Metagenome sequencing

USDA ARS Microbiome Workshop
August 30, 2017
Adam R. Rivers

Hello!

The Genomics and Bioinformatics Unit

Brian Scheffler – Research leader
Stoneville, Mississippi
Brian.Scheffler@ars.usda.gov

Adam Rivers – SY Microbiomes and microbial genomes
Gainesville, Florida
Adam.Rivers@ars.usda.gov

Justin Vaughn – SY Crop Genetics
Athens, Georgia
Justin.Vaughn@ars.usda.gov

Amanda Hulse-Kemp – SY Complex and Polyploid Genomes
Raleigh, North Carolina
Amanda.Hulse-Kemp@ars.usda.gov

Learning Objectives

- By the end the training you should

1. What questions metagenomics can address
2. Know how metagenome samples are processed
3. Understand an assembly-based bioinformatics workflow
4. Know the types of analyses are possible with metagenomics.

1.

What is metagenome sequencing?

Metagenome sequencing

"Metagenome sequencing is the random sequencing of genomic material from a mixed group of organisms, including natural communities, enrichments and co-cultures."

Short read
- PE Illumina
- SE Illumina
- Nextera
- Swift ss/dsDNA
- Reverse transcription for RNA viruses
- MDA

Long read
- PacBio
- X10 Genomics
- Oxford Nanopore
- Dovetail Genomics

Fractionated
- SIP
- Crosslinked - Phase genomics

Metagenomics and Hypothesis driven science

"Which comes first the experiment or the model is the same as which came first the chicken or the egg. The answer is they co-evolve."

-Neil Lawrence paraphrasing Karl Popper's from *The problem on Induction*

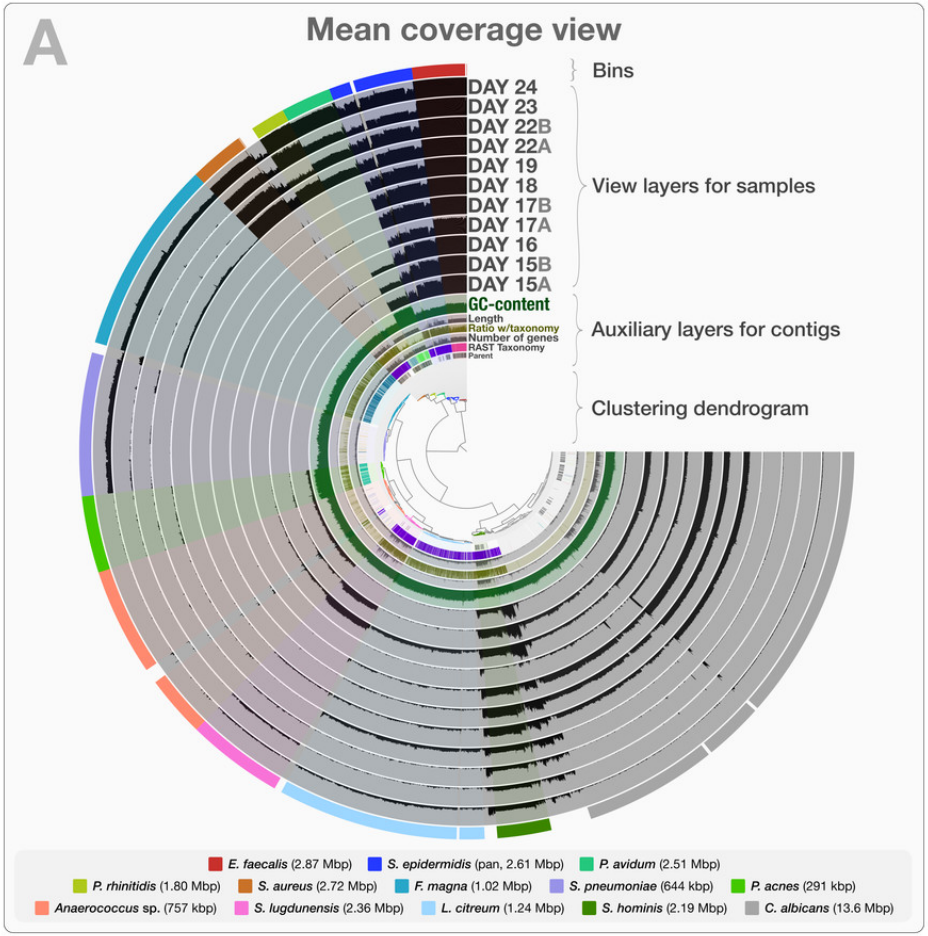
Questions drive the collection of metagenomics data but hypotheses are often generated and tested after data collection.

2.

What can metagenomics answer?

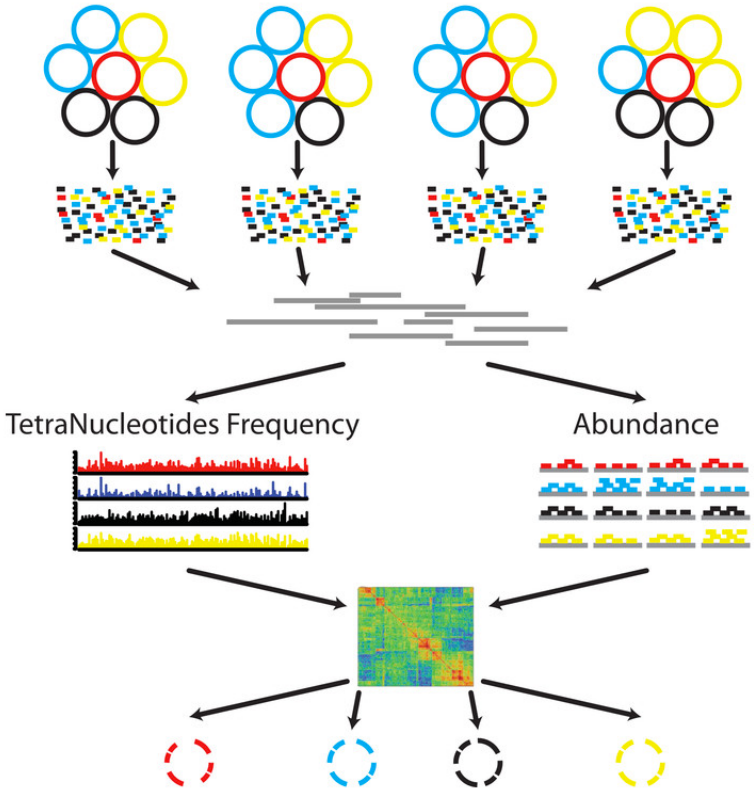
Who lives here and what can they do?

Taxonomic composition



Eren et al. 2015

Genome reconstruction



Preprocessing

1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

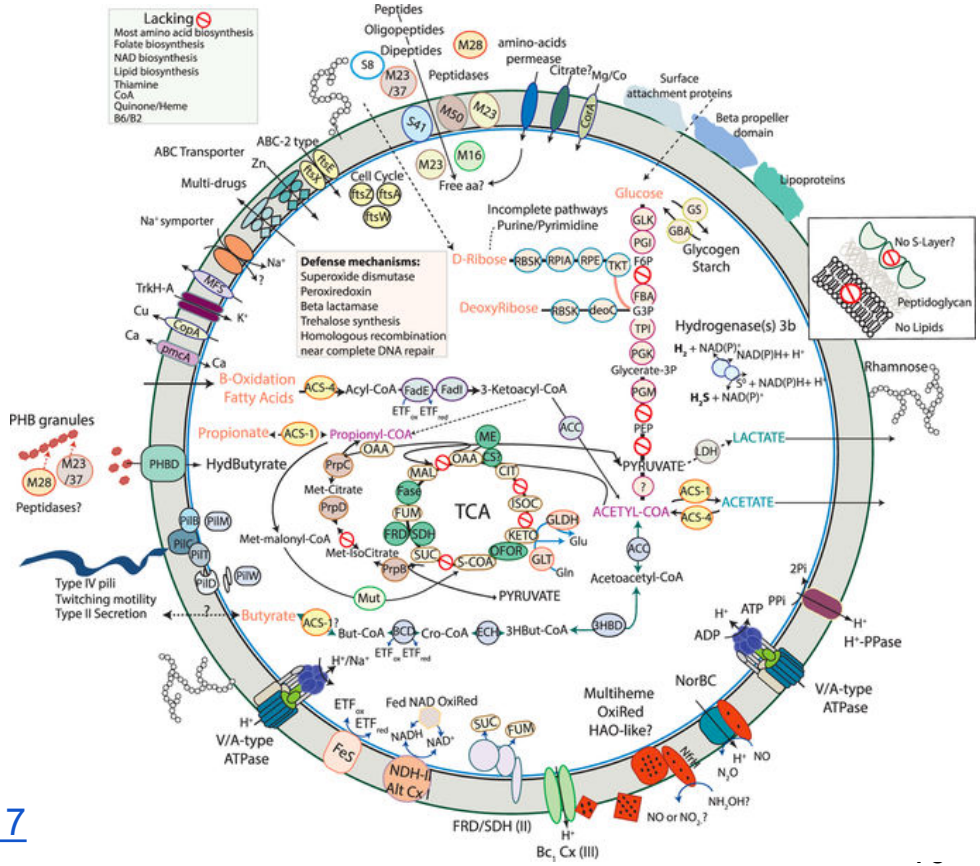
MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

