QIIME2

Hands On Command Line Tutorial

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

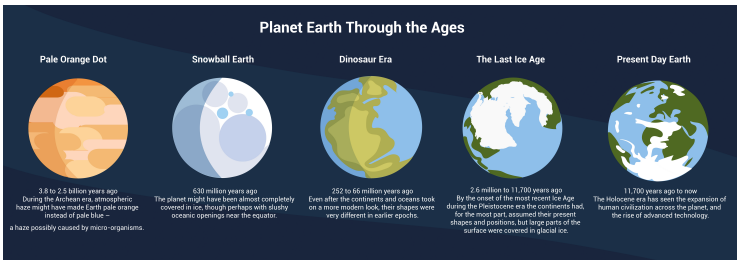

Figure 1.1: Earth. Image credit: NASA/JPL-Caltech/Lizbeth B. De La Torre

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

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

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

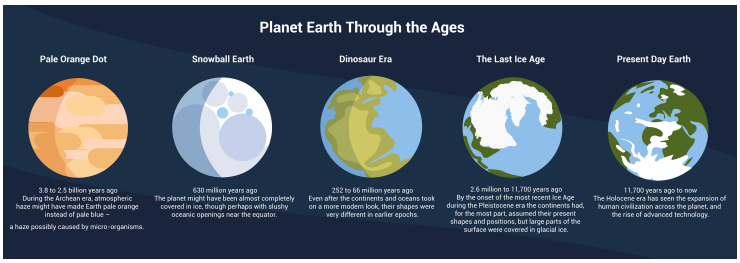

Figure 1.2: Earth. Image credit: NASA/JPL-Caltech/Lizbeth B. De La Torre

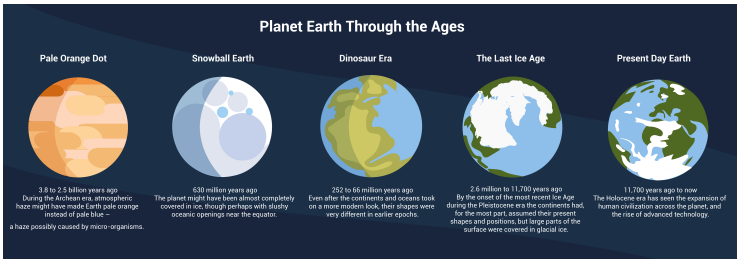

Figure 1.3: Earth. Image credit: NASA/JPL-Caltech/Lizbeth B. De La Torre

6. 200 MYA mammals emerged
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

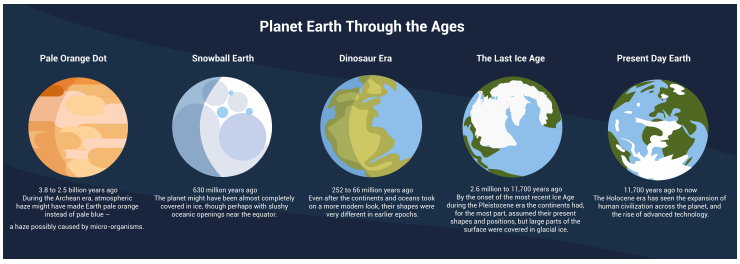

Figure 1.4: Earth. Image credit: NASA/JPL-Caltech/Lizbeth B. De La Torre

