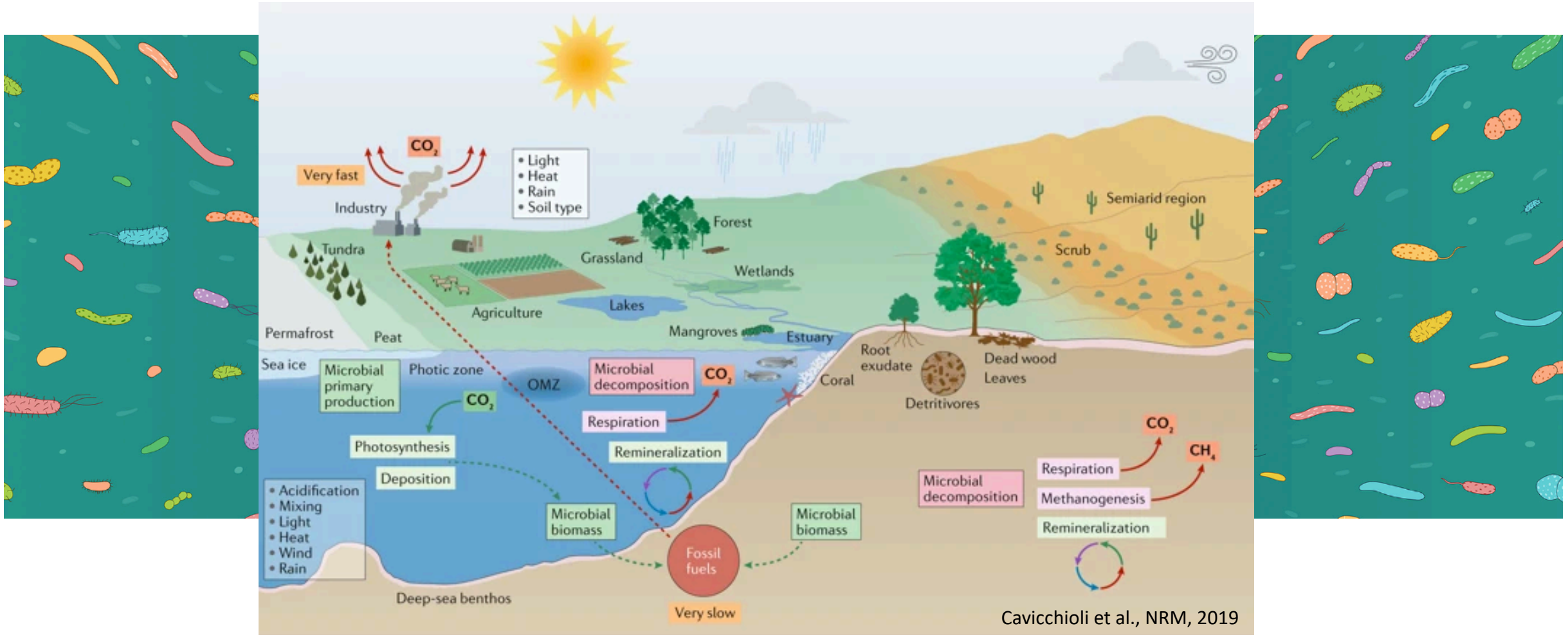
BIO390 - Introduction to Bioinformatics: Metagenomics

October 20th, 2020

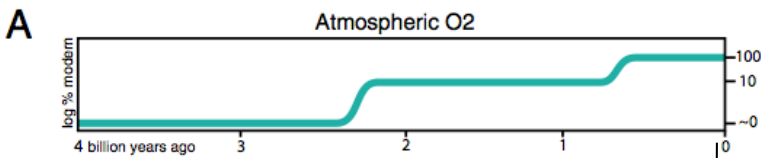
Shinichi Sunagawa
Microbiome Research Group
Institute of Microbiology, D-BIOL, ETH Zürich