Kang et al. 2015

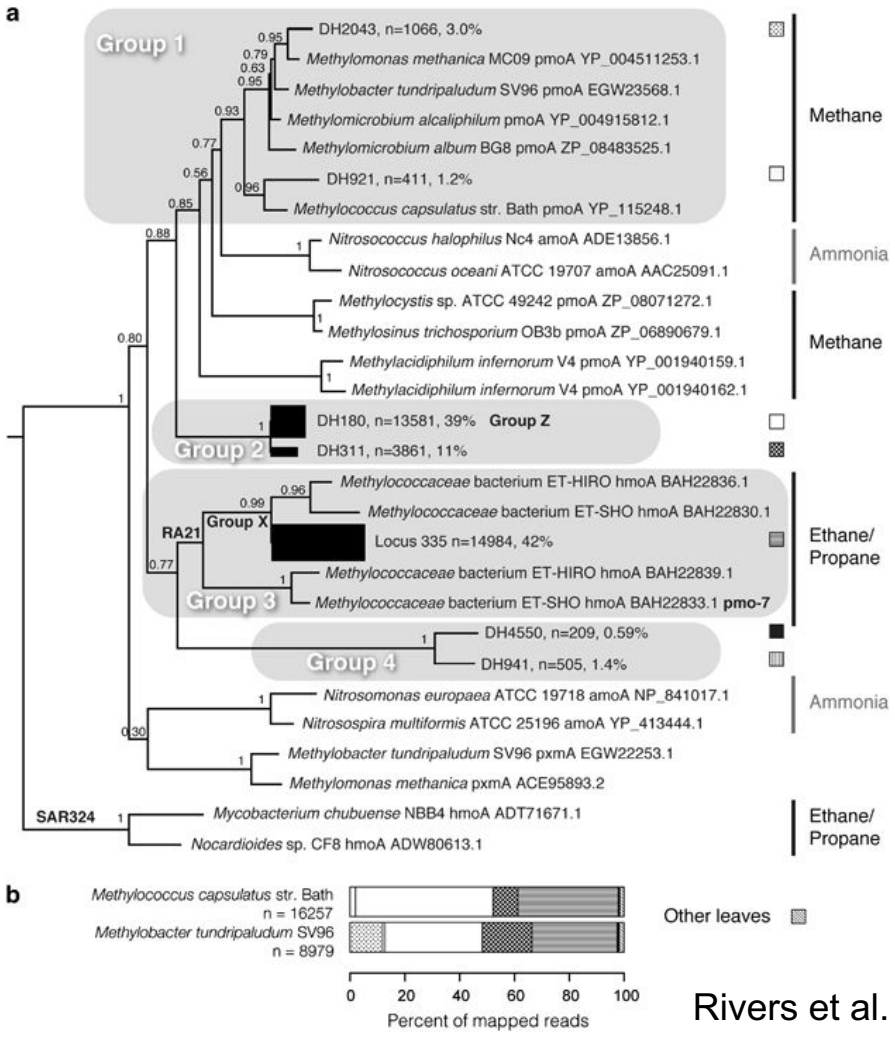
Functional annotation and Metabolic pathway construction

Metabolic capacity for Candidate phyla radiation bacterium ***Parcunitrobacter nitroensis***



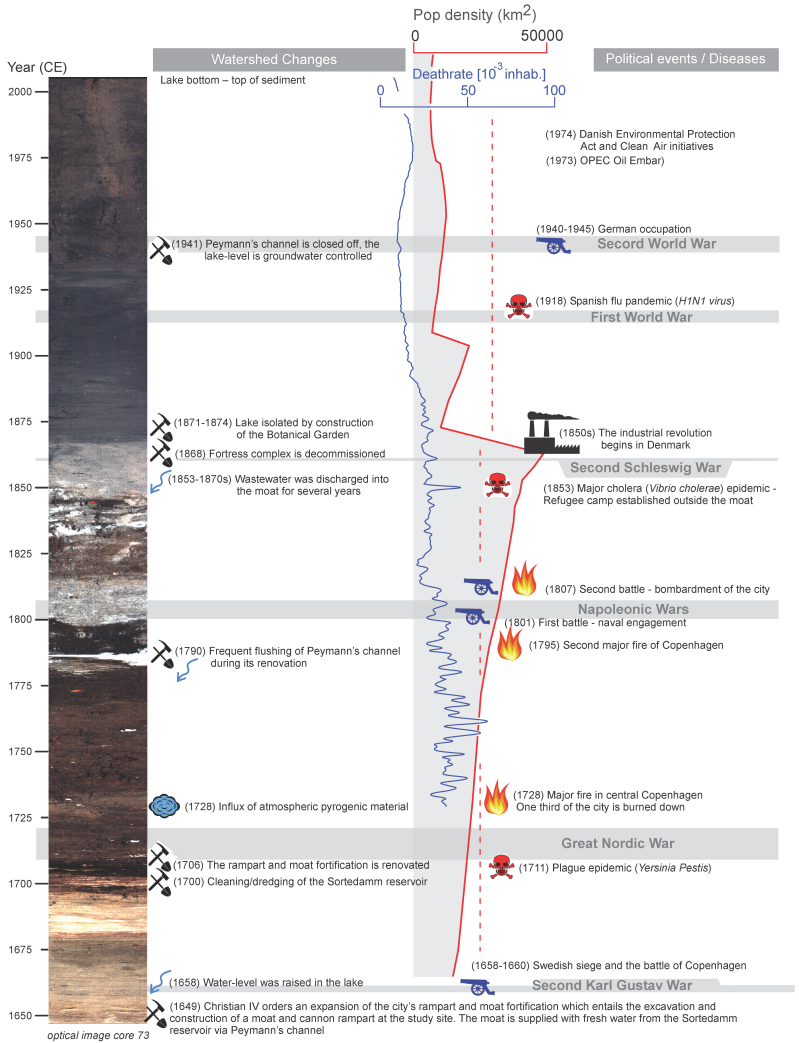
Castelle et al. 2017

Functional gene phylogeny



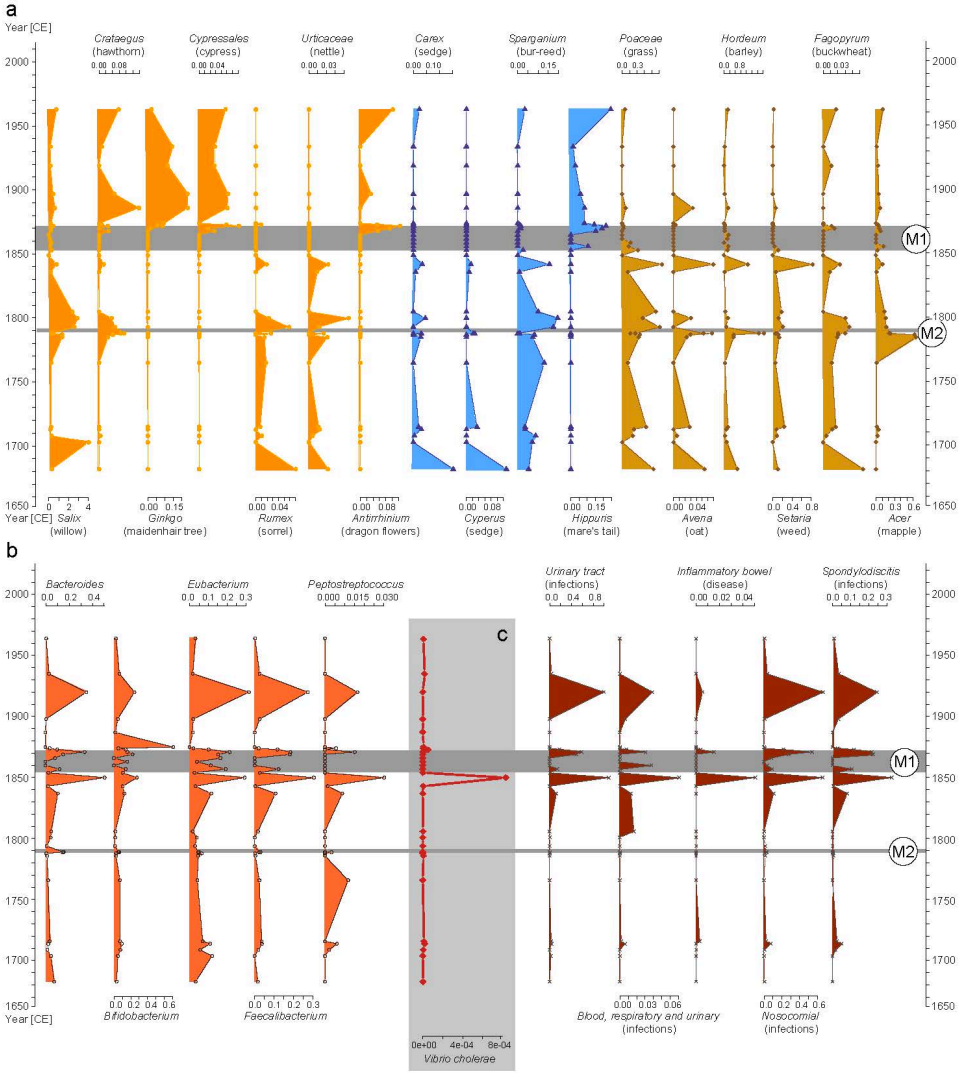
