Introduction to Metagenomics

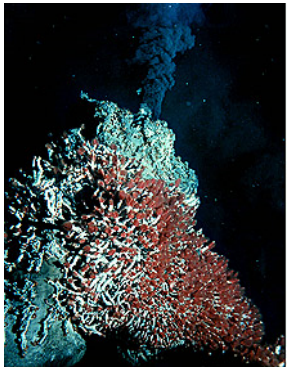
Bhavya Papudeshi
bhnala@iu.edu

What initial steps would you take to study microbes in an environmental sample?

- Collect the sample
- Culture the microbes in the sample- petri plate, liquid media.
- Count the number of viable cells cultured in the media
- Focus on a select cultures based on your research question

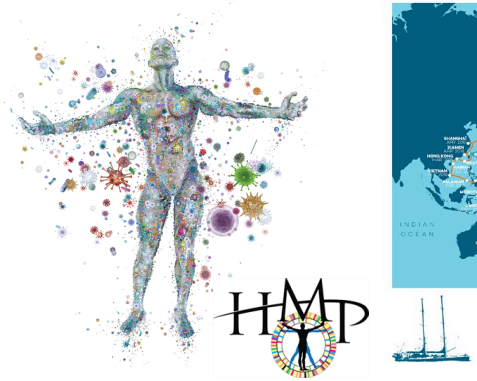
Why metagenomics?

- Less than 1% of the microbes can be cultured in lab
- Metagenomics is culture independent, providing an unbiased view of the microbial community
- Any environment can be sampled provided that nucleic acids can be extracted, a minimal amount of ~5ng



Hydrothermal vents

Acid mine drainage



Human Microbiome Project

TARA ocean sampling

Generally speaking on “omics” studies

- Who is there? - 16S amplicon sequencing
- What are they capable of doing? - Whole genome sequencing
- What are they actually doing at a given time and place? - Metatranscriptomes
- What are they actually doing at a given time and place in protein level? - Metaproteomics

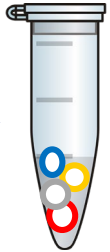
Sample collection to sequencing

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Sample collection to sequencing

- Analysis of DNA directly obtained from an environmental sample (Hugenholtz and Tyson, 2008)

Sample environment



Sample collection

DNA Extraction

Sequencing preparation

Sequencing

Library preparation

Healthy coral reefs

Coral bleaching (unhealthy coral)

Destroyed coral reefs

Sample collection

DNA extraction and adaptor/indices addition

Mix all the samples together and sequence on one lane

Why is all this relevant?

Now if you notice a set of samples with poor quality or your data doesn’t seem right,

1. were they sequenced on the same lane or samples collected from the same environment?
2. accidentally cross contaminated during sequencing preparation
3. mislabeled samples
4. the noticed poor quality is actually biologically relevant

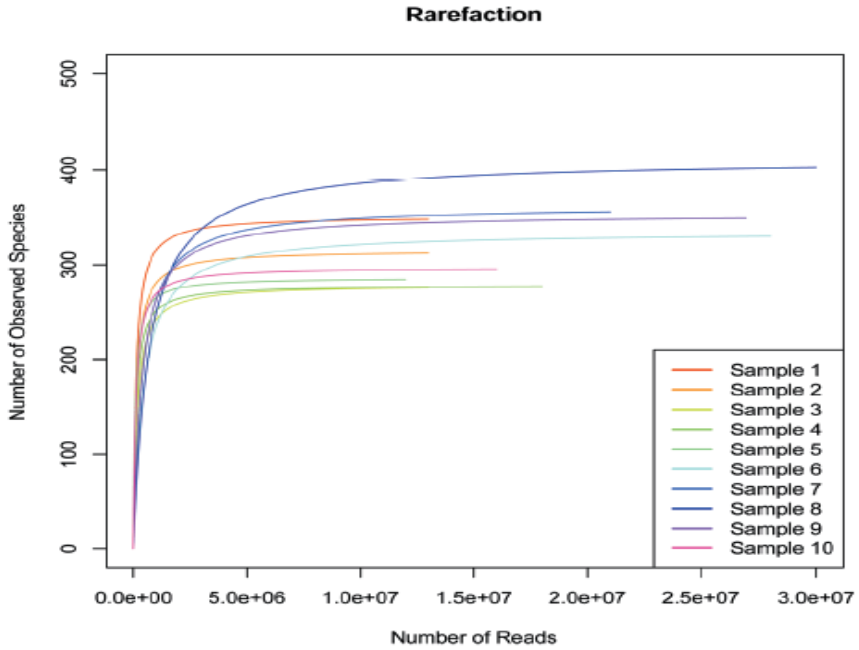
Bioinformatic Analysis

Steps after getting sequencing data

- Archive your raw data
- Check the quality of sequences- FastQC, Prinseq, Trimmomatic
- Remove PhiX sequences – PhiX bacteriophage is added for calibration control and to improve diversity (https://support.illumina.com/bulletins/2016/07/what-is-nucleotide-diversity-and-why-is-it-important.html)
- Host genome sequence for example, human genome sequences
- Sequencing depth of the sample- how much of the environmental diversity did you capture in your sample?