The world is run by microbes



Evolution and significance of microbiomes

From the origin of life to today



Microorganisms

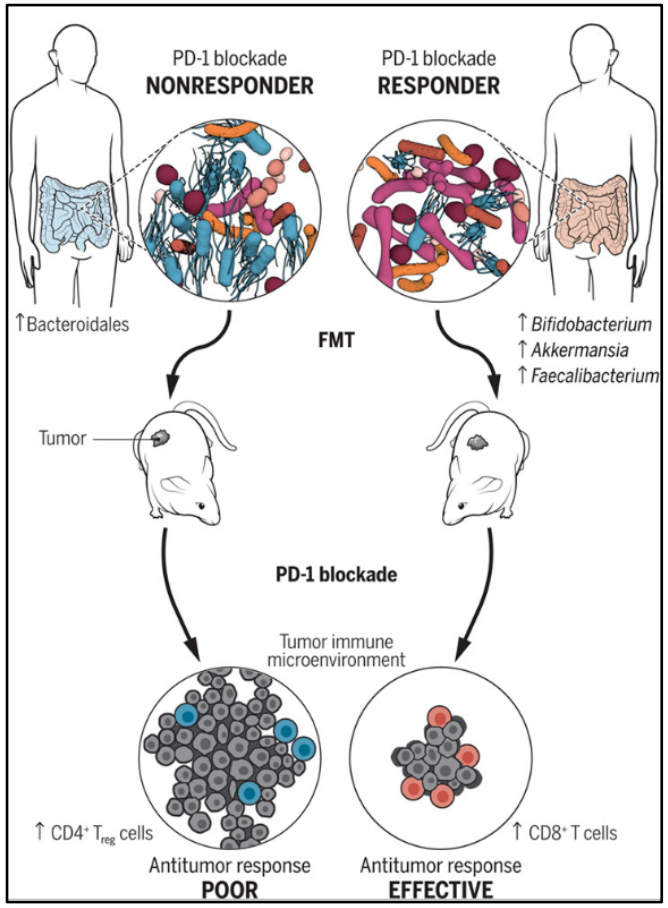
- originated some 3.8 billion years ago
- drive biogeochemical cycles of elements (C, N, P, S, etc.)
- transform energy and biomass

Significance (examples):

- biogeochemistry: e.g., photosynthesis by microbes, carbon fixation/export, nitrogen fixation
- health: help us digest food, provide essential vitamins, train the immune system

McFall-Ngai et al., PNAS, 2013; Venn et al., Science, 2014

Why should I care about formally describing microbial communities?



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

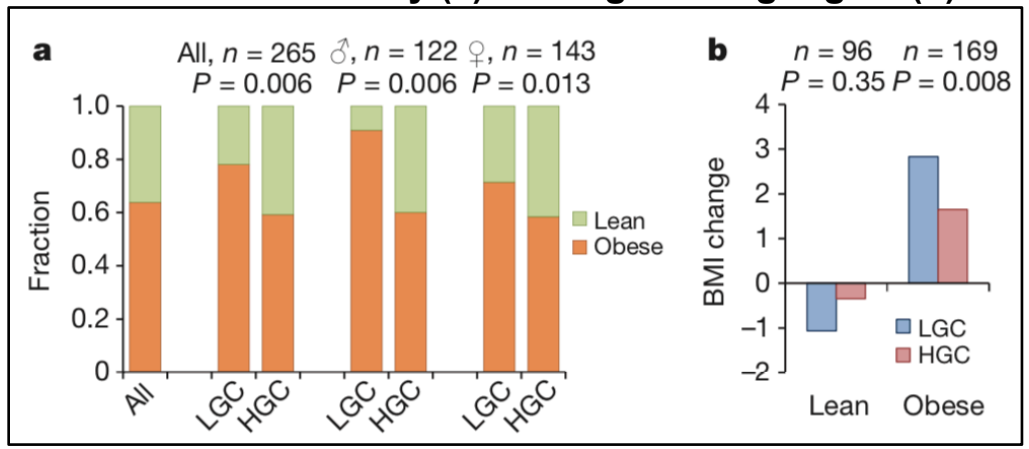
- Compositions of microbial communities are important to characterize, because many host-associated microbes are increasingly implicated in diseases (and personalized medicine)

→ Enrichment of specific microbial taxa may influence the response to cancer immunotherapy

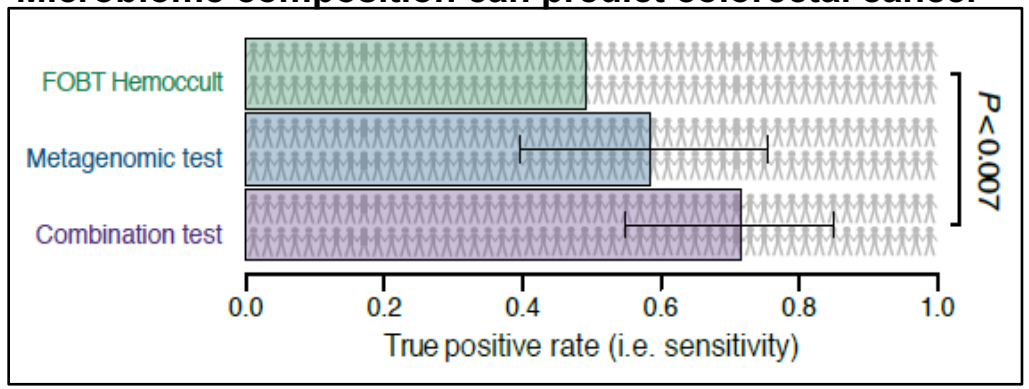
Routy et al., Gopalakrishnan et al., and Matson et al. Science 2018

Why should I care about formally describing microbial communities?

Less diverse microbiomes (LGC=low gene count) are associated with obesity (a) and higher weight gain (b)



Microbiome composition can predict colorectal cancer



- Many diseases are associated with low microbial diversity (e.g., obesity)
- Microbial community compositions can be indicative for disease (e.g., colorectal cancer)

Le Chatellier et al., Nature, 2014; Zeller et al., MSB, 2014

LO 1: Today, you will learn how to formally describe the composition and diversity of, and differences between microbial communities

Why should I care about reconstructing, annotating and assessing the quality of microbial genomes?