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

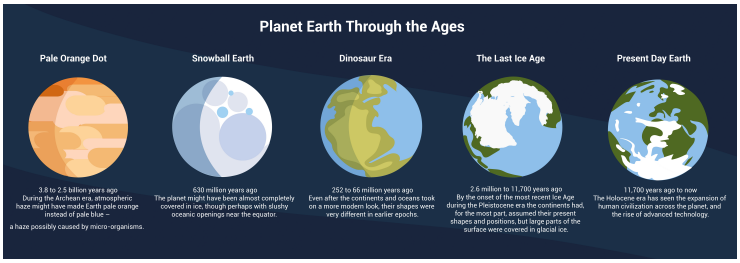

Figure 1.5: Nitrogen Cycle

1.1.1.2 Carbon Cycle

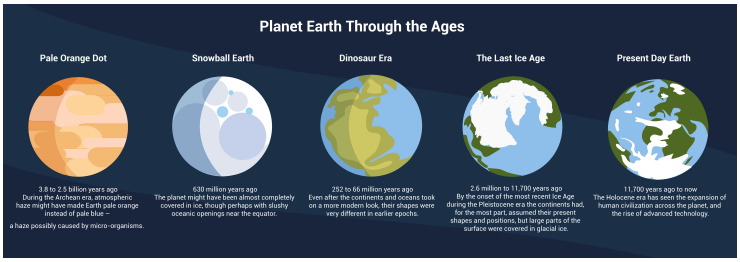

Figure 1.6: Carbon Cycle

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

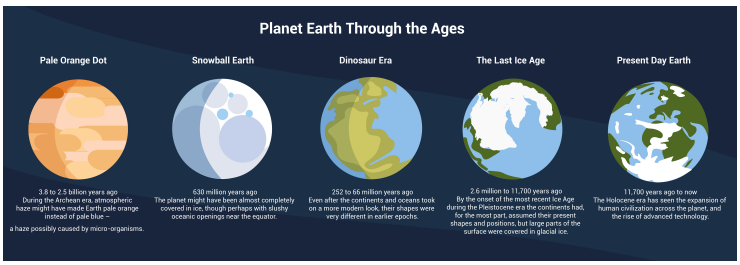

Figure 1.7: Carbon Cycle

2. 2000 yrs later

- Edward Hitchcoc
  - 1840

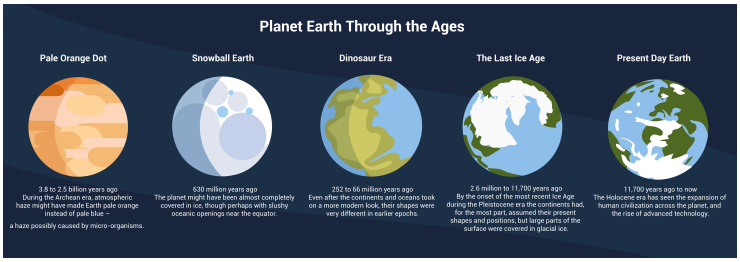

Figure 1.8: Carbon Cycle

- Ernst Haeckel
  - 1879

- 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

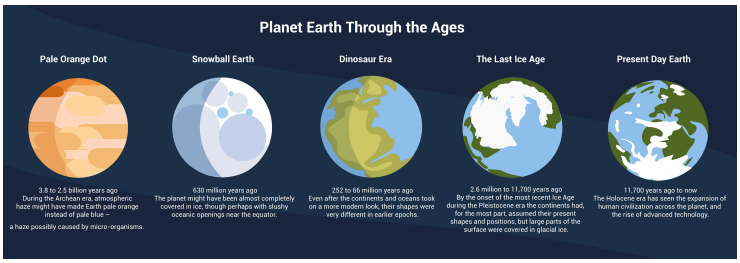

Figure 1.9: Carbon Cycle

1.3 What happened when we found out about microbes?

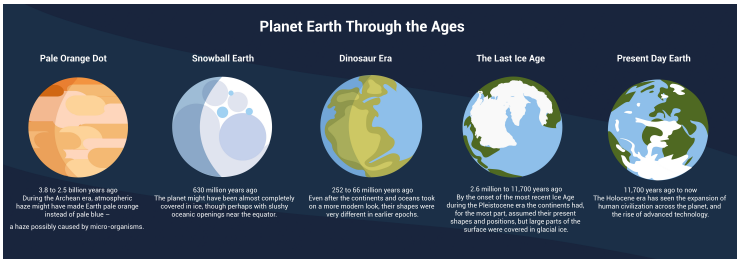

Figure 1.10: Carbon Cycle

Roadmap to where we are now with determining microbial diversity

1. Leeuwenhoek

- Father of microbiology
- Late 1600's
- Microscope

2. Robert Koch

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

3. Discovery of DNA structure

- Rosiland Franklin
  - 1951

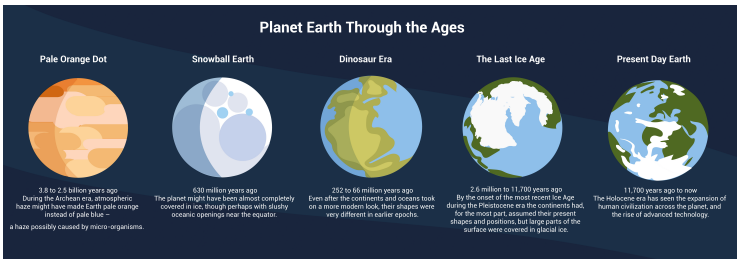

Figure 1.11: Leeuwenhoek

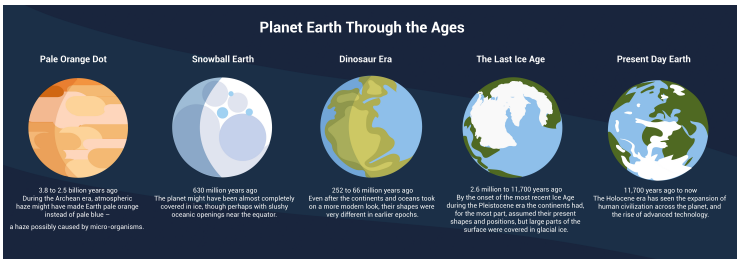

Figure 1.12: Robert Koch

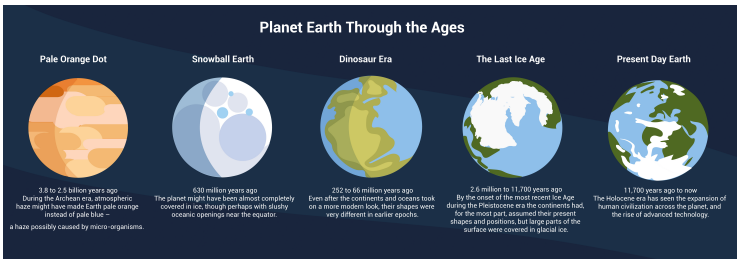

Figure 1.13: Rosiland Franklin

- Frederick Sanger
  - 1975

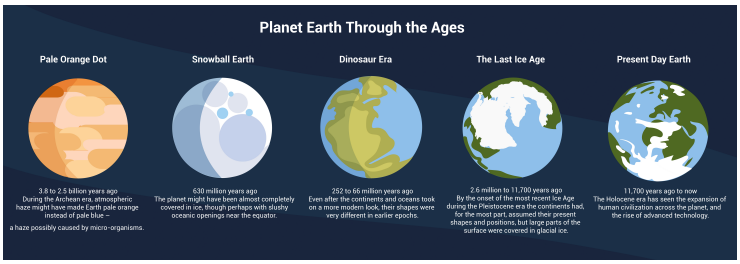

Figure 1.14: Frederick Sanger

- Carl Woese
  - 1977

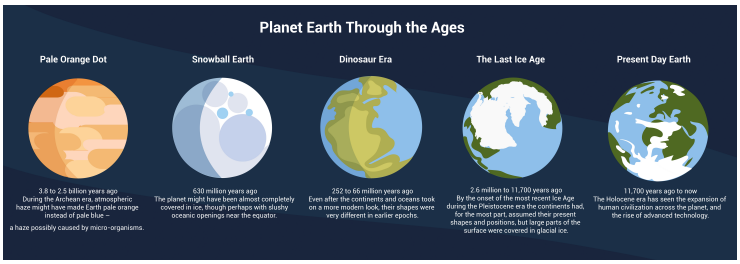

Figure 1.15: Carl Woese

DNA Structure

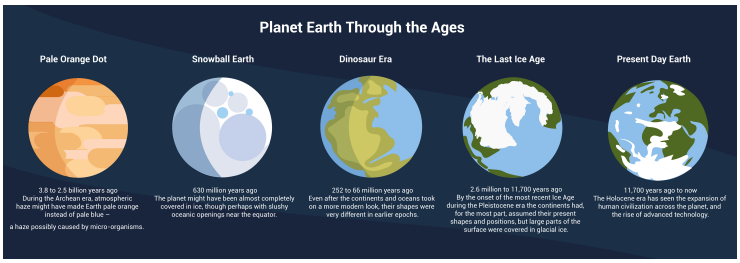

Figure 1.16: DNA Structure

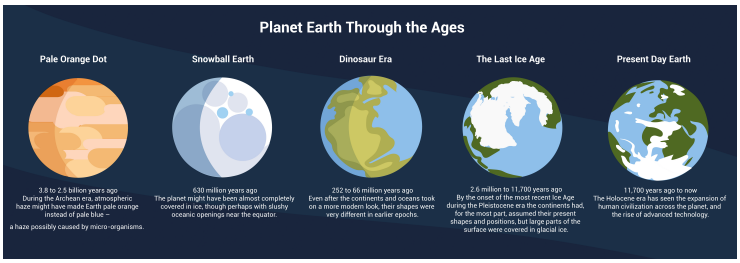

Figure 1.17: Tree of Life

1.4 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!

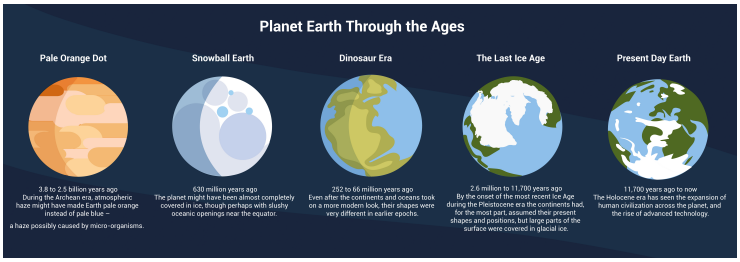

Figure 1.18: Comparing Trees of Life

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.

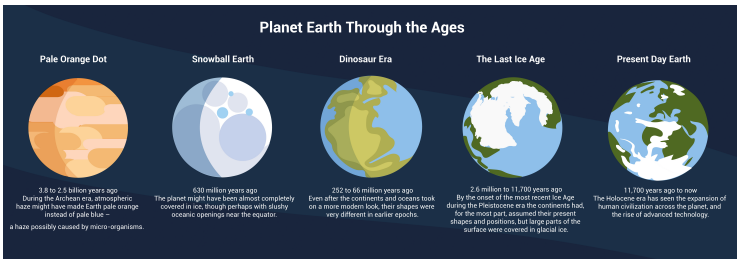

Figure 1.19: Grand Prismatic Spring

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

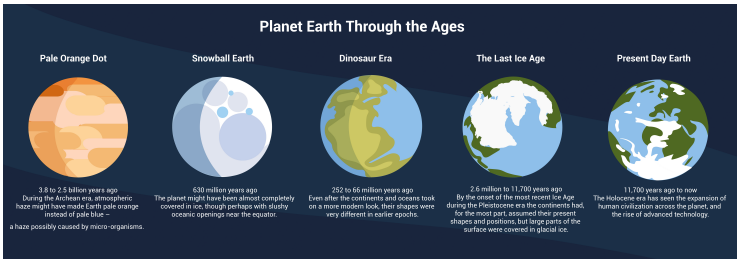

Figure 1.20: Plate Count Anomaly

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

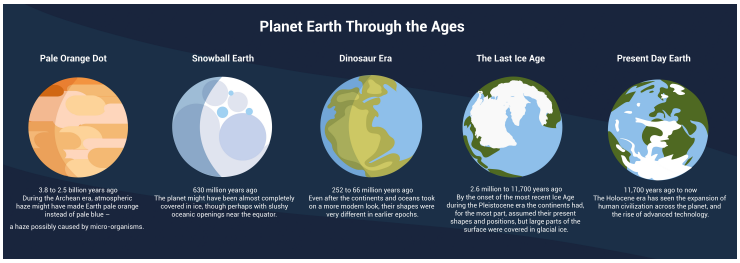

Figure 1.21: Plate Count Anomaly

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

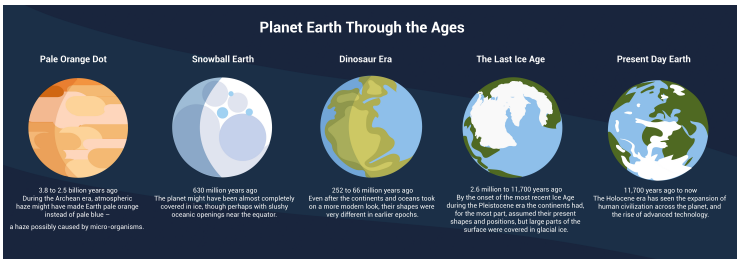

Figure 1.22: Diversity of the Microbiome, Trends in Genetics

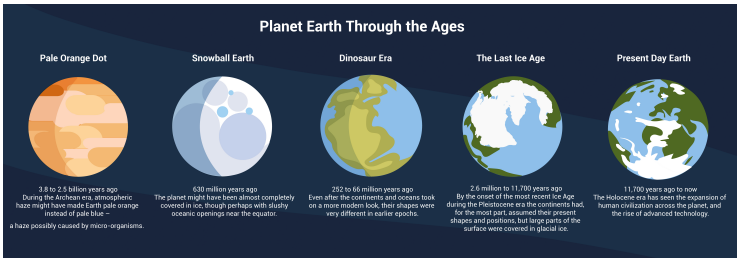

Figure 1.23: Cell 2019 Western

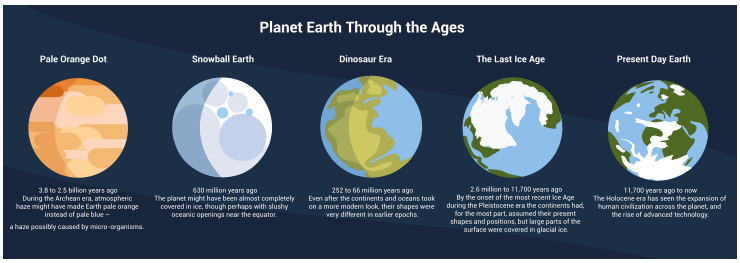

Figure 1.24: Cell 2019 Western

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

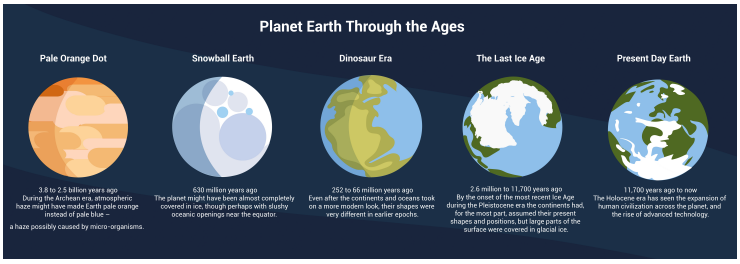

Figure 1.25: International Journal of Genomics, 2018

1.8.1 16S rRNA as an evolutionary chronometer

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

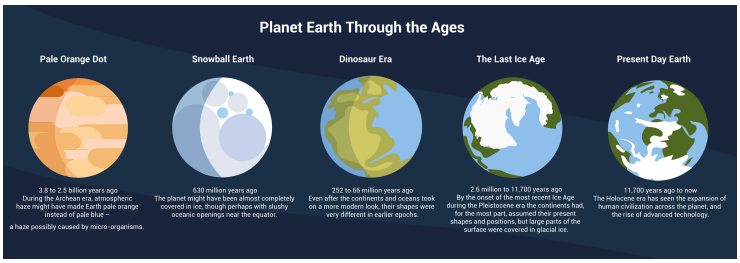

Figure 1.26: Kjellerberg et al, Microbial Ecology, 2007

5. Limited exchange – limited examples of rRNA gene sharing between organisms

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

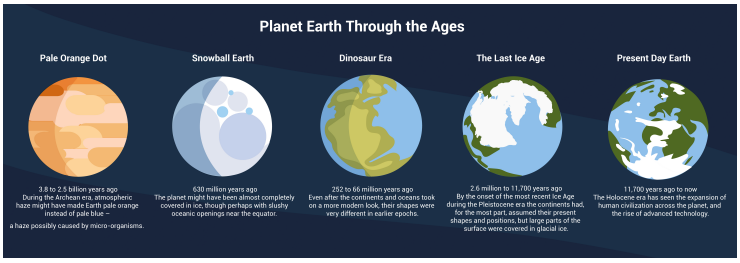

Figure 1.27: 2Kjellerberg et al, Microbial Ecology, 2007

1.8.2.1 16S rRNA hypervariable regions

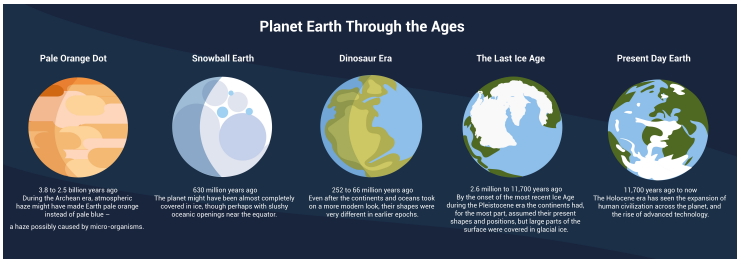

Figure 1.28: Microbiome.com

Illustration of different hypervariable regions of 16S rRNA

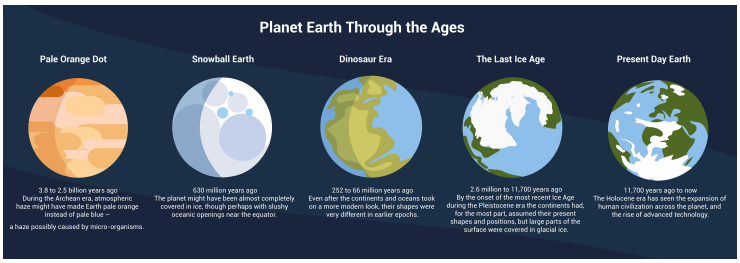

Figure 1.29: BMC Bioinf, 2016

1.9 Basic Workflow for 16S Gene Based Sequencing

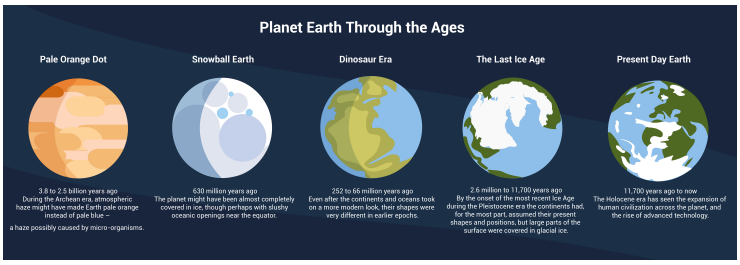

Figure 1.30: J. Investigative Dermatol, 2016

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)

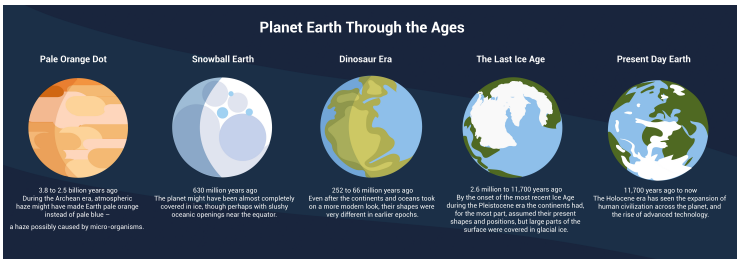

Figure 1.31: V4 Region

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

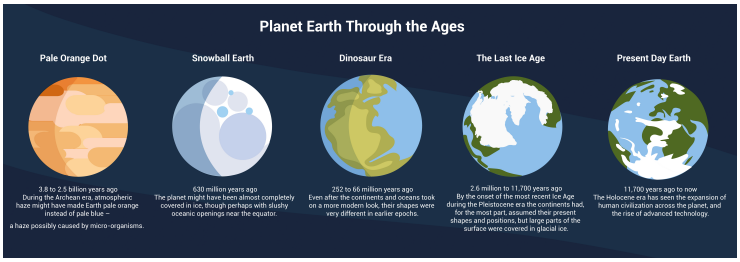

Figure 1.32: Illumina.com

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

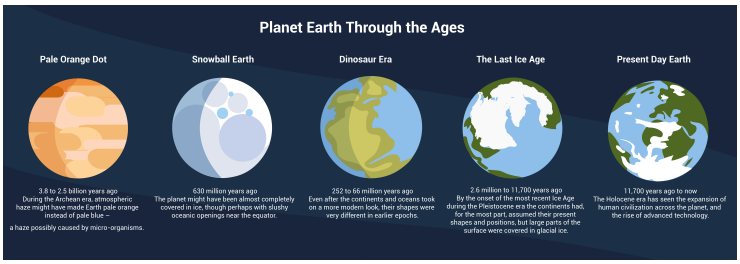

Figure 1.33: Bead beating

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

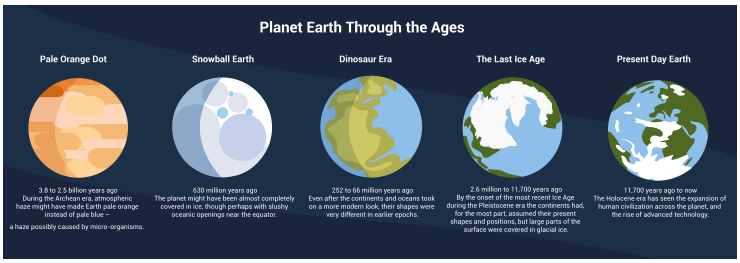

Figure 1.34: Plos

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
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)

