Metagenomics Workshop NCGR

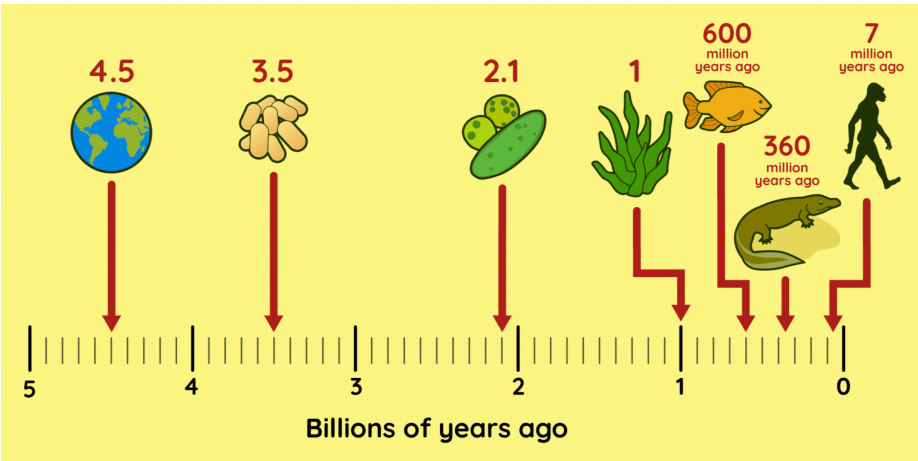
Contents

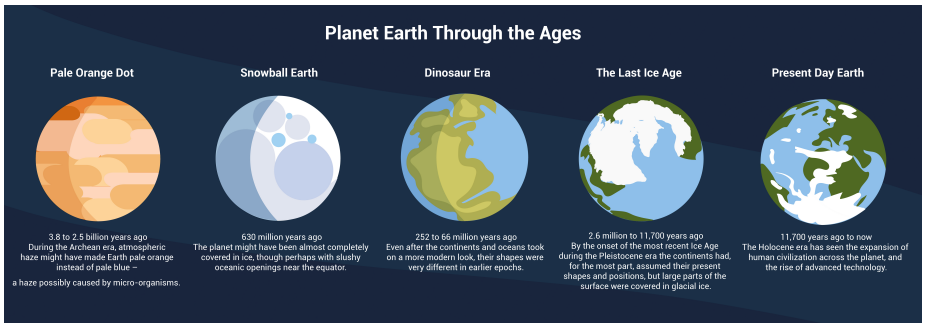
Chapter 1

Community Profiling and Metagenomics

1.1 Microbes were the first life forms on this planet

1. Earth declares its independence about 4600 MYA





2. First photosynthetic bacteria 3.4 billion years ago (BYA)

- Used sunlight for energy to create biomass
- Anaerobic (anoxic photosynthesis)

Figure 1.1: 3.4 billion year old Stromatolite fossil

3. 2.7 BYA first oxygen producers emerge

- Oxygen as waste product during respiration
- Most of the oxygen was sequestered and not readily available

4. 2.3 BYA atmosphere has oxygen
5. 500 million year ago (MYA) first terrestrial plants
6. 200 MYA mammals emerged

Figure 1.2: Modern Stromatolites https://en.wikipedia.org/wiki/Stromatolite

7. 13 MYA one of us makes all of us proud by learning how to fly
8. 10 MYA the branch of life currently called homo emerges
9. 400 years ago humans observe the first microbe under a simple scope

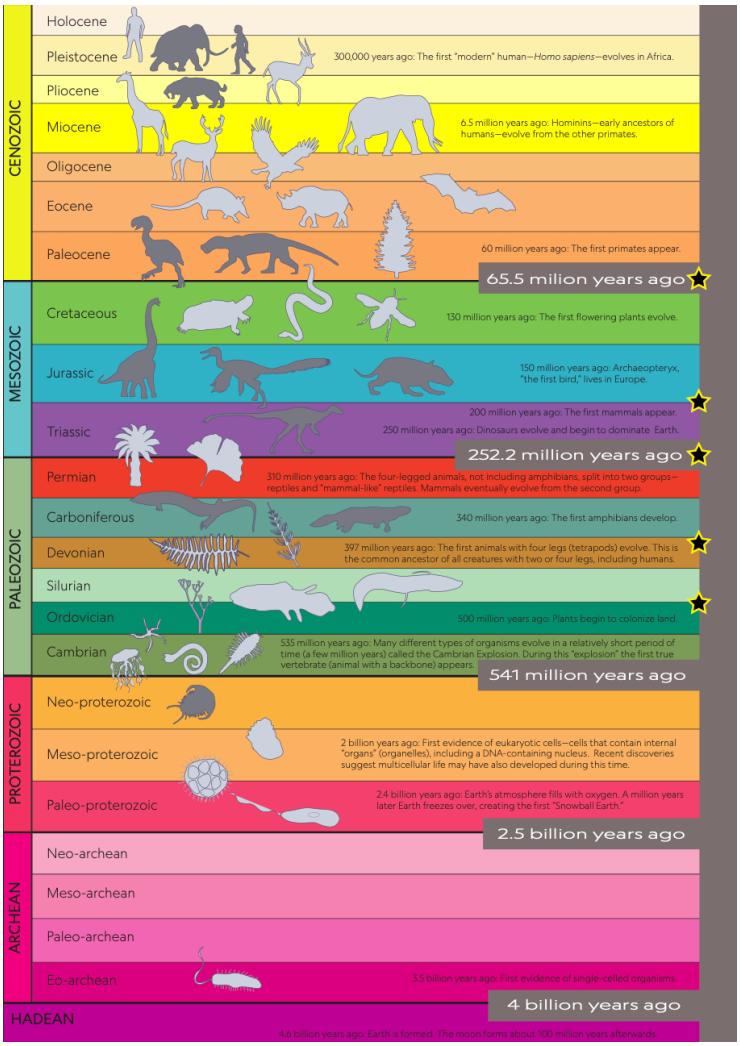


Figure 1.3: History of life on earth

https://education.nationalgeographic.org/resource/age-earth/#undefined

THERE WOULD BE NO LIFE WITHOUT MICROBES

1.1.1 Microbes enable habitability on Earth by catalyzing reactions of biogeochemical cycles

1. The amount or % of elements on Earth remains constant
2. Recycling of these elements, flux, and bio-availability is largely taken care of by microbes
3. Best example to illustrate – nitrogen

- 78% of Earths atm is N2
- Required for important biological processes
- In gaseous form it is unavailable
- In fact many processes are N2 limited
- Making N2 bioavailable in a form that can be by eukaryotes is completely on the shoulders of microbes

1.1.1.1 Nitrogen Cycle

https://cdn.britannica.com/37/6537-050-CF14602B/ammonia-Nitrogen-fixation-nitrogen-form-means-nitrates-1909.jpg

1.1.1.2 Carbon Cycle

https://www.pmel.noaa.gov/co2/story/Carbon+Cycle

How many microbes??

1. 40 million microbes in a gram of soil
2. One million microbes in a ml of fresh water
3. One trillion in a human body

MICROBES ARE ABUNDANT.....AND EXTREMELY DIVERSE!

