion (or lack thereof). We do not tolerate harassment of participants in any form. Sexual language and imagery is generally not appropriate for any STAMPS venue.

If you are being harassed, notice that someone else is being harassed, or have any other concerns, please contact Tracy Teal or Mihai Pop immediately. If either is the cause of your concern, please see the MBL's Equal Employment Opportunity Coordinator at 508-289-7378 or eeo@mbl.edu.

https://github.com/mblstamps/stamps2019/wiki/Code-of-Conduct

T-shirt

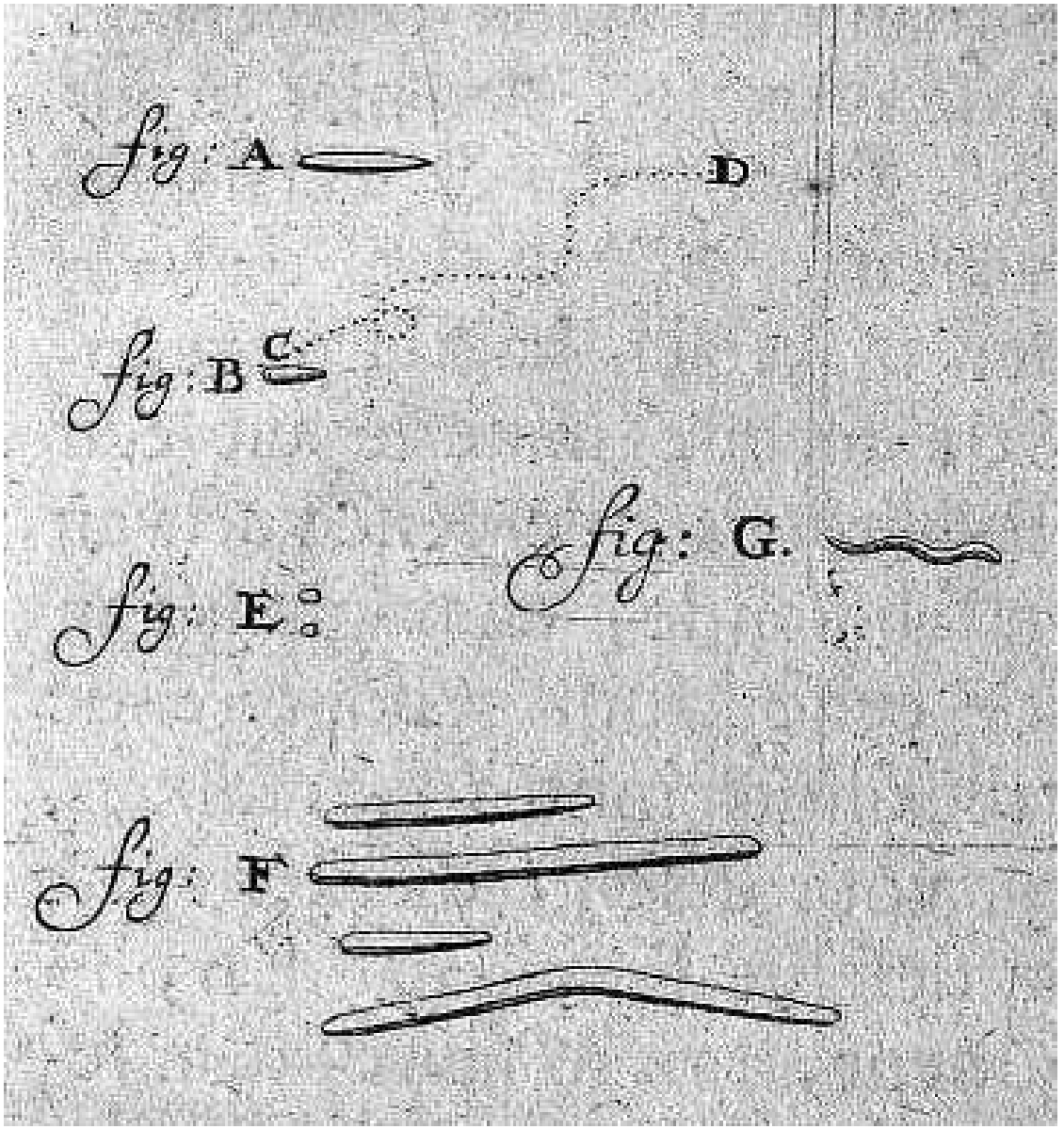


Introductions!

- TAs
- Faculty

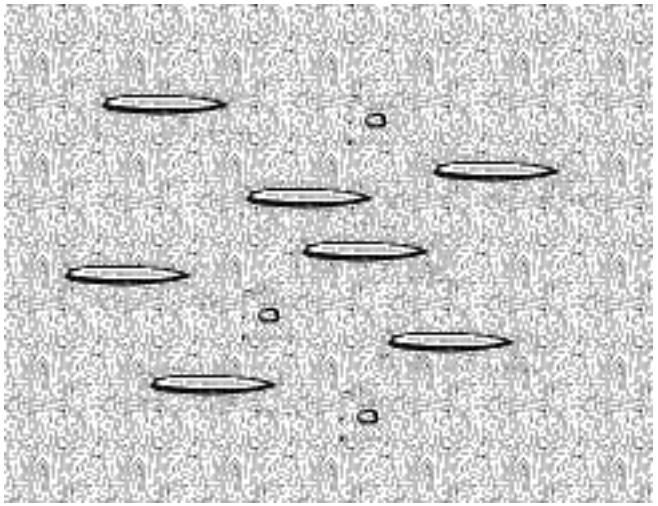
Let’s get started!

17th century biology

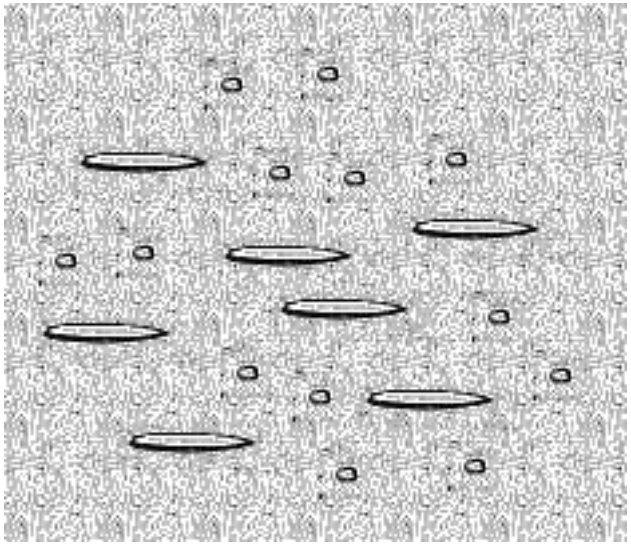


What causes disease?

Healthy



Sick



21st century biology

```
>F4BT0V001CZSIM rank=0000138 x=1110.0 y=2700.0 leng
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAA
>F4BT0V001BBJQS rank=0000155 x=424.0 y=1826.0 lengt
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAA
>F4BT0V001EDG35 rank=0000182 x=1676.0 y=2387.0 leng
ACTGACTGCATGCTGCCTCCCGTAGGAGTCGCCGTCCTCGACNC
>F4BT0V001D2HQQ rank=0000196 x=1551.0 y=1984.0 leng
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCGTCCCTCGAC
>F4BT0V001CM392 rank=0000206 x=966.0 y=1240.0 lengt
AANCAGCTCTCATGCTCGCCCTGACTTGGCATGTGTTAAGCCTGTAGGCTA
>F4BT0V001EIMFX rank=0000250 x=1735.0 y=907.0 lengt
ACTGACTGCATGCTGCCTCCCGTAGGAGTGTCGCGCCATCAGACTG
>F4BT0V001ENDKR rank=0000262 x=1789.0 y=1513.0 leng
GACACTGTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAA
>F4BT0V001D91MI rank=0000288 x=1637.0 y=2088.0 leng
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAA
>F4BT0V001D0Y5G rank=0000341 x=1534.0 y=866.0 length=75
GTCTGTGACATGCTGCCTCCCGTAGGAGTCTACACAAGTTGTGGCCCAGAACCACTGAGCCAGGATCAAACTCTG
>F4BT0V001EMLE1 rank=0000365 x=1780.0 y=1883.0 length=84
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAATGCTGCATGCTGCTCCCTGAGCCAGGATCAAACTCTG
```

Quick aside – sequencing technologies

- Caveat: Slide is already obsolete!

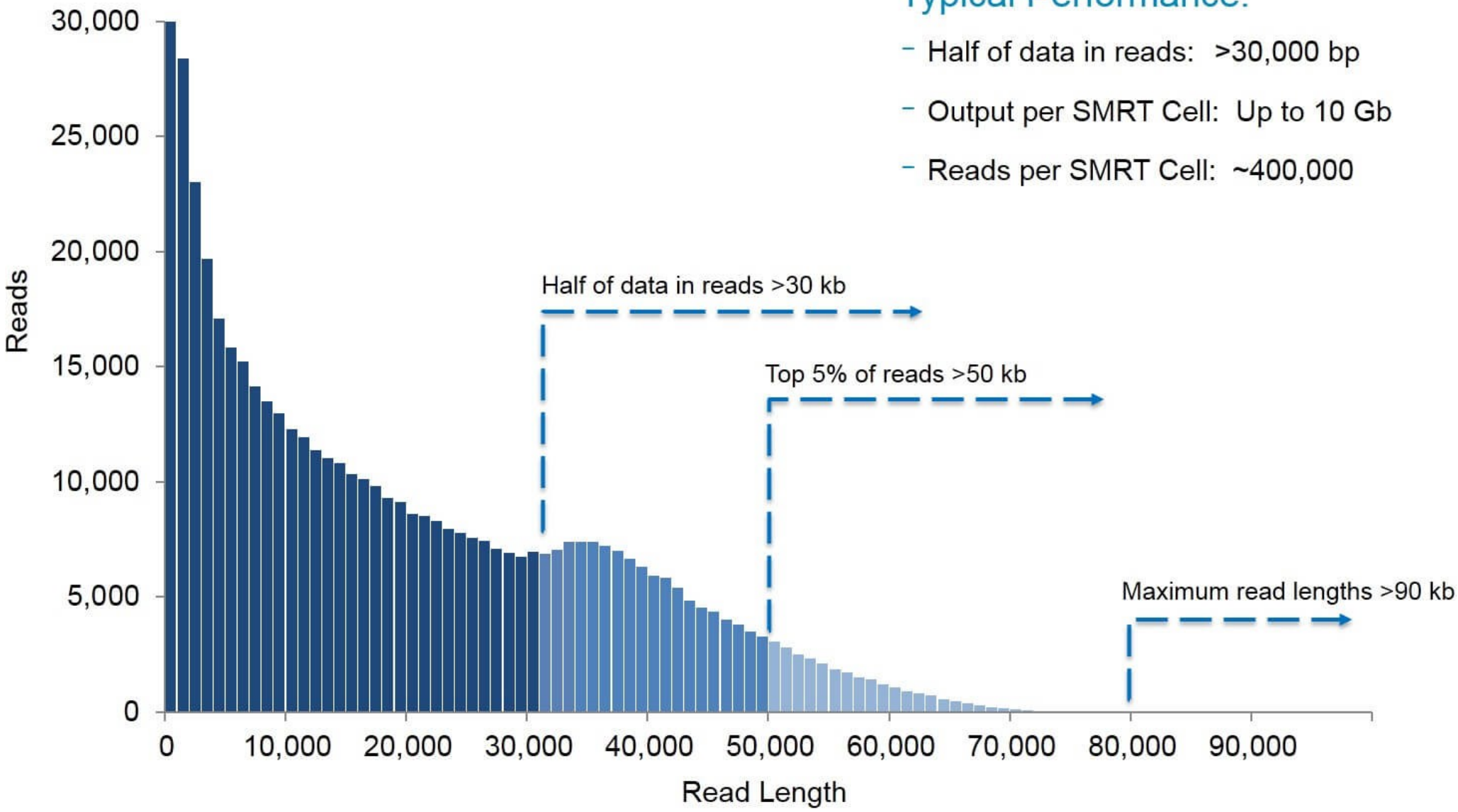
Technology	Throghput (reads/run)	Read length (bp)	Native read pairs	Benchtop	Comments
MiSeq	25M	250-300	Y	Y	<1% error, mostly substitutions
HiSeq	5G	100-150	Y		<1% error, mostly substitutions
NextSeq	400M	100-150	Y	Y	<1% error, mostly substitutions
IonTorrent	10M	400-600	N	Y	~1-2% error, indels, homopolymers
Pacific Biosciences	~400k ~10Gbp	up to 20kbp (or more)	N		~14% error, many indels, single molecule
Oxford Nanopore	~20Gbp	up to 1Mbp? ~20kbp more likely	N	Y	~14% error, many indels, single molecule, real time, can detect DNA modifications

NOTE: Length matters for sequence reconstruction!

SEQUEL SYSTEM PERFORMANCE: GENOMIC LIBRARY

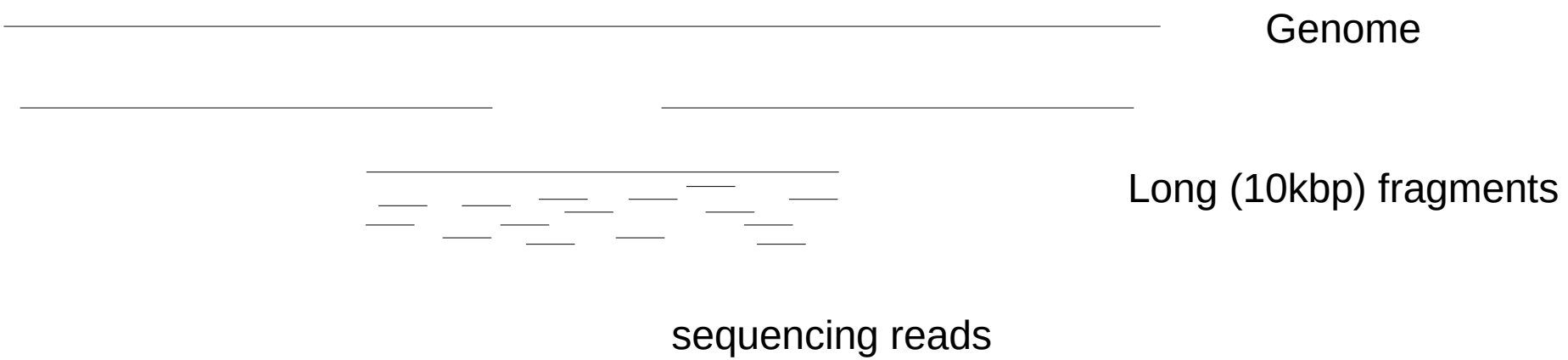
Typical Performance:

- Half of data in reads: >30,000 bp
- Output per SMRT Cell: Up to 10 Gb
- Reads per SMRT Cell: ~400,000



