Metagenomics Workshop NCGR

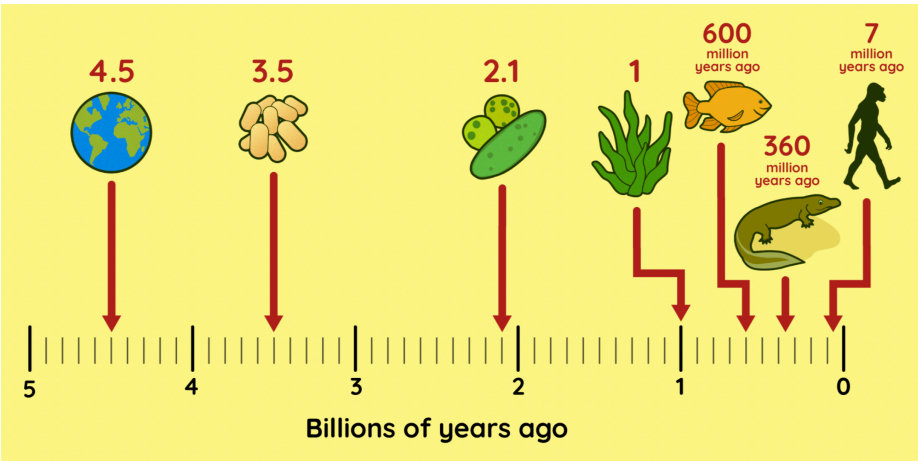
Contents

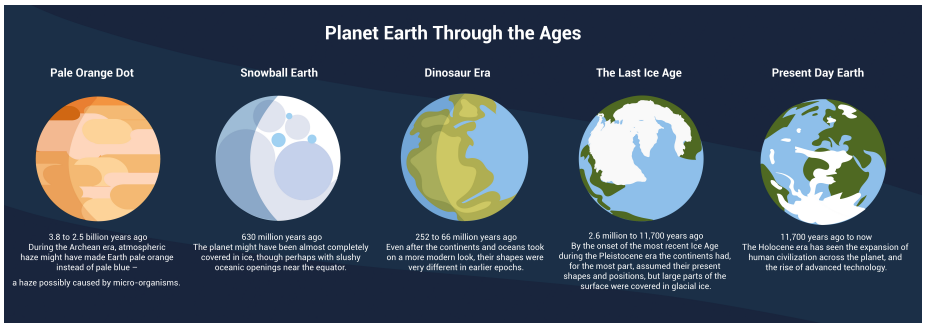
Chapter 1

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

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

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

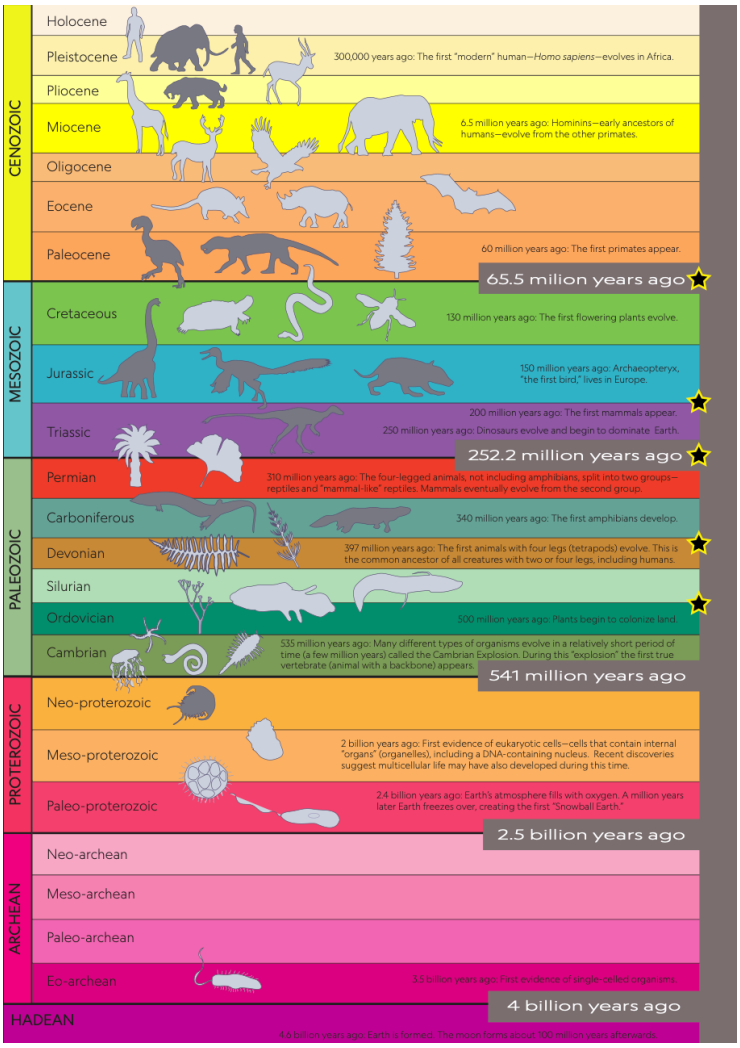


Figure 1.3: History of life on earth

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**

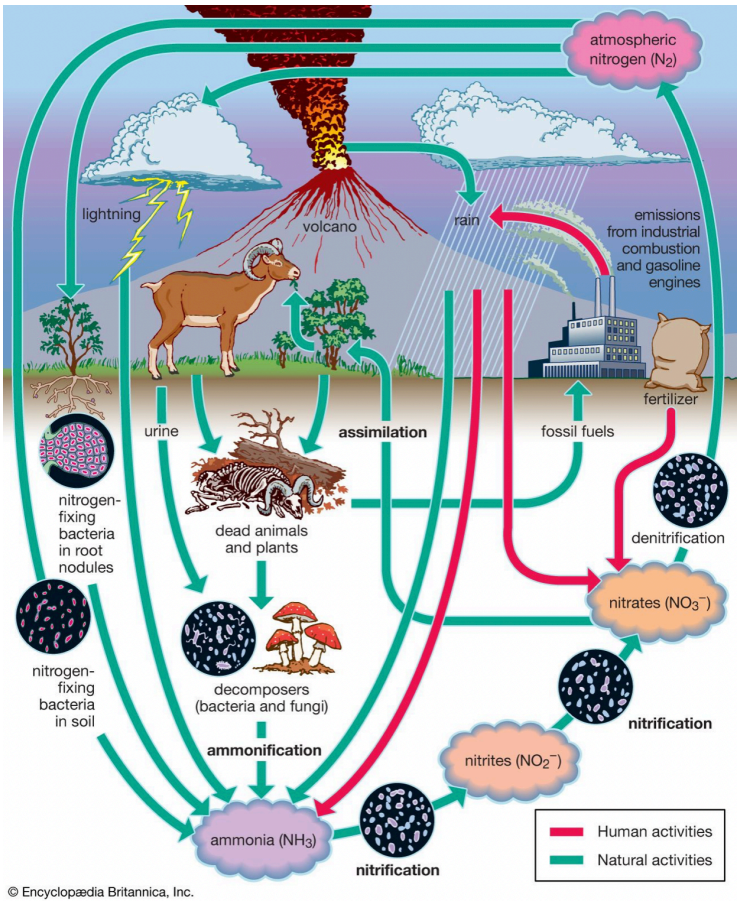


Figure 1.4: Nitrogen Cycle

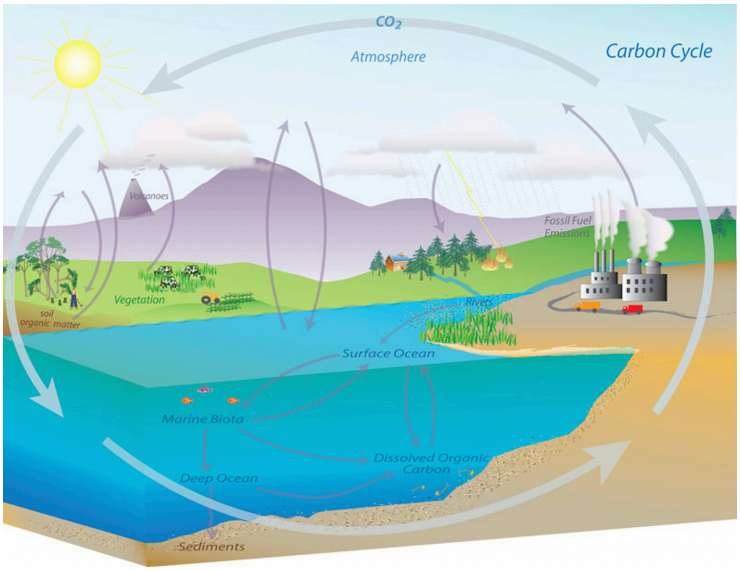


Figure 1.5: Carbon Cycle

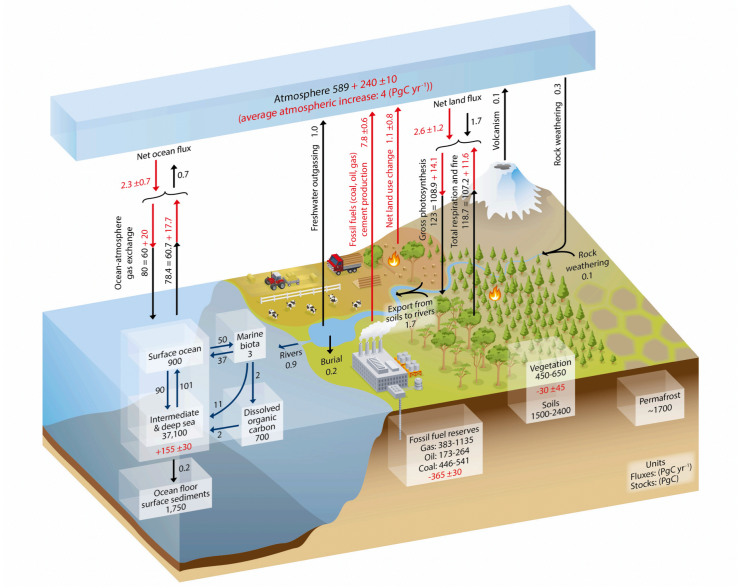
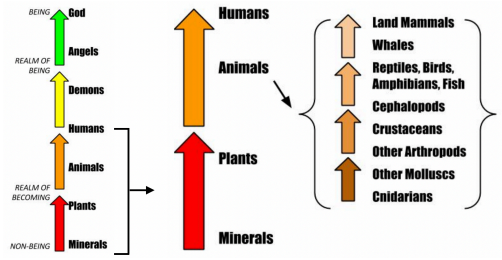