Sequencing depth of the sample

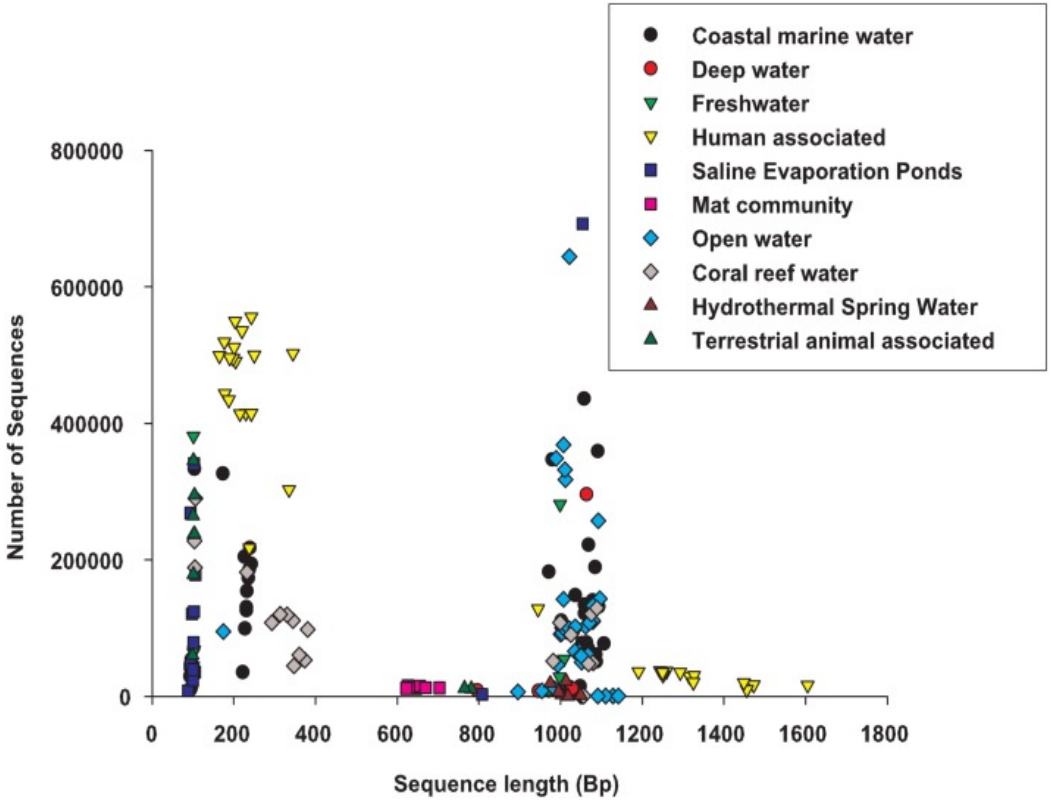
- Is your metagenome sample(s) a good representation of microbial community?
- Rarefaction curves plot the species richness in a sample, the graph initially increases exponentially as new species are being identified per sample, and slowly beings to plateau as the number of new species added decrease
- Bottlenecks- does not provide input to species abundance, does not account for rare taxa



Zheng et al., 2016
doi: 10.11979/idtm.201602005

Visualizing read data

- Especially important when using publicly available data
- Sequencing length vs number of sequences
- In the figure, we can conclude environments were represented by two or more sequencing technologies



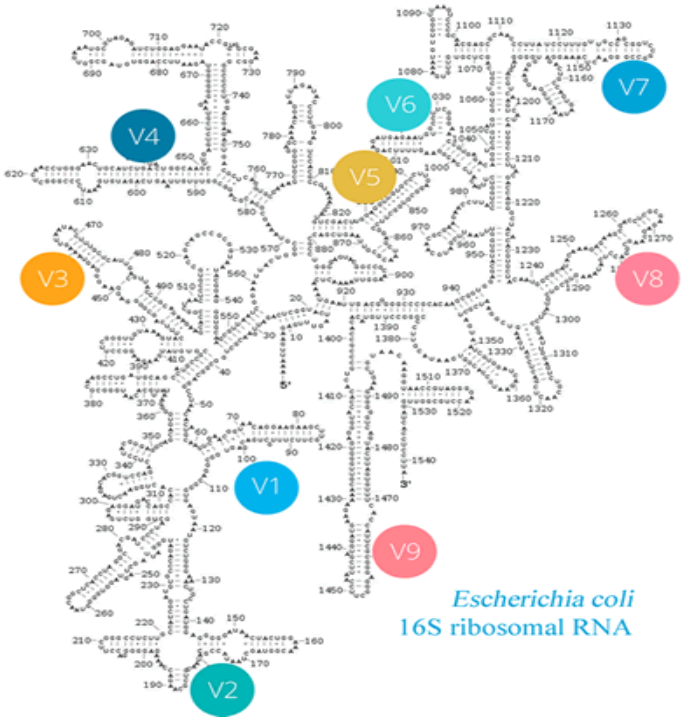
10.3389/fgene.2013.00041

16S Amplicon sequencing

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Amplicon sequencing

- Characterize the microbes by targeting the conserved regions 16S in the sample
- PCR amplification of the hypervariable regions in conserved genes
- Different variable regions provide distinct microbial composition, selection of primers is therefore critical for different environments
- Limited resolution of the microbial community
- Operational taxonomic units (OTU) are based on identified 16s surveys