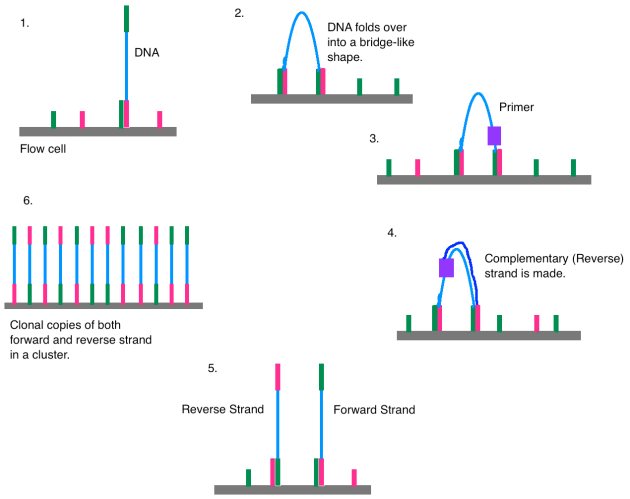
Other interesting ideas

- Sub-cloning
  - Illumina TruSeq Synthetic Long Read (TSLR)
  - 10X Genomics



Paired reads

- "artifact" of some sequencing technologies

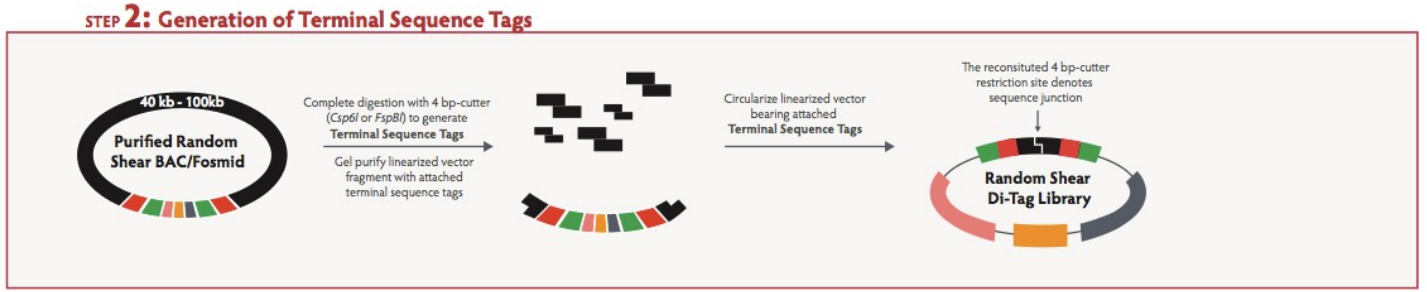


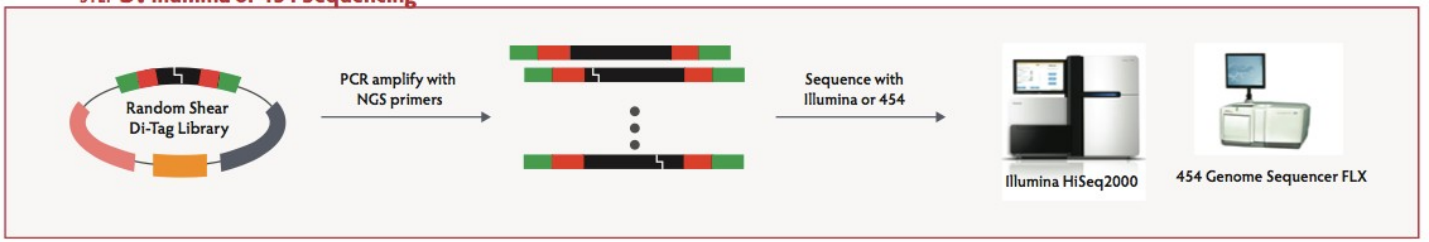

```
F: GATTACATACCAGAGGGCAGACGT
                   CGCTGACGTTGACCCAGAGGAT :R

Joined: GATTACATACCAGAGGGCTGACGTTGACCCAGAGGAT
```

Aside: quality values...

- but can also be generated





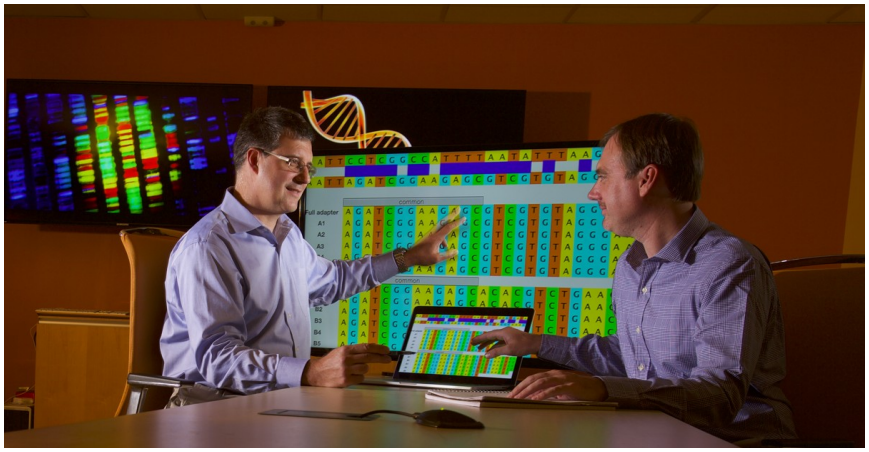
A "standard" experiment

Collect sample

Extract DNA

Sequence

Analyze



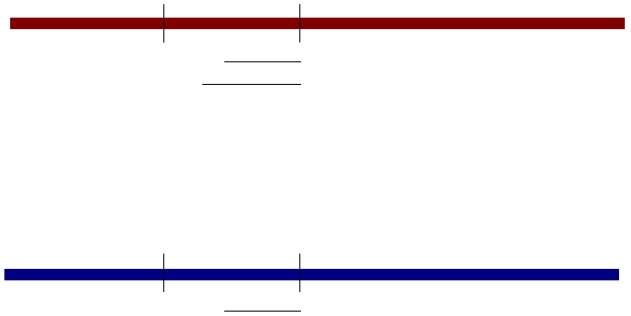
Publish

"Analyze"

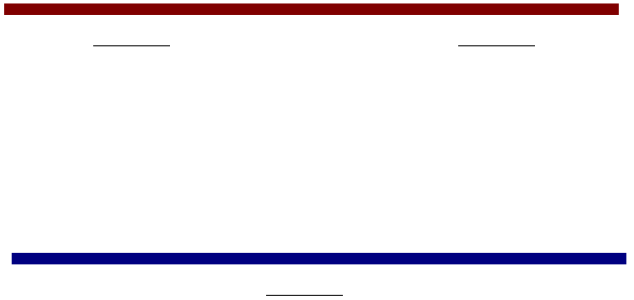
- Identify the "building blocks"
- Figure out how they relate to each other
- Figure out what they do
- Quantify them
- Understand how their abundance relates to:
  - each other
  - environmental parameters
  - time
  - disease status
  - etc.

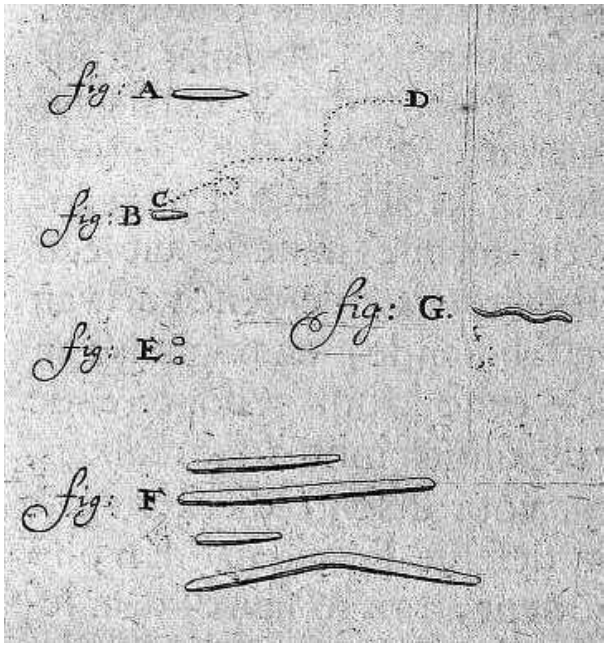
Microbiome data: same versus different

16S



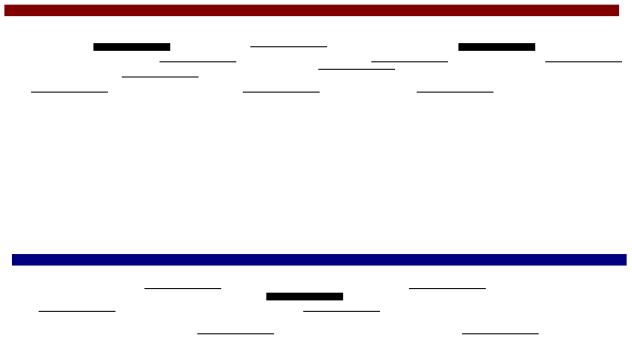
WGS





meta-genome assembly

WGS



Operational taxonomic units

- Sequencing run = ~26 million sequences
- Irrespective of number of organisms
- Which sequences come from which (unknown) organism?

CCCCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAAT
CCCCGCCTACTACCTAATCAGACGCATCCCATCCATCACCGATAATC
TTTACCCCCGCCGAACTAAACGTACTAACTACGAACGACCGACTACC
CCCCGCCTACTACCTAATCAGACGCATCCCATCCATCACCGATAATC
CCCCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAAT

O1
CCCCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAAT
CCCCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAAT

O2
CCCCGCCTACTACCTAATCAGACGCATCCCATCCATCACCGATAATC
CCCCGCCTACTACCTAATCAGACGCATCCCATCCATCACCGATAATC

O3
TTTACCCCCGCCGAACTAAACGTACTAACTACGAACGACCGACTACC

More from Tracy

Multiple alignment

```
CCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAAT
CCCCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAT
CCCCGCCTACTACCTATCAGACGCATCCCATCCATCACCGATAATC
CCCCGCCTACTACCTAATCAGACGCATCATCCATCACCGATAATC
TTTACCCGCCGAACTACGTACTAACTACGAACGACCGACTACC
```

```
CC--GCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAAT
CCCCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGAT-AAT
CCCCGCCTACTACCTA-TCAGACGCATCCCATCCATCACCGATAATC
CCCCGCCTACTACCTAATCAGACGCATC--ATCCATCACCGATAATC
TTTA--CCCGCCGAACTA--CGTACTAACTACGAACGACCGACTACC
```

Computationally difficult (takes time and it's never perfect), but important
Oligotyping tries to escape from needing to do this...

More from Tandy Warnow

Another way of thinking about it....

- Denoising or sequence inference

Given a "true" sequence

CCCCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAT

--CCGCCTACTACCTAATGGAACGCATCCCATCGTACACCGATAAAT

Can the others be derived from it with small changes?

CCCCGCCTACTACCTATCAGACGCATCCCATCCATCACCGATAATC
CCCCGCCTACTACCTAATCAGACGCATCATCCATCACCGATAATC
TTTACCCGCCGAACTACGTACTAACTACGAACGACCGACTACC

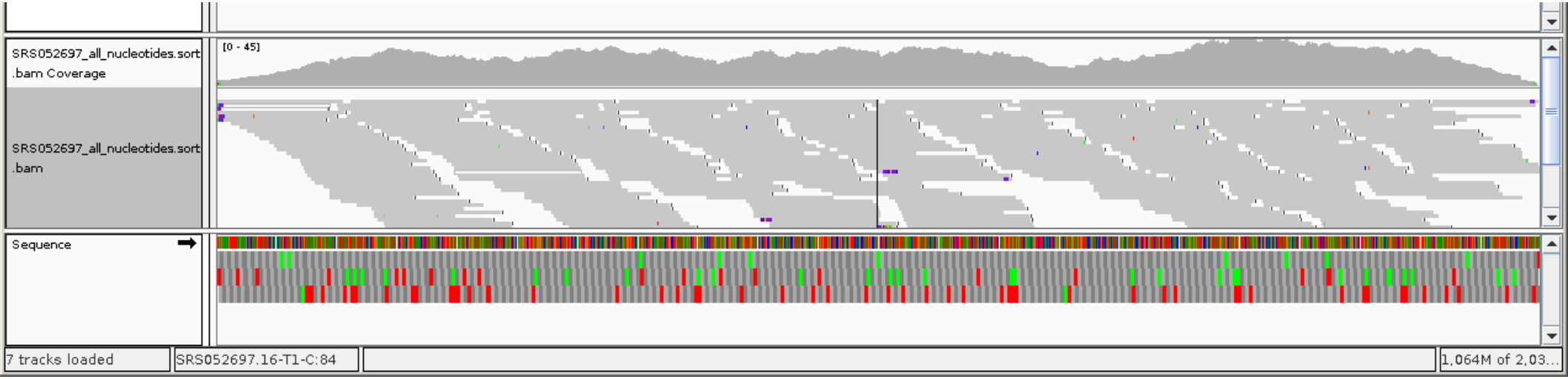
Or... What's the most likely set of sequences from which all the others have been derived through a likely path of mutations?

More from Ben Callahan, Susan Holmes

Metagenome assembly

- "stitch" the reads together to reconstruct contiguous segments from the genomes (contigs)

- unlike clustering, reads don't pile up on top of each other



- Usually structured as a graph problem

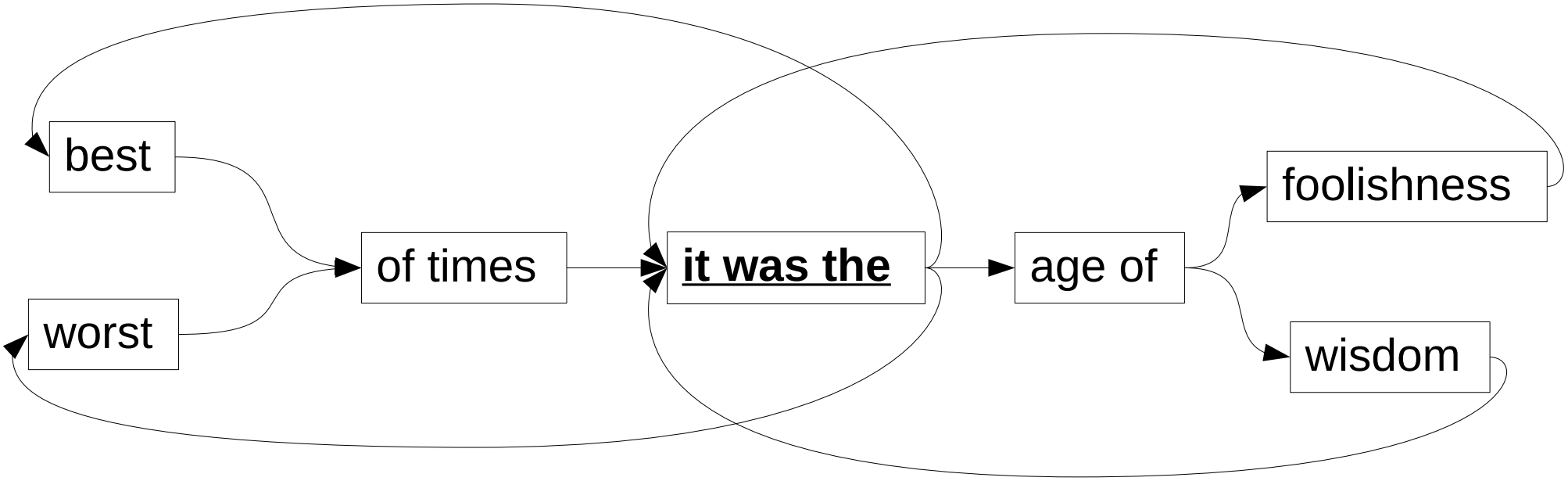
More from Titus Brown, Adina Howe, Todd Treangen

Repeats are the problem

Read = 3 “words” (<= length of repeat)

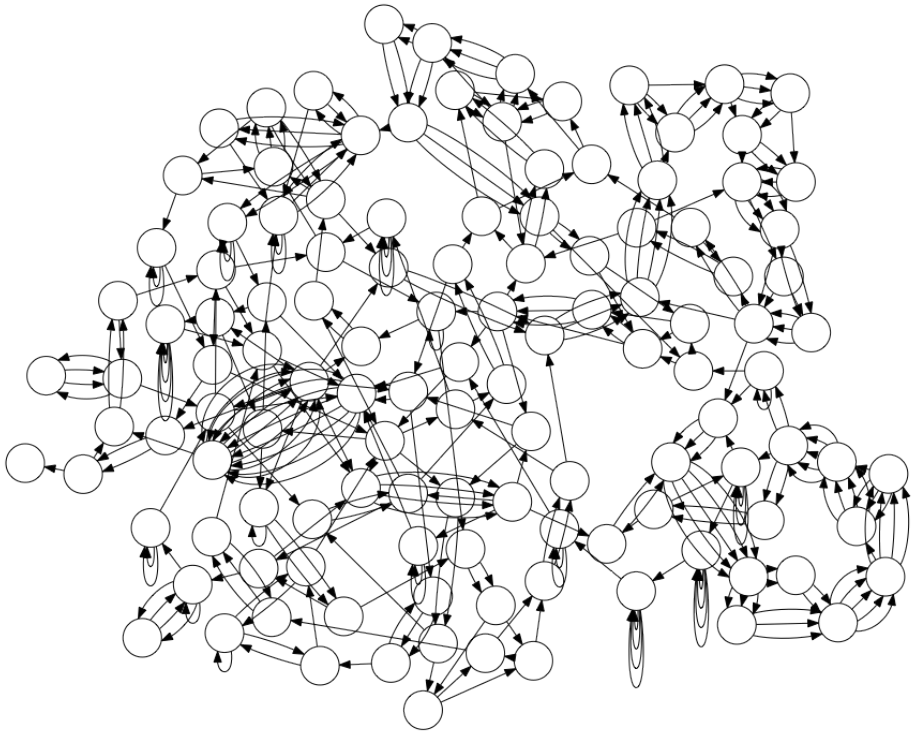
it was the best *of times* **it was the** worst *of times*
it was the *age of* wisdom **it was the** *age of* foolishness

it was the, was the best, was the worst, was the age
the age of,...