1.2 How many kinds of living beings are there?

1. Aristotle's Scala naturae = **350 BC**

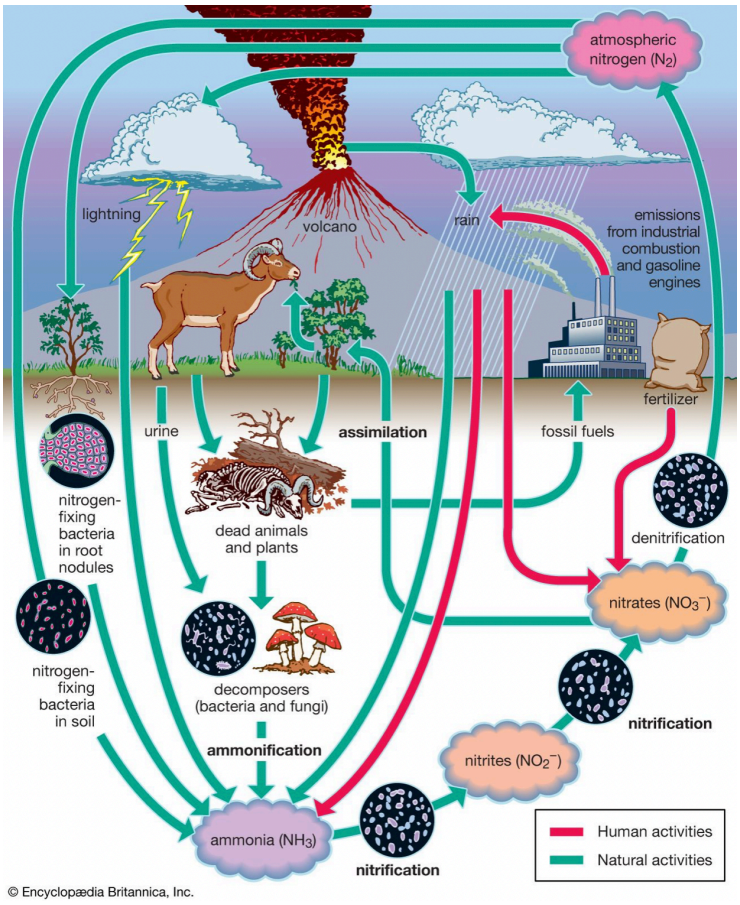


Figure 1.4: Nitrogen Cycle

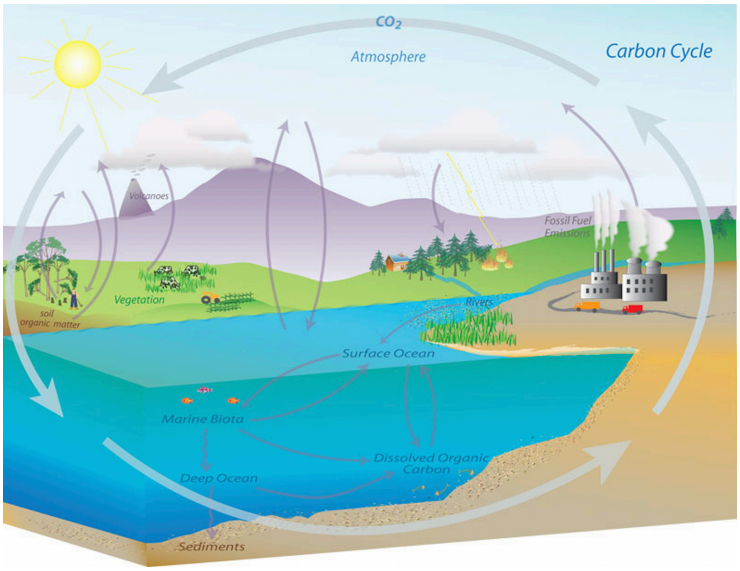


Figure 1.5: Carbon Cycle

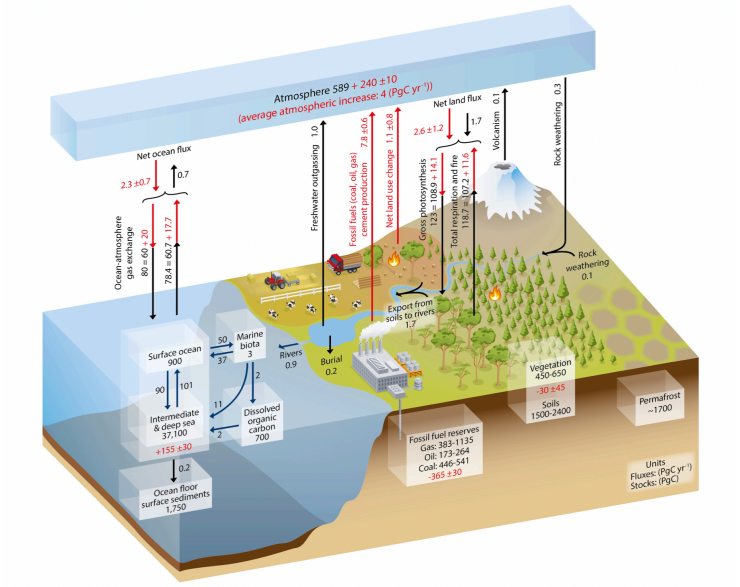
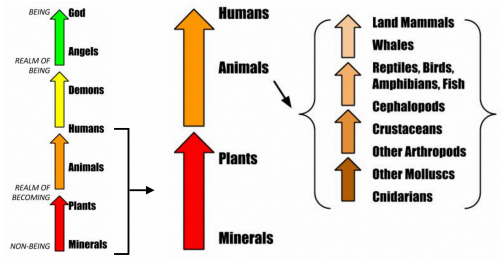


Figure 1.6: Carbon Cycle



	ability to grow and reproduce	ability to move	ability to think rationally
Humans	X	X	X
Animals	X	X	
Plants	X		
Minerals			

https://sites.google.com/site/aristotlethebiologist/aristotle-s-biology/great-chain-of-being

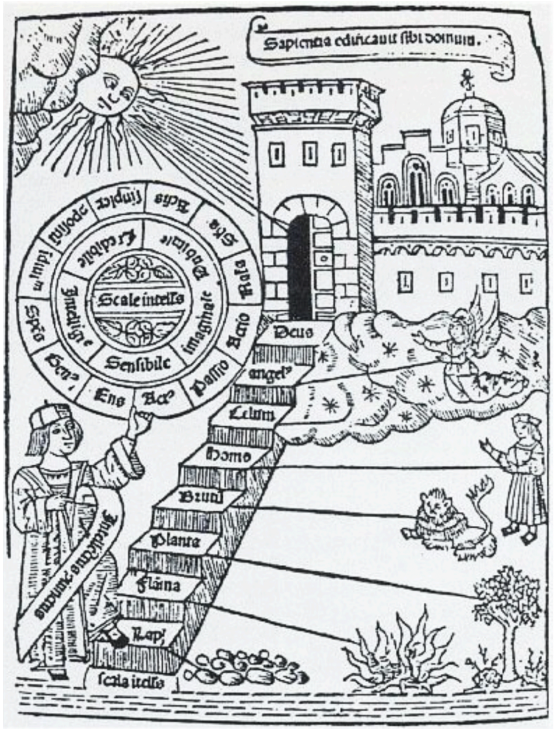


Figure 1.7: Ladder of Ascent and Descent of the Mind, 1305

https://upload.wikimedia.org/wikipedia/commons/e/e9/Die_Leiter_des_Auf-_und_Abstiegs.jpg

2. 2000 yrs later

- Edward Hitchcock
  - 1840
-

https://upload.wikimedia.org/wikipedia/commons/8/8f/Edward_Hitchcock

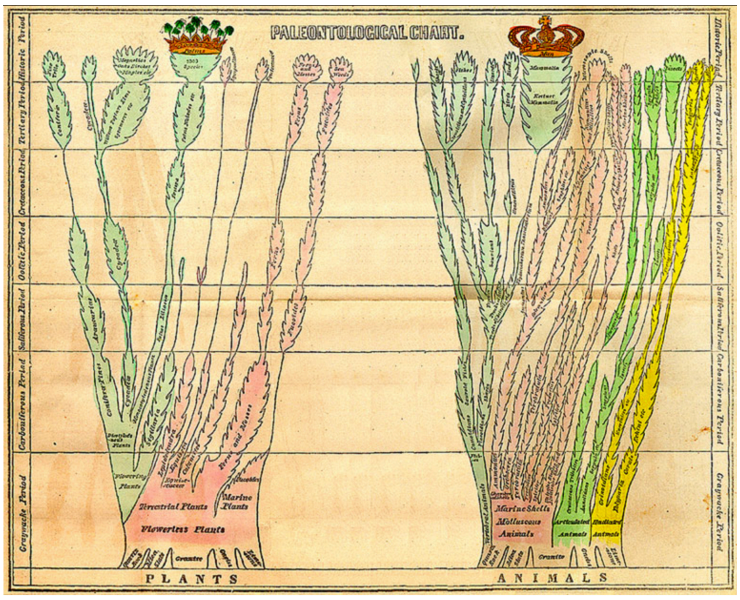


Figure 1.8: Tree of Life

_Paleontological_Chart.jpg

- Ernst Haeckel
  - 1879

https://upload.wikimedia.org/wikipedia/commons/d/de/Tree_of_life_by_Haeckel.jpg

- Charles Darwin
  - 1837
  - The idea that species could have evolved from an ancestor
  - This could have happened through transmutations
  - Premise for trees today
  - ALL METHODS DEPEND ON **OBSERVABLE MORPHOLOGICAL TRAITS** FOR CATEGORIZATION