Figure 1.6: Carbon Cycle



	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

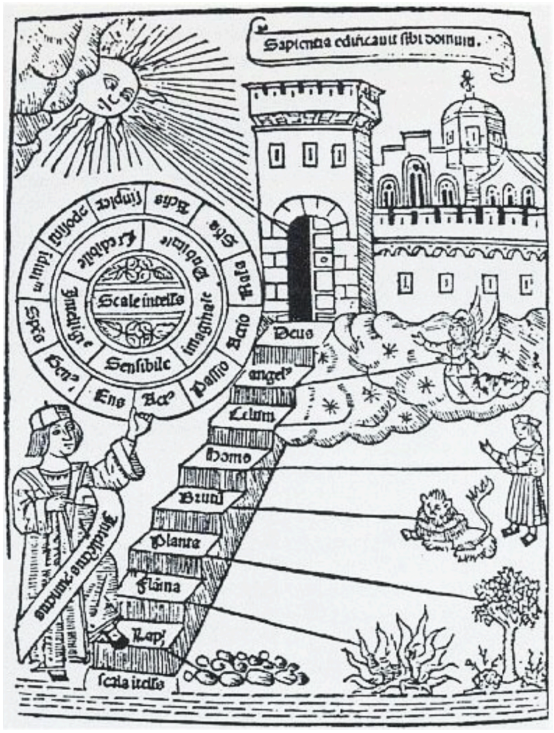


Figure 1.7: Ladder of Ascent and Descent of the Mind, 1305

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

2. 2000 yrs later

- Edward Hitchcock
  - 1840

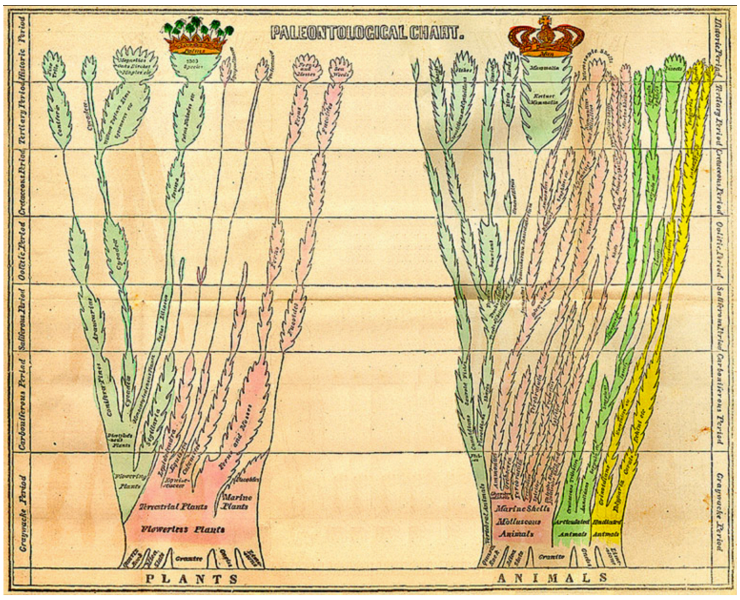


Figure 1.8: Tree of Life

•

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

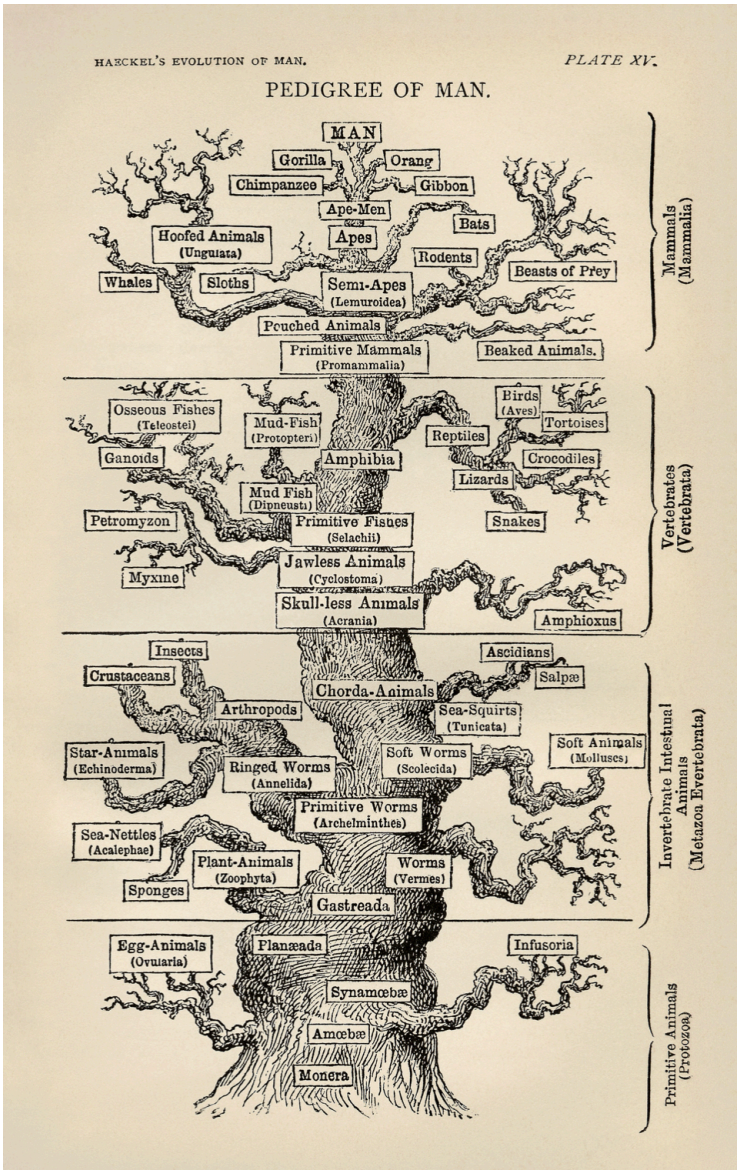


Figure 1.9: Another Tree of Life

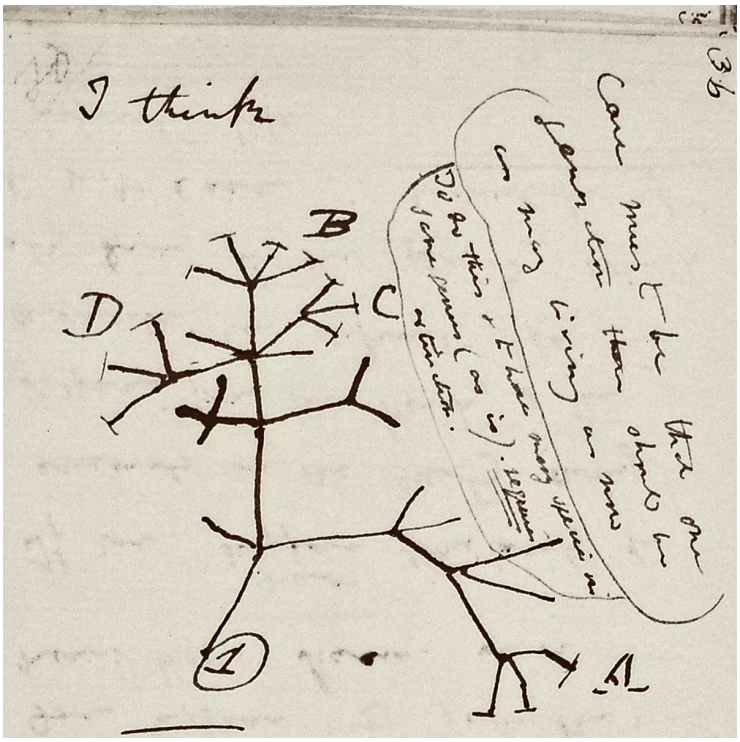


Figure 1.10: Tree of Life sketch

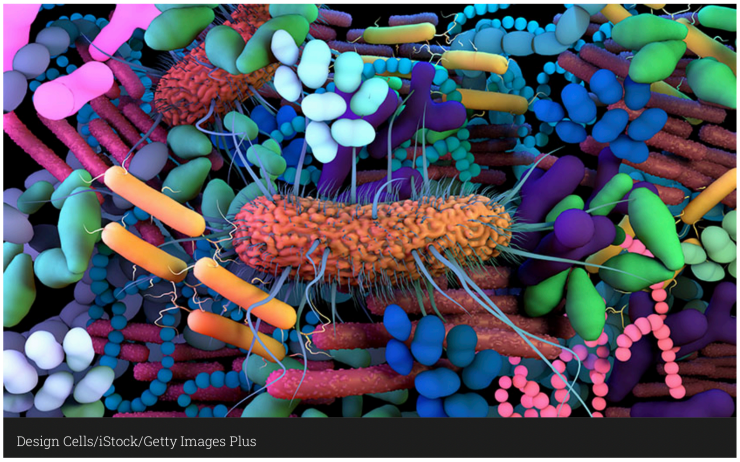


Figure 1.11: Microbes

1.3 What happened when we found out about microbes?

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

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

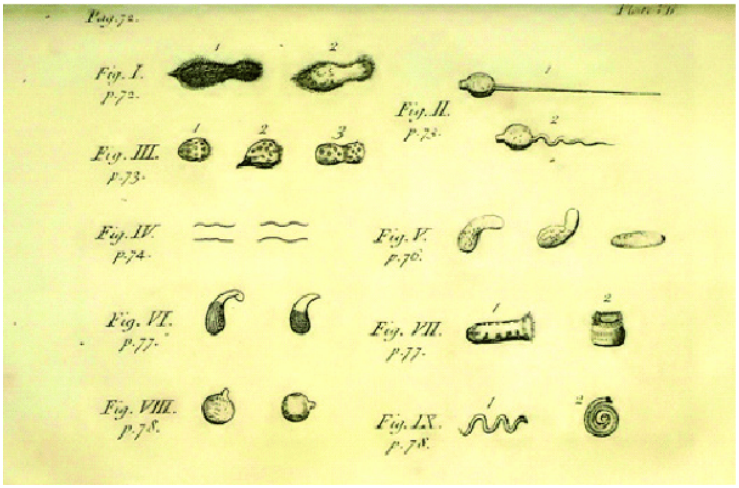
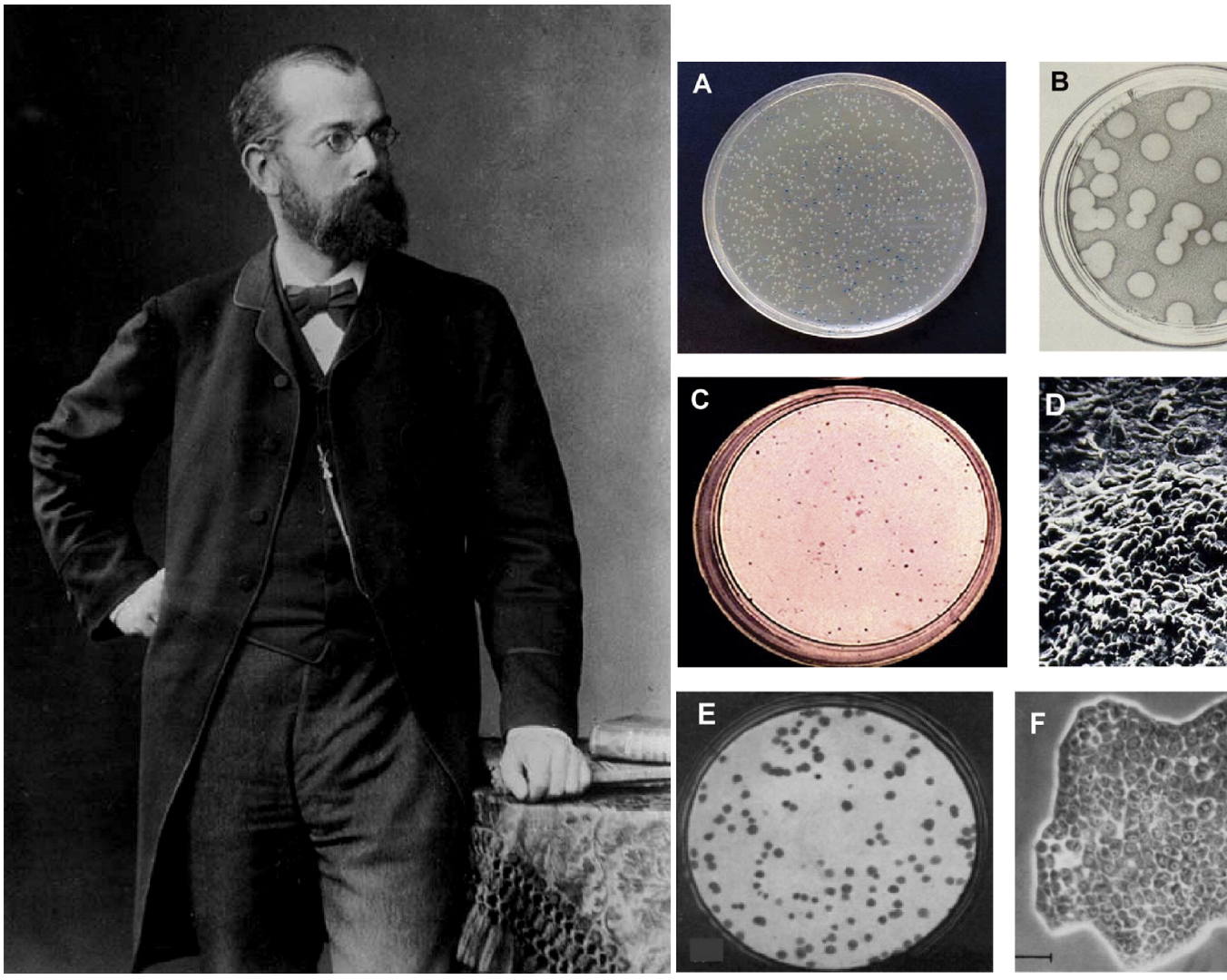