Yang et al., 2016
doi: 10.1186/s12859-016-0992-y

Operational Taxonomic Units (OTU)

- A cluster of similar sequence variants (more than 97%) are called OTU, to represent a bacteria, this can mean genera or species depending on the sample
- Unidentified OTU‘s are still pretty common (we have only been studying 1 % of the microbes)

- Programs that help identify OTU’s

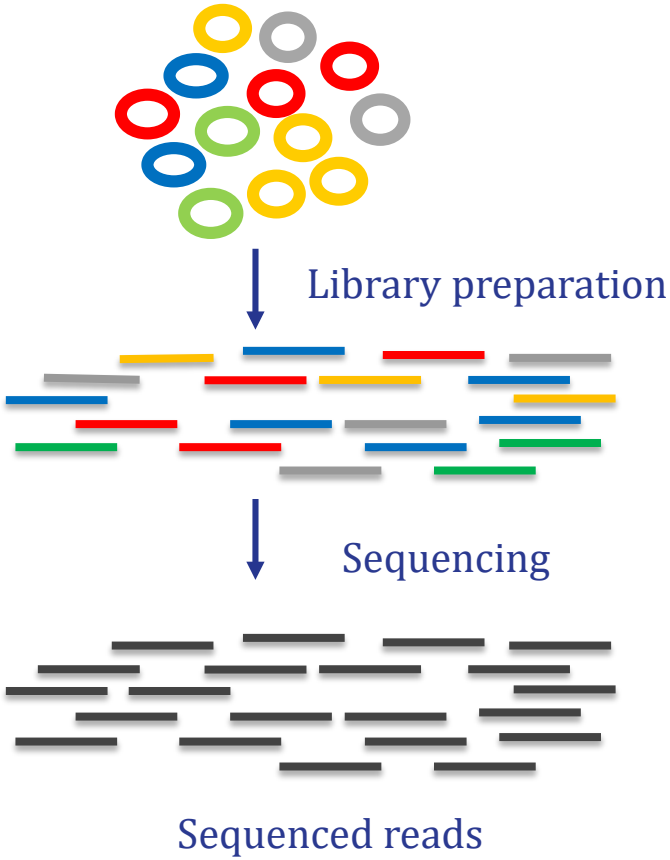
➢ **uclust** / **usearch**
➢ **QIIME 1** (internally using usearch for 16S OTU clustering) – available on Jetstream cloud
➢ **DADA2** to analyze exact 16S sequence variants (no OTU cluster)
➢ **CD-HIT** – installed and available on Carbonate cluster

Whole genome metagenomic sequencing (WGS)

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Whole genome metagenomics

- PCR independent method
- Sequences the whole genome, providing insights into both the taxa and functional characteristics of the community
- Higher resolution and sensitivity in characterizing microbes and in calculating microbial diversity
- A certain portion of the metagenome remain uncharacterized due to limited number of sequences available in the databases



Identify taxa and function

- Sequenced reads are aligned against a reference database, identifying the taxa and functional genes
- Under or overrepresentation of specific species or genes can provide insights into the role of microbes in that particular environment