1.3 What happened when we found out about microbes?

https://hms.harvard.edu/news/diet-gut-microbes-immunity

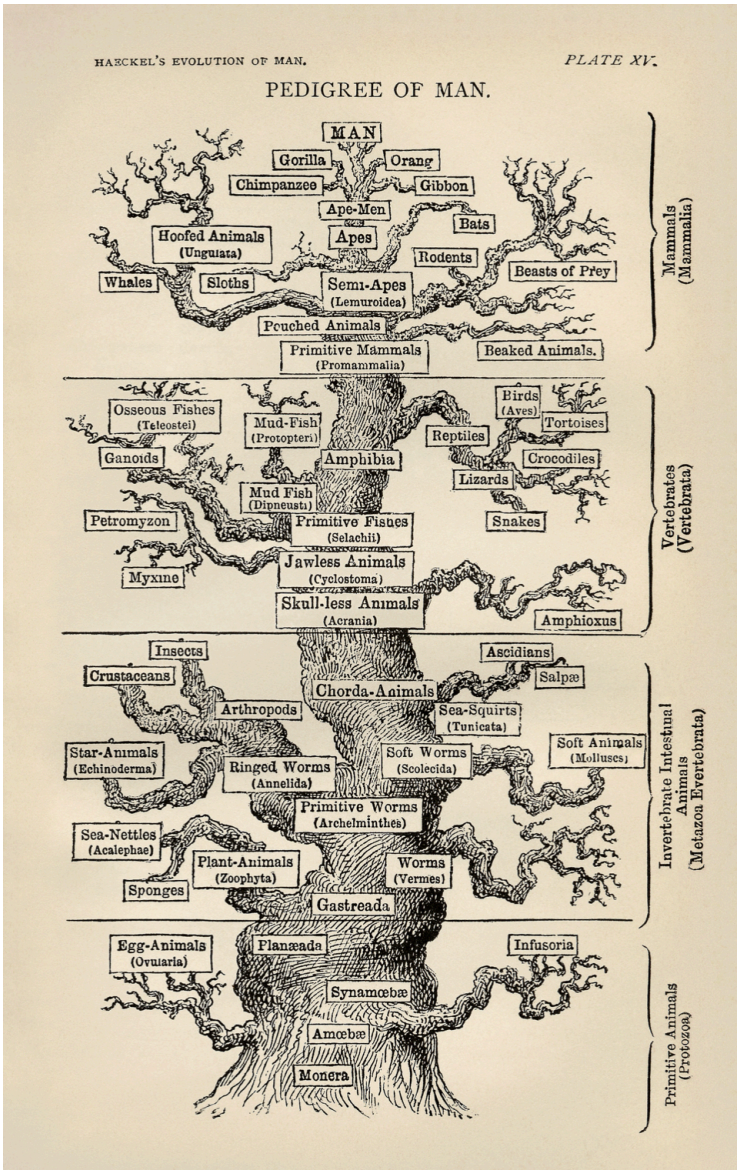


Figure 1.9: Another Tree of Life

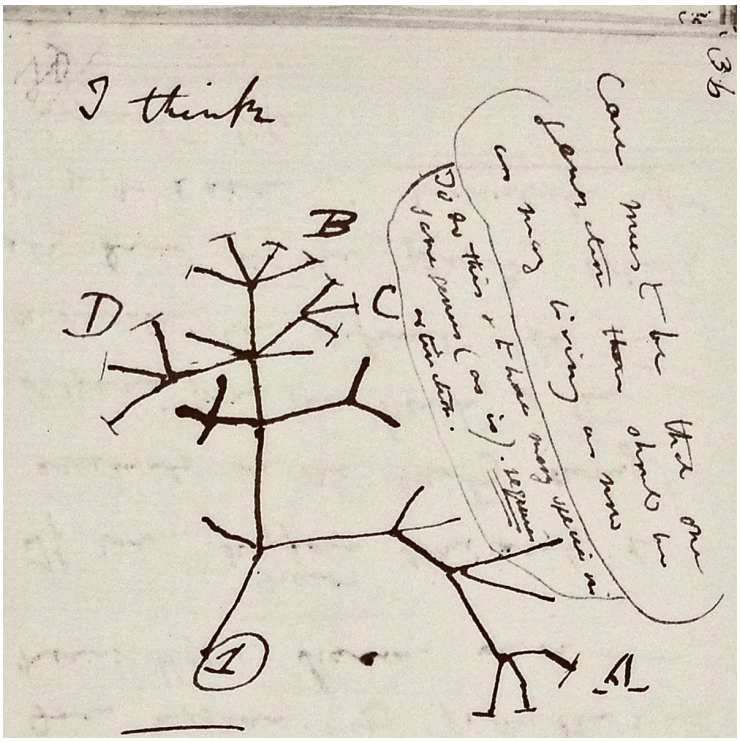


Figure 1.10: Tree of Life sketch

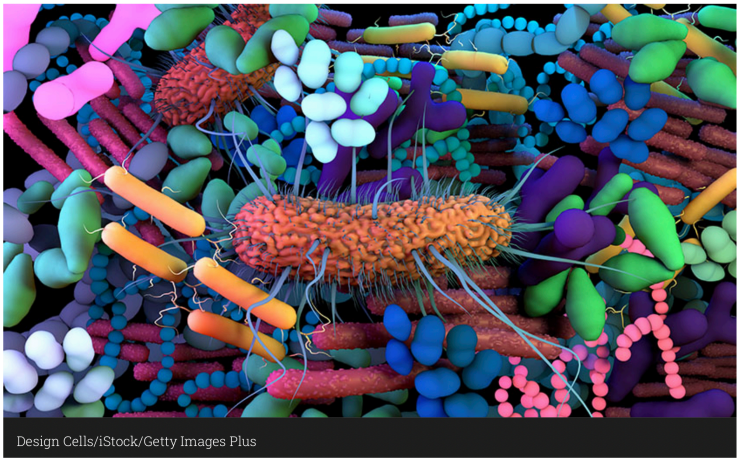


Figure 1.11: Microbes

Roadmap to where we are now with determining microbial diversity

1. Leeuwenhoek

- Father of microbiology
- Late 1600's
- Microscope

Figure 1.12: Original Microscope

2. Robert Koch

- 1890
- First time bringing microbes to the lab
- Cultivation of microbes

3. Discovery of DNA structure

- Rosiland Franklin
  - 1951

- Frederick Sanger
  - 1975

- Carl Woese
  - 1977

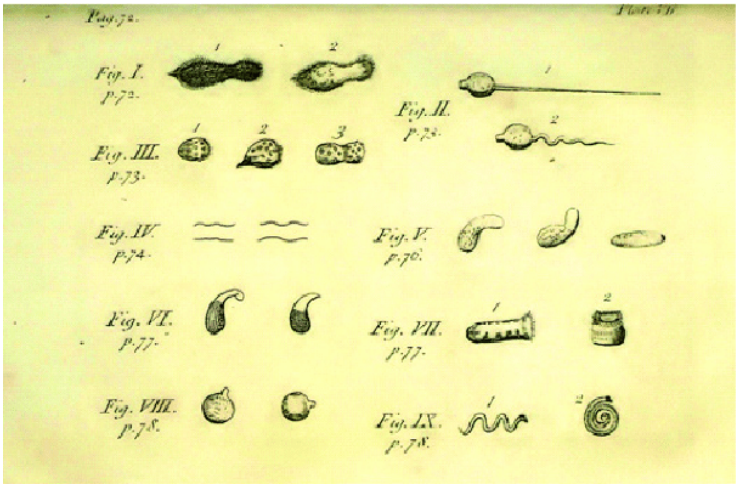


Figure 1.13: First Sketch of Microbes

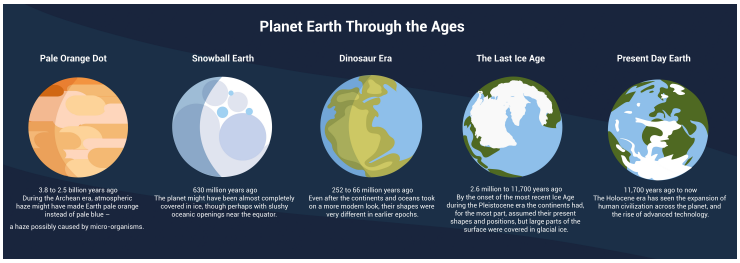


Figure 1.14: Robert Koch

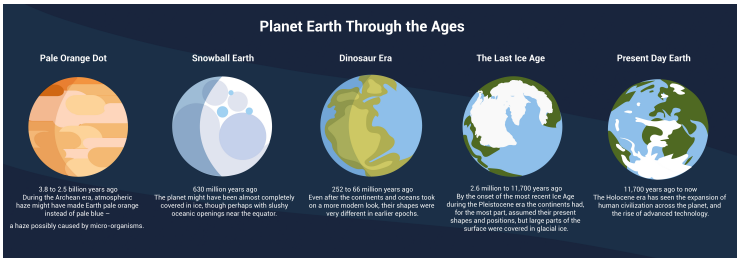


Figure 1.15: Rosiland Franklin

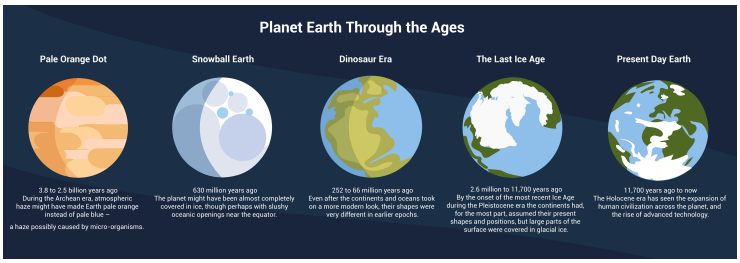


Figure 1.16: Frederick Sanger

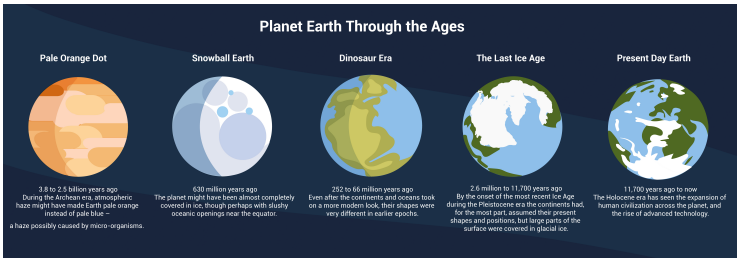


Figure 1.17: Carl Woese

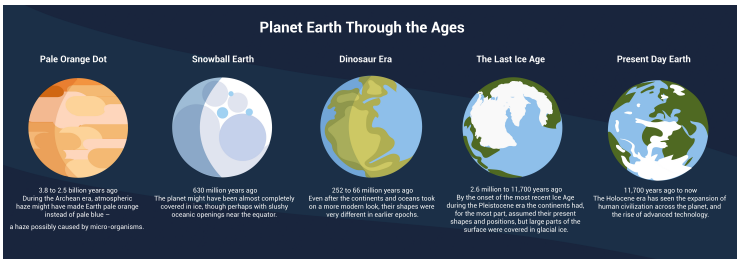


Figure 1.18: DNA Structure

DNA Structure

1.4 Tree of Life

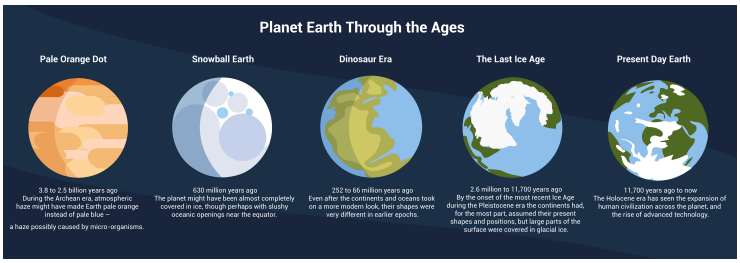


Figure 1.19: Tree of Life

"Visible organisms represent the smallest sliver of life's diversity. Bacteria are the true lords of the world. They have been on this planet for billions of years and have irrevocably changed it, while diversifying into endless forms most wonderful and most beautiful." (The Atlantic)

Life just got weird!