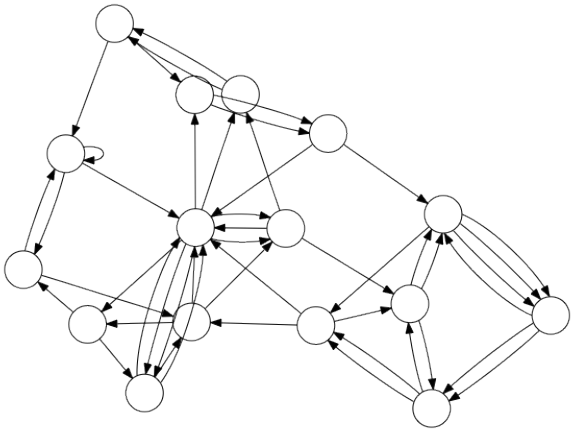


Read length matters for assembly

k = 50



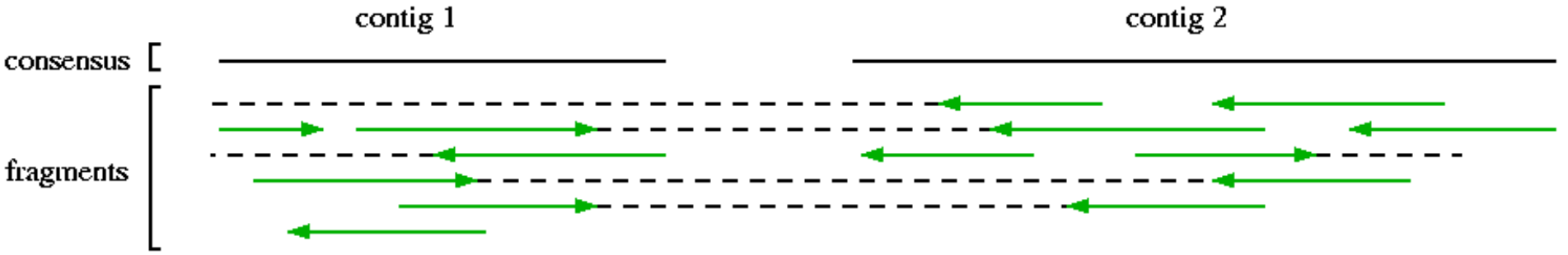
k = 1,000



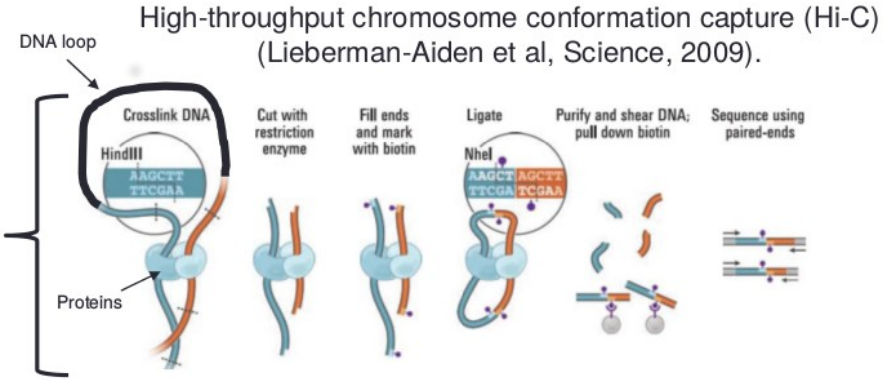
k = 5,000

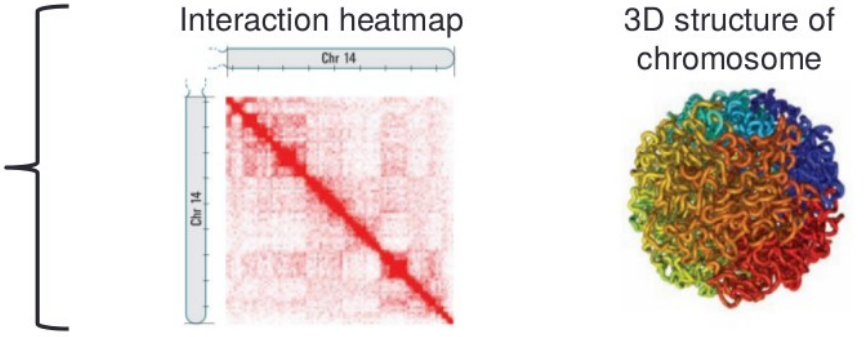
Scaffolding/using linking information

Mate-pairs, paired-ends



Hi-C





from Raphael Mourad

Basic idea:

Figure out which pieces of DNA come from the same DNA segment

Can also use optical maps, long reads, etc.

Summary so far...

Collect sample

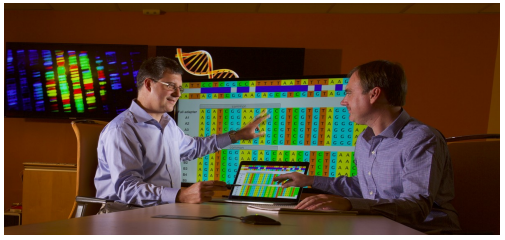
Extract DNA

Sequence

Assemble or cluster

"Bucket" of sequences OTUs or contigs

Analyze

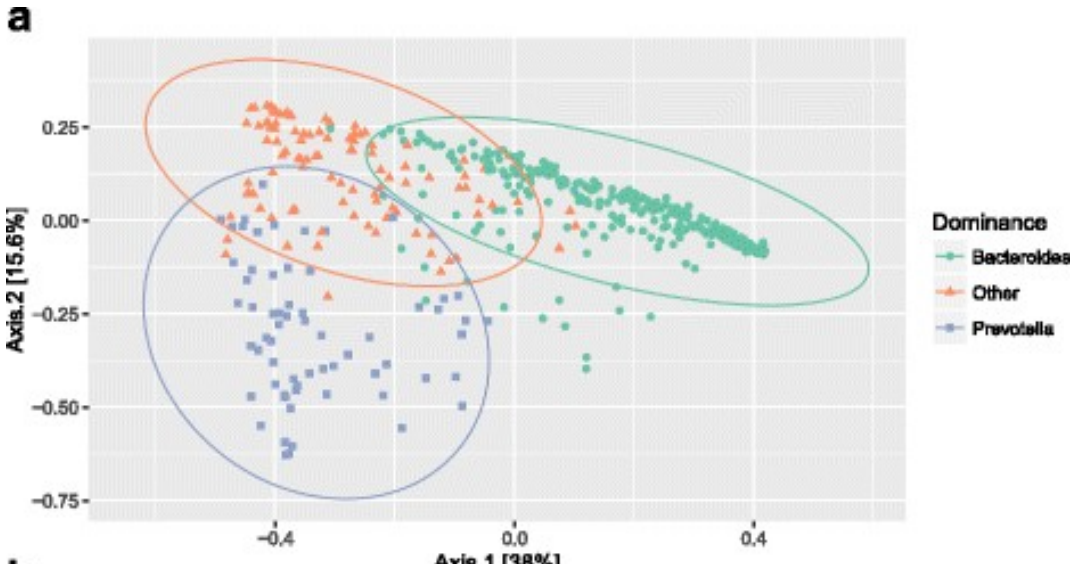


Statistical questions...

Alpha diversity

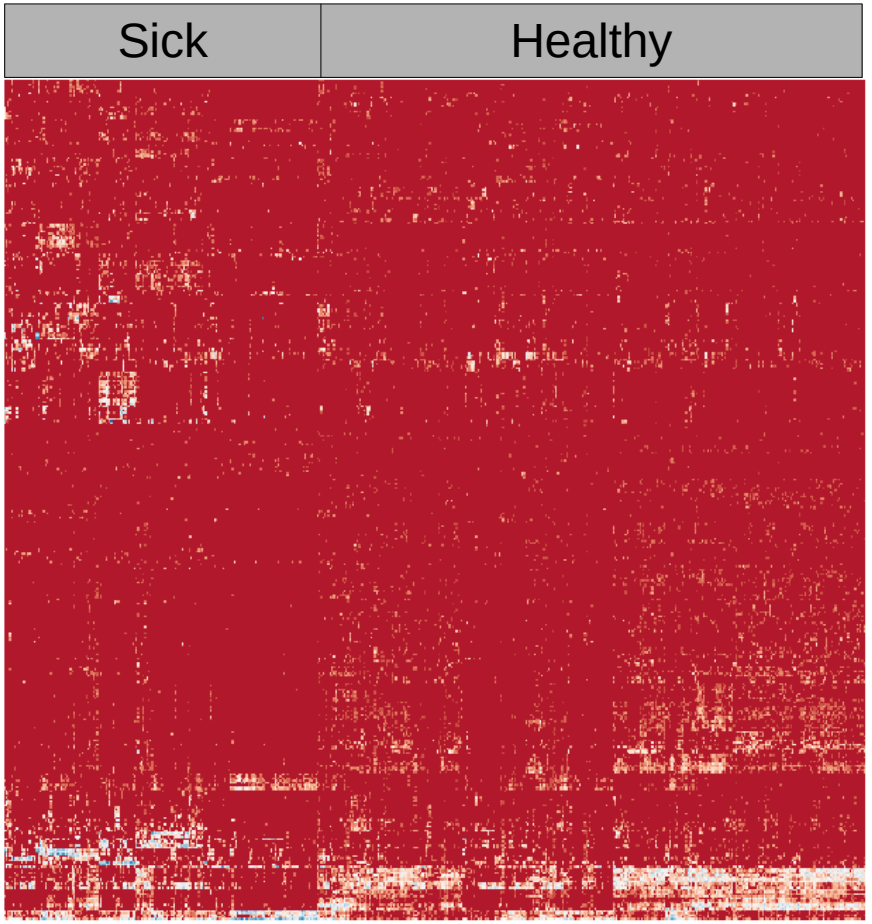
18	25	3
9	10	4
3	0	3
1	0	2
1	0	5

Beta diversity



Differential abundance

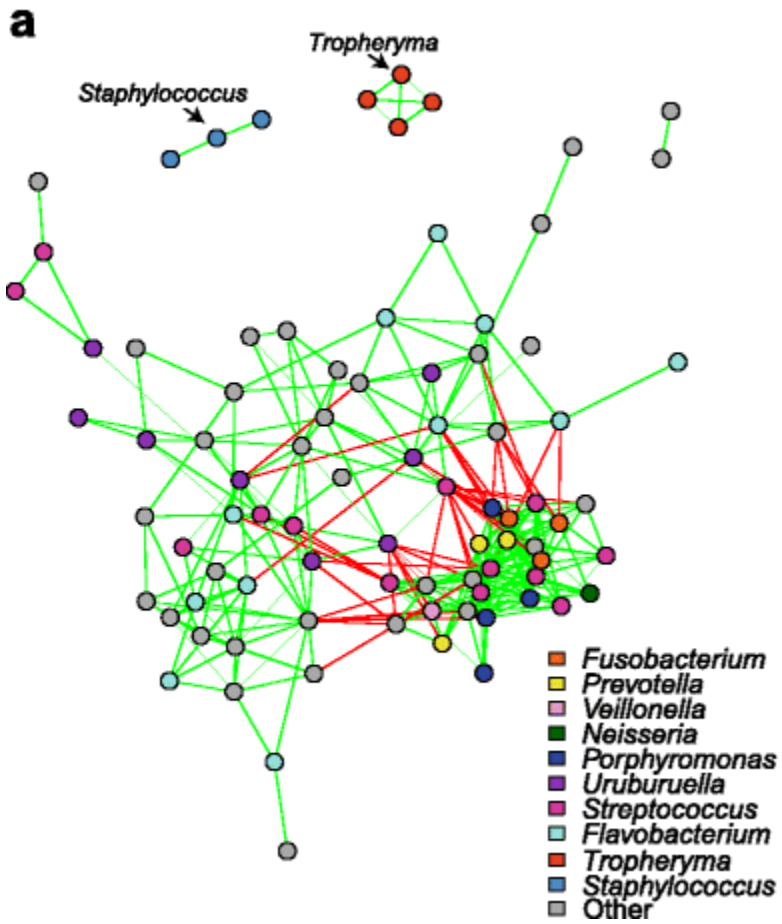
?