Rivers et al. 2013

Historical reconstruction



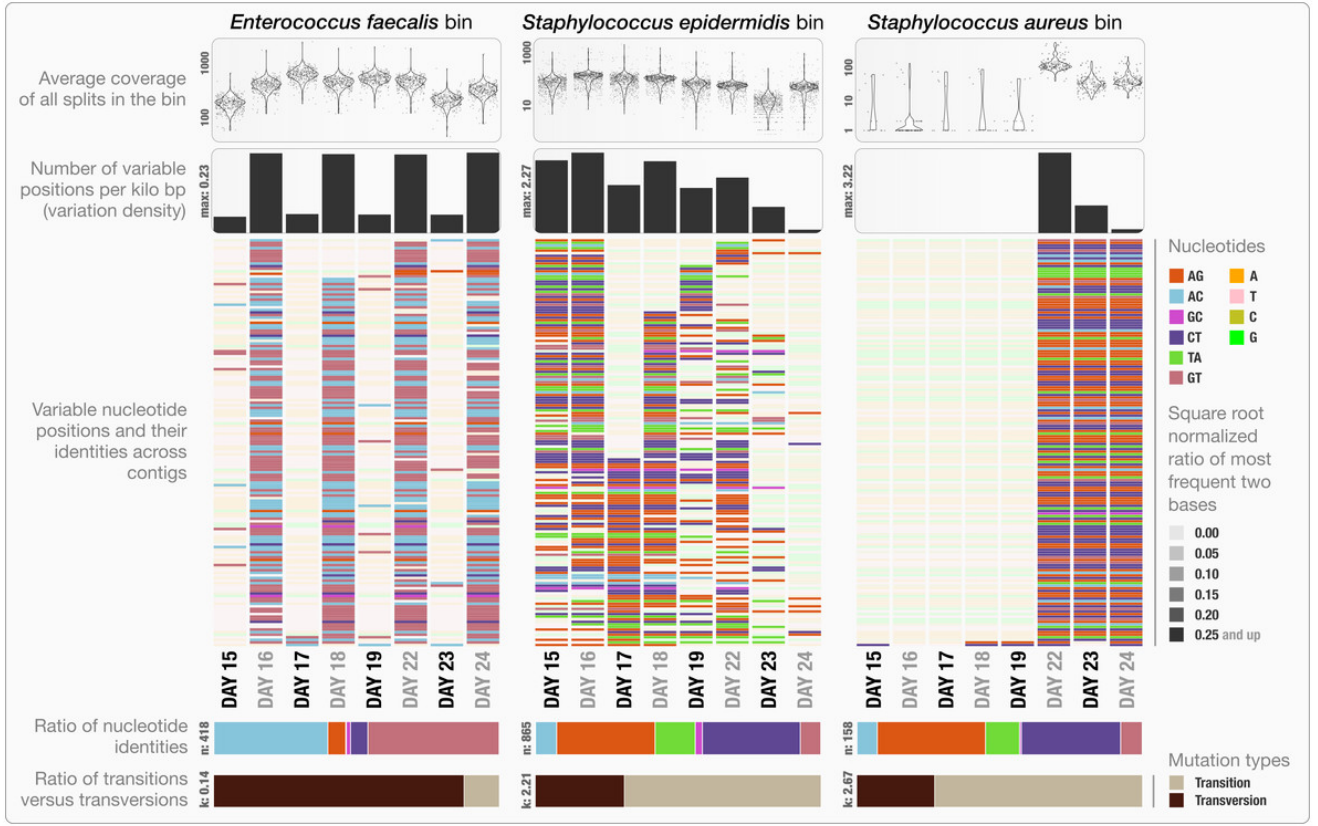
(Pedersen et al. In review)