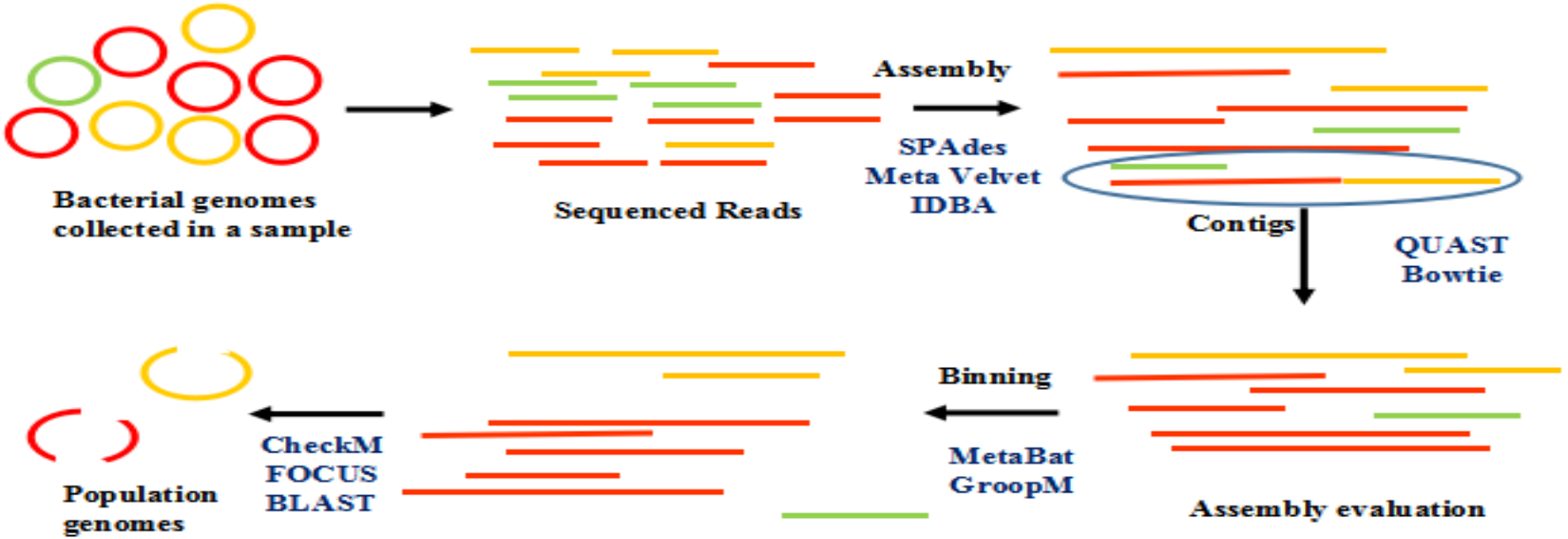
Sequenced reads

Aligned against a reference (nr/nt, Refseq) database

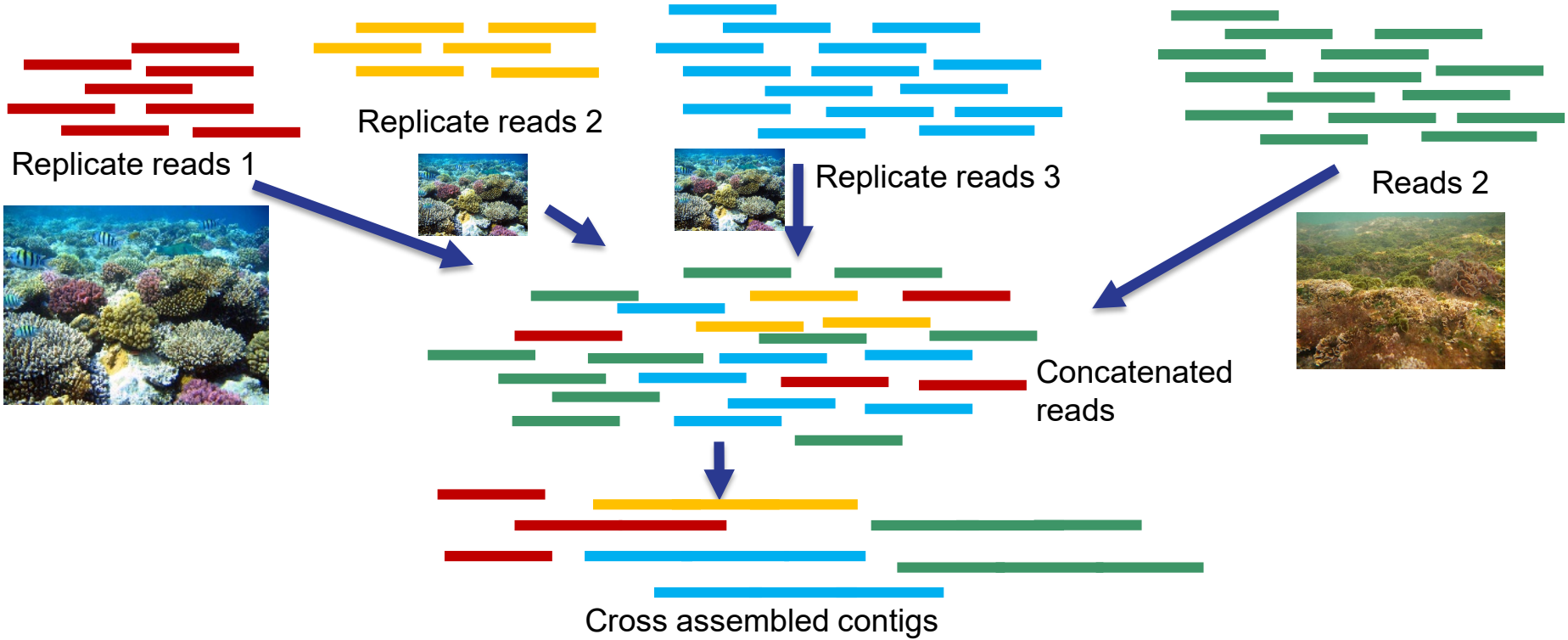
Taxa/Function identified

Genome centric approach

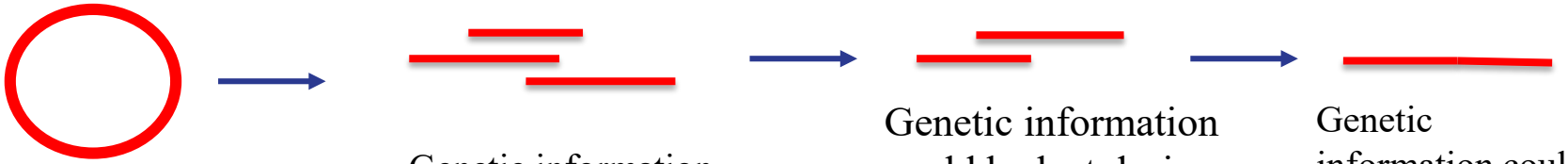
- Link function to taxa, and infer metabolic process that are absent or present in this population within an environment
- Reconstruct genomes from dominant population in the sequenced reads called, metagenome assembled genomes (MAG)



Cross assembly/ Co-assembly



Why cross assembly/ co-assembly



For e.g, bacteria X from metagenome 1

Genetic information could be lost during sequencing prep

Genetic information could be lost during sequencing or pre processing (low quality)

Genetic information could be lost during assembling