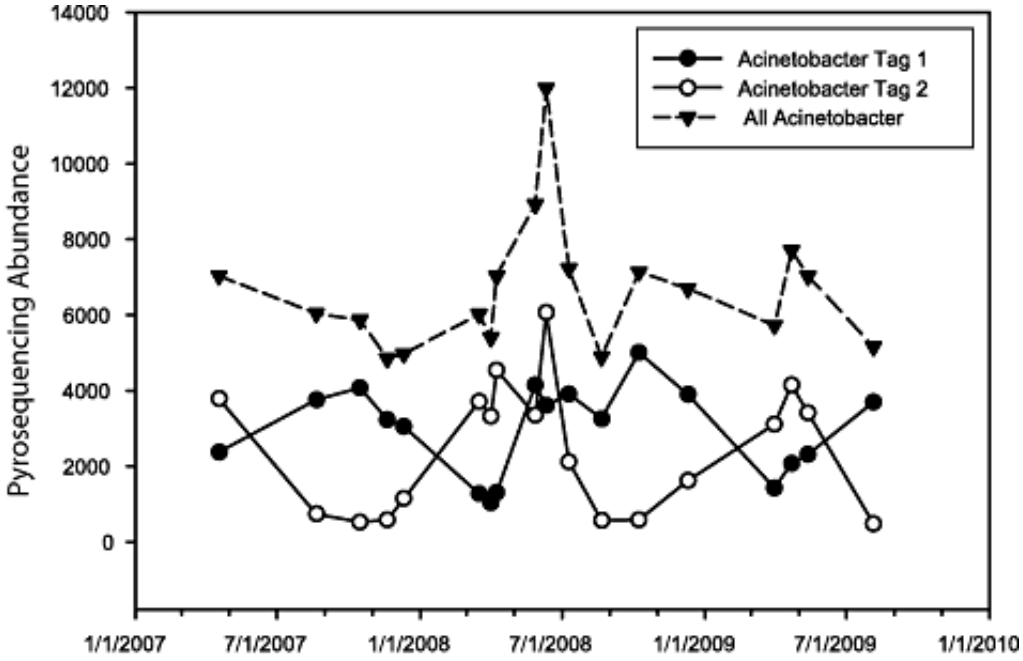


More from Amy Willis and Susan Holmes

"Statistical" interactions

Abundance as a proxy for function

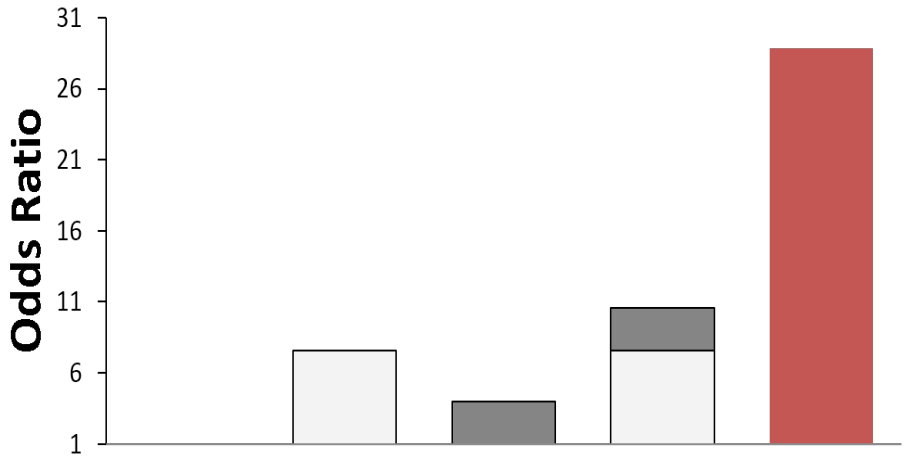


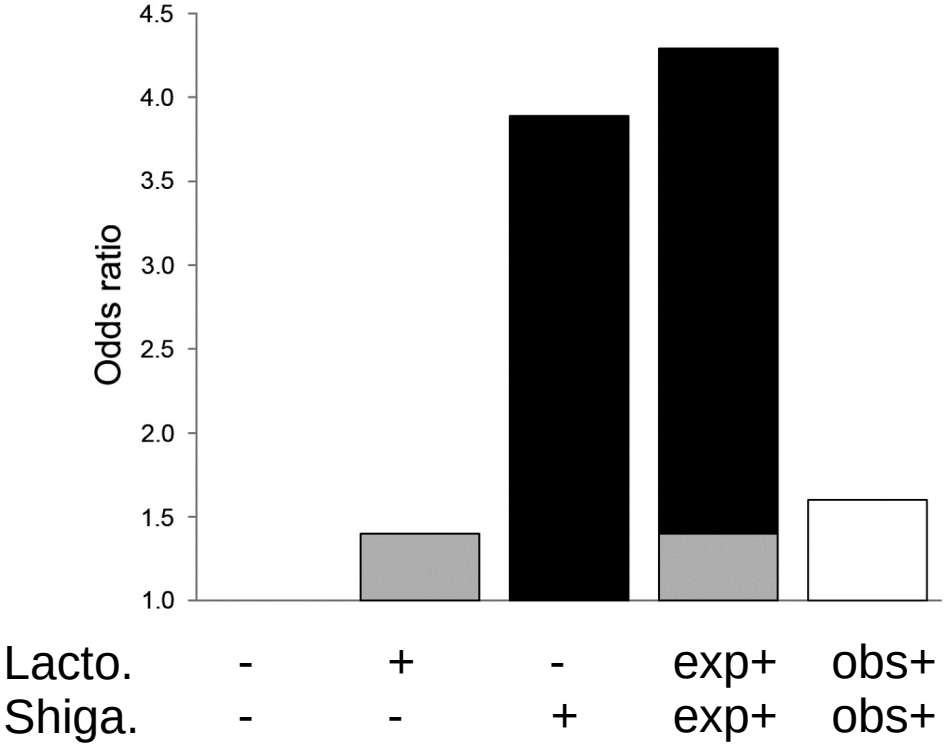


from VandeWalle et al. Environ Microbiol. 2012 Sep; 14(9): 2538–2552.

More from Christian Müller

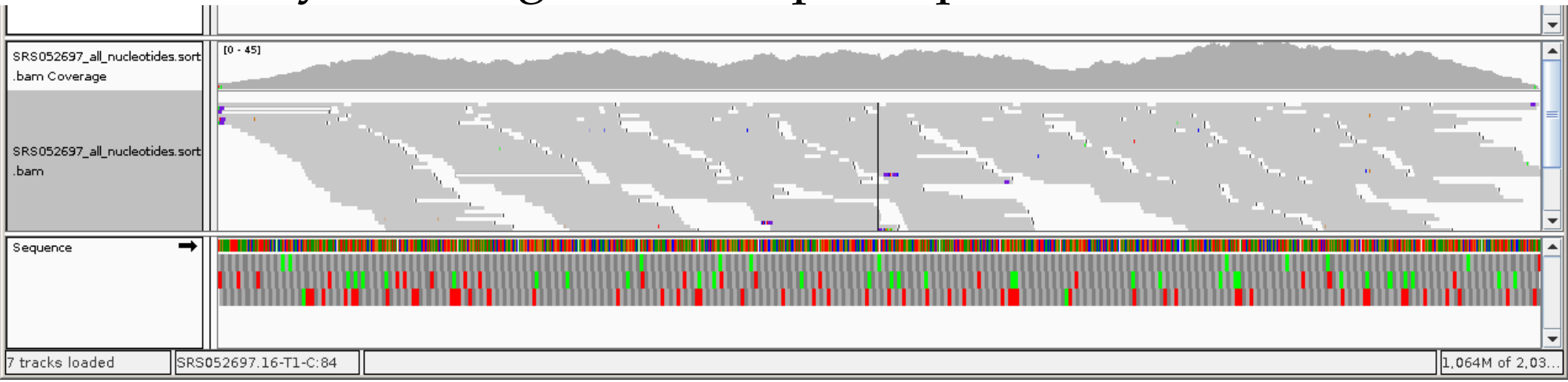
Ecological interactions





Estimating abundances

- Clusters – number of sequences in cluster
- Assembly – average/median pile-up



Normalization

Raw

	A1	A2	B1	B2
O1	24	22	11	12
O2	18	8	4	8
O3	12	10	5	6
sing	7	18	4	10
Total	61	58	24	36

Relative

	A1	A2	B1	B2
O1	.39	.37	.45	.33
O2	.29	.13	.16	.22
O3	.19	.17	.20	.16
sing	.11	.31	.16	.27
	1	1	1	1

Rarified

	A1	A2	B1	B2
O1	8	7	9	7
O2	6	3	3	4
O3	4	3	4	3
sing	2	7	4	6
Total	20	20	20	20

Ask Amy Willis
and Susan Holmes

Aside...dealing with confounding

- For every observation

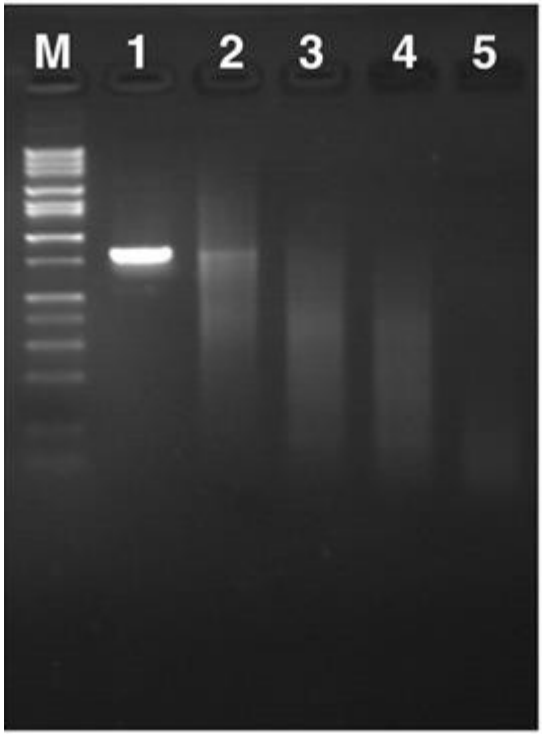
- Find non-biological reason for the behavior of the data
  - uneven amount of sequencing
  - experimental biases/batch effects
  - edge effects
  - etc., etc.
- Only if all alternative explanations fail, you may have something

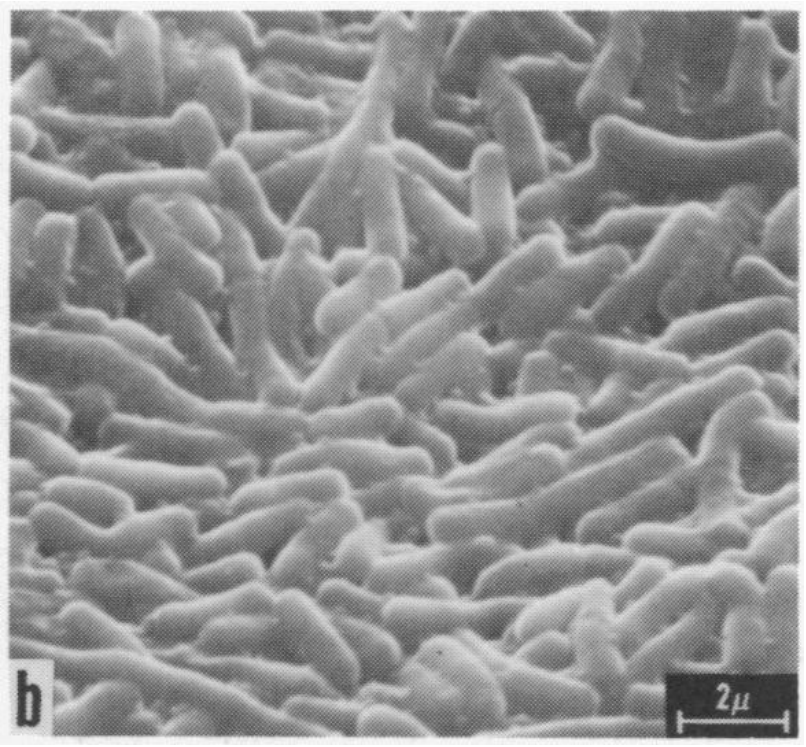
O1:	0	1	0	0	0	1	1	0	0	1	1	1	0
O2:	1	0	1	1	1	0	0	1	1	0	0	0	1

Strong negative "interaction" - Are O1 and O2 competitive organisms?

Other explanations?

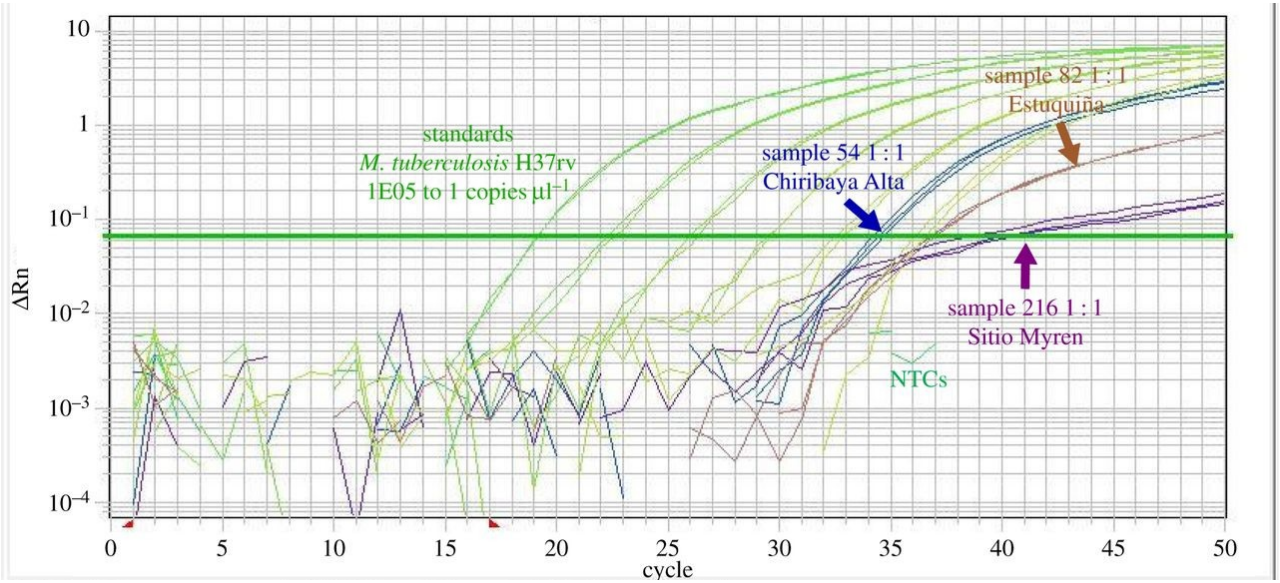
More from Amy Willis, Susan Holmes, Tandy Warnow



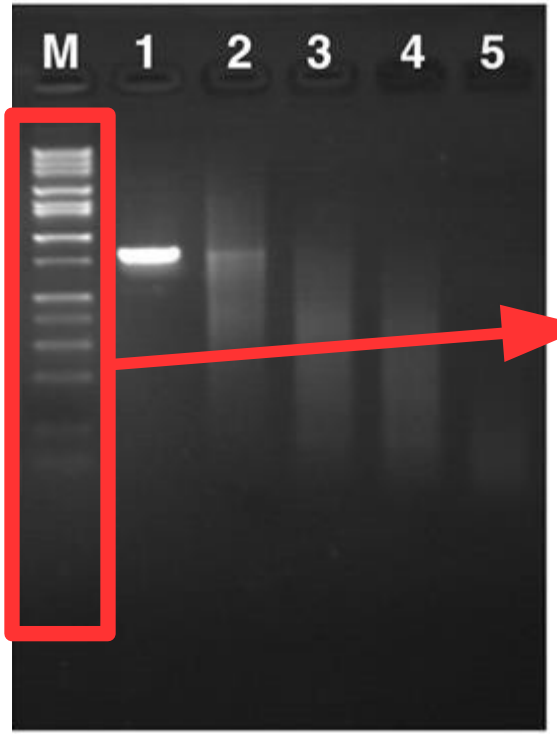


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC379936/

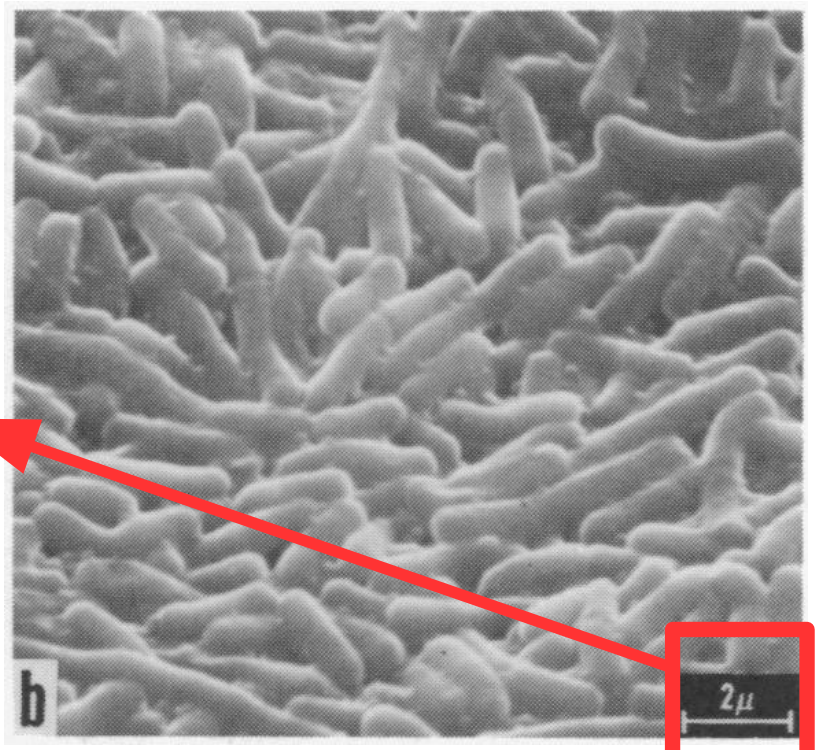
https://academic.oup.com/nar/article/30/24/e139/1077910/Random-DNA-fragmentation-with-endonuclease-V