Historical reconstruction



(Pedersen et al. In review)

Population structure



3.

Field to sequencer – the nuts and bolts.

Sample preparation and sequencing

Collect DNA

Metagenome DNA sampling is more forgiving than RNA sampling but care must be taken to prevent degradation, especially for long read technologies.

4ºC Short-term
-20ºC Long-term
(Rubin et al. 2013)

Collect and store
within 1 hour
Extract as soon as practical

Consider collecting RNA and storing it in RNAlater (saturated ammonium sulfate) at 4ºC until extraction. Metatransctiptomics can answer new questions.

Collect Metadata

"There is no such thing as metadata, everything is data."
- Susan Holmes

- Sample collection is the time to record environmental data
- The GSC has created environmental and sequence data standards, MIxS. Use them as a guide for your collections.
- Store environmental data in NCBI or ENA Biosamples databases or Gold database. Do it now, while you still remember what you did. You can link sequence data later.

DNA extraction

Target amount of DNA for Illumina sequencing:
1 µg at 20-50 ng/ul in 10-50 ul
About 200ng is needed for Truseq.
10 ng for Netxera and Swift.*

Target amount of DNA for long read sequencing:
1-5 µg at 50-500 ng/ul in 10-100 ul

* These kits go to ~50 pg but but going below a few ng is a great way to accidentally sequence yourself or your pets

DNA extraction

Extraction kits vary by sample type

- MoBio (now Qiagen) kits have been very popular
- Phenol-chloroform and TRIsol extractions yield good results in experienced hands

Long read sequencing requires special extractions

- Fosmid methods
- PFGE
- No sonication, French press or bead beating!

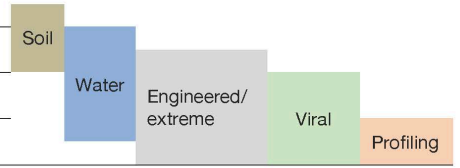
All extractions are selective

- There is a tradeoff between DNA quality and bias
- Multi-enzyme cocktails are now available

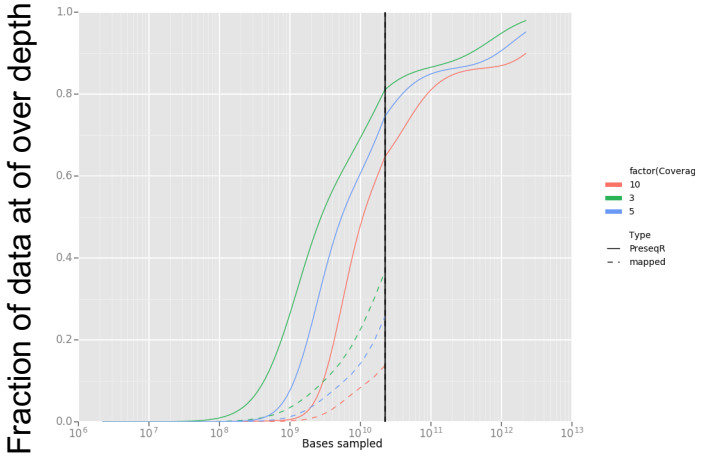
How much sequence is needed?

Table 2 Common pooling degrees, yields and environments sequenced with the HiSeq 2000 1T.

Degree of pooling	Bases (GB)	Read pairs (M)	Typical applications
1	64	212	
2	32	107	
3	21	71	
4	16	53	
6	11	36	
8	8	27	
12	5.3	14	
64 (uncommon)	1.0	3.3	



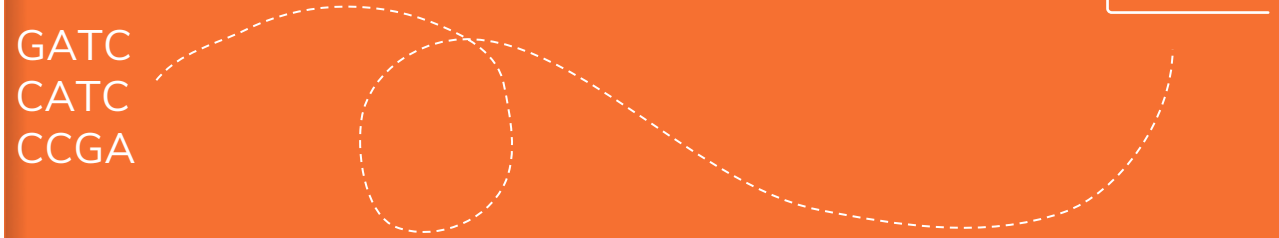
Kmer based estimates of coverage can be made



4.

Analysis of metagenome data

What to do with all those fastq.gz files.



Metagenomic sequencing

Read-Based

GATC

Focused on functional and taxonomic composition

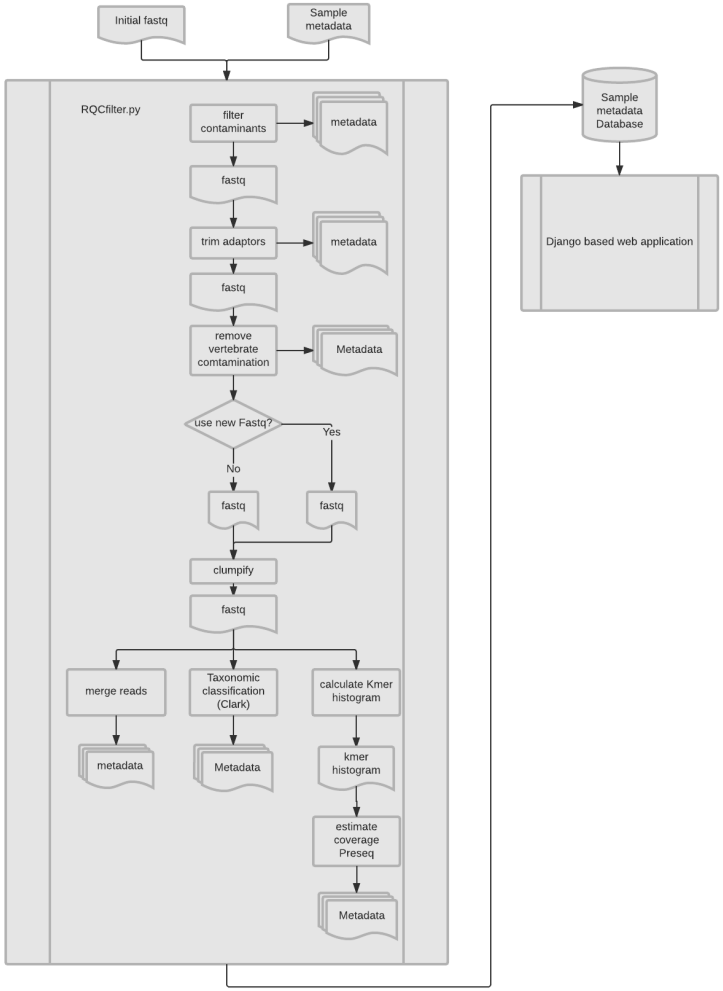
Assembly-Based

Useful for pathway analysis, phylogenetics

Detection-based

Primarily for pathogen detection

Quality Control



[Demo of QC application]

Taxonomic composition

Read identification

Read based approaches require rapid methods

- Kmer based matching
  - Clarke, Kraken
- BWA/FM indexed based
  - Centrifuge
- Accelerated gapped alignment
- Diamond/ Megan

- Database selection is key
- Precision-recall should be considered

Gene based identification

First gene calling is run
Metagenemark, Prodigal

Full length genes increase accuracy

Data reduction allows for more sensitive methods:
Hmmer,

Gapped alignment (LAST, RapSearch)

Determining the level to classify at is a challenge for all algorithms

Metagenome binning – genome reconstruction

Metagenome binning is the clustering of contigs into bins that originate from a single genome.