- Through cross assembly, you are therefore improving the coverage of the microbe, increasing the chances of reconstruction of a **Metagenome Assembled Genomes (MAG).**

Anvi'o platform

- Open source, metagenomics visualization platform from University of Chicago
- Some features of this program

1. Reconstruct metagenome assembled genomes
2. Phylogeneomic analysis of MAG's or invidual genomes
3. Pangenomic analysis of individual genomes or MAG's to identify conserved/core genes

- All the above mentioned methods are available as pipelines (http://merenlab.org/software/anvio/) under tutorials with detailed walkthrough on running the analysis
- Anvi'o is available as a preconfigured image on Jetstream cloud.

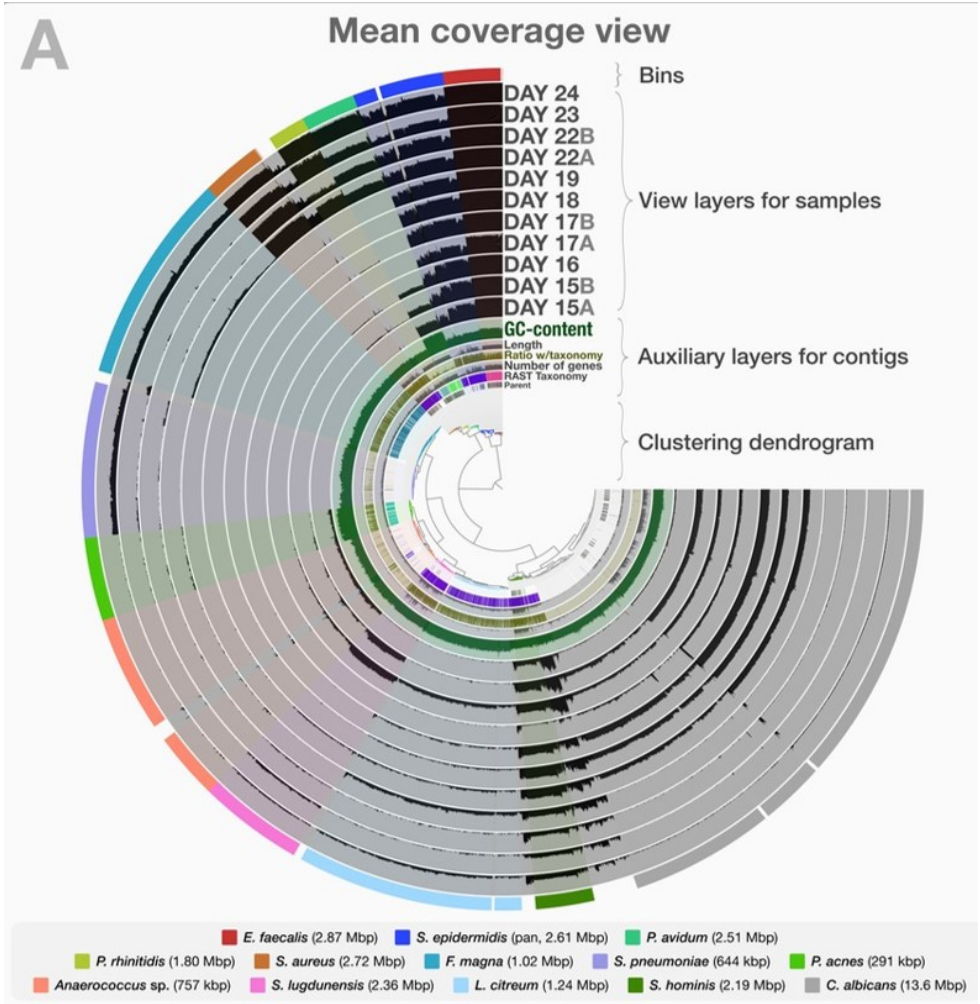
Anvi’o images

The figure on the right shows clustering of contigs (assembled reads) based on sequence composition (GC content) – **clustering dendrogram**