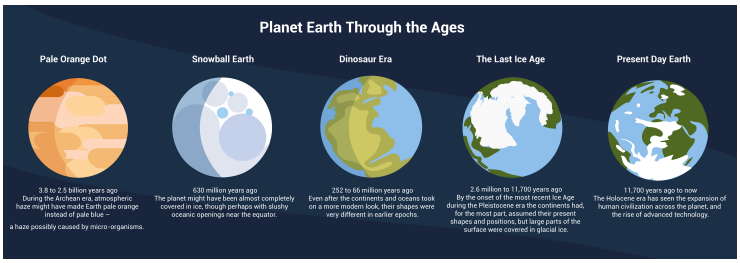


Figure 1.20: Comparing Trees of Life

1.5 What Makes Microbes so Special?

1. -15oC/4oF to 130oC/266oF temperatures
2. 0 to 12.8 pH acidity
3. More than 200 atm pressure
4. 4 miles deep into Earth's crust
5. Up to 5kGy radiation

Grand Prismatic Spring – YNP – 183oC

1. Validates the importance of microbes and sums up life on Earth with boundaries.

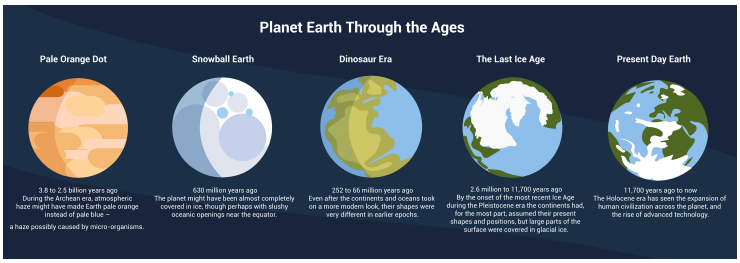


Figure 1.21: Grand Prismatic Spring

2. Microbes are constantly trying to evolve and get deeper and deeper into the hot springs

3. Eukaryotes only surround this spring – cannot survive close to the hot spring

1.5.1 The great "plate count" anomaly

1. Cultivation based cell counts are orders of magnitude lower than direct microscopic observation

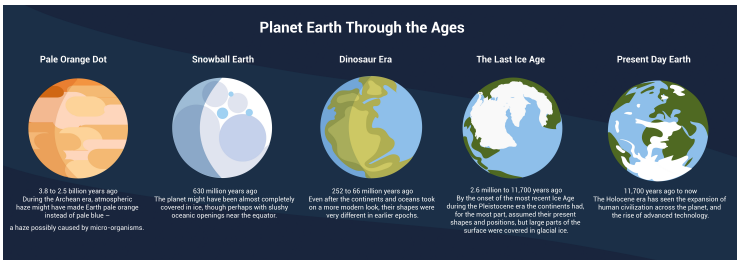


Figure 1.22: Plate Count Anomaly

2. As microbiologists, we are able to cultivate only a small minority of naturally occurring microbes

3. Our nucleic acid derived understanding of microbial diversity has rapidly outpaced our ability to culture new microbes

1.5.2 Total number of genomes at NCBI

1. Haploid genome
2. Single circular chromosome, plasmids
3. Metabolic diversity
4. Genetic malleability
5. No nucleus
6. Easy interspecies gene transfer

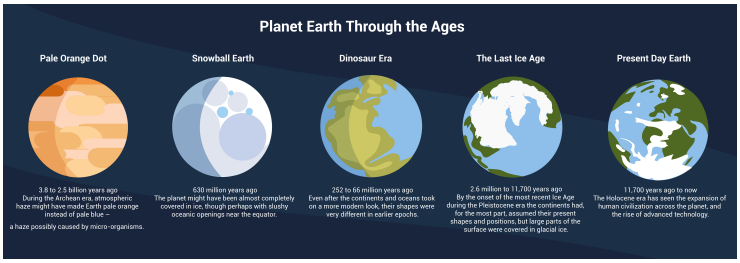


Figure 1.23: Plate Count Anomaly

https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/

1.6 Roadmap to Culture Independent Techniques

1. rRNA as an evolutionary marker

- 1977
- (Woese and Fox, PNAS)

2. Polymerase Chain Reaction

- 1985
- (K. Mullis, Science)

3. “Universal Primers” for rRNA sequencing

- 1985
- (N. Pace, PNAS)

4. PCR amplification of 16S rRNA gene

- 1989
- (Bottger, FEMS Microbiol)

5. Curation and hosting of RDP

- Early 1990’s

- (rRNA database) FTP

6. Term 'microbiome'

- 2001
- coined by Lederberg and McCray

1.7 Microbiomes and their significance

- Microbes do not work or function as a single entity
- Most microbial activities are performed by complex communities of microorganisms
  - **Microbiome**

1.7.1 What is a microbiome

1. Totality of microbes in a defined environment, and their intricate interactions with each other and the surrounding environment

- A population of a single species is a culture(monoculture), extremely rare outside of lab and in some infections
- A microbiome is a mixed population of different microbial species
- MIXED COMMUNITY IS THE NORM!

1.7.2 Why Study Microbiomes

1. Microbes modulate and maintain the atmosphere

- Critical elemental cycles (carbon, nitrogen, sulfur, iron,…)
- Pollution control, clean up fuel leaks

2. Microbes keep us healthy

- Protection from pathogens
- Absorption/production of nutrients in the gut
- Role in chronic diseases (obesity, Crohn's/IBD, arthritis…)

3. Microbes support plant growth and suppress plant disease

- Most complex communities reside in soil
- Crop productivity

1.7.3 Why is Microbiome Research New?

1. Bias for microbes (especially pathogens) that are cultivable

- Culture-based methods do not detect majority of microbes
- Only pathogens are easily detected
- And most microbes are not pathogens

2. Availability of tools

- Discovery of culture independent techniques
- Amplicon sequencing and DNA sequencing

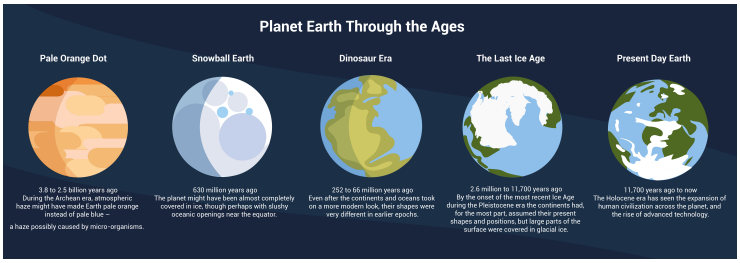


Figure 1.24: Diversity of the Microbiome, Trends in Genetics

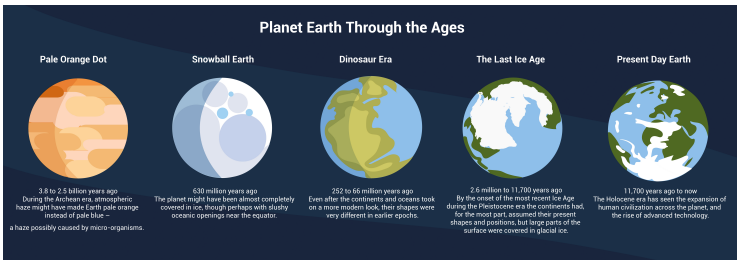


Figure 1.25: Cell 2019 Western

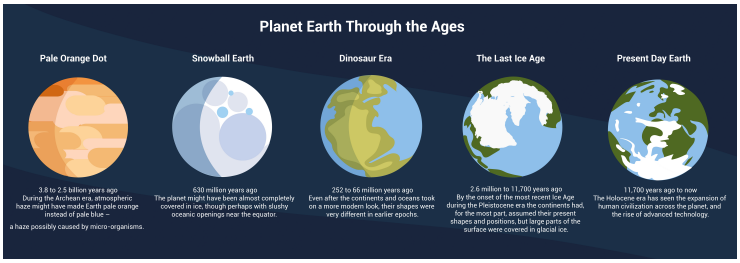


Figure 1.26: Cell 2019 Western

1. Recovered over 150,000 microbial genomes from ~10,000 metagenomes
2. 70,178 genomes assembled with higher than 90% completeness
3. 3,796 SGBs (species-level genome bins) identified -77% of the total representing species without any publicly available genomes

1.7.4 Microbiome Projects and Databases

1. American Gut Project
2. Earth microbiome Project
3. Human Oral Microbiome Database
4. CardioBiome
5. Human Microbiome Studies – JCVI
6. MetaSub – Metagenomics and metadesign of Subways and Urban Biomes
7. Gut microbiota for Health
8. NASA: Study of the impact of long term space travel in the Astronaut's microbiome
9. Michigan microbiome project
10. Coral microbiome project
11. Seagrass microbiome project

1.8 Structural and Functional Approaches to study microbiomes

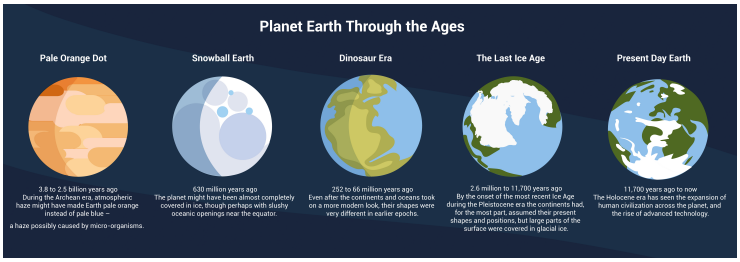


Figure 1.27: International Journal of Genomics, 2018

1.8.1 16S rRNA as an evolutionary chronometer

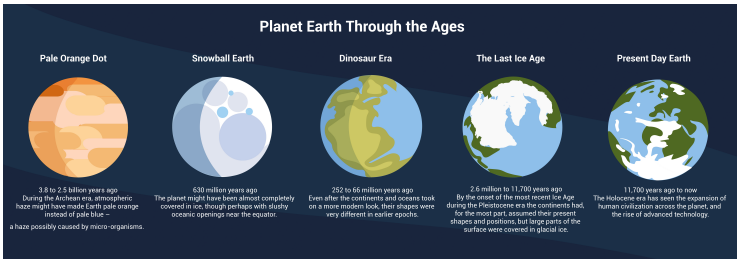


Figure 1.28: Kjellerberg et al, Microbial Ecology, 2007

1. Ubiquitous – present in all known life (excluding viruses)
2. Functionally constant wrt translation and secondary structure
3. Evolves very slowly – mutations are extremely rare
4. Large enough to extract information for evolutionary inference
5. Limited exchange – limited examples of rRNA gene sharing between organisms

1.8.2 16S rRNA vs rpoB (RNA polymerase subunit gene)