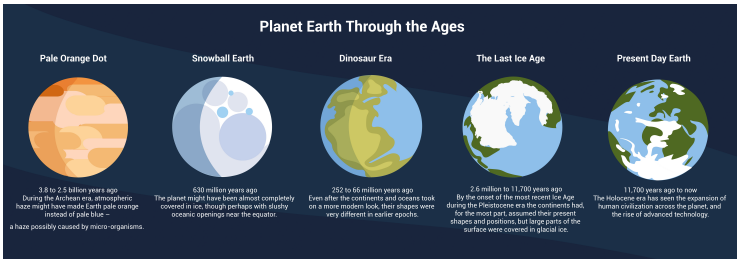

Figure 1.35: HHS

1.11.5 Purification of Amplicons

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

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

Amplicon Sequencing Library Prep - PacBio

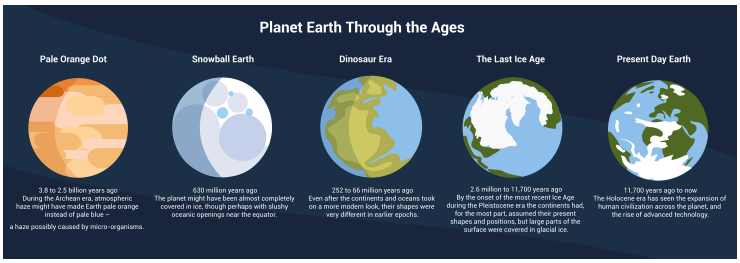

Figure 1.36: Ampure

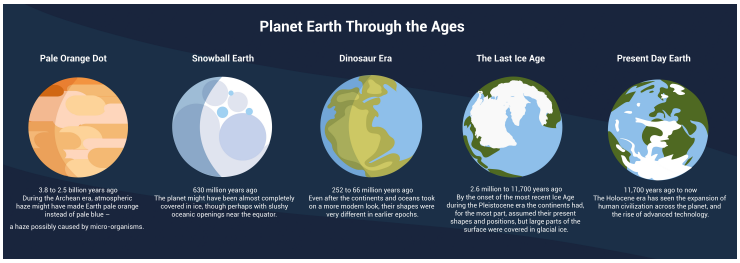

Figure 1.37: 16S Summary

1.11.5.1 Overview of generic amplicon workflow

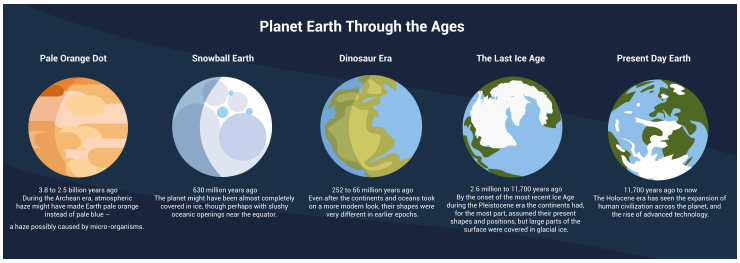

Figure 1.38: Amplicon Workflow

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!