http://rstb.royalsocietypublishing.org/content/370/1660/20130622

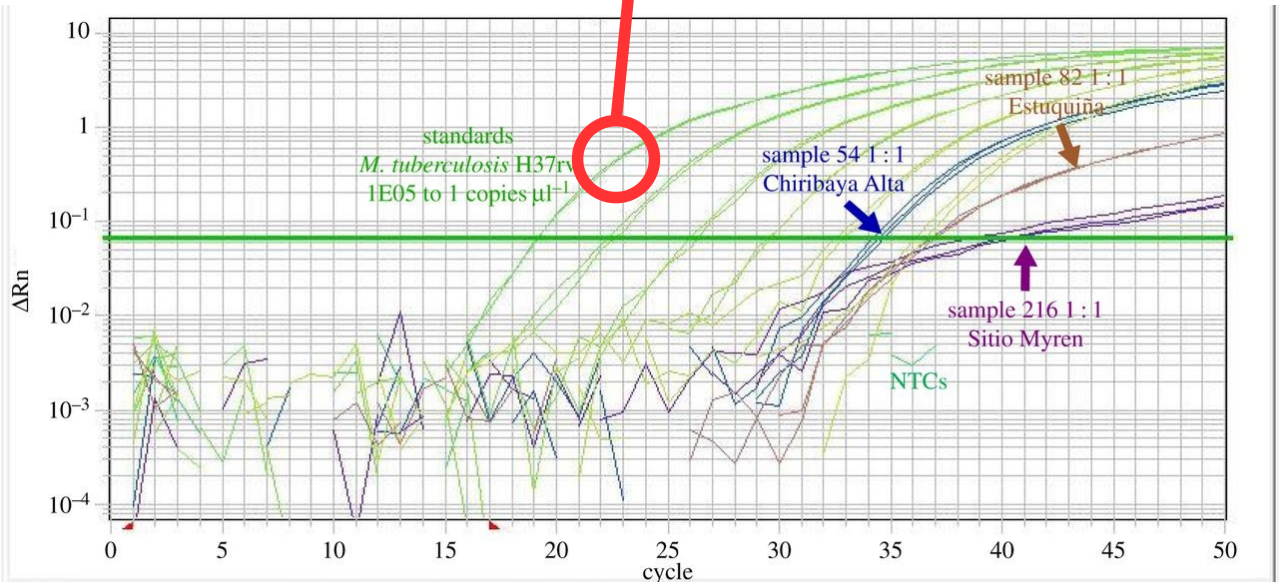


STANDARDS



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC379936/

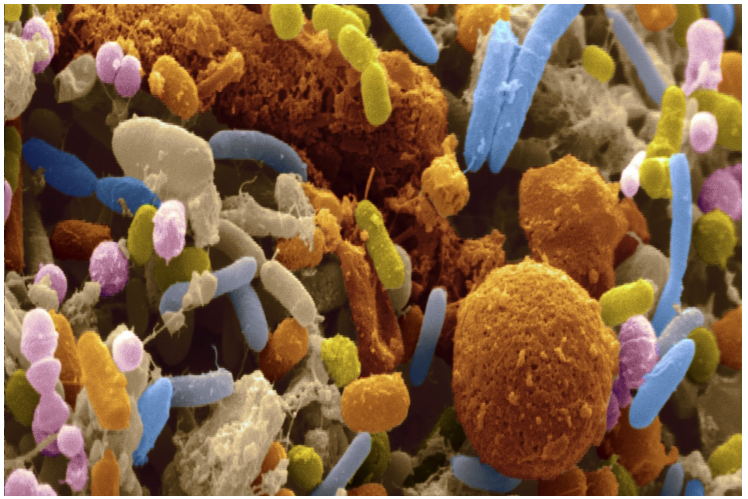
https://academic.oup.com/nar/article/30/24/e139/1077910/Random-DNA-fragmentation-with-endonuclease-V



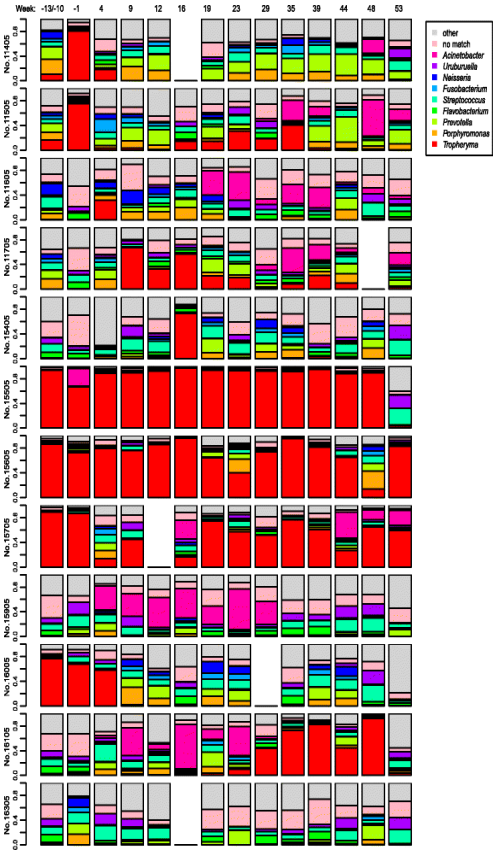
http://rstb.royalsocietypublishing.org/content/370/1660/20130622

Standards in metagenomics

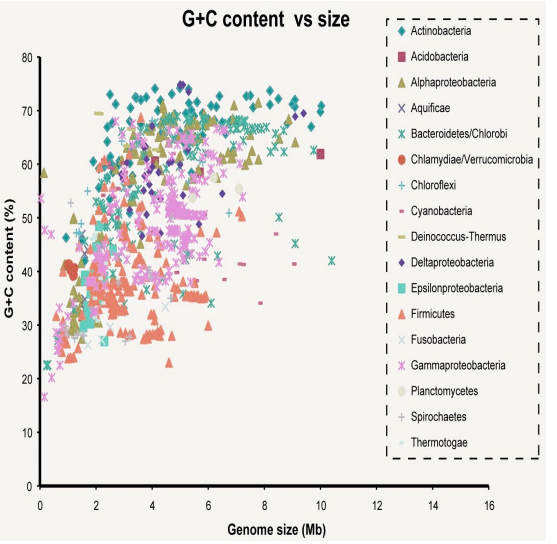
- What is a standard microbe?
- What is a standard microbial community?
- What is a standard measurement of abundance?



from themocracy.com

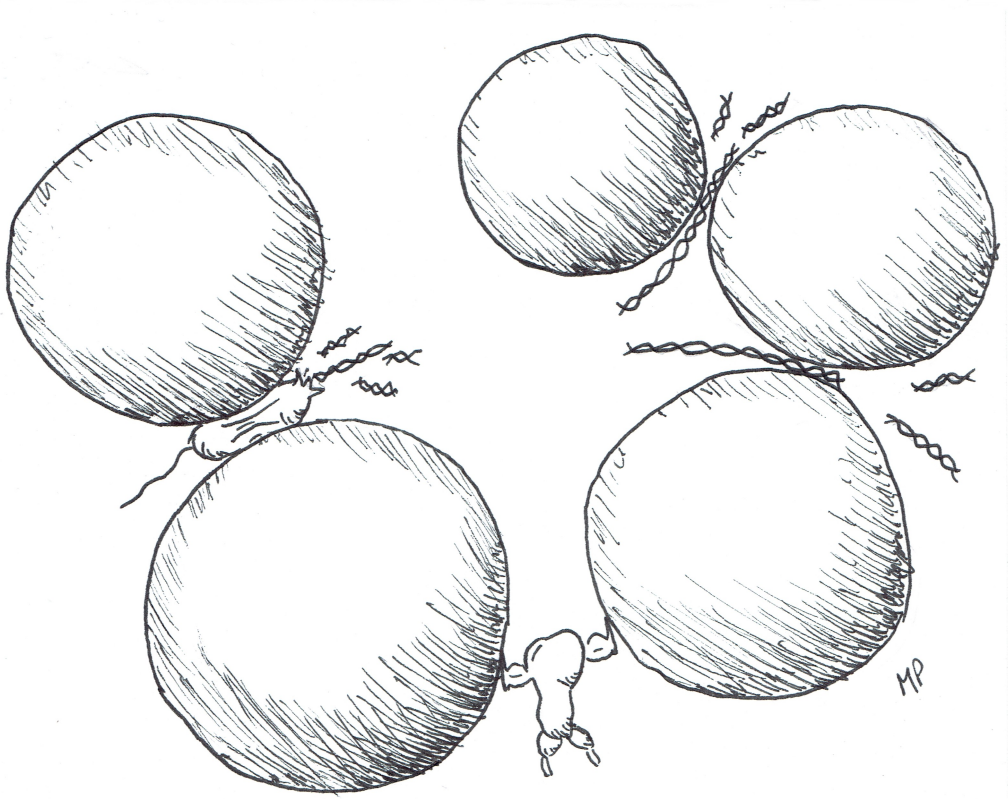


Morris et al. Microbiome 2016

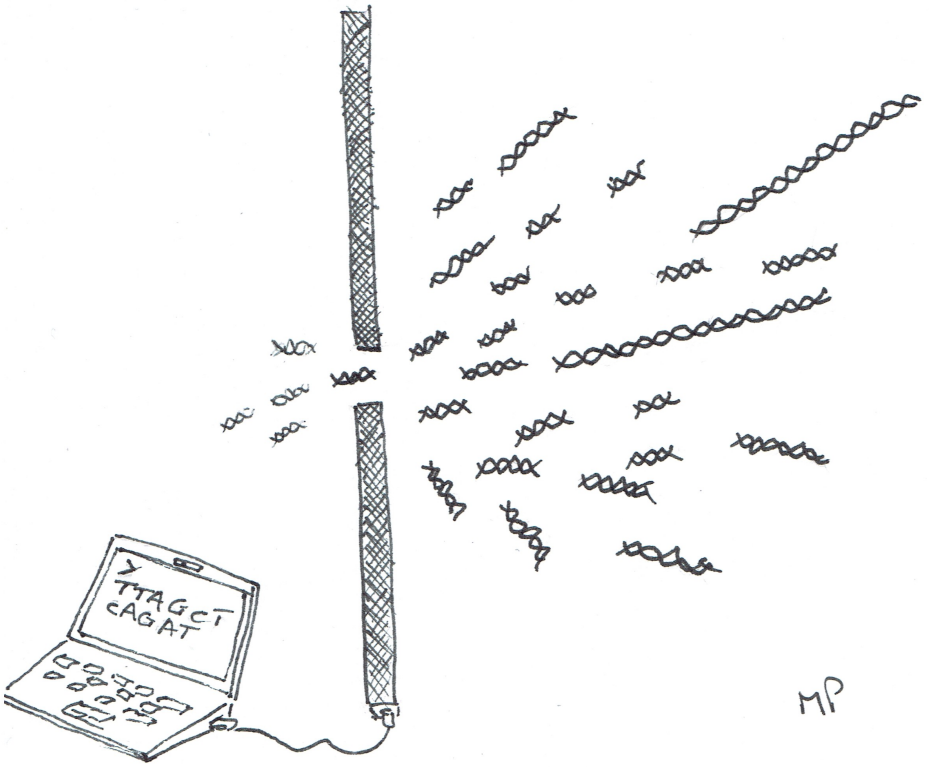


from asserttrue.blogspot.com

Biased cell lysis



Biased sequencing



Everything is biased

- Absolute abundances are impossible to determine
- Measuring relative abundances is biased as well
- Every step in the process leads to errors
- Including bioinformatics (your computer is **not** smarter than a 5th grader)

```
x = 0.3
y = 0.1 + 0.1 + 0.1

if (x == y) :
  print "Hello"
```

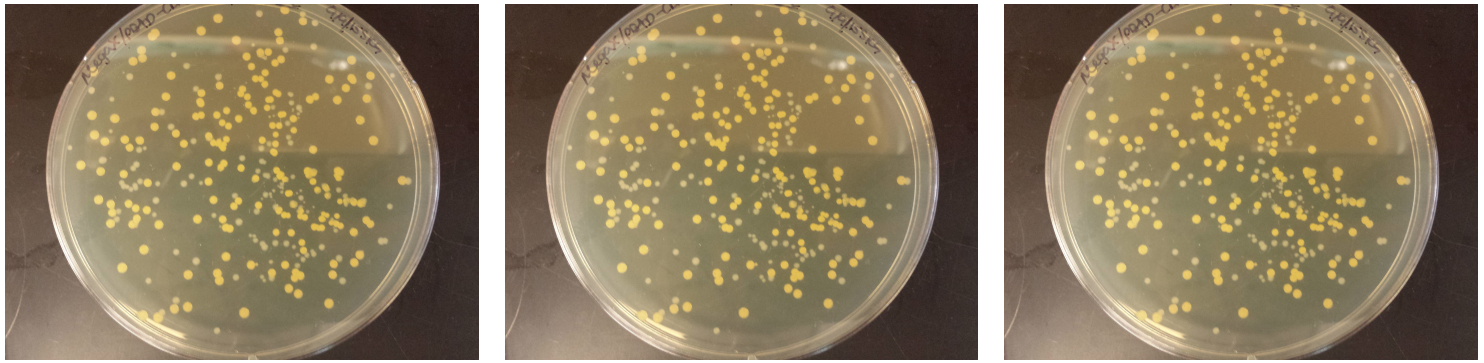
- EXPERIMENTAL DESIGN IS KEY!
- SO IS VALIDATION!

What is your question?

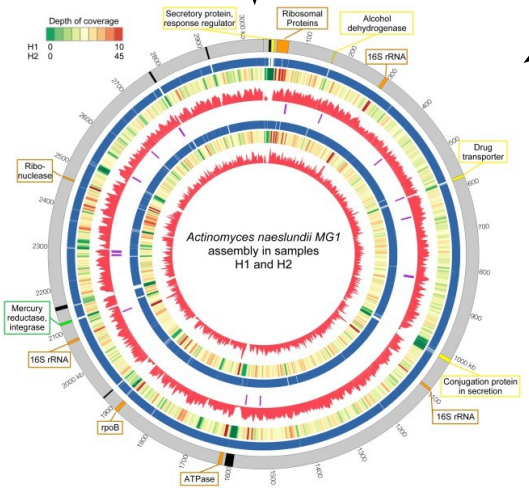
- Data generation goes hand in hand with the question you ask!