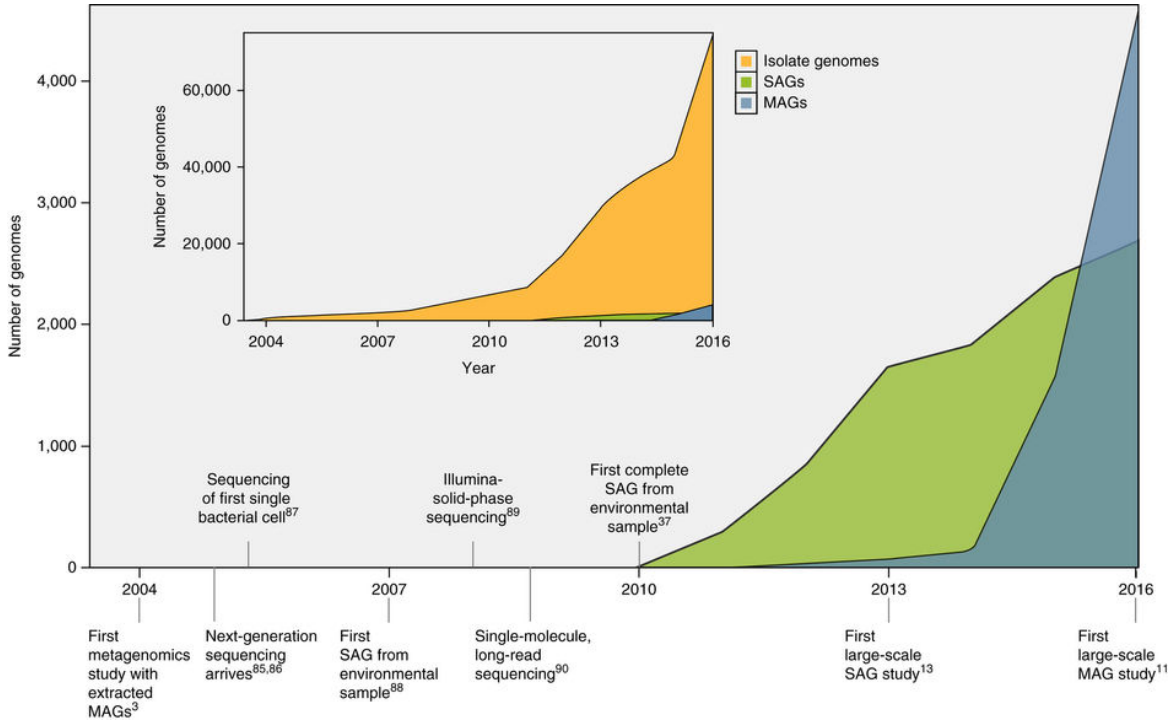
Input data used:
- 4-mer frequencies
- Abundance profiles

Common tools:
- Metabat
- Maxbin2
- Concoct
- ABAWACA
- ESOM

Quality assessment
- CheckM

Metagenome binning – genome reconstruction

Growth in metagenome assembled genomes (MAG's)



Quality standards are [now available]

Functional annotation

Functional databases/ontologies

Protein families
Pfam
Superfamily
eggNog

Metabolic pathways
Kegg
SEED
MetaCyc

Antibiotic Resistance
CARD
ARDB

Virulence
MvirDB
PATRIC
vFAM
VFDB
Victors

Carbohydrates
dbCAN

Mobile genetic elements
ICEberg
Prophages

Cross-mapping does seem to work (Kultima et al. 2016)

Functional Gene phylogeny

- Select reference gene family
- Search for contig homologs (HMMER)
- Build MSA (MAFFT) and HMM
- Select regions (TrimAL)
- Test phylogenetic models
- Build ML Tree
- Search then Align Reads with HMMER
- Read placement (Pplacer, EPA)

Tutorial time

Assembly-based metageomics workflows can be found at the ARS microbiome workshop website.

Tutorial 1: Raw data to assembly and mapping
Tutorial 2: Annotation, binning and exploration with Anvi'o

Photo credits

- 1: DNA - Made By MadeByOliver from Flaticon.com
- 1: Tilled field Photo by Zbysiu Rodak on Unsplash
- 3. Rice paddies Photo by Doan Tuan on Unsplash
- Crowded Photo by denis ng on Unsplash