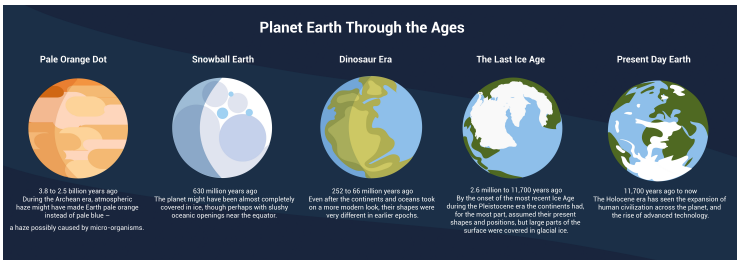


Figure 1.29: 2Kjellerberg et al, Microbial Ecology, 2007

1.8.2.1 16S rRNA hypervariable regions

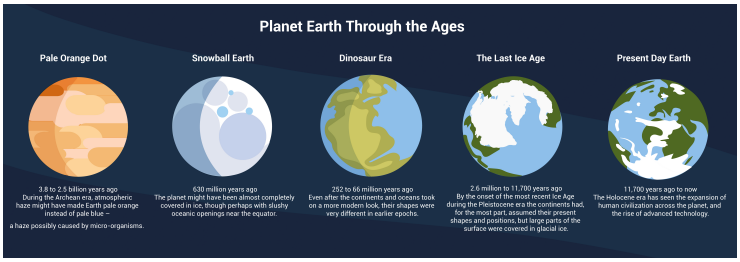


Figure 1.30: Microbiome.com

Illustration of different hypervariable regions of 16S rRNA

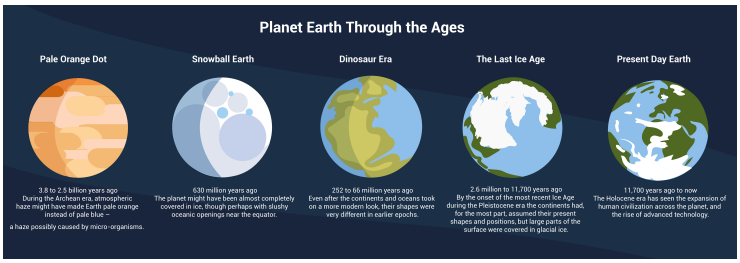


Figure 1.31: BMC Bioinf, 2016

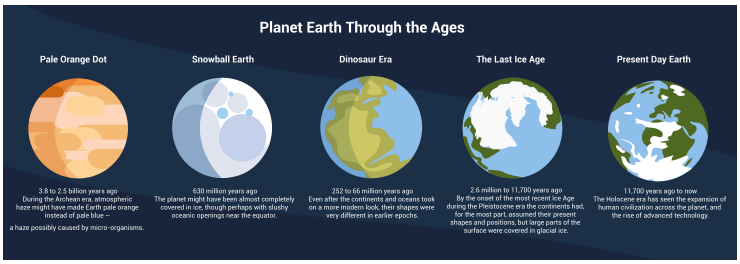


Figure 1.32: J. Investigative Dermatol, 2016

1.9 Basic Workflow for 16S Gene Based Sequencing

1.10 Addressing the 'fine print' while generating 16S rRNA gene amplicon libraries

1. Sample Collection

- Sample collection significantly influences the microbiome profiler after sequencing
- Sample storage

2. DNA isolation

- Template concentration
- Template extraction protocol

3. PCR amplification

- PCR bias and inhibitors
- Amplification of contaminants

J. Microbiol Methods (2018), App. Environ. Microbiol. (2014), Microbiome (2015)

1.11 Steps Involved

1. Experimental Design: How many samples can be included in the sequencing run?

- By using barcoded primers, numerous samples can be sequenced simultaneously (multiplexing)

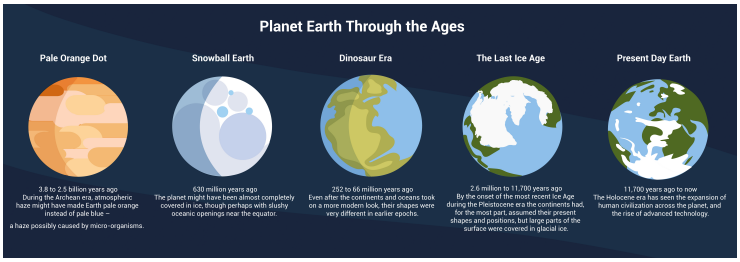


Figure 1.33: V4 Region

1.11.1 Samples

1. More the number of samples, more cost effective the run (sequencing depth will be compromised)

Comparison of multiplexing capacity by sequencing system

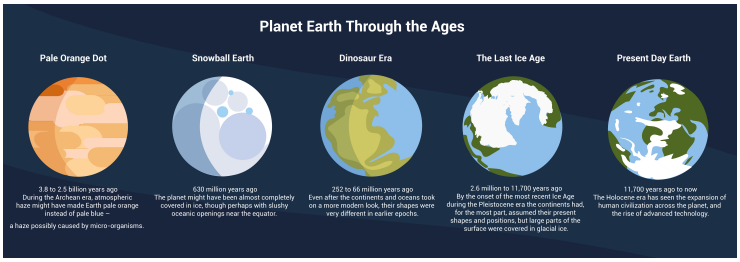


Figure 1.34: Illumina.com

2. It is critical to have a 'library prep manifest' to document the position of each sample with its associated barcode along with additional metadata information

1.11.2 Include Controls

1. Between run repeat (process any sample in duplicate per run to measure reproducibility across runs)

2. Within run repeat (process any sample in duplicate per plate to measure reproducibility)
3. Water used during PCR (water blank- to determine if any contaminant was introduced during PCR reaction)
4. Water spiked with known bacterial DNA (mock bacterial communities- enables quantification of sequencing errors, minimizes bias during sampling and library preparation)

1.11.3 DNA extraction protocol

1. Effect of mechanical lysis methods for extraction
2. Presence of inhibitors such as organic matter, humic acid, bile salts, polysaccharides
3. DNA yield post extraction and reproducibility

Effect of bead beating was larger than sampling time over 5 months

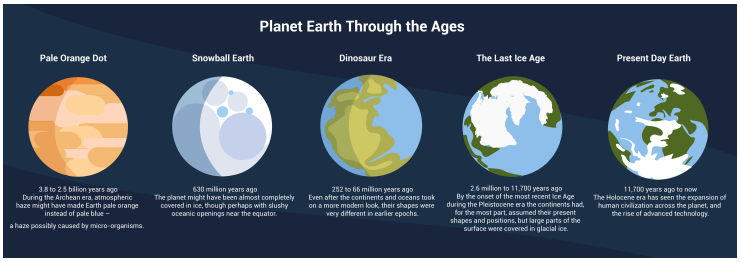


Figure 1.35: Bead beating

A. Percentage read abundance of the 11 most abundant phyla as a result of bead beating intensity B. PCA of samples with different bead beating intensities vs. samples taken at different dates

1.11.4 Selection of primers and region of 16S gene influence microbial profile

V2, V4, V6-V7 regions produced consistent results

1. V2, V3 and V6 contain maximum nucleotide heterogeneity
2. V6 is the shortest hypervariable region with the maximum sequence heterogeneity
3. V1 is best target for distinguishing pathogenic S aureus
4. V2 and V3 are excellent targets for speciation among Staph and Strep pathogens as well as Clostridium and Neisseria species
5. V2 especially useful for speciation of Mycobacterium sp. and detection of E coli O157:H7

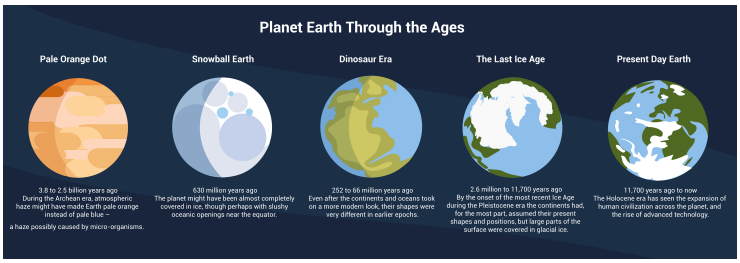


Figure 1.36: Plos

6. V3 useful for speciation of Haemophilus sp
7. V6 best target for probe based PCR assays to identify CDC select agents (bio-terrorism agents)

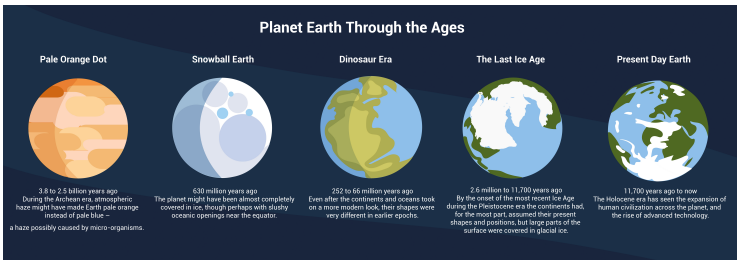


Figure 1.37: HHS

1.11.5 Purification of Amplicons

After one –step or two-step PCR, products are cleaned up using AMpure beads

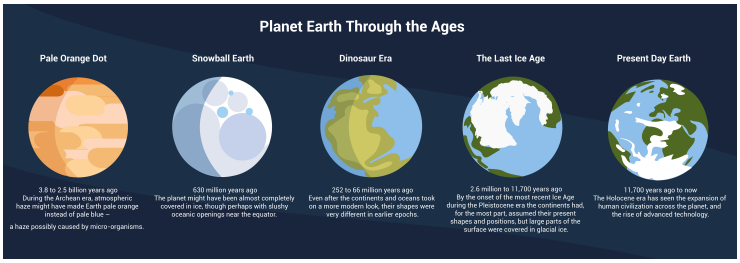


Figure 1.38: Ampure

1. Gel Electrophoresis and quantification of cleaned amplicon products

- Qubit

2. Sample pooling – equimolar concentrations (how many samples do you want to pool? How many reads per sample?
3. Gel extraction of pooled product
4. Final clean up (Qiagen kit) and QC