- Exploratory analysis of previously collected data (usually) leads to hypotheses, not publishable results.

n * metagenomes > genomes * n

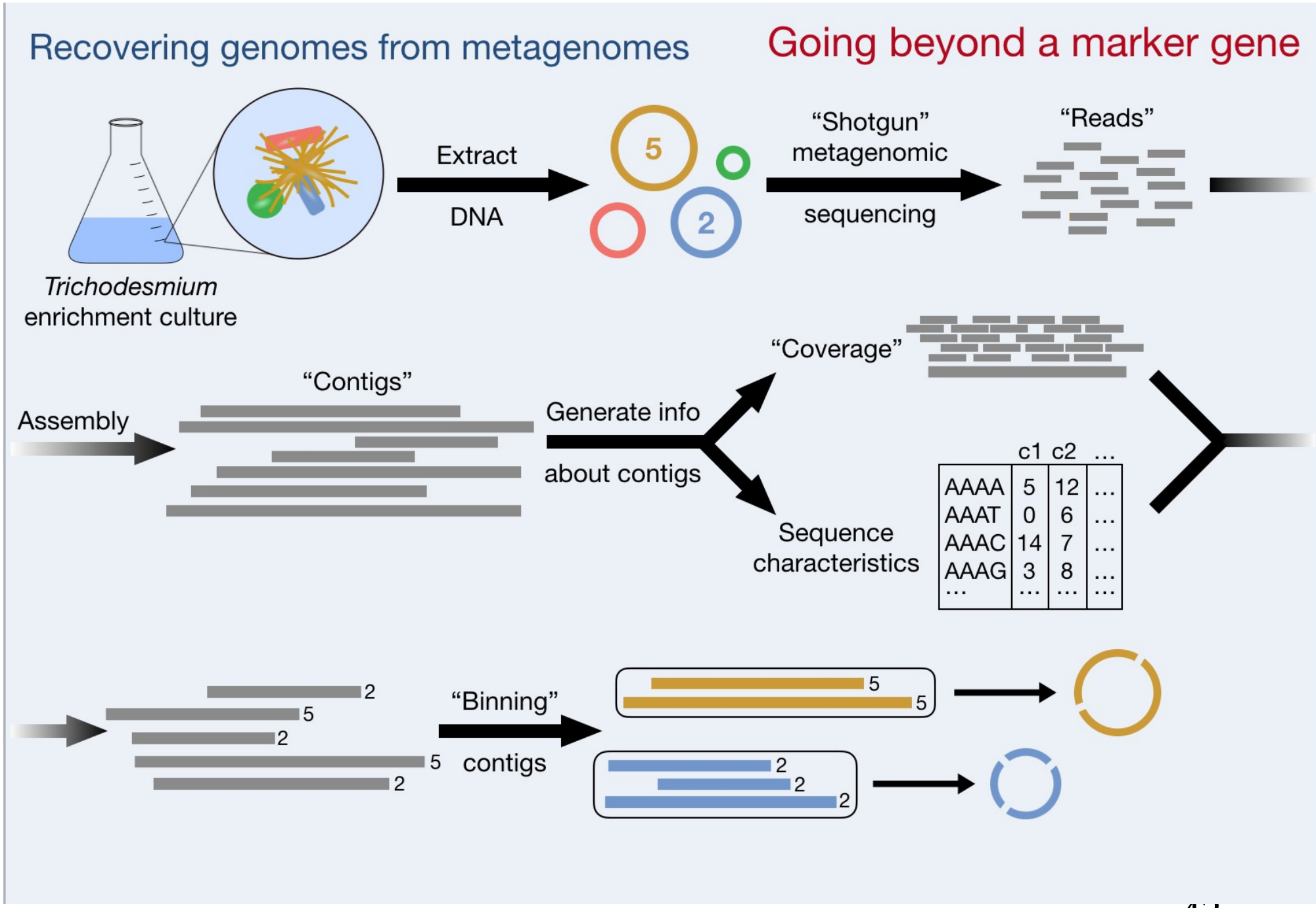


...

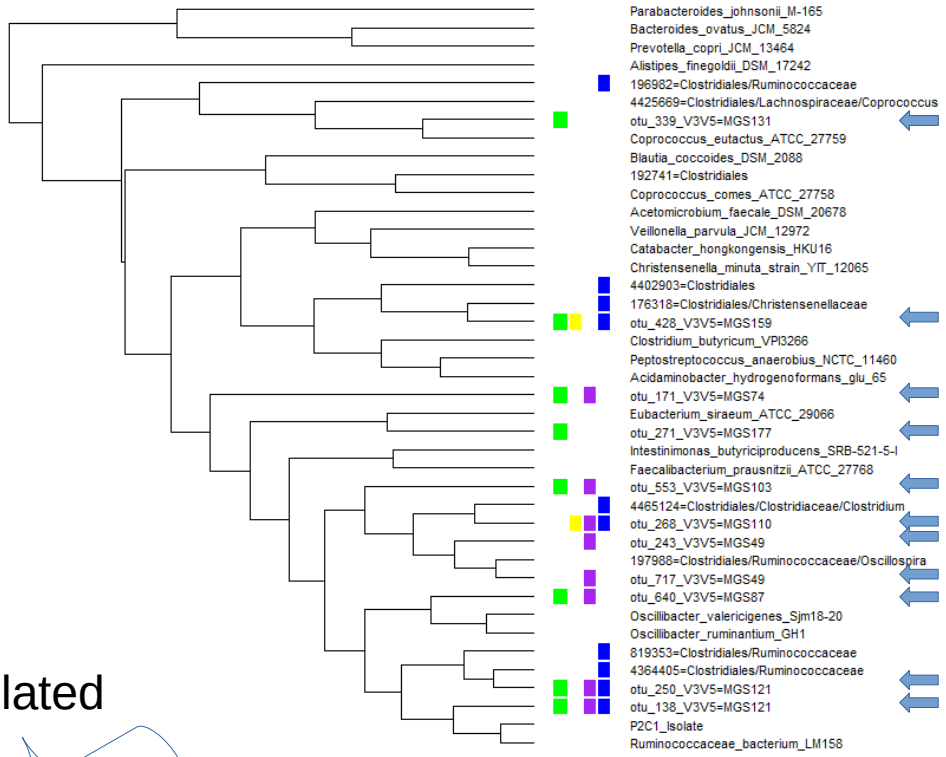


Binning: correlation across samples

Big picture overview...

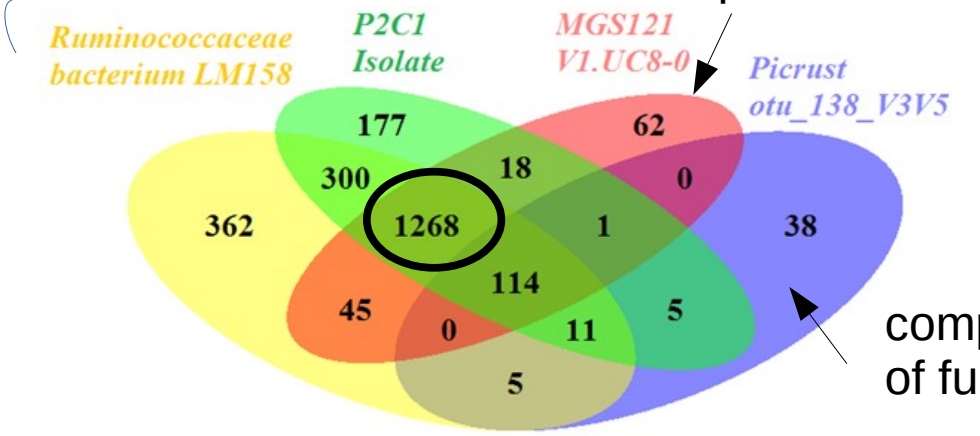


Associating most-wanted OTUs with genes



laboriously isolated genomes

our prediction based on correlation



More from Rob Knight on PiCrust ?

computational prediction of function

Almeida et al. ISME J. 2016

What does it all mean?

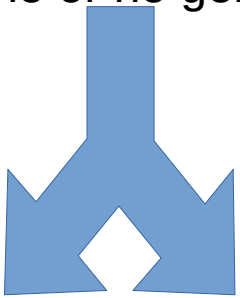
Functional annotation

ATTAGATGGTATTGAGACCACTGGCACAAAGATAATTTGTAC

ATTAG ATG GTA TTG AGA CCA CTG GCA CAA AGA TAA TTTGTAC

fancy machine learning algorithms

Gene or no gene?



more fancy machine learning

Database searches

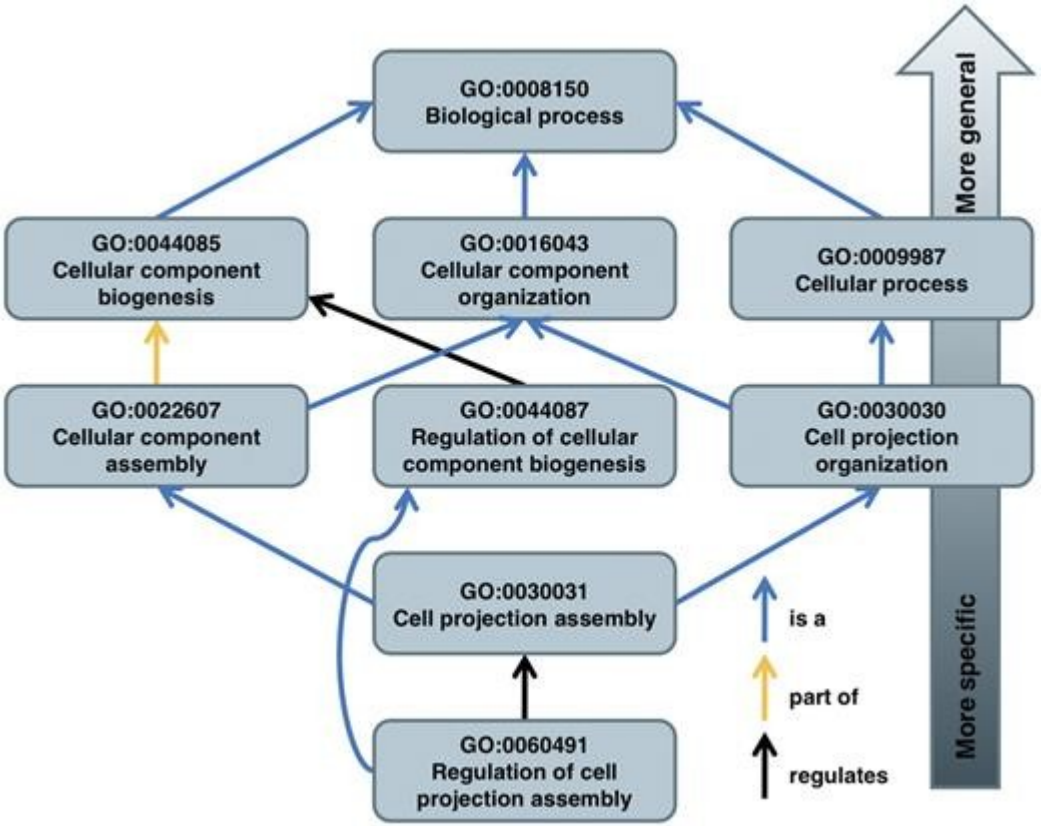
Does it look like a known class of proteins?

Does it look similar to something out there?

More from Curtis Huttenhower

More on functional annotation

- Can be done on genes and on fragments
- Controlled vocabulary annotations - gene ontology
  - Why? Because computers are stupid



Special types

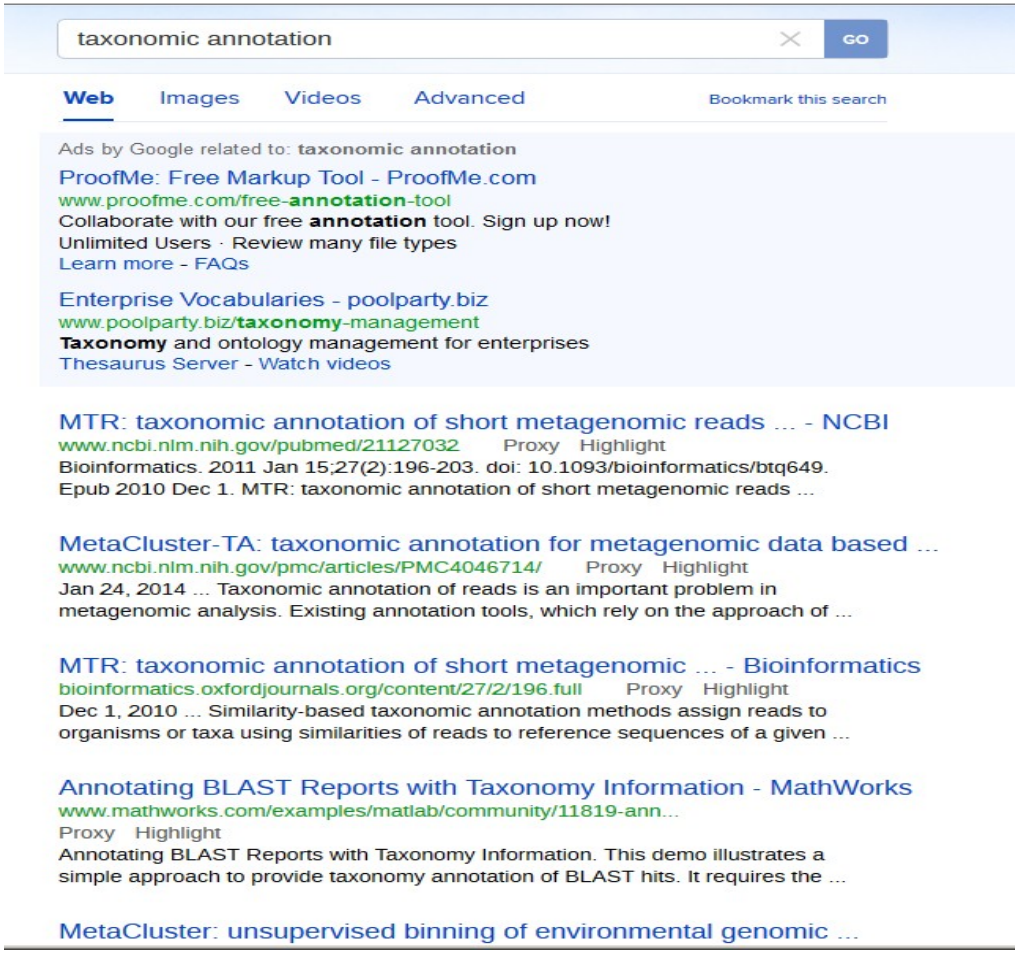
- Antibiotic resistance
  - see CARD database at McMaster University (Andrew McArthur)
- Virulence factors
  - see VFDB
- etc. etc.

Ask Yana Bromberg, Todd Treangen and Curtis Huttenhower

Taxonomic annotation

Google: "taxonomic annotation"

- Database of known pages
- Report all that contain keyword



- Ranking important (which of the thousands is most relevant)

The solution

- Organize the database (taxonomy)

Kingdom;Phylum;Class;Order;Family;Genus;Species;Strain

(and levels in between?)

- Use search procedure that can generalize from existing knowledge

BLAST top hit

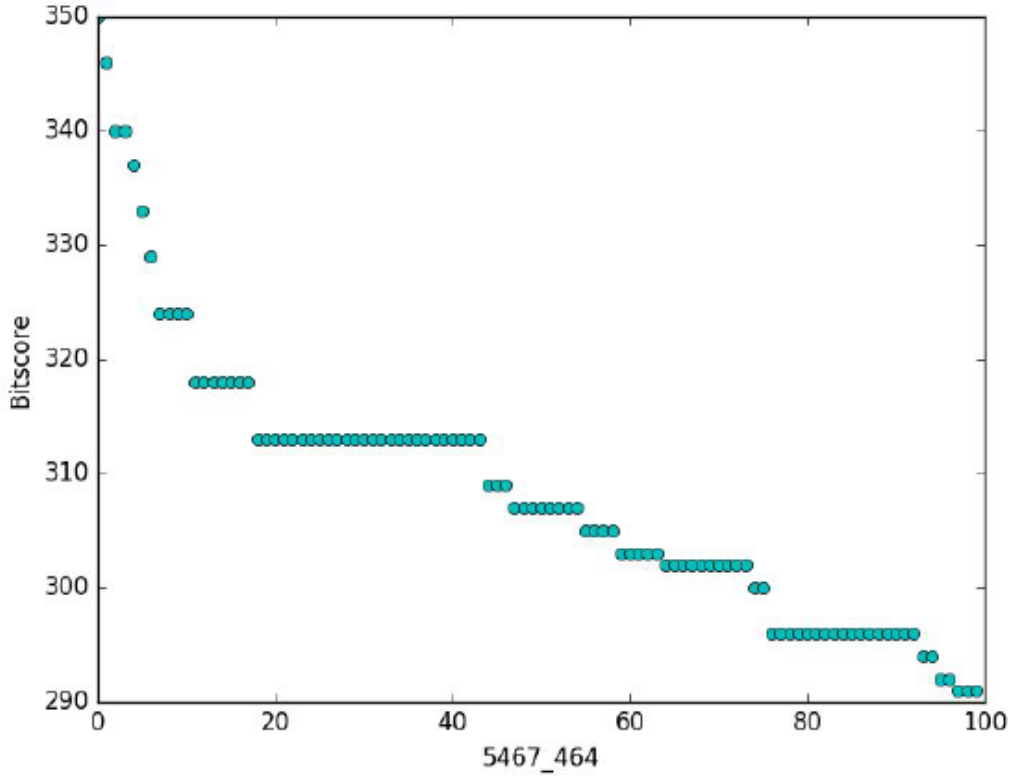
5467_464 HM038000.1.1446 E-value: 6e-96 Bit score: 350

Bacteria;Cyanobacteria;Melainabacteria;Vampirovibrionales;Vampirovibrio chlorellavorus

bit score – the "information" contained in the alignment
E-value – how many random alignments one expects for the same bit score

BLAST...more hits

5467_464 HM038000.1.1446 Identity: 80.00% E-value: 6e-96 Bit score: 350



top 100 hits sorted by bit score

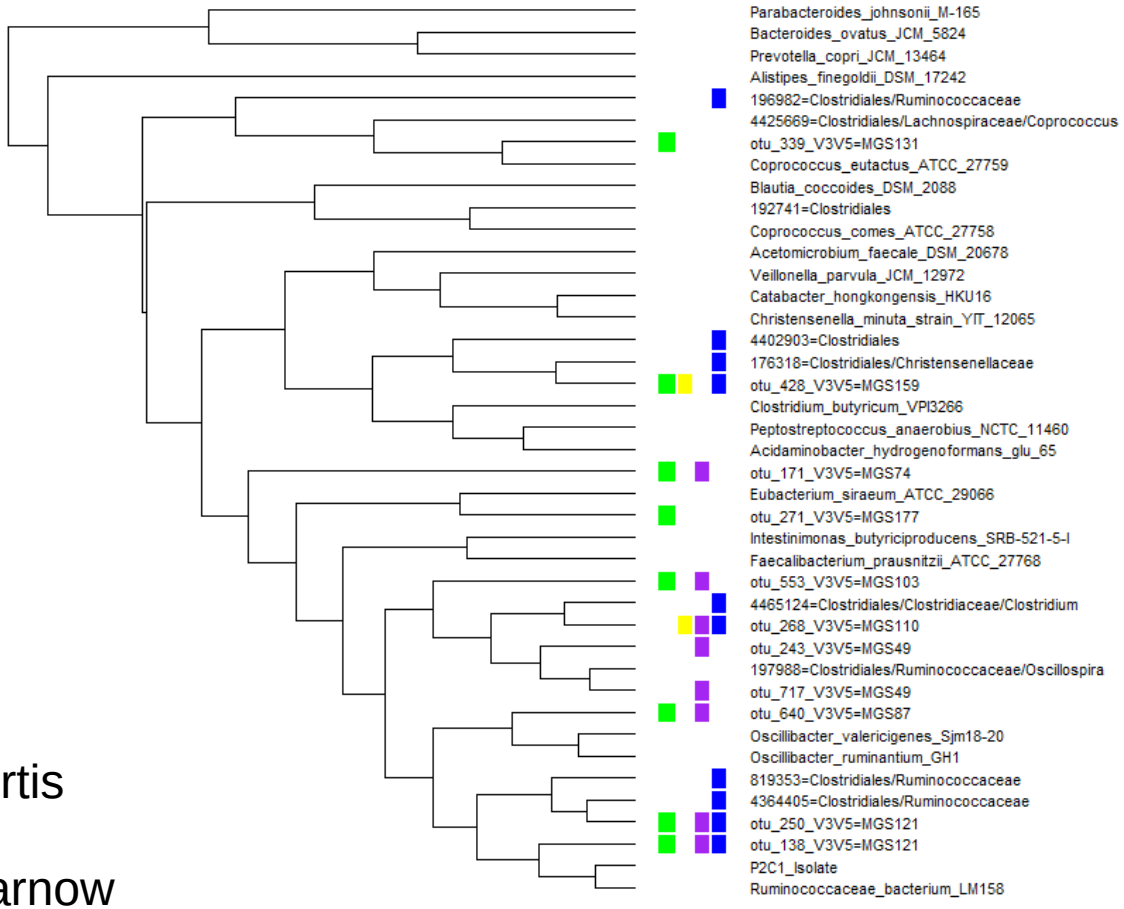
Bacteria;Cyanobacteria;Melainabacteria;Vampirovibrionales;Vampirovibrio chlorellavorus
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas;
Brevundimonas mediterranea

Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas;
Brevundimonas bacteroides

Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Butyricicoccus;Butyricicoccus pullicaecorum

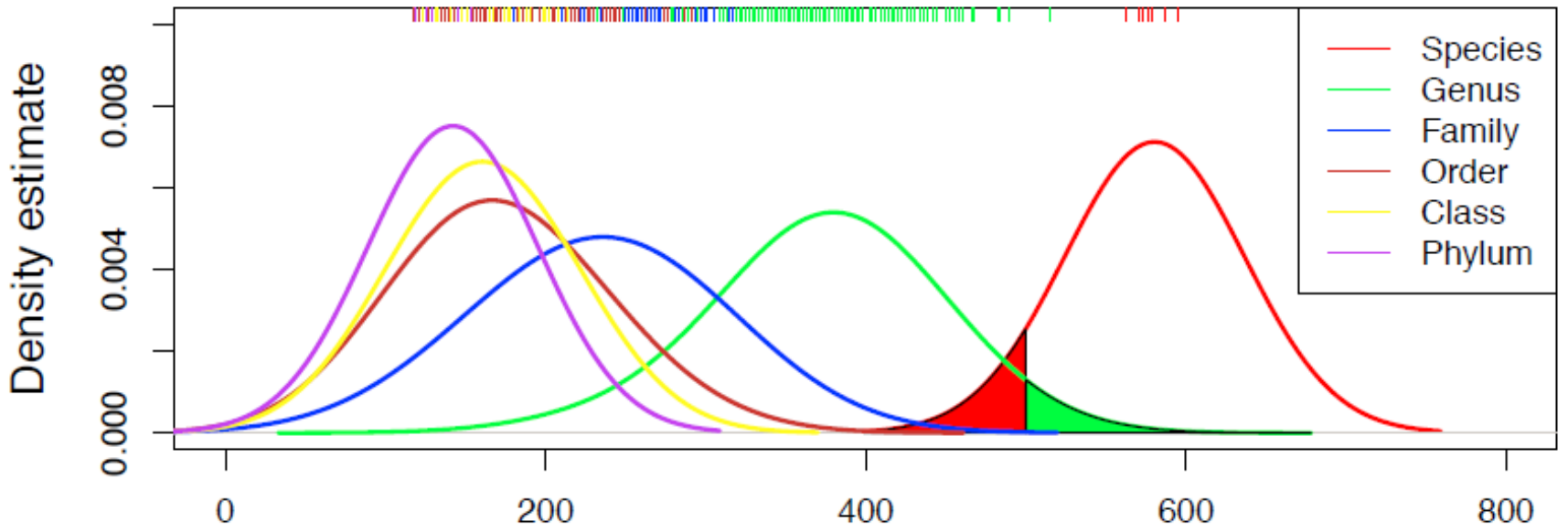
To generalize you need to make assumptions

- Taxonomically related sequences are similar – sequence similarity searches
  - Blast
  - Metaphyler
  - Megan
  - Usearch, etc.
- Taxonomically related sequences have similar k-mer frequencies – k-mer spectrum methods
  - Tetra
  - RDP classifier
  - kraken
- Simulate evolution – phylogenetic methods
  - pplacer
  - TIPP
  - mOTU



More from Curtis Huttenhower and Tandy Warnow

The classification problem

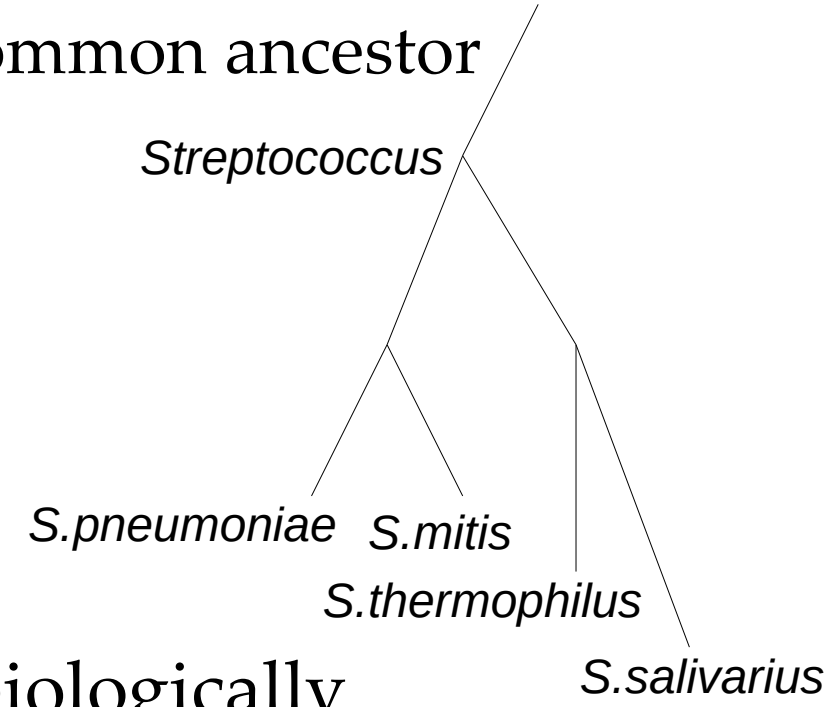


- Identify what is common to sequences with the same label and different from sequences with a different label
- Important: labels must be consistent with the definition of similarity (operational definition of taxonomy not necessarily consistent with "true" taxonomy)

More from Curtis Huttenhower

Some caveats

- Dealing with ambiguities
  - common solution: most recent common ancestor



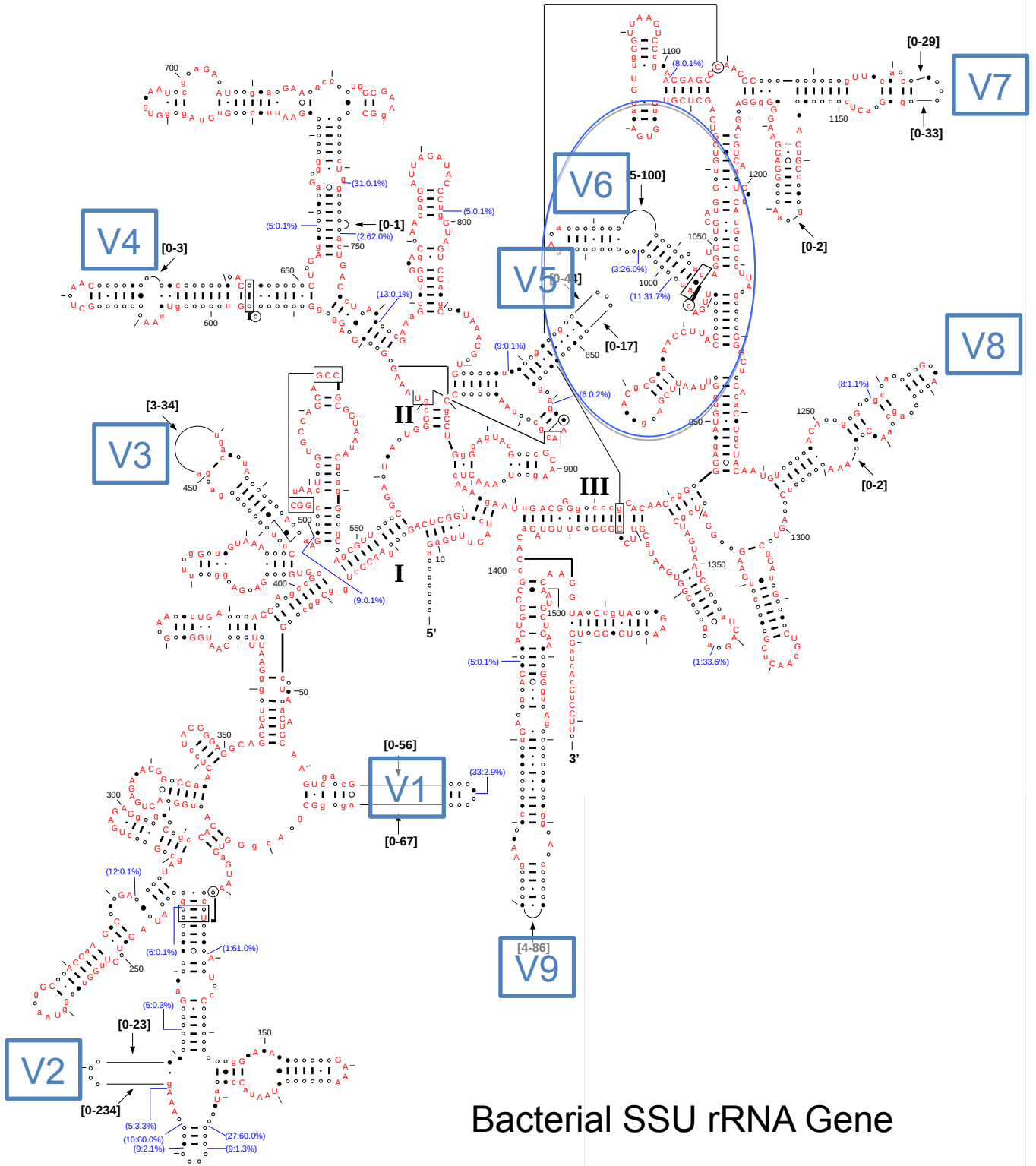
- Discriminant sequences may be biologically meaningless
  - e.g., phage protein used as discriminator in MetaPhlAn

- Database incompleteness leads to errors

Need for marker genes...

- Compare apples to apples
- Avoid lateral gene transfer
- Can (in theory) extrapolate based on evolutionary principles
- Can target experimentally ($s matter)

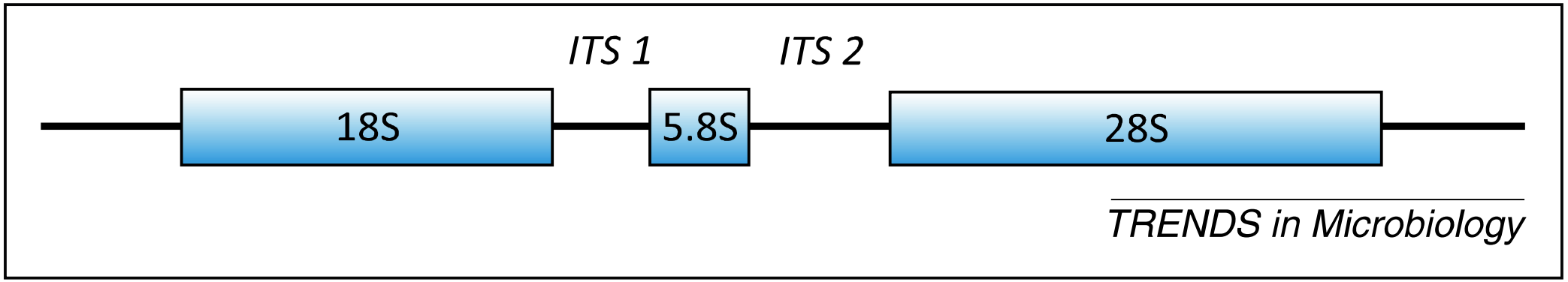
- Some choices:
  - Segments of the ribosomal RNA (16S, 18S, ITS, etc.)
  - Housekeeping genes (rpoA, rpoB, ribosomal proteins, tRNA synthetases, elongation factors, etc.)



Bacterial SSU rRNA Gene

from Sue Huse

Internal Transcribed Spacer (ITS) often used for fungal taxonomy



from Sue Huse

Hufnagle and Noverr, 2013

“Official” Taxonomic Names

- Bergey’s Taxonomic Outline – manual of taxonomic names for bacteria
- List of Prokaryotic names with Standing in the Nomenclature (vetting process)
- **BASED ON CULTURE!**
- NCBI – similar taxonomy, but multiple “subs” (subclass, suborder, subfamily, tribe)
- Fungi – UNITE curated database
- Archaea – a work in progress

from Sue Huse

Runtime

k-mer counting
(e.g., kraken)

k-mer stat
model
(e.g., RDP)

sequence alignment
(blast, usearch, diamond)

phylogenetic

Similarity search

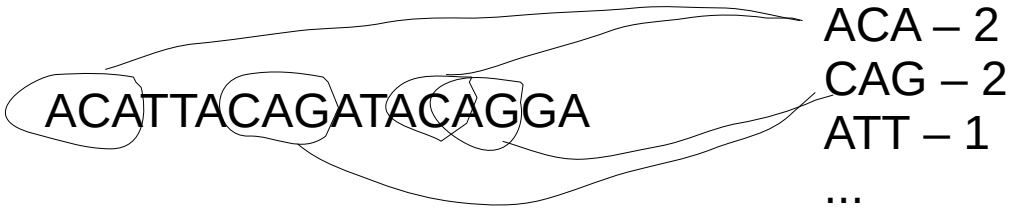
```
Query   ACCATAG-GCCGTCAGACCTAGACTAGA
DB      AC-ATAGAGCCGTCAGA-CTATACTAGA
```

- Finds exact matches
- Handles sequencing errors
- May handle evolutionary divergence
- May provide statistical guarantees (is this a random hit?)

- MANY tools exist for doing the search!
- Differ by
  - assumptions about data
  - similarity cutoff
  - heuristics to speed up search (incl. memory/speed trade-off)

RDP classifier, PhymmBl, NBC, etc.

- Like Metaphyler but...
- instead of alignment, use k-mer spectra

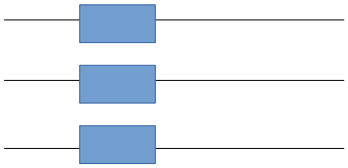


- Different statistical frameworks for comparing count distributions
  - PhymmBL – interpolated Markov Models
  - RDP, NBC – Bayesian classifiers
  - ...

MetaPhlAn, Kraken

- Find segments common to a group of organisms but not found in other organisms
- Database coverage/completeness critical for good accuracy

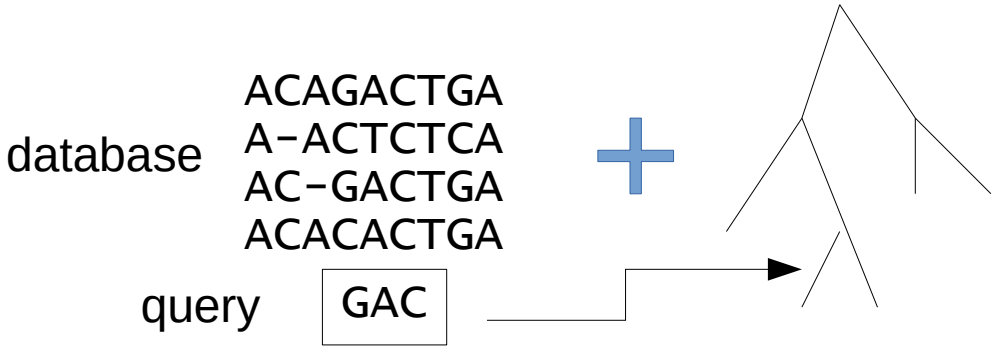
target taxon



unrelated organism not in database - > false positive

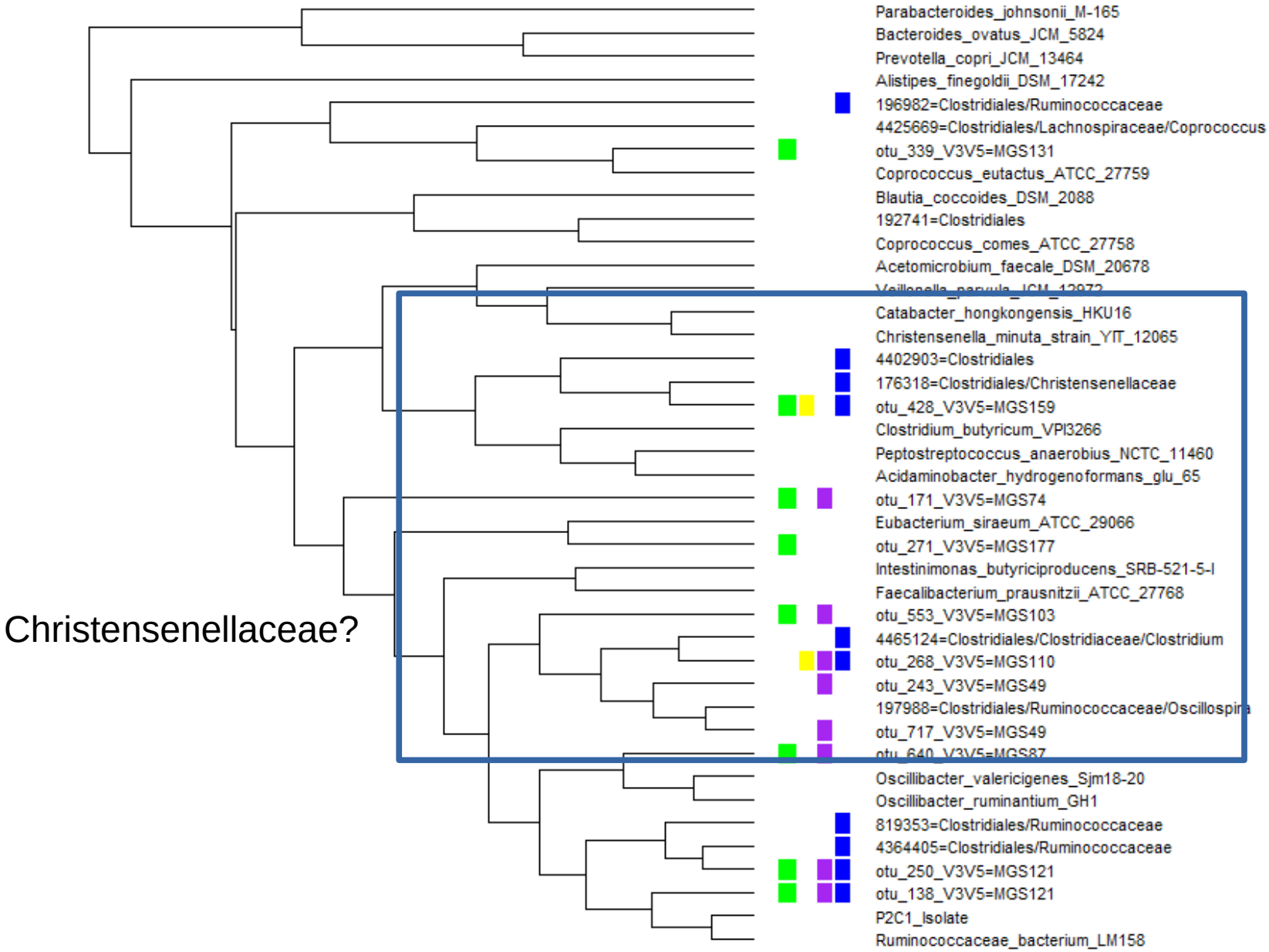
Phylogenetic approaches (TIPP, mOTU, etc)

- Computationally expensive
- Can (in theory) handle sparse database

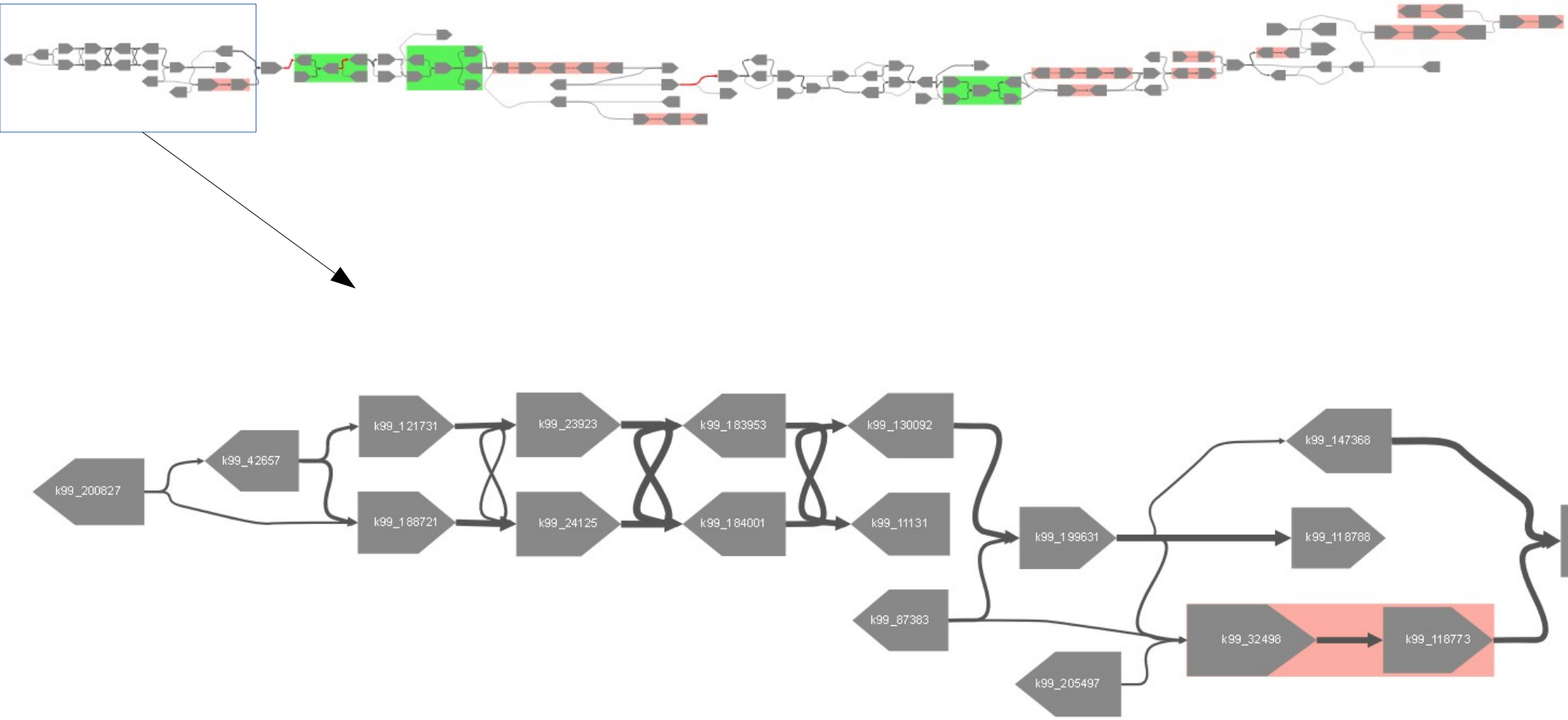


Validation?

- What do the names mean after all?



Strains matter



More from Titus Brown, Todd Treangen, Curtis Huttenhower

STAMPS is just the start

- (meta-)transcriptomics
- multi-(meta-)omics
- machine learning
- etc...

16S/Amplicon sequencing

Overview of generic amplicon workflow

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → fastq files → demultiplex (split samples by barodes) → quality filter/trim (remove adapters/**primers**) → fasta files

might be done by sequencing facility

fastq files:
@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G:<G<AGGII...

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

fastqc/multiqc plotQuality (dada2)

Some tools:
- trimmomatic
- fastq_filter (usearch/vsearch)
- bbduk.sh (bbtools suite of tools)
- filterAndTrim (dada2)

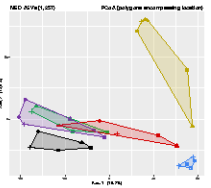
fasta files:
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

dereplication → chimera removal → generate OTUs / resolve ASVs → Standard outputs: count table, fasta file, taxonomy

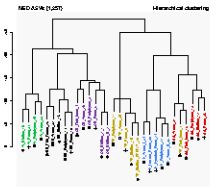
	Sample_A	Sample_B	. . .
Seq_1	0	428	. . .
Seq_2	306	323	. . .
Seq_3	217	1	. . .
. . .	. . .	. . .	. . .

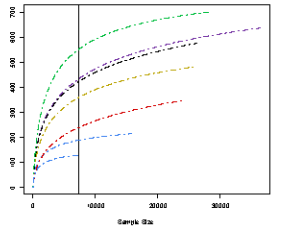
This is generic; specific tools can vary on the order of steps here and how they are done.

Analysis

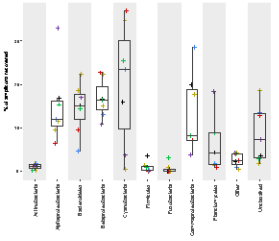


Beta diversity





Alpha diversity



Taxonomic summaries

Some tools that provide whole workflows:

dada2 runs within R (ASVs)
usearch/vsearch runs at the command line (ASVs and OTUs)
mothur runs at the command line (OTUs only currently)
qiime2 runs at the command line (provides an environment that employs other processing tools like those above, and provides visualization capabilities)

from Mike Lee

Metagenomics

Overview of generic metagenomics workflow

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → fastq files

@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

might be done by sequencing facility

→ demultiplex (split samples by barodes)

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

→ quality filter/trim (remove adapters/**primers**) (fastqc/multiqc)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

→ fasta files

>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

→ read-based analysis

Some tools:
- metaphlan2
- humann2

no assembly path

assembly path → consider testing assemblies with and w/o digital normalization

Some tools:
- bbnorm
- diginorm

MetaQUAST is a great tool for comparing assemblies

(co)-assembly

Some assemblers:
- SPAdes
- Megahit
- idba-ud

map individual sample reads to (co)-assembly → Generate coverage information (mappnig)

Some tools:
- bowtie2
- bwa

Gene calling Functional/taxonomic profiling

Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

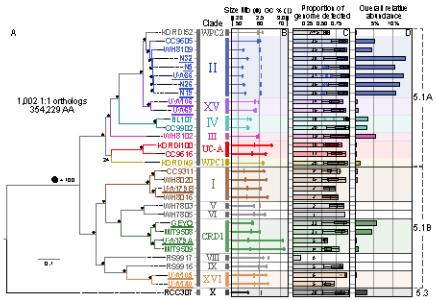
Recovering MAGs (metagenome-assembled genomes)

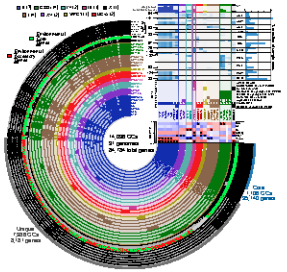
Some tools:
- anvi'o (interactive manual curation of bins)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- MetaBAT2 (kmer-based and coverage-based)
- BinSanity (primarily coverage-based, optional second round kmer-based; memory intensive)
- checkm (taxonomy; estimate % completion/redundancy)

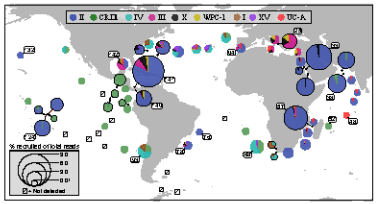
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignerment)







from Mike Lee