Auxiliary layers – information on each genomes, taxonomy, GC content, number of genomes

DAY *- represent infant gut metagenomes collected from different days

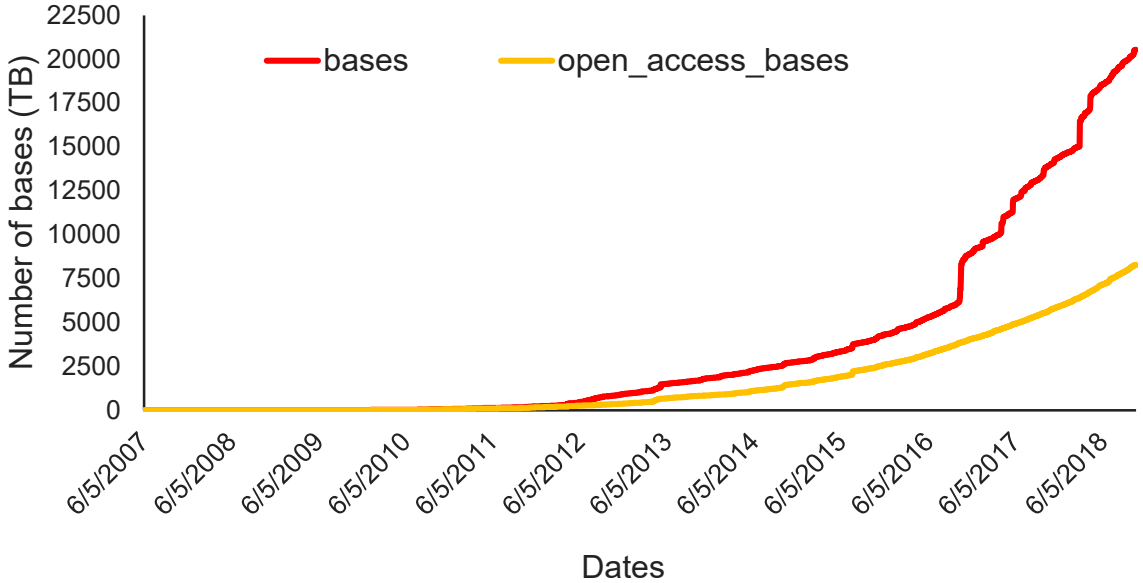
Bins – Clustering of regions across genomes based on presence or absence of specific genes.



Sequence Read Acrhive (SRA)

Sequence Read Archive (SRA)

- SRA is a NCBI public database that hosts next generation sequencing data (NGS) with metadata
- This data is publically available, including sensitive data (human data) which is accessible through a via dbGap (controlled access)



Curently hosts more than 1000 tera bases of data (as of October, 2018)

https://www.ncbi.nlm.nih.gov/core/assets/sra/files/Factsheet_SRA.pdf

Mining the SRA

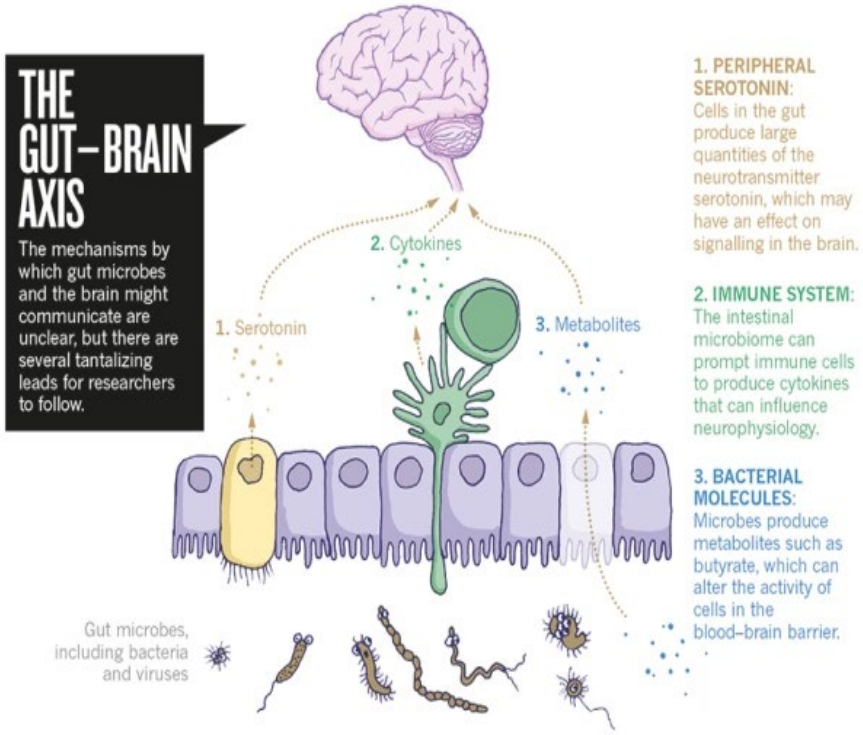
- You can search the SRA on the NCBI website https://www.ncbi.nlm.nih.gov/sra, with SRA IDs (specific study, sample, experiment), organism, author, sequencing platform. For more documentation (https://www.ncbi.nlm.nih.gov/sra/docs/srasearch/)
- Sratoolkit is downloaded and available to download and analyze SRA sequences on Carbonate (HPC)
- If you are interested in searching SRA to look for a sequence of interest (particular genome or gene), https://www.searchsra.org/ platform is available. This is a tool developed by Rob Edwards, a collaborator of NCGAS