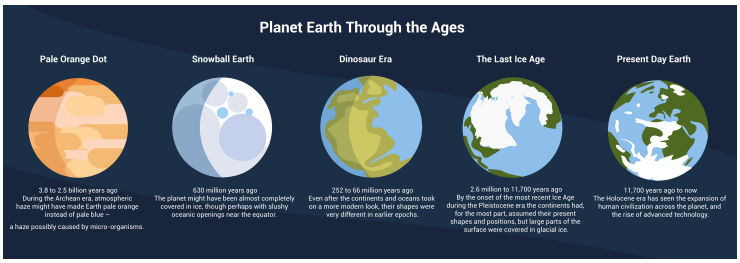

Figure 1.39: FastQC

3. Removal of adapters is critical for downstream analysis

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

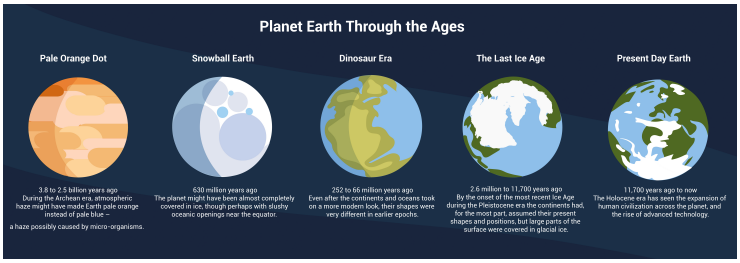

Figure 1.40: Chimera

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

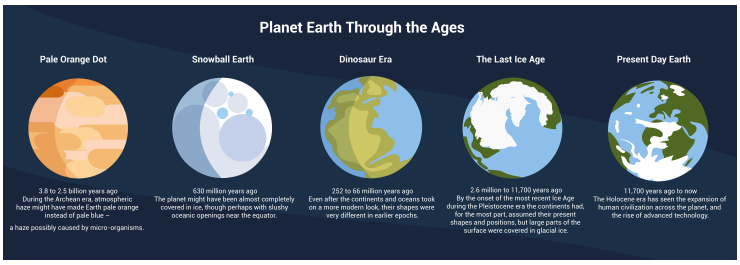

Figure 1.41: Clustering

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

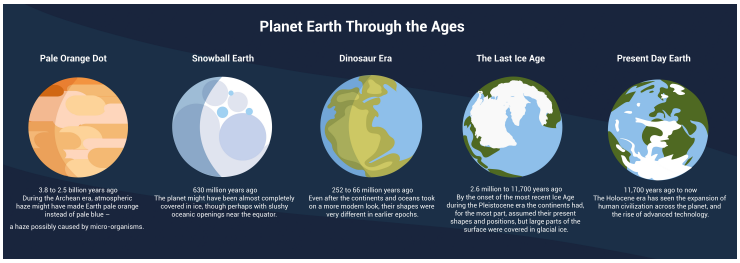

Figure 1.42: OTUs

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

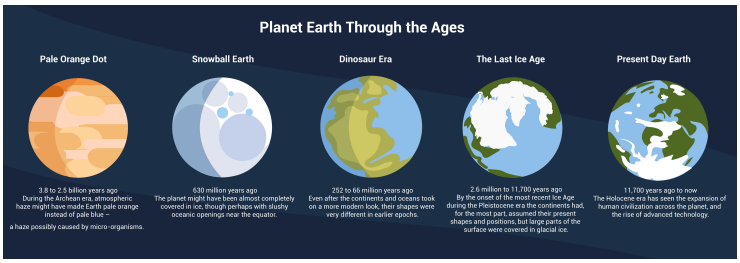

Figure 1.43: Database

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

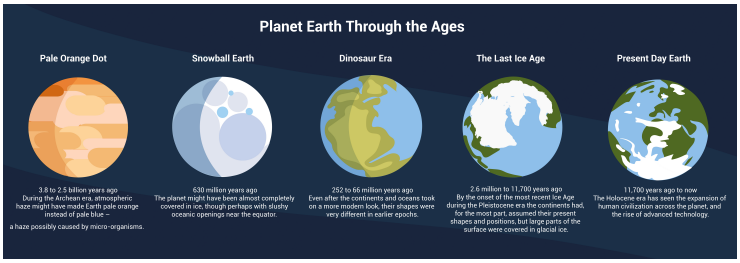

Figure 1.44: Human Mol. Genet., 2013

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

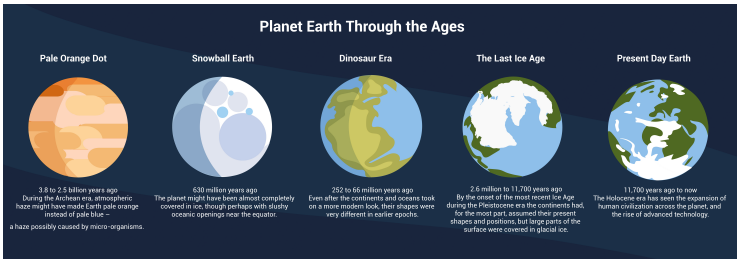

Figure 1.45: PLoS Computational Biol.,2012

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

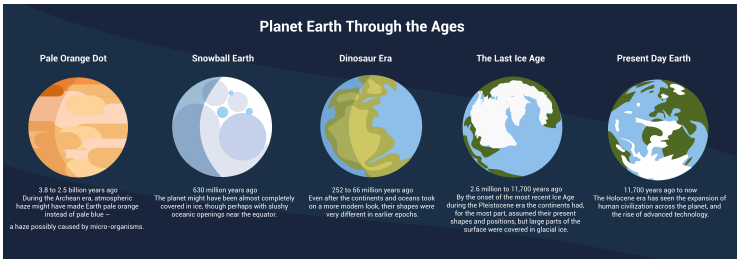

Figure 1.46: Expectation vs. Reality

1.15 Beta Diversity - UniFrac

1. Measures how different two samples' component sequences are

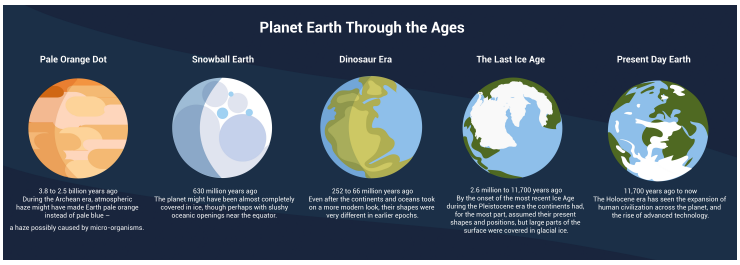

Figure 1.47: Unifrac

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?