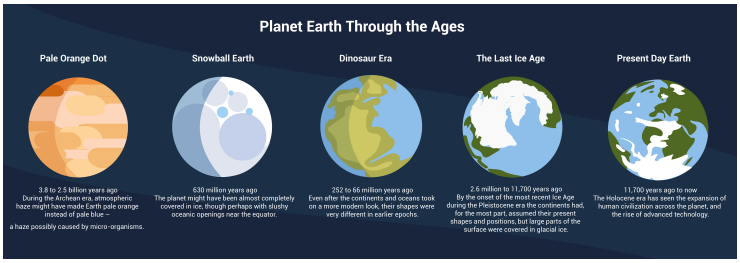


Figure 1.39: 16S Summary

Amplicon Sequencing Library Prep - PacBio

1.11.5.1 Overview of generic amplicon workflow

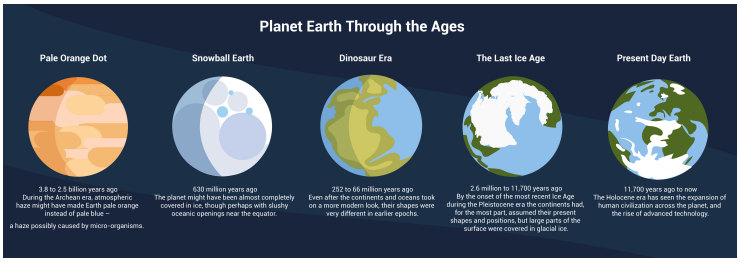


Figure 1.40: Amplicon Workflow

Data Processing

1.11.6 FastQC

1. Many tools/options to filter and trim data
2. Trimming does not always improve things as valuable information can be lost!
3. Removal of adapters is critical for downstream analysis

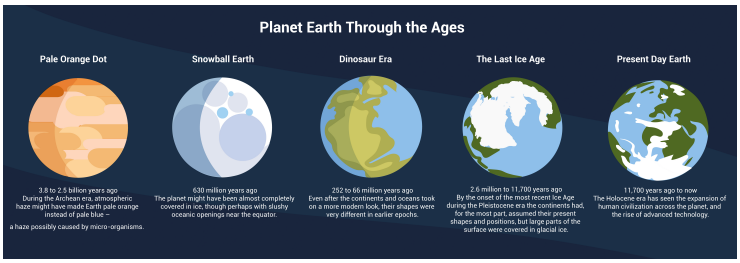


Figure 1.41: FastQC

1.11.7 Dereplication

1. In this process all the quality-filtered sequences are collapsed into a set of unique reads, which are then clustered into OTUs
2. Dereplication step significantly reduces computation time by eliminating redundant sequences

1.11.8 Chimera detection and removal of non-bacterial sequences

Chimeras as artifact sequences formed by two or more biological sequences incorrectly joined together

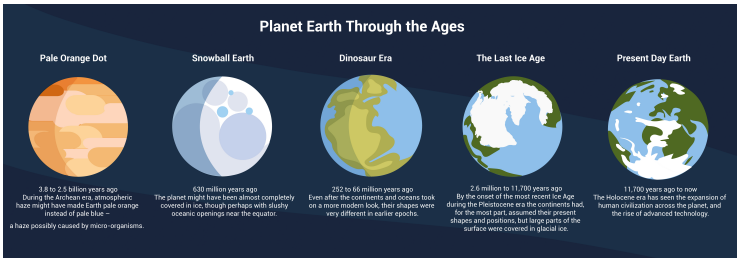


Figure 1.42: Chimera

Incomplete extensions during PCR allow subsequent PCR cycles to use a partially extended strand to bind to the template of a different, but similar, sequence. This partially extended strand then acts as a primer to extend and form a chimeric sequence.

1.11.9 Clustering

1. Analysis of 16S rRNA relies on clustering of related sequences at a particular level of identity and counting the representatives of each cluster

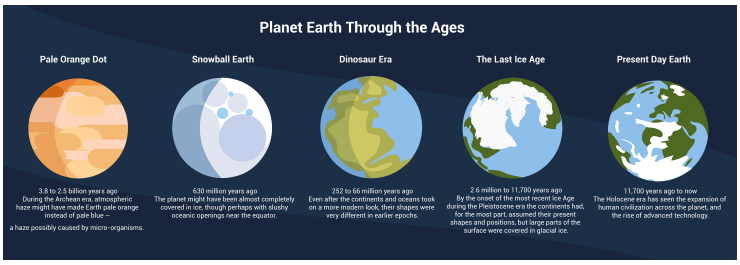


Figure 1.43: Clustering

Some level of sequence divergence should be allowed – 95% (genus-level, partial 16S gene), 97% (species-level) or 99% typical similarity cutoffs used in practice and the resulting cluster of nearly identical tags (assumedly identical genomes) is referred to as an OTU (Operational Taxonomic Unit)

1.11.10 Create OTU tables

OTU table is a matrix that gives the number of reads per sample per OTU

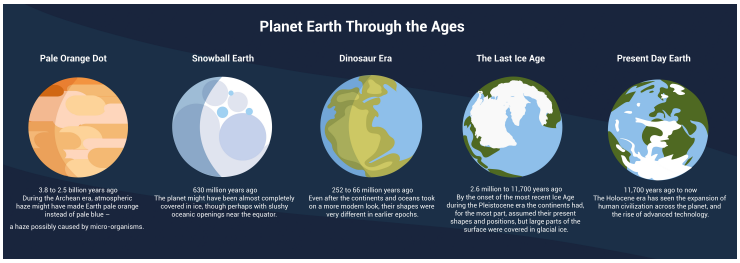


Figure 1.44: OTUs

1.11.11 Bin OTUs into Taxonomy (assign taxonomy)

1. Accuracy of assigning taxonomy depends on the reference database chosen

- Ribosomal Database Project
- GreenGenes
- SILVA

2. Accuracy depends on the completeness of databases

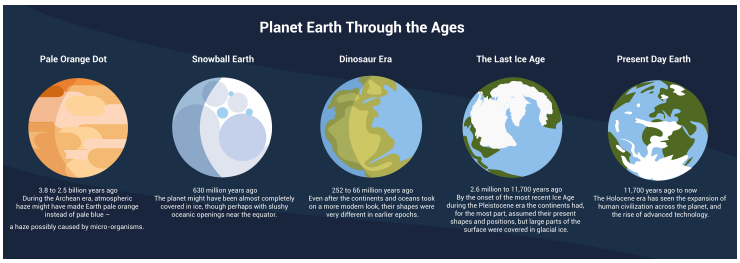


Figure 1.45: Database

1.11.12 Assess Population Diversity: alpha diversity

1. Assessment of diversity involves two aspects

- Species richness (# of species present in a sample)
- Species evenness (distribution of relative abundance of species)

2. Total community diversity of a single sample/environment is given by alpha-diversity and represented using rarefaction curves

3. Quantitative methods such as Shannon or Simpson indices measure evenness of the alpha- diversity

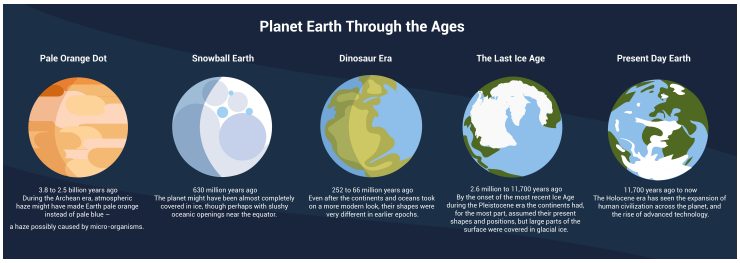


Figure 1.46: Human Mol. Genet., 2013

1.11.13 Assess Beta Diversity

1. Beta-diversity measures community structure differences (taxon composition and relative abundance) between two or more samples

- For example, beta-diversity indices can compare similarities and differences in microbial communities in healthy and diseases states

2. Many qualitative (presence/absence taxa) and quantitative(taxon abundance) measures of community distance are available using several tools

- LIBHUFF, TreeClimber, DPCoA, UniFrac (QIIME)

1.11.14 Measuring Population Diversity: alpha and beta diversity

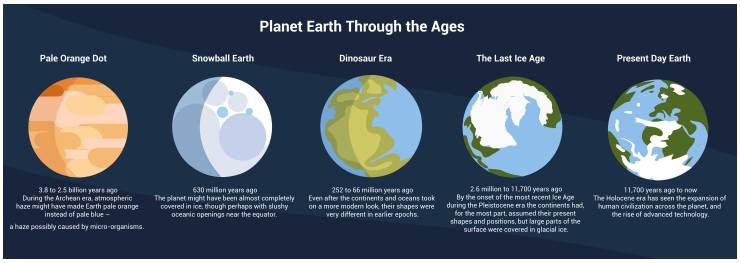


Figure 1.47: PLoS Computational Biol.,2012

1.12 Diversity Measurements with 16s rRNA sequencing

1. Overall Benefits

- Cost effective
- Data analysis can be performed by established pipelines
- Large body of archived data is available for reference

2. Overall Limitations

- Sequences only a single region of the genome
- Classifications often lack accuracy at the species level
- Copy number per genome can vary. While they tend to be taxon specific, variation among strains is possible
- Relative abundance measurements are unreliable because of amplification biases
- Diversity of the gene tends to overinflate diversity estimates

3. FastQC for 16S rRNA dataset

- Extremely biased per base sequence content
- Extremely narrow distribution of GC content
- Very high sequence duplication levels
- Abundance of overrepresented sequences
- In cases where the PCR target is shorter than the read length, the sequence will read through into adapters

1.13 QIIME 2

Importing data Demultiplexing Running Quality Control Creating a feature table Building a phylogenetic tree Calculating core diversity metrics Testing alpha diversity group significance and correlation Performing beta diversity ordination Testing beta diversity group significance Assigning taxonomies Performing differential abundance analysis with ANCOM and/or gneiss

1.14 Taxonomy: Expectation vs Reality

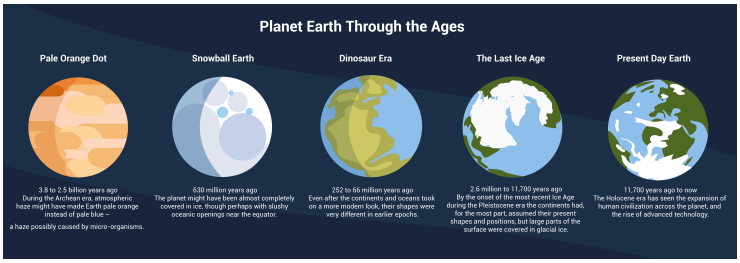


Figure 1.48: Expectation vs. Reality

1.15 Beta Diversity - UniFrac

1. Measures how different two samples' component sequences are

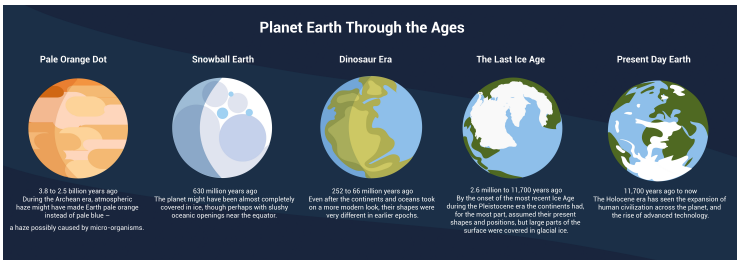


Figure 1.49: Unifrac

2. Weighted Unifrac: takes abundance of each sequence into account

1.16 Results from Paper

1. Main phyla: Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, Fusobacteria with differences bw samples
2. Sputum (patient) samples had highest diversity followed by oropharynx samples followed by nasal
3. Healthy controls (N and O) more diverse than samples from TB patients
4. Between-group comparisons?
5. Phyla differences?

```
knitr::write_bib(c(
  .packages(), 'bookdown', 'knitr', 'rmarkdown'
), 'packages.bib')
```

Chapter 2

Filter out Red Alder Sequencing Reads

In order to focus on sequencing reads from the microbes in the nodule, we will filter out reads that align to the red alder genome as follows:

1 Align the fastq-formatted reads to the red alder genome using minimap2. 2 Extract reads that do not align to red alder and sort them using samtools. 3 Create a fastq file with only the unaligned reads using samtools bam2fastq. 4 Compress the fastq file using gzip.

- Activate the environment that contains minimap2 and samtools

```
conda activate seqtools
```

- Make a directory and go into it

```
mkdir ~/microbe_fastq
cd ~/microbe_fastq
```

- Link to the merged minion reads

```
ln -s /home/cjb/minion/2022/data/minit4/data/AlderNodule3469-3/3469-3/20220701_2144_MC-113445_FAS
```

- Run Minimap2 to align the MinION reads to the red alder genome
  - he -x map-ont parameter (allows ~10% error + divergence)

```
minimap2 -x map-ont -L -t 8 -a \
/home/cjb/minion/2019/indexes/red_alder_genome/consensus.fasta \
3469-3.all.fastq > 3469-3-minionxredalder.mm2.sam
```

- Convert the unmapped reads in the alignment file (sam) to a fastq file
  - The -f4 includes only reads with the 4 flag (unmapped)

```
samtools fastq -f4 3469-3-minionxredalder.mm2.sam > 3469-3.microbe.fq
```

- Compress the new fastq file
  - (note that it will automatically add the extension .gz)

```
gzip 3469-3.microbe.fq
```

Now run these steps with 4956-3 (minit3)

Reads are here:

/home/cjb/minion/2022/data/minit3/data/aldernodule4956-3/4956-3/20220701_2154_MC-113286_FAS37509_f3119554/fastq_pass/all.fastq

Chapter 3

Meta 3

Chapter 4

Incremental Graphs with *minigraph*

4.1 Minigraph Functions

Constructs graphs

Maps sequences to graphs

https://github.com/lh3/minigraph

4.2 Minigraph Overview

Finds approximate positions of diverged sequences and adds them into the graph

1-to-1 orthogonal regions

4.3 Minimizers and Minimap2

Minigraph is built off of ideas and code from Minimap2.

Minimizers

From the minimap2 usage: "A minimizer is the smallest k-mer in a window of w consecutive k-mers."

Essentially a minimizer tags a sequence window with the kmer that is first alphabetically.

A nice tutorial:
+ https://homolog.us/blogs/bioinfo/2017/10/25/intro-minimizer/

Minimap2

Minimap2 is a fast sequence aligner. It can align short or long reads or assemblies against a reference using the seed-chain-align approach that many aligners employ. It finds exact matches (anchors) between query minimizers (seeds) and indexed reference minimizers. It links colinear anchors together (chains). For nt-level alignment it fills in regions between anchors within chains and between chains (align).

https://github.com/lh3/minimap2
https://academic.oup.com/bioinformatics/article/34/18/3094/4994778

4.4 Pipeline

1. Prepare the input
2. Build graphs

- Build a single genome graph and incrementally add more sequences
- Or build a graph for all sequences at once

3. View with Bandage

4.5 Yeast Assemblies

12 Mb
16 chromosomes

Yeast Population Reference Panel (YPRP) https://yjx1217.github.io/Yeast_PacBio_2016/data/

12 Yeast PacBio Assemblies (Chromosome level)

1. ~100-200x PacBio sequencing reads
2. HGAP + Quiver polishing

3. ~200-500x Illumina (Pilon correction)

4. Manual curation

5. Annotation

Make sure your chromosome names are unique across all samples and that they contain the sample name. We're using <strain name>.<chromosome> (>S288C.chrVIII)

4.6 Prepare the Input

1. Make sure you're working in a **screen**
2. Make Directory

```
mkdir --parents ~/pangenomics/minigraph
```

3. Navigate to the Directory

```
cd ~/pangenomics/minigraph
```

4. Link to data

```
ln -s /home/pangenomics/data/yprp/assemblies/*.fa .
```

4.7 Graphical Fragment Assembly (GFA) format

Tab-delimited text

Lines start with one of the following types:

Type	Description	Explanation
H	Header	
S	Segment	A continuous sequence or subsequence
L	Link	Segment overlaps (basepairs & orientations)
J	Jump	Jumps link sequences across gaps
C	Containment	Segment contained in another segment
P	Path	An ordered & oriented list of linked segments
W	Walk	An ordered & oriented list of segments w/o overlaps
#	Comment	

Optional Fields TAG:TYPE:VALUE
http://gfa-spec.github.io/GFA-spec/GFA1.html

Let's go to the specification to look at optional fields and an example:
http://gfa-spec.github.io/GFA-spec/GFA1.html

4.8 *reference* Graphical Fragment Assembly (rGFA)

https://github.com/lh3/gfatools/blob/master/doc/rGFA.md

- Strict subset of GFA
- Tags that trace origin
- Stable coordinates

4.9 Build Graphs

1. The first sequence in the graph is used as a "reference":

```
minigraph -xggs reference.fa -t 20 > ref.minigraph.gfa
```

- **-xggs**
  - build a graph using a simple (ggs) algorithm
- **-t 20**
  - use 20 threads

2. Incrementally add strains to graph:

```
minigraph -xggs ref.minigraph.gfa strain_1.fa > strain_1.minigraph.gfa

minigraph -xggs strain_1.gfa strain_2.fa > strain_2.minigrpah.gfa
...

minigraph -xggs string_N-1.gfa strain_N.fa > yprp.minigraph.gfa
```

3. Or build a graph for all strains at once:

```
minigraph -xggs ref.minigraph.gfa strain_*.fa > yprp.minigraph.gfa
```

4.10 Reference Graph

Activate the environment

```
source activate pangenomics
```

Now that you know how to build a minigraph, try to create a graph for the reference (S288C) and answer the following questions.

1. Make the reference graph

2. How many lines in the gfa file?

3. What type of lines are they?

Reference Graph Commands

1. Make the reference graph

```
minigraph -xggs S288C.genome.fa > ref.minigraph.gfa
```

2. How many lines in the gfa file?

```
wc -l ref.minigraph.gfa
```

3. What type of lines are they?

```
cut -f 1 ref.minigraph.gfa | sort | uniq -c
```

Reference Graph Bandage Visualization

Download your gfa file onto your computer and upload it into Bandage.

4.11 YPRP Graphs

1. Add in the rest of the lines

- we'll do this alphabetically
- capture the stderr

```
minigraph -xggs ref.minigraph.gfa $(ls yprp/assemblies/*fa | grep -v S288C) > yprp.minigraph.gfa
```

Note: We can simply use the reference fasta instead of a gfa

```
minigraph -xggs /home/pangenomics/data/yprp/assemblies/S288C.genome.fa $(ls yprp/assemblies/*fa |
```

Try to answer the following questions:

1. How many lines are in the gfa file?

2. What type of lines are they?

3. How many yeast assemblies have inversions compared to S288C (hint: look in the stderr)?

YPRP Commands

1. How many lines are in the gfa file?

```
wc -l yprp.minigraph.gfa
```

2. What type of lines are they?

```
cut -f 1 yprp.minigraph.gfa | sort | uniq -c
```

3. How many yeast assemblies have inversions compared to S288C (hint: look in the stderr)?

```
grep inv yprp.minigraph.err
```

YPRP Graph Statistics

```
export PATH=$PATH:/home/vesw/gfatools/
gfatools stat yprp.minigraph.gfa
```

Number of segments: 2729

Number of links: 3864

Number of arcs: 7728

Max rank: 11

Total segment length: 13243550

Average segment length: 4852.895

Sum of rank-0 segment lengths: 12157149

Max degree: 7

Average degree: 1.416

YPRP Graph in Bandage

Take a look at the YPRP graph in Bandage. Your might be rendered differently.

4.12 Structures in the graph

Insertions and Diverged Regions

Zoom in on segment s1054.

Trace the S288C path (hint: the S288C segments are numbered sequentially).

Identify insertions and regions that have diverged.

Group Exercise

1. Find a simple and a complex region
2. Discuss it in your group
3. Share it with everyone
4. Keep track of the segments

Inversions

Here are some examples of inversions.