There are currently a lot of tools under active development to make this data more readily available to the public

Conclusions from metagenome studies

Gut-brain axis

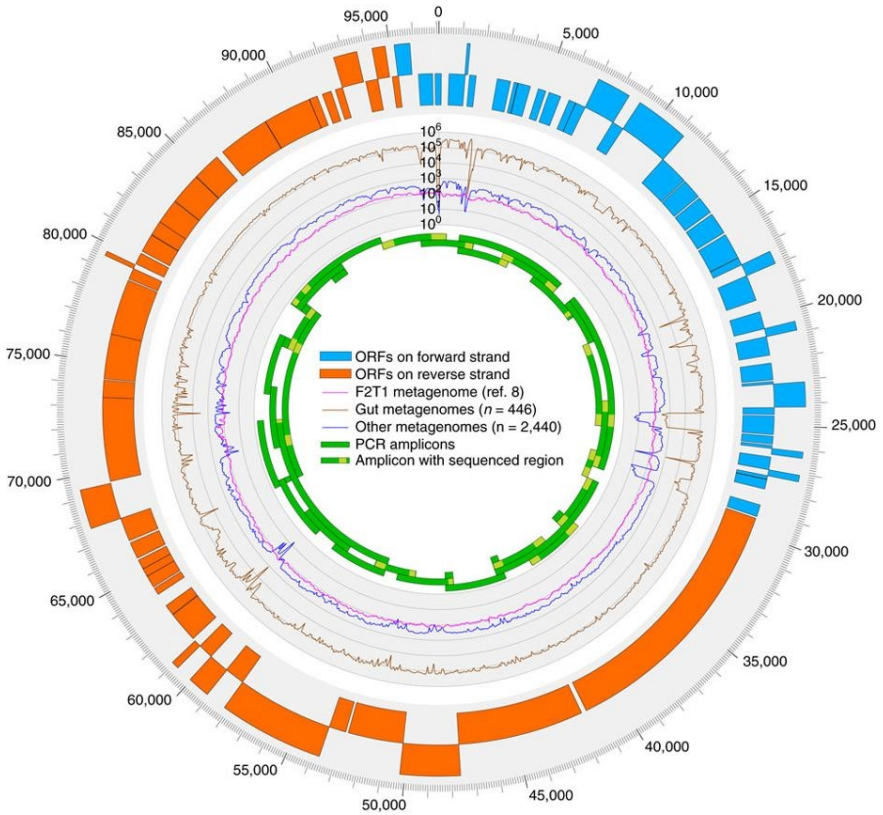
- The biochemical signaling between the gut flora (GI tract) and brain (CNS)
- Gut microbiome promotes serotonin production in the lining of the colon, was shown in mice studies through fecal transplant. However the cascade of molecular events are still unknown
- In other studies the formation of fatty sheathing that insulates nerve fibres was found to be influenced by gut microbes, to help treat multiple sclerosis.