Figure 1.13: First Sketch of Microbes



3. Discovery of DNA structure

- Rosiland Franklin
  - 1951

Figure 1.14: https://en.wikipedia.org/wiki/Rosalind_Franklin#/media/File:Rosalind_Franklin_(1920-1958).jpg

- Frederick Sanger
  - 1975

- Carl Woese

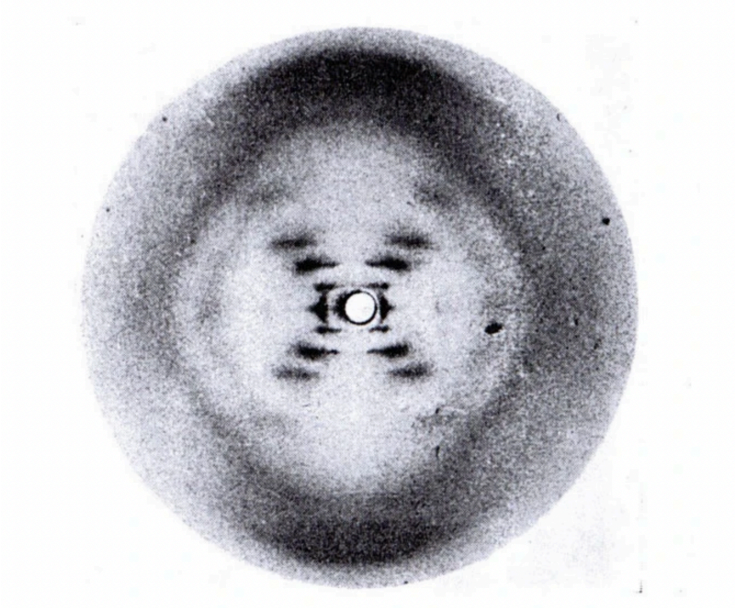


Figure 1.15: — J. D. Bernal, 1958. Franklin's X-ray diagram of the B form of sodium thymonucleate (DNA) fibres, published in Nature on 25 April 1953, shows "in striking manner the features characteristic of helical structures"

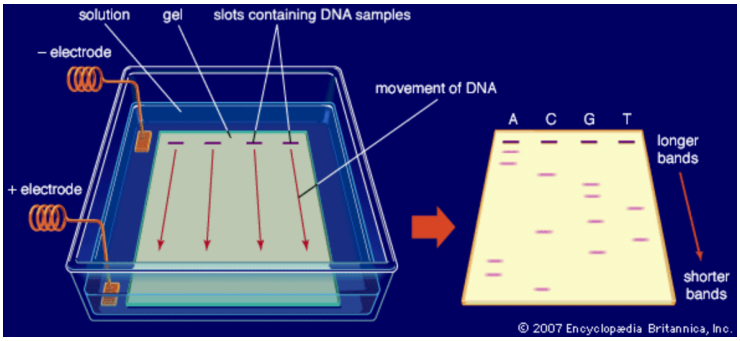


Figure 1.16: https://www.britannica.com/science/DNA-sequencing#/media/1/422006/40224

– 1977
– Archaea
– Phylogenetic tree based on Woese et al. rRNA analysis. The vertical line at bottom represents the last universal common ancestor (LUCA).

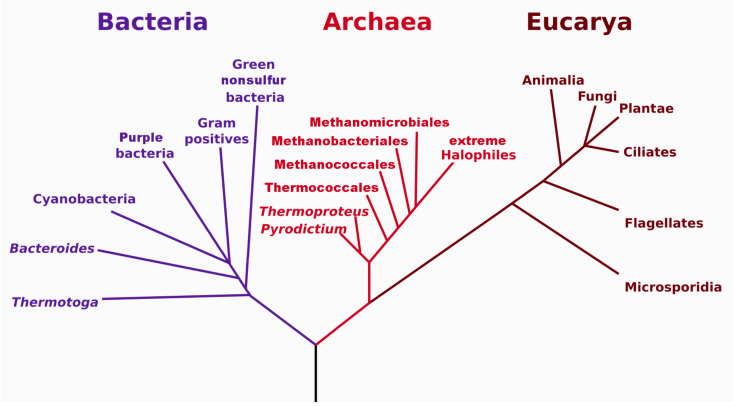


Figure 1.17: Maulucioni, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons

DNA Structure

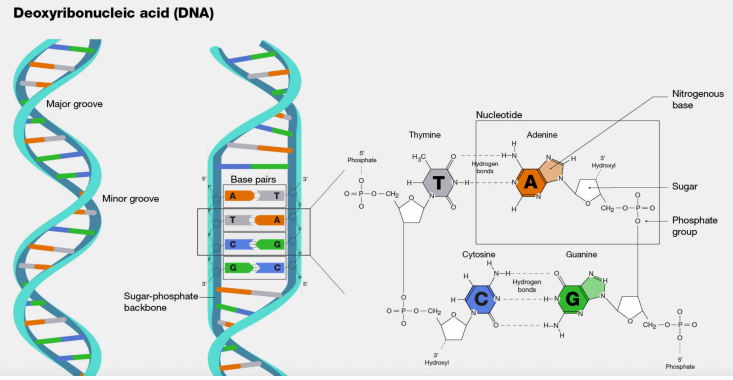


Figure 1.18: https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid

https://youtu.be/L9NriBoubWE

1.4 Tree of Life

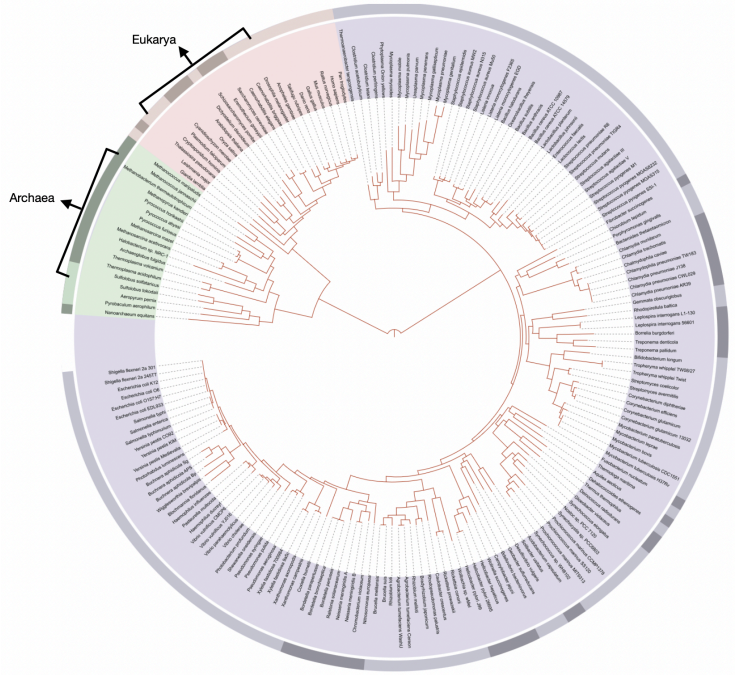


Figure 1.19: Ivica Letunic: Iletunic. Retraced by Mariana Ruiz Villarreal: LadyofHats, Public domain, via Wikimedia Commons

"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!

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

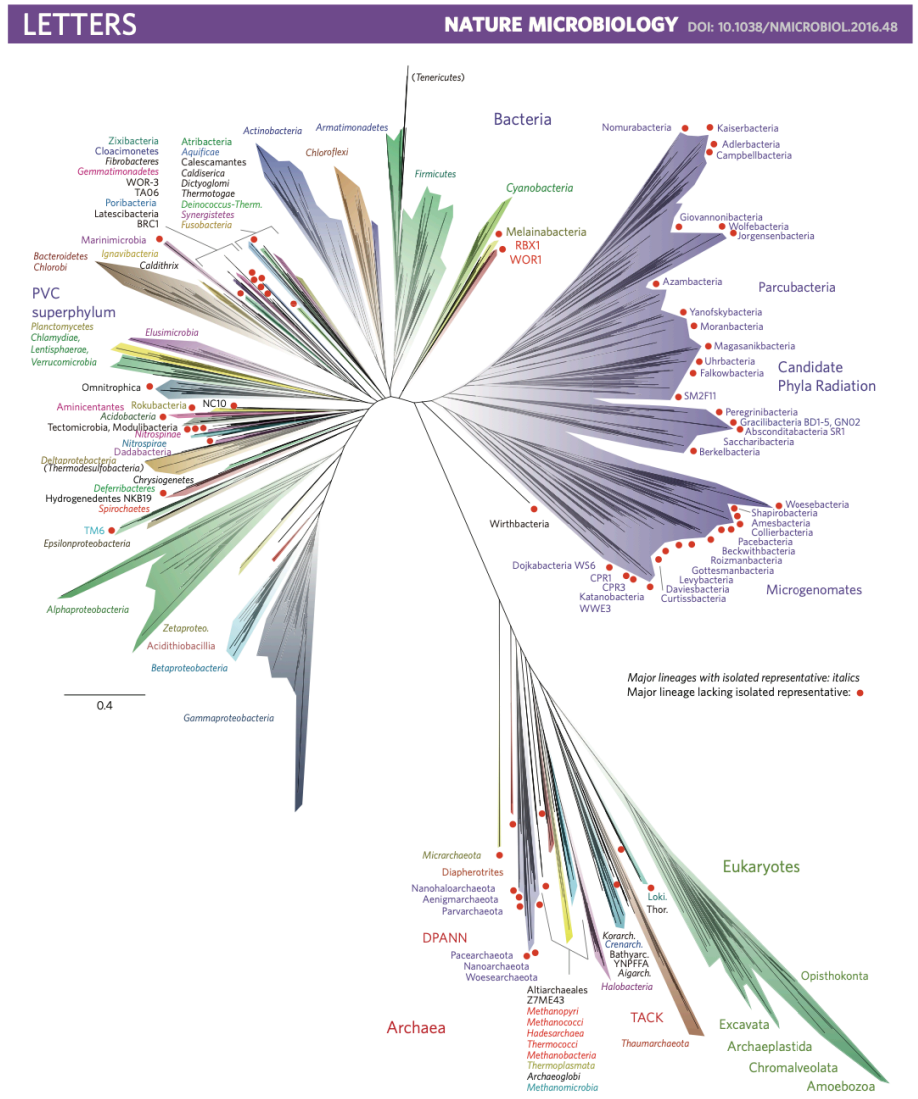


Figure 1 | A current view of the tree of life, encompassing the total diversity represented by sequenced genomes. The tree includes 92 named bacterial phyla, 26 archaeal phyla and all five of the Eukaryotic supergroups. Major lineages are assigned arbitrary colours and named, with well-characterized lineage names, in italics. Lineages lacking an isolated representative are highlighted with non-italicized names and red dots. For details on taxon sampling and tree inference, see Methods. The names Tenericutes and Thermodesulfobacteria are bracketed to indicate that these lineages branch within the Firmicutes and the Deltaproteobacteria, respectively. Eukaryotic supergroups are noted, but not otherwise delineated due to the low resolution of these lineages. The CPR phyla are assigned a single colour as they are composed entirely of organisms without isolated representatives, and are still in the process of definition at lower taxonomic levels. The complete ribosomal protein tree is available in rectangular format with full bootstrap values as Supplementary Fig. 1 and in Newick format in Supplementary Dataset 2.