Let's find them in bandage.

s1289 LN:i:24089 SN:Z:S288C.chrXIV SO:i:567595

s1672 LN:i:257 SN:Z:DBVPG6044.chrV SO:i:446168

Click on the links (black) to see the direction that paths can travel.

Do a web-blast in Bandage to see what they might code for.

Inversions in the GFA

I found the inversions by searching the graph GFA file for pairs of segments that have two links between them.

```
grep '^L' yprp.minigraph.gfa | awk '{print $2 "\t" $4}'|sort|uniq -c|awk '$1>=2{print}'
```

2 s1288 s1289
2 s1289 s1290
2 s1671 s1672
2 s1672 s1673

1. Pull out s1289, its adjacent segments and the links connecting them

```
gfatools view -l s1289 -r 1 yprp.minigraph.gfa
```

-l STR/@FILE. segment list to subset [ ]
-r INT. subset radius (effective with -l) [0]

Showing just the links here:
L s1288 + s1289 + 0M SR:i:0 L1:i:4105 L2:i:24089
L s1288 + s1289 - 0M SR:i:5 L1:i:4105 L2:i:24089
L s1289 + s1290 + 0M SR:i:0 L1:i:24089 L2:i:1511
L s1289 - s1290 + 0M SR:i:5 L1:i:24089 L2:i:1511

2. Now do the same for s1672

3. Try extending the surrounding region by increasing the -r parameter

4.13 Minigraph Blog

Heng Li http://lh3.github.io/2021/01/11/minigraph-as-a-multi-assembly-sv-caller

4.14 Bonus Questions

1. What is the longest segment in the graph? [Hint: Parse out the number from the 4th field of the segment line]
2. What is the shortest segment in the graph?
3. What cigar strings exist for the overlaps in the links? [Hint: Field 6 of the link line]
4. How many segments are attributed to each genome? [Hint: Parse out field 5 of the segment line]

Commands for Bonus Questions

1. What is the longest segment in the graph?

```
grep "^S" yprp.minigraph.gfa |cut -f 4|sed 's/.\+://'|sort -n | tail -1
```

2. What is the shortest segment in the graph?

```
grep "^S" yprp.minigraph.gfa |cut -f 4|sed 's/.\+://'|sort -n | head -1
```

3. What cigar strings exist for the overlaps in the links? [Hint: Field 6 of the link line]

```
grep "^L" yprp.minigraph.gfa |cut -f 6|sort -u
```

4. How many segments are attributed to each genome?

```
grep "^S" yprp.minigraph.gfa |cut -f 5|sed 's/SN:Z://'|sed 's/\..\+//'|sort|uniq -c
```

Note: There are lots of ways to do this.

4.15 Graph to Fasta

We can convert the gfa graph file to a fasta file the represents the sequence of the pangenome.

> Fasta format:
> >header
> ACGCGCTAGCGCGAC
> ACGGCGTAGGGGCAG
> ACGGCT

```
gfatools gfa2fa -s yprp.minigraph.gfa > minigraph.stable.fa
```

FASTA questions

Answer the following questions:

1. How many sequences?

2. Take a look at the headers

FASTA Commands

1. How many sequences?

```
grep -c '>' minigraph.stable.fa
```

2. Take a look at the headers

```
grep '>' minigraph.stable.fa|less
```

4.16 GAF format

"The only visual difference between GAF and PAF is that the 6th column in GAF may encode a graph path like >MT_human:0-4001<MT_orang:3426-3927 instead of a contig/chromosome name."

https://github.com/lh3/minigraph

Let's look at PAF format https://lh3.github.io/minimap2/minimap2.html

4.17 Read Mapping

Align reads from SK1 to the minigraph

21,906,518 paired Illumina reads
Read length = 151 nts

```
minigraph -x sr yprp.minigraph.gfa /home/pangenomics/data/yprp/reads/SK1.illumina.fast
```

-x sr map short reads (sr)

4.18 Read Mapping Stats

Ideally we would convert from GAF to GAM using vg convert the calculate stats with vg stats but it doesn't work.

Count the number of primary alignments

```
grep -c "tp:A:P" SK1.mapped.gaf
```

18092858 primary alignments

Calculate the percent of reads that had alignments

18092858/21906518 = 82.59% of reads aligned

4.19 Structural Variant Calling

Call structural variants with gfatools (doesn't work with VG graphs):

```
gfatools bubble yprp.minigraph.gfa > yprp.minigraph.structural.bed
```

https://github.com/lh3/minigraph

Structural Variant Stats

1. Total number of variants:

```
wc -l yprp.minigraph.structural.bed
```

2. Indels (the shortest path is 0)

```
awk '$7==0{print}'
yprp.minigraph.structural.bed|wc -l
```

3. Inversions

```
awk '$6==1{print}' yprp.minigraph.structural.bed | cut -f 1-12
```

4.20 CUP1

Visualize the CUP1 region

10 working copies + 1 pseudogene in S288C

1. Find the region in the graph based on its S288C coordinates
   S288C.chrVIII:213045-233214

```
gfatools view -R S288C.chrVIII:213045-233214 yprp.minigraph.gfa > cup1.gfa
```

2. Create a .csv to bring in the segment names
   Note that you need a header

```
 cat <(echo "Segment,Name") <(grep "^S" cup1.gfa | awk '{print $2 "," $5}') > cup1.csv
```

3. Load the graph and the .csv file into Bandage

This compares S288C and SK1.

If you have blast installed on your computer, you can blast the two gene sequences to their positions. CUP1 is the smaller one. Gene sequences are in: /home/pangenomics/data/yprp/genes/

CUP1 Paths in Y12

Let's find the Y12 paths through the graph for all bubbles in the CUP1 graph file.

```
minigraph -xasm -l100 --call cup1.gfa /home/pangenomics/data/yprp/assemblies/Y12.genome.fa > Y12.
```

Output S288C.chrVIII 213609 233593 >s720 >s722 >s2512>s2052>s2513>s2054>s2514>s2056>s2515>s2058:11410:

alignment path through the bubble:path length:mapping strand:the contig name:approximate contig start:approximate contig end

CUP1 Paths in all yeast genomes

Let's do all the samples:

```
for i in /home/pangenomics/data/yprp/assemblies/*.fa; do
   bn=`basename $i .fa`
   minigraph -xasm -l100 --call cup1.gfa $i > $bn.call.bed
done
```

Compare to the Bandage Graph

4.21 Minigraph Pros and Cons

Pros

Captures length variation
Efficient
Easy to add new genomes

Cons

Lack of base level alignment + sample input order dependency =
+ suboptimal mappings
+ suboptimal local graphs
Needs collinear chains so it doesn't work well with many short segments such as rare SNPs.

https://github.com/lh3/minigraph#limitations

"Please do not overinterpret complex subgraphs. If you are interested in a particular subgraph, it is recommended to extract the input contig subsequences involved in the subgraph with the –call option and manually curated the results."

4.22 Blog Battle

Heng Li (Minigraph)

https://lh3.github.io/2019/07/08/on-a-refere nce-pan-genome-model

Erik Garrison (VG)

https://ekg.github.io/2019/07/09/Untangling -graphical-pangenomics

Heng Li (Minigraph)

- https://lh3.github.io/2019/07/12/on-a-refere nce-pan-genome-model-part-ii

4.23 Exercises

Start with another reference

1. What reference did you choose?
2. What order are the other samples in?
3. How does the graph compare?
4. How does read mapping compare?
5. How do structural variant calls compare?
6. How does the cup1 region compare?
7. Any other interesting differences?

Another Yeast Dataset

A subset of yeast genomes from: https://www.nature.com/articles/s41586-018-0030-5.pdf

Data are in: /home/pangenomics/data/1011yeast/assemblies/*fa.gz

1. How many sequences in each assembly? Min? Max?
2. Make and characterize a minigraph

- Choose 13 lines to match the number of genomes we ran earlier
- Try all 127 assemblies

3. How do these graphs compare to our previous yeast graph?
4. Pick a region from one of the graphs and make and characterize a subgraph.

Human GFA

lipoprotein(a) - LPA

See if you can pull out and visualize the LPA region pictured below from two human GFA files from different versions of the human reference (GRCh38.p13 and CHM13). BLAST the LPA gene against the graphs in Bandage.

GFA files are in /home/pangenomics/data/1011yeast/assemblies/. The LPA gene sequence is in /home/pangenomics/data/human/genes/

Approximate Positions:

- GRCh38.p13 chr6 160000000-161000000 complement
- CHM13 chr6 161200000-162200000

Your results should look something like this.

Convert to VG and call variants

Convert minigraph to vg (<1min):

```
vg convert -g yprp.minigraph.gfa -v -t 20 > yprp.minigraph.vg
```

-g input GFA graph **-v** output VG graph **-t 20** use 20 threads

Make vertices small enough (<=1024bp) for indexing (<1min):

```
vg mod -X 256 yprp.minigraph.vg -t 20 > yprp.minigraph.mod.vg
```

-X max node size **-t 20** use 20 threads

NOTE: Converting to VG isn't required if not calling variants, i.e. you can index and map directly on GFA.

1. Use vg to index the VG graph (2min)
2. Use vg to map SK1 reads to minigraph GFA (17min)
3. Use vg to call variants on read mapping GAM
    a. pack (20min)
    b. call (<1min)
    c. don't do augment; run-time too long!

Chapter 5

Meta 5

Chapter 6

Meta 6

Chapter 7

Meta 7

Chapter 8

Meta 9

Chapter 9

QIIME2

Command Line Tutorial

9.1 Logging on to the server

Make it a practice to start a screen before you begin analysis.

```
ssh -p 44111 <USERNAME>@gateway.training.ncgr.org

screen -S QIIME2
```

9.2 Setting up a working directory for QIIME2 analysis

```
cd /home/$USER/

mkdir QIIME2_Analysis

cd QIIME2_Analysis
```

9.3 Dataset for analysis

For this tutorial, we will use data from the following published manuscript: ht
tps://microbiomejournal.biomedcentral.com/articles/10.1186/2049-2618-2-29