- Discovery of many new microbial species, even entire new phyla
- Requirement to advance from: “Who is there?” to “What can they do?”
- New microbes may have health, ecologically and/or economically relevant functions
- Taxonomic information alone does not provide any information about the genomic capacity

- Genome-resolved analyses can explain many biologically relevant differences between very closely related organisms
  - Compare: differences between two human individuals: same 18S rRNA sequence, 99.6% genome similarity

LO 2: Today, you will also learn about the process, value and challenges of reconstructing microbial genomes from metagenomes

Overview

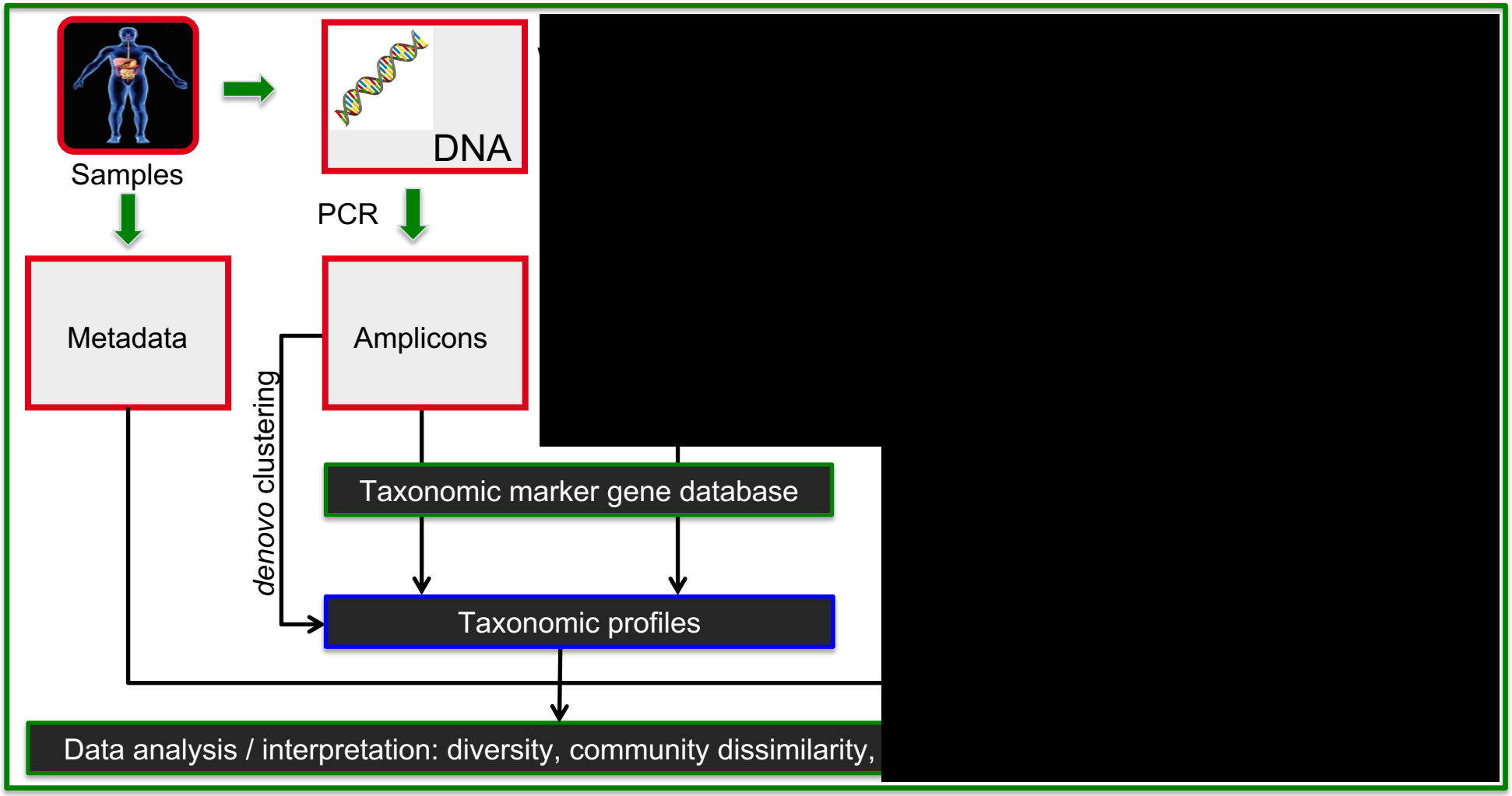
Part I - Microbial community structure

- definition of genotype, taxonomic resolution and operational taxonomic units (OTUs)
- analysis of microbial diversity, richness, evenness
- comparison of microbial community compositions

Part II – Reconstruction and annotation of microbial community genomes

- assembly of individual genomes and metagenomes
- binning of metagenomic assemblies
- annotation of metagenomes and metagenome assembled genomes

Overview – Part 1