Figure 1.20: Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... & Banfield, J. F. (2016). A new view of the tree of life. Nature microbiology, 1(5), 1-6.

Grand Prismatic Spring – YNP – 183oC

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

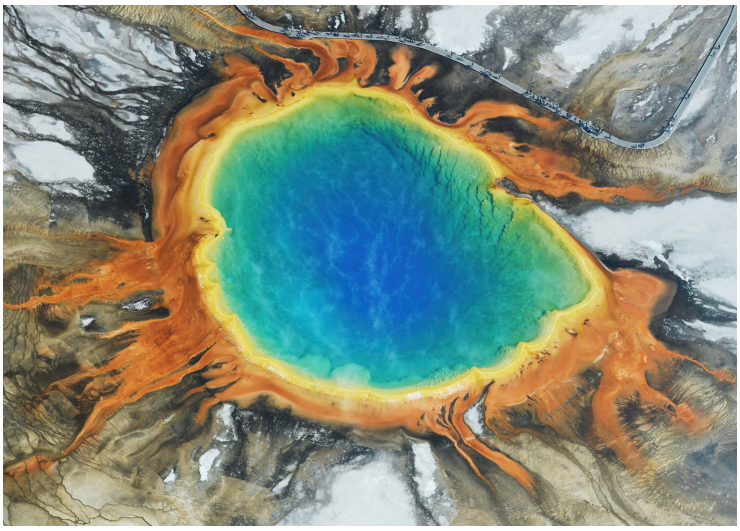


Figure 1.21: Carsten Steger, CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons

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

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

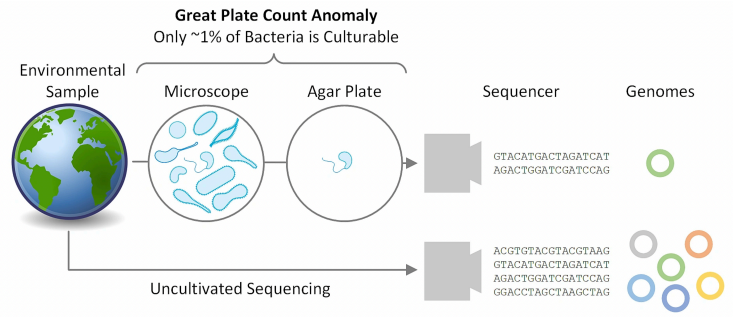


Figure 1.22: Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., … & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature microbiology, 2(11), 1533-1542.

1.5.2 Total number of genomes at NCBI

Most Prokaryotes:

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

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)

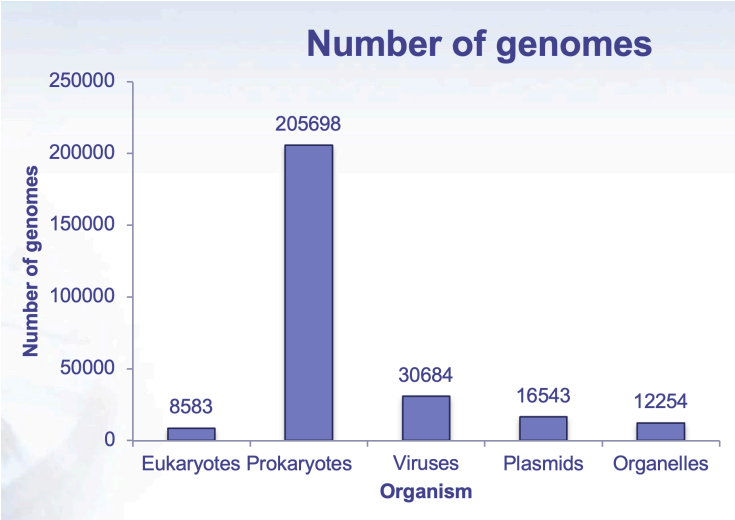


Figure 1.23: Plate Count Anomaly

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

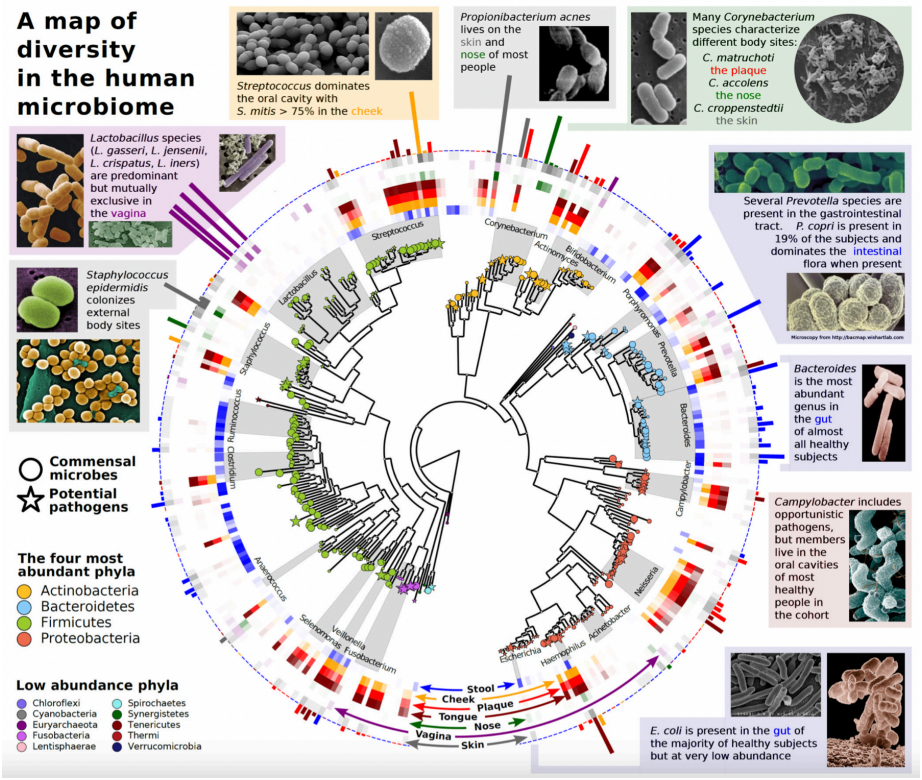
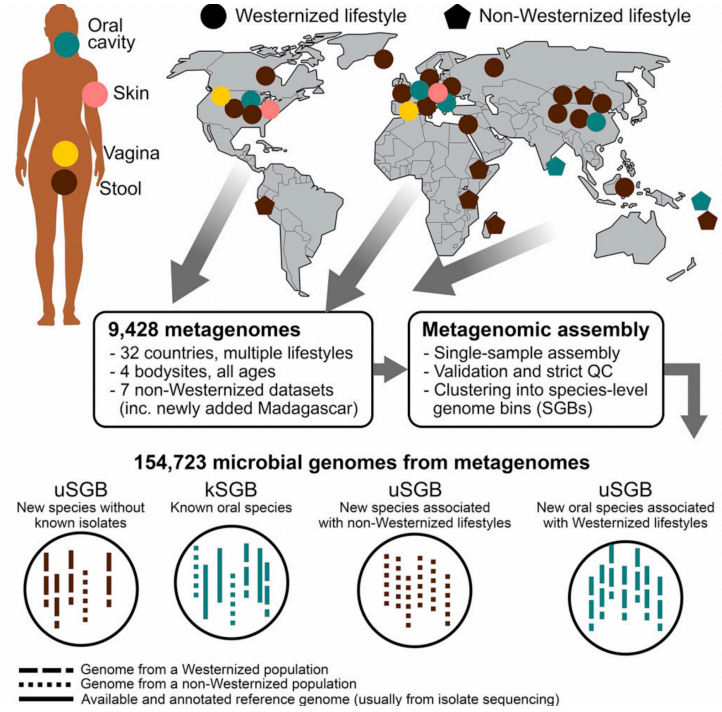
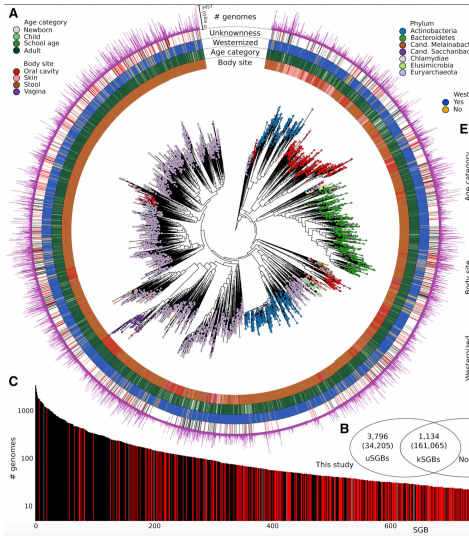


Figure 1.24: Morgan, X. C., Segata, N., & Huttenhower, C. (2013). Biodiversity and functional genomics in the human microbiome. Trends in genetics, 29(1), 51-58.

1.7.4 Biodiversity and functional genomics in the human microbiome





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

1.7.5 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

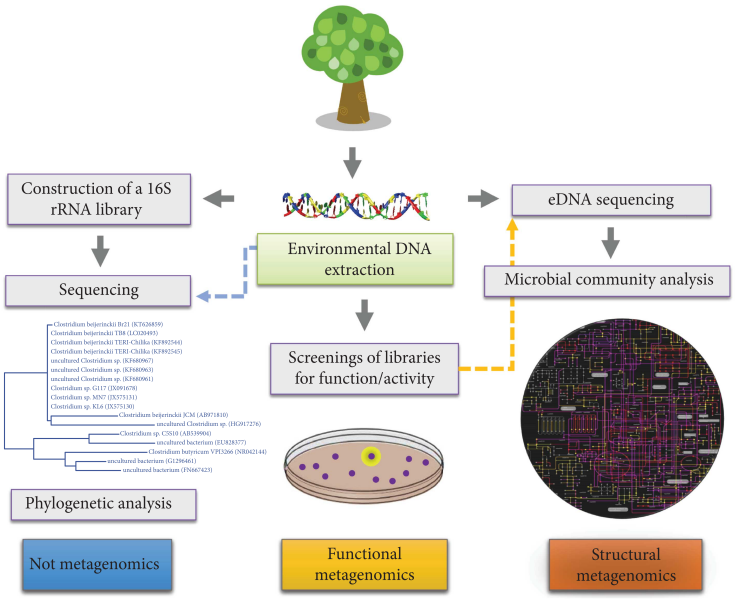


Figure 1.25: Alves, L. D. F., Westmann, C. A., Lovate, G. L., de Siqueira, G. M. V., Borelli, T. C., & Guazzaroni, M. E. (2018). Metagenomic approaches for understanding new concepts in microbial science. International journal of genomics, 2018.

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

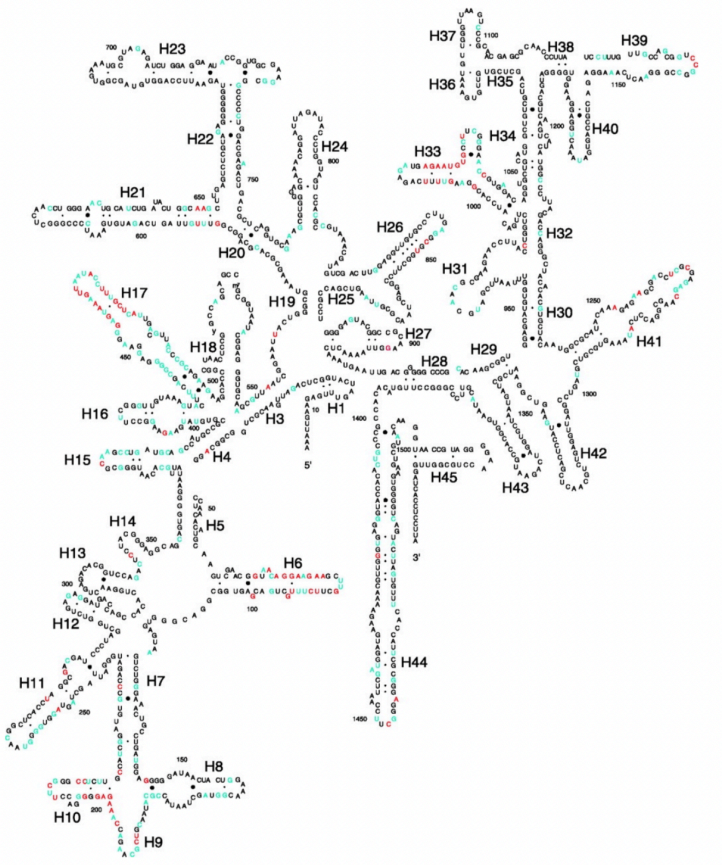


Figure 1.26: Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and environmental microbiology, 73(1), 278-288.

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

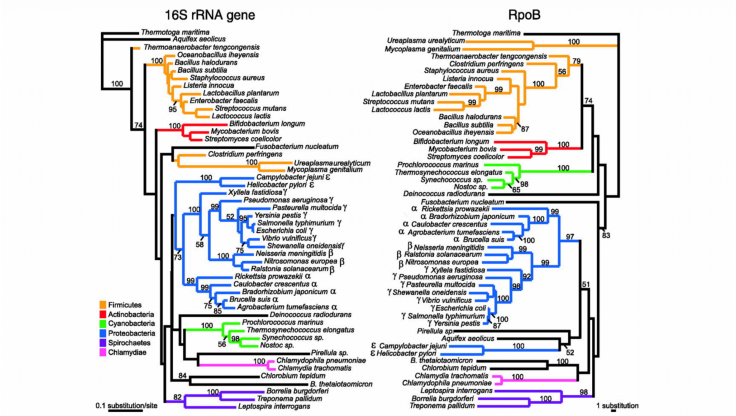


FIG. 3. Comparison of the best maximum likelihood trees for the 16S rRNA gene and the RpoB protein for the domain *Bacteria*.

Figure 1.27: Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and environmental microbiology, 73(1), 278-288.

1.8.2.1 16S rRNA hypervariable regions

Illustration of different hypervariable regions of 16S rRNA

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

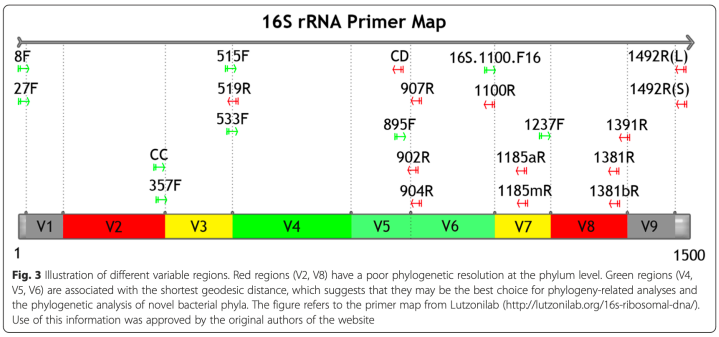


Fig. 3 Illustration of different variable regions. Red regions (V2, V8) have a poor phylogenetic resolution at the phylum level. Green regions (V4, V5, V6) are associated with the shortest geodesic distance, which suggests that they may be the best choice for phylogeny-related analyses and the phylogenetic analysis of novel bacterial phyla. The figure refers to the primer map from Lutzonilab (http://lutzonilab.org/16s-ribosomal-dna/). Use of this information was approved by the original authors of the website

Figure 1.28: Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC bioinformatics, 17(1), 1-8.

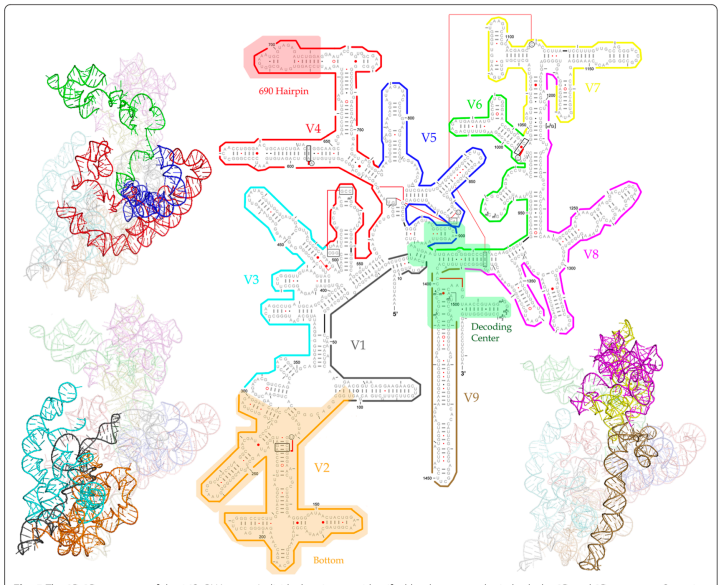


Fig. 5 The 2D-3D structures of the 16S rRNA gene. Individual regions are identified by the same color in both the 2D and 3D structures. Some important structures are colored with blocks

Figure 1.29: Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC bioinformatics, 17(1), 1-8.

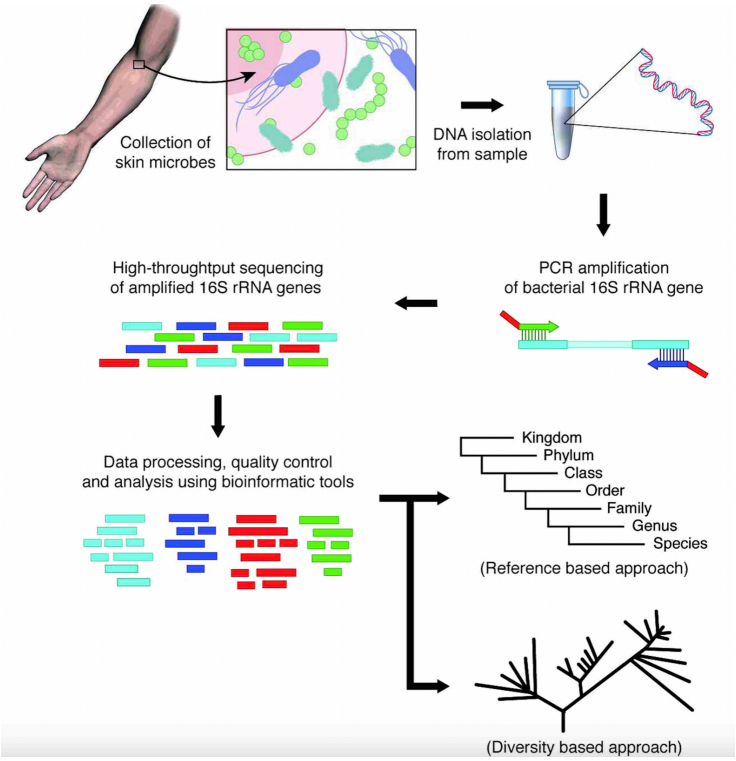


Figure 1.30: Jo, J. H., Kennedy, E. A., & Kong, H. H. (2016). Research techniques made simple: bacterial 16S ribosomal RNA gene sequencing in cutaneous research. Journal of Investigative Dermatology, 136(3), e23-e27.

- 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)

Sample ID	Barcode
AB123_newborn	GAACCAAAGGAT
AB121_one_month	TACACGATCTAC
AB_145_two_months	GCGATATATCGC
AB_134_birth	CAGTGCATATGC
AB_189_birth	TCCAAAGTGTTC
AB_170_birth	GGCCACGTAGTA

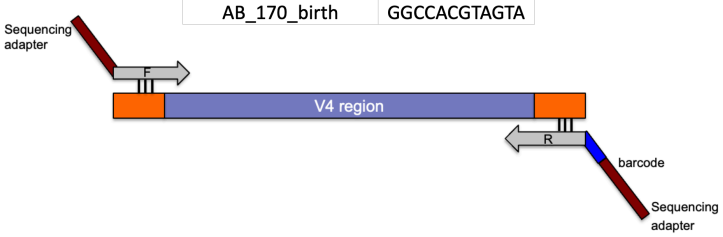


Figure 1.31: V4 Region

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

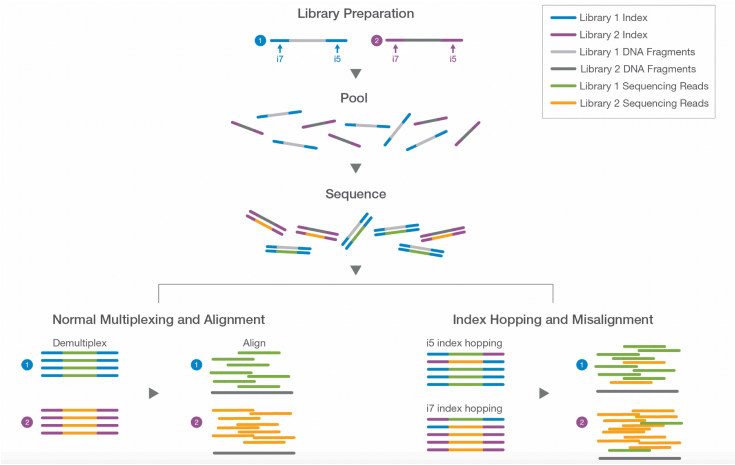


Figure 1.32: https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf

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

https://www.youtube.com/watch?v=3SEz-i517Oo&t=5s

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

The effect of bead beating on the observed microbial community composition:

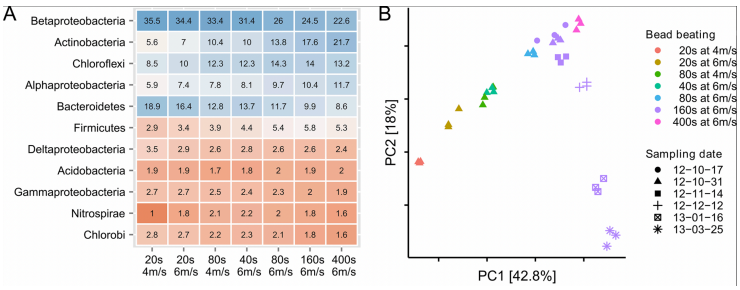


Figure 1.33: Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H., & Nielsen, P. H. (2015). Back to basics–the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. PloS one, 10(7), e0132783.

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

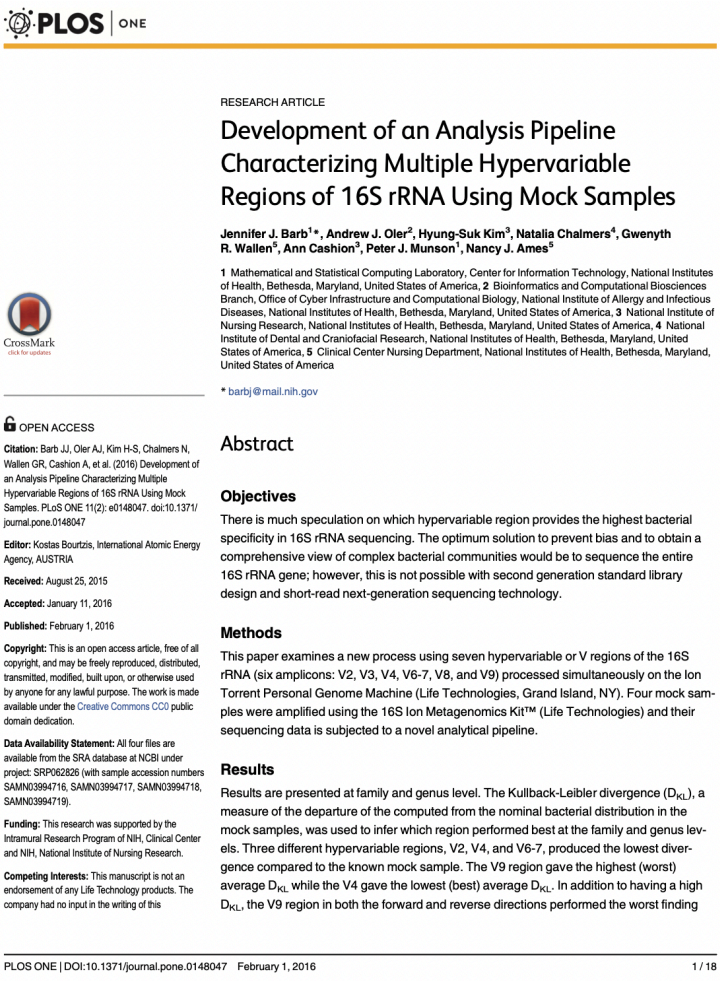
1.11.4 Selection of primers and region of 16S gene influence microbial profile

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
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)

1.11.5 Purification of Amplicons

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



PLOS | ONE

RESEARCH ARTICLE

Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples

Jennifer J. Barb[1]*, Andrew J. Oler[2], Hyung-Suk Kim[3], Natalia Chalmers[4], Gwenyth R. Wallen[5], Ann Cashion[3], Peter J. Munson[1], Nancy J. Ames[5]

1 Mathematical and Statistical Computing Laboratory, Center for Information Technology, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, United States of America, **3** National Institute of Nursing Research, National Institutes of Health, Bethesda, Maryland, United States of America, **4** National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, Maryland, United States of America, **5** Clinical Center Nursing Department, National Institutes of Health, Bethesda, Maryland, United States of America

* barbj@mail.nih.gov

OPEN ACCESS

Citation: Barb JJ, Oler AJ, Kim H-S, Chalmers N, Wallen GR, Cashion A, et al. (2016) Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. PLoS ONE 11(2): e0148047. doi:10.1371/journal.pone.0148047

Editor: Kostas Bourtzis, International Atomic Energy Agency, AUSTRIA

Received: August 25, 2015

Accepted: January 11, 2016

Published: February 1, 2016

Data Availability Statement: All four files are available from the SRA database at NCBI under project: SRP062826 (with sample accession numbers SAMN03994716, SAMN03994717, SAMN03994718, SAMN03994719).

Funding: This research was supported by the Intramural Research Program of NIH, Clinical Center and NIH, National Institute of Nursing Research.

Competing Interests: This manuscript is not an endorsement of any Life Technology products. The company had no input in the writing of this

Abstract

Objectives

There is much speculation on which hypervariable region provides the highest bacterial specificity in 16S rRNA sequencing. The optimum solution to prevent bias and to obtain a comprehensive view of complex bacterial communities would be to sequence the entire 16S rRNA gene; however, this is not possible with second generation standard library design and short-read next-generation sequencing technology.

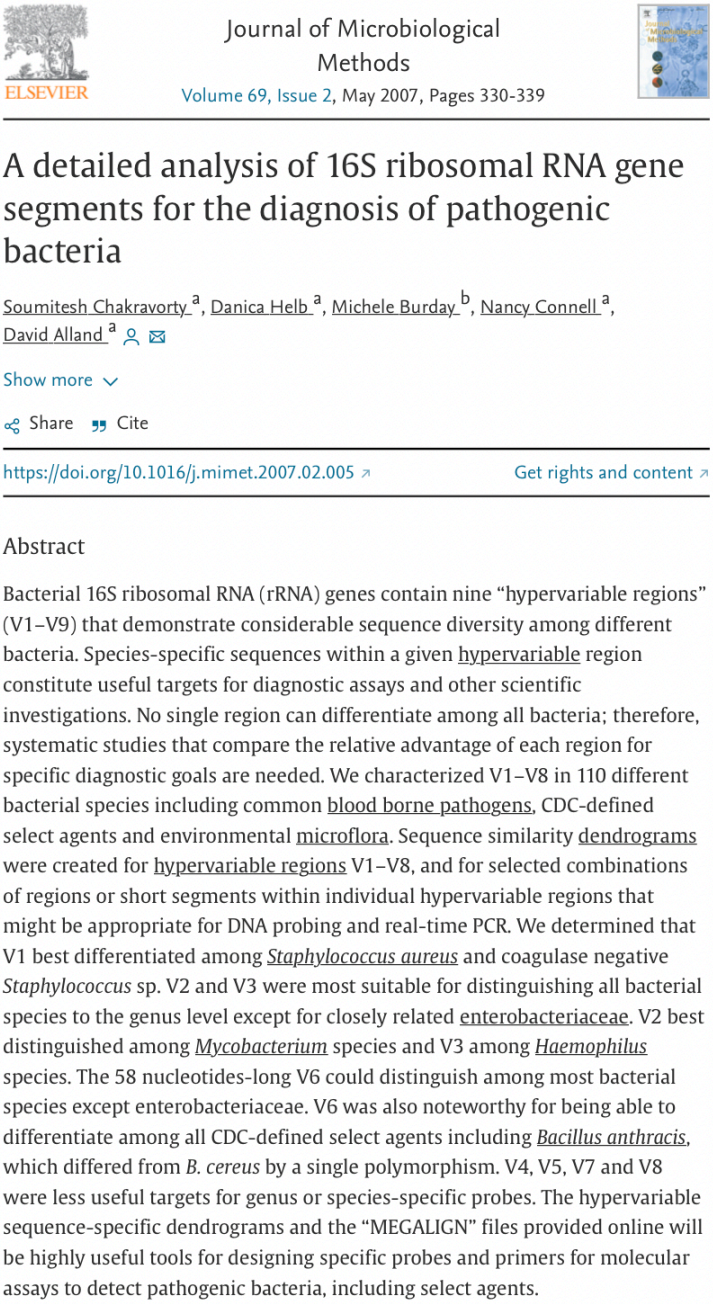
Methods

This paper examines a new process using seven hypervariable or V regions of the 16S rRNA (six amplicons: V2, V3, V4, V6-7, V8, and V9) processed simultaneously on the Ion Torrent Personal Genome Machine (Life Technologies, Grand Island, NY). Four mock samples were amplified using the 16S Ion Metagenomics Kit™ (Life Technologies) and their sequencing data is subjected to a novel analytical pipeline.

Results

Results are presented at family and genus level. The Kullback-Leibler divergence (D_{KL}), a measure of the departure of the computed from the nominal bacterial distribution in the mock samples, was used to infer which region performed best at the family and genus levels. Three different hypervariable regions, V2, V4, and V6-7, produced the lowest divergence compared to the known mock sample. The V9 region gave the highest (worst) average D_{KL} while the V4 gave the lowest (best) average D_{KL}. In addition to having a high D_{KL}, the V9 region in both the forward and reverse directions performed the worst finding

Figure 1.34: Barb, J. J., Oler, A. J., Kim, H. S., Chalmers, N., Wallen, G. R., Cashion, A., ... & Ames, N. J. (2016). Development of an analysis pipeline characterizing multiple hypervariable regions of 16S rRNA using mock samples. PLoS One, 11(2), e0148047.



Journal of Microbiological Methods
Volume 69, Issue 2, May 2007, Pages 330-339

A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria

Soumitesh Chakravorty [a], Danica Helb [a], Michele Burday [b], Nancy Connell [a], David Alland [a]

Show more

Share Cite

https://doi.org/10.1016/j.mimet.2007.02.005 Get rights and content

Abstract

Bacterial 16S ribosomal RNA (rRNA) genes contain nine "hypervariable regions" (V1–V9) that demonstrate considerable sequence diversity among different bacteria. Species-specific sequences within a given hypervariable region constitute useful targets for diagnostic assays and other scientific investigations. No single region can differentiate among all bacteria; therefore, systematic studies that compare the relative advantage of each region for specific diagnostic goals are needed. We characterized V1–V8 in 110 different bacterial species including common blood borne pathogens, CDC-defined select agents and environmental microflora. Sequence similarity dendrograms were created for hypervariable regions V1–V8, and for selected combinations of regions or short segments within individual hypervariable regions that might be appropriate for DNA probing and real-time PCR. We determined that V1 best differentiated among *Staphylococcus aureus* and coagulase negative *Staphylococcus* sp. V2 and V3 were most suitable for distinguishing all bacterial species to the genus level except for closely related enterobacteriaceae. V2 best distinguished among *Mycobacterium* species and V3 among *Haemophilus* species. The 58 nucleotides-long V6 could distinguish among most bacterial species except enterobacteriaceae. V6 was also noteworthy for being able to differentiate among all CDC-defined select agents including *Bacillus anthracis*, which differed from *B. cereus* by a single polymorphism. V4, V5, V7 and V8 were less useful targets for genus or species-specific probes. The hypervariable sequence-specific dendrograms and the "MEGALIGN" files provided online will be highly useful tools for designing specific probes and primers for molecular assays to detect pathogenic bacteria, including select agents.

Figure 1.35: Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods, 69(2), 330-339.

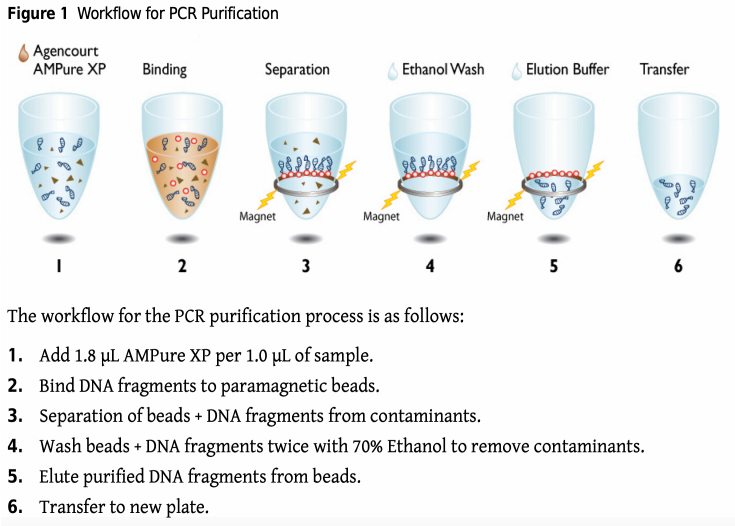


Figure 1.36: https://research.fredhutch.org/content/dam/stripe/hahn/methods/mol_biol/Agencourt%20AMPure%20XP.pdf

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

1.11.5.1 Overview of generic amplicon workflow

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

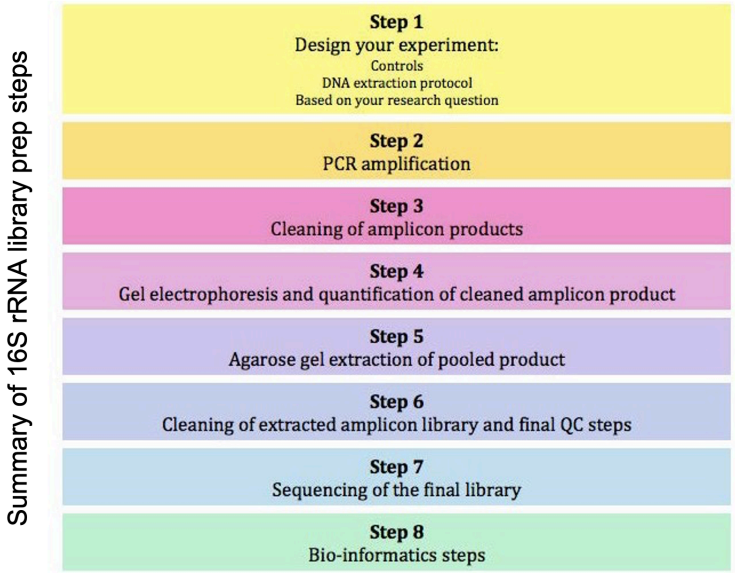


Figure 1.37: 16S Library Prep

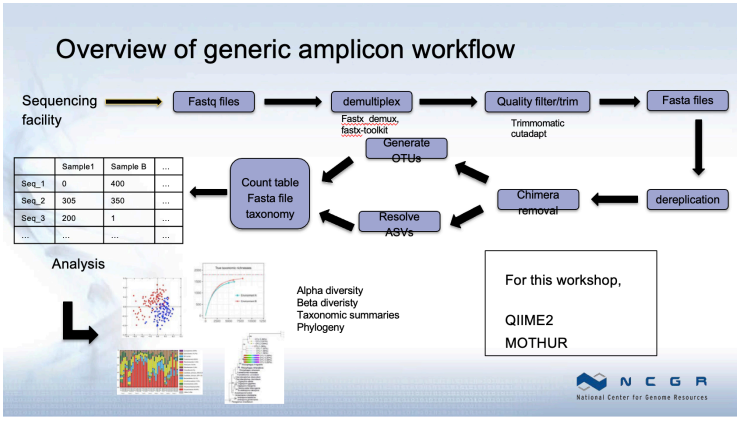


Figure 1.38: Amplicon Workflow

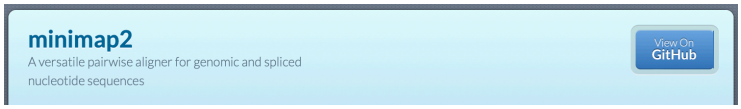
Chapter 2

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.

2.1 Activate the environment that contains minimap2




```
conda activate minimap2
```

- Make a directory and go into it

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

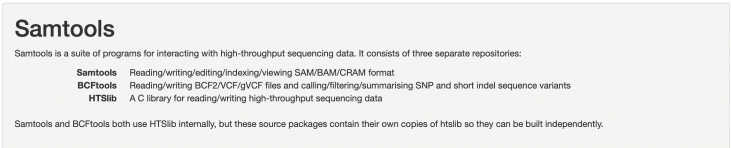
- Link to the merged minion reads

```
ln -s /home/cjb/minion/2022/data/AlderNodule3469-3/3469-3/20220701_2144_MC-113445_FAS2
```

- 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/indexes/red_alder_genome/consensus.fasta \
3469-3.all.fastq > 3469-3-minionxredalder.mm2.sam
```

2.2 Activate the environment that contains samtools




```
conda activate samtools
```

- 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

Reads are here:

```
/home/cjb/minion/2022/data/Aldernodule4956-3/4956-3/20220701_2154_MC-113286_FAS37509_f3119554/fas
```

Chapter 3

Data Processing

3.1 FastQC

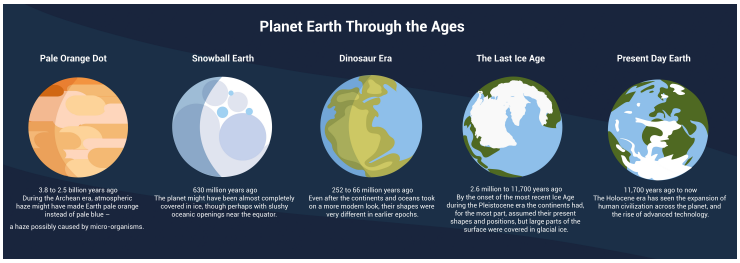


Figure 3.1: FastQC

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

3.2 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

3.3 Chimera detection and removal of non-bacterial sequences

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

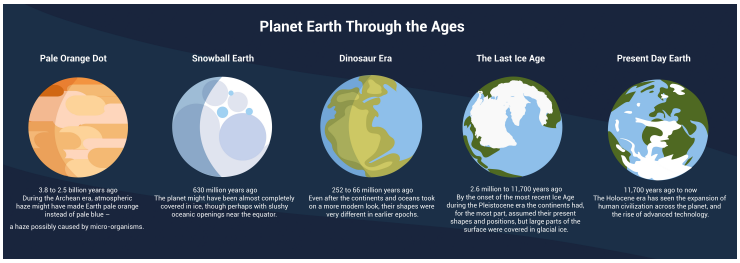


Figure 3.2: Chimera

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.

3.4 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

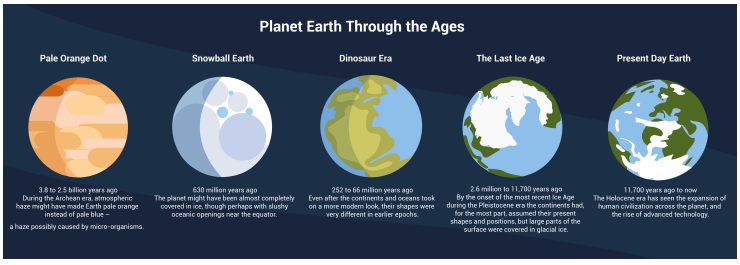


Figure 3.3: Clustering

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)

3.5 Create OTU tables

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

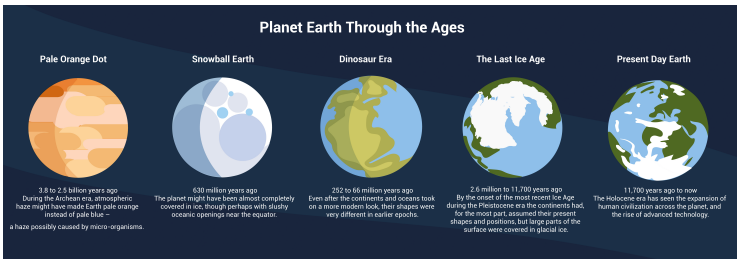


Figure 3.4: OTUs

3.6 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

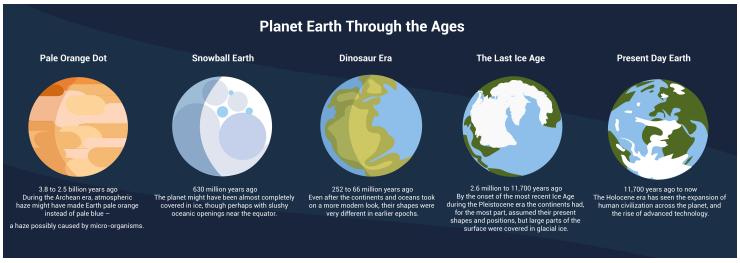


Figure 3.5: Database

3.7 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

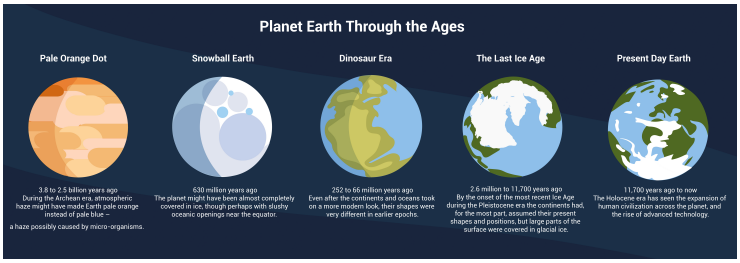


Figure 3.6: Human Mol. Genet., 2013

3.8 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)

3.9 Measuring Population Diversity: alpha and beta diversity

3.10 Diversity Measurements with 16s rRNA sequencing

1. Overall Benefits

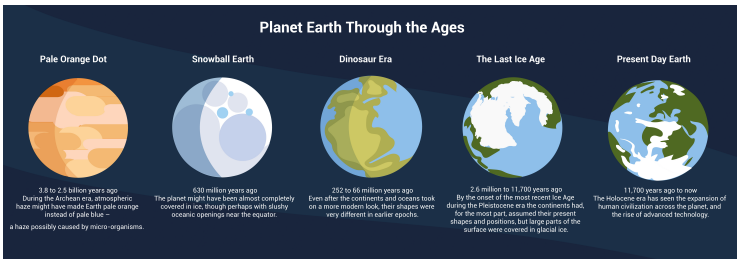


Figure 3.7: PLoS Computational Biol.,2012

- 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

3.11 Taxonomy: Expectation vs Reality

3.12 Beta Diversity - UniFrac

1. Measures how different two samples' component sequences are

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

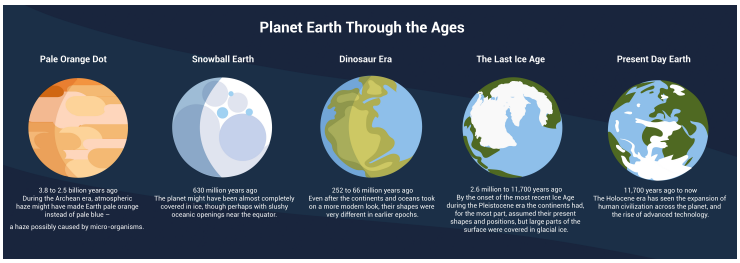


Figure 3.8: Expectation vs. Reality

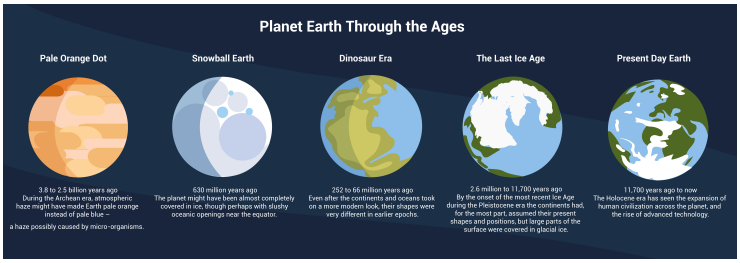


Figure 3.9: Unifrac

3.13 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?

Chapter 4

PycoQC

Quality Control

- "PycoQC computes metrics and generates interactive QC plots for Oxford Nanopore technologies sequencing data" (https://a-slide.github.io/pycoQC/)

1. What do we need in order to run a Quality Control Check?

- Sequencing Summary File
  - Automatically produced by the MinIon basecaller.
- PycoQC package + dependencies
  - Already downloaded for you in Logrus.
- Line of code to produce .html file.
- Line of code to secure copy file to your computer.

4.1 Sequencing Summary File

Where is it?

```
/home/cjb/minion/2022/data/AlderNodule3469-3/3469-3/20220701_2144_MC-113445_FAS21661_134c02ac/seq
```

4.2 PycoQC package & code

1. Where is it? How to activate it?

- Log in to Logrus
- Stay in your home directory (check with pwd)
- Type the following:

```
source activate seqtools
pycoQC
pycoQC -f inputfilename.txt -o outputfilename.html
```

4.3 Secure Copy

1. What terminal to copy from? What is the code?

- Open new Terminal window (Not Logrus, but keep Logrus terminal window open)
- Type:

```
scp username@logrus.training.ncgr.org:/home/username/outputfile.html ~/Desktop/
```

4.4 Open your URL

Find your file on your desktop.

Double-click to open, or right-click to select browser

4.4.1 Normalization

With normalization we are trying to get the correct relative gene expression abundances between cells.

Gene expression between cells is based on count data.

What does a count in a count matrix represent?

- mRNA Capture
- Reverse transcription of mRNA
- sequencing of a molecule of mRNA

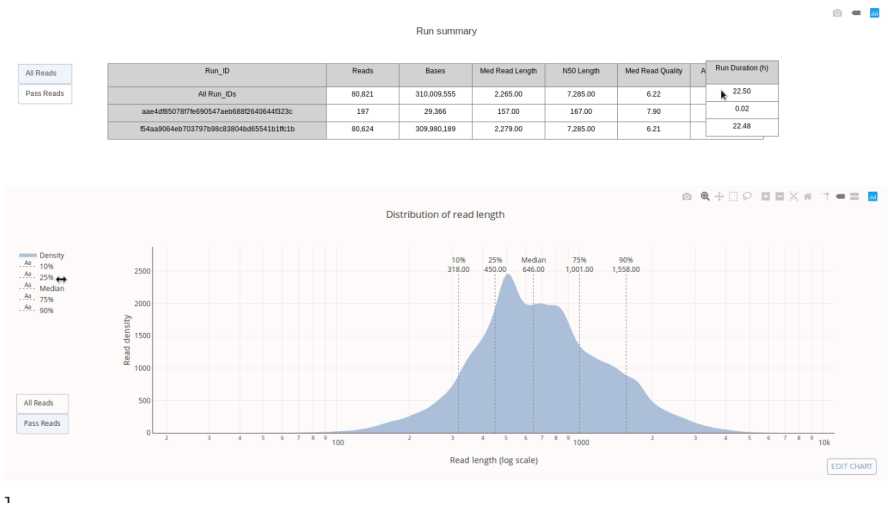


Figure 4.1: https://a-slide.github.io/pycoQC/

The most common normalization protocol is:

- count depth scaling
- aka CPM or counts per million
- it assumes that all cells in the dataset initially contain an equal number of mRNA molecules
- it assumes that count depth differences arise from sampling

Normalize complete

- But wait!
- We still have unwanted variability in the data.
- What kind of unwanted variability?
- What is the solution? Data Correction.

4.4.2 Data correction and integration

Biological Covariates

- Cell-Cycle effects
- Batch
- Dropout

Which Covariates to Consider?

- Depends on downstream analysis
- Correct for biological and technical to be considered separately
- Corrections are used for different purposes
- Each approach to correction presents unique challenges

What are the Correction methods?

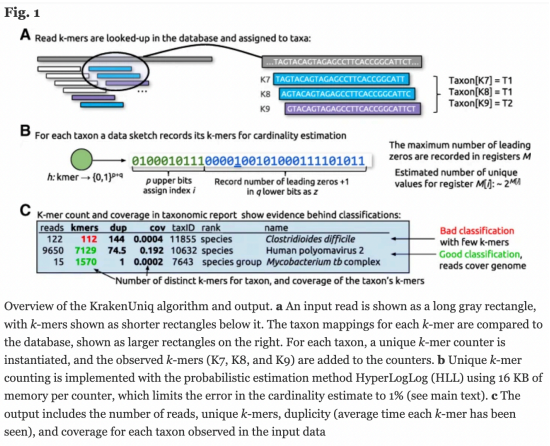
- Regressing out biological effects
- Regressing out technical effects
- Batch effects and data integration
- Expression recovery

Chapter 5

KrakenUniq

"confident and fast metagenomics classification using unique k-mer counts"

Default kmer = 31



Overview of the KrakenUniq algorithm and output. **a** An input read is shown as a long gray rectangle, with *k*-mers shown as shorter rectangles below it. The taxon mappings for each *k*-mer are compared to the database, shown as larger rectangles on the right. For each taxon, a unique *k*-mer counter is instantiated, and the observed *k*-mers (K7, K8, and K9) are added to the counters. **b** Unique *k*-mer counting is implemented with the probabilistic estimation method HyperLogLog (HLL) using 16 KB of memory per counter, which limits the error in the cardinality estimate to 1% (see main text). **c** The output includes the number of reads, unique *k*-mers, duplicity (average time each *k*-mer has been seen), and coverage for each taxon observed in the input data

Figure 5.1: KrakenUniq Overview

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1568-0

https://github.com/fbreitwieser/krakenuniq

https://gitlab.umiacs.umd.edu/derek/krakenuniq/-/blob/master/MANUAL.md

5.1 Kraken/Centrifuge Family

Kraken 1 is obsolete.

Centrifuge was written to improve Kraken's memory issues. It is completely new code with a different classification index. Its databases are also much smaller. Centrifuge can assign a sequence to multiple taxa. We'll use centrifuge later in the course.

KrakenUniq is based on Kraken 1. Adds an efficient algorithm to assess coverage of unique kmers, running at the same speed and with only slightly more memory than Kraken 1. KrakenUniq distinguishes low abundance organisms from false positives.

Kraken 2 uses "probabilistic data structures" to reduce memory and increase speed at the expense of lower accuracy that leads to false positives during the classifications ("a few dozen out of millions").

KrakenUniq was updated in May 2022 (v0.7+) that helped to deal with the large databases on computers without enough RAM to load them into memory. Databases can be read in chunks that fit in RAM.

Mode	Running time
`default (memory mapping)`	48 hours
`--preload`	14 min
`--preload-size 8G`	47 min
`--preload-size 16G`	32 min
`--preload-size 32G`	25 min
`--preload-size 64G`	19 min

Table 1: Running times for classifying 9.4 million reads (from a human eye metagenome, SRR12486990) with 8 threads using KrakenUniq in different modes. The database size was 392 GB, and it consisted of all complete bacterial, archeal, and viral genomes in RefSeq from 2020, 46 selected eukaryotic human pathogens [LS18]), as well the human genome and a collection of common vector sequences. In the database chunking experiments (using –preload-size) KrakenUniq loaded the database into main memory in 49, 25, 13 and 7 chunks (respectively).

Figure 5.2: Kraken preload

KrakenUniq gives you k-mer coverage information, reporting the number and percentage k-mers hit by reads. This helps to differentiate between false positives and good classifications.

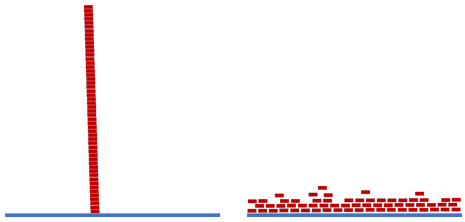


Figure 5.3: Alignment

https://www.biorxiv.org/content/10.1101/2022.06.01.494344v1.full.pdf

http://ccb.jhu.edu/software/choosing-a-metagenomics-classifier/#:~:text=However%2C%20while%20K

https://github.com/fbreitwieser/krakenuniq/blob/master/MANUAL.md

5.2 KrakenUniq databases

KrakenUniq requires Kraken1 not Kraken2 databases

We won't build a database since it can take several days, but let's look at the documentation:
https://github.com/fbreitwieser/krakenuniq#database-building

You can also get prebuilt databases:
https://benlangmead.github.io/aws-indexes/k2

We will use a bacterial databases located here:
/home/cjb/indexes/krakenuniq/db_052422

Link to it.

```
ln -s /home/cjb/indexes/krakenuniq/db_052422/ .
```

The taxa it contains are here:
/home/cjb/indexes/krakenuniq/db_052422/library/bacteria/library_headers.orig

Take a look at the file. How many sequences are there?

How many genera?

```
awk '{print $2}' /home/cjb/indexes/krakenuniq/db_052422/library/bacteria/library_headers.orig | s
```

And how many of each genera?

```
awk '{print $2}' /home/cjb/indexes/krakenuniq/db_052422/library/bacteria/library_headers.orig | s
```

5.3 Run KrakenUniq

Make a working directory and go into it.

```
mkdir ~/krakenuniq
cd ~/krakenuniq
```

Activate the KrakenUniq environment.

```
source activate krakenuniq
```

Run Kraken on sample 3469-3.

```
time krakenuniq \
    --db /home/cjb/indexes/krakenuniq/db_052422 \
    --threads 32 \
    --report-file 3469-3.report \
    --unclassified-out 3469-3.unclassified.fna \
    --classified-out 3469-3.classified.fna \
    ~/microbe_fastq/3469-3.microbe.fq.gz \
    > 3469-3.krakenuniqoutput
```

Make sure you got the following files: 3469-3.classified.fna 3469-3.krakenuniqoutput 3469-3.report 3469-3.unclassified.fna

This command will also bring up the input file as well (3469-3.microbe.fq.gz)

```
ls 3469*
```

Use the less command to look at each of them.

Now run it on the other samples.

5.4 Pavian

Copy the reports to your computer. Let's open them in Pavian.

Information on installing and some things you can do in Pavian is in the Centrifuge chapter.

Chapter 6

QIIME2

6.1 Hands On Command Line Tutorial

=======

Chapter 7

QIIME2

Command Line Tutorial

7.1 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
```

7.2 Setting up a working directory for QIIME2 analysis

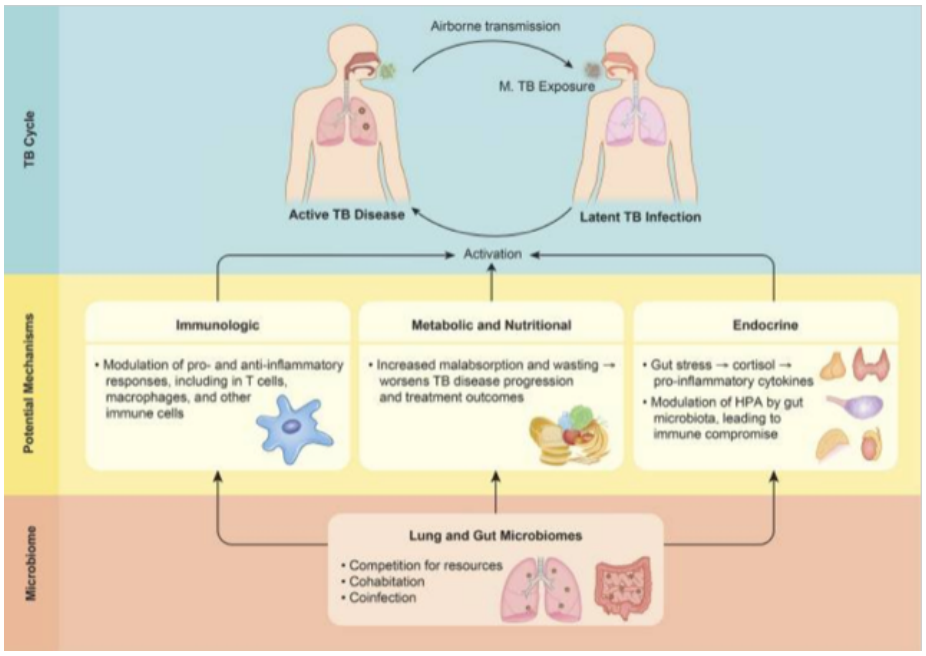
```
cd /home/$USER/

mkdir QIIME2_Analysis

cd QIIME2_Analysis
```

7.3 Dataset for analysis

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

Respiratory tract clinical sample selection for microbiota analysis in patients with pulmonary tuberculosis

Pulmonary Tuberculosis (The Human Microbiome in the Fight Against Tuberculosis, 2017)



Why study microbiome in relation to TB? 1. Microbiome composition is a potential risk factors for TB infection and disease susceptibility, disease progression, and treatment outcomes 2. TB infection and disease can affect microbiome, subsequently impacting health via alteration of immune responses

This pilot study was carried out using sputum, oropharynx, and nasal respiratory tract samples collected from patients with pulmonary tuberculosis and healthy control individuals, in order to compare sample types and their usefulness in assessing changes in bacterial and fungal communities.

7.4 Data download

Data was downloaded from the National Center for Biotechnology Information (NCBI) based on the SRA accession numbers reported in the manuscript - PRJNA242354. A program called fastq-dump version 2.10.0 was used to download the data directly from NCBI SRA database. Tbe following exercise will help you use fastq-dump for your own projects too.