Smith 2015, Nature

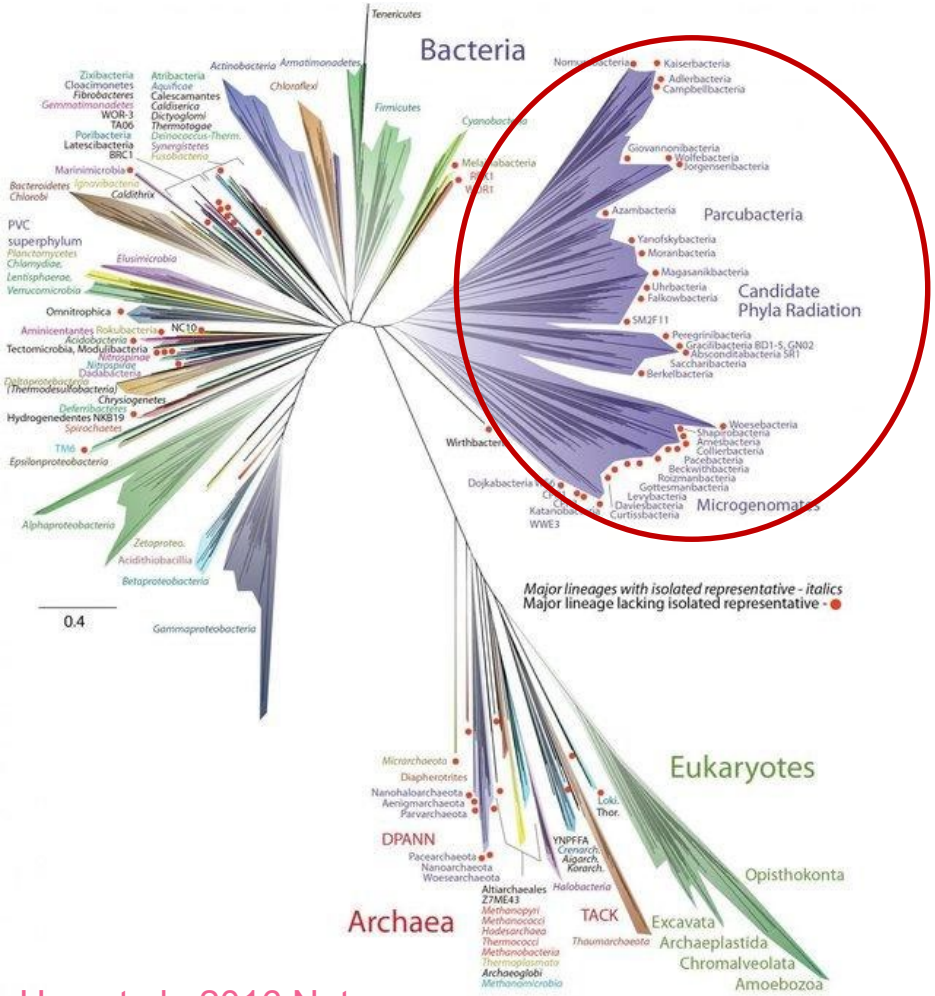
Abundant bacteriophage in human guts

- Discovered in 2014, while binning the human microbiome data, and found in most of the samples (Dulith et al., 2014)
- It was especially interesting, most of the genes we unidentified, no homologs were detected with very few evidence to the biology of the phage
- Since, they have been identified to infect Bacteroidetes and can be classified to the order Caudovirales



Improved resolution of tree of life

- With improved sequencing methods and being able to capture more than 1 % of cultured microbes. In 2016 there are 3,083 organisms in total in the tree of life
- Archaea are more closely related to Eukaryotes
- The branch that splits near the base, are small in size with simple metabolism, predicted to be an early evolving group.



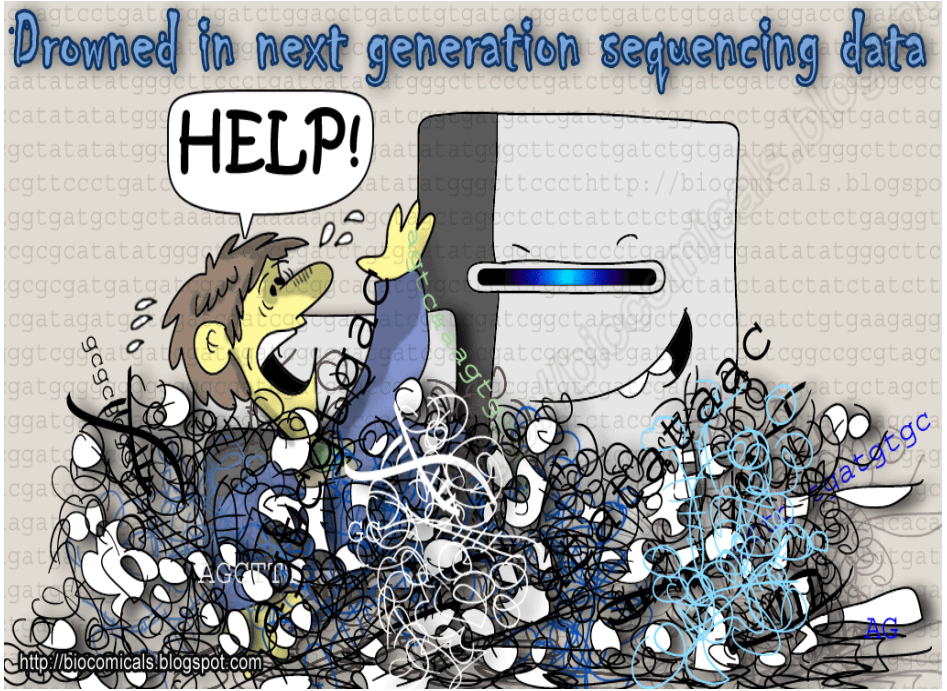
Hug et al., 2016 Nature

More “omics” studies

Other “omics” studies

To capture different signals within an environmental sample

- Meta-transcriptomics
- Meta- proteomics
- Meta-metabolomics



How can we help

- Help with experimental setup and planning
- Software is available on IU High Performance clusters. If there is a program you are interested in, we can help you install it
- Pipeline for reconstructing genomes from metagenome(WGS) is available
- Meta- transcriptome pipeline is under development
- Blogposts - https://ncgas.org/NCGAS_Blog.php#NCGAS%20Blog

Contact information – help@ncgas.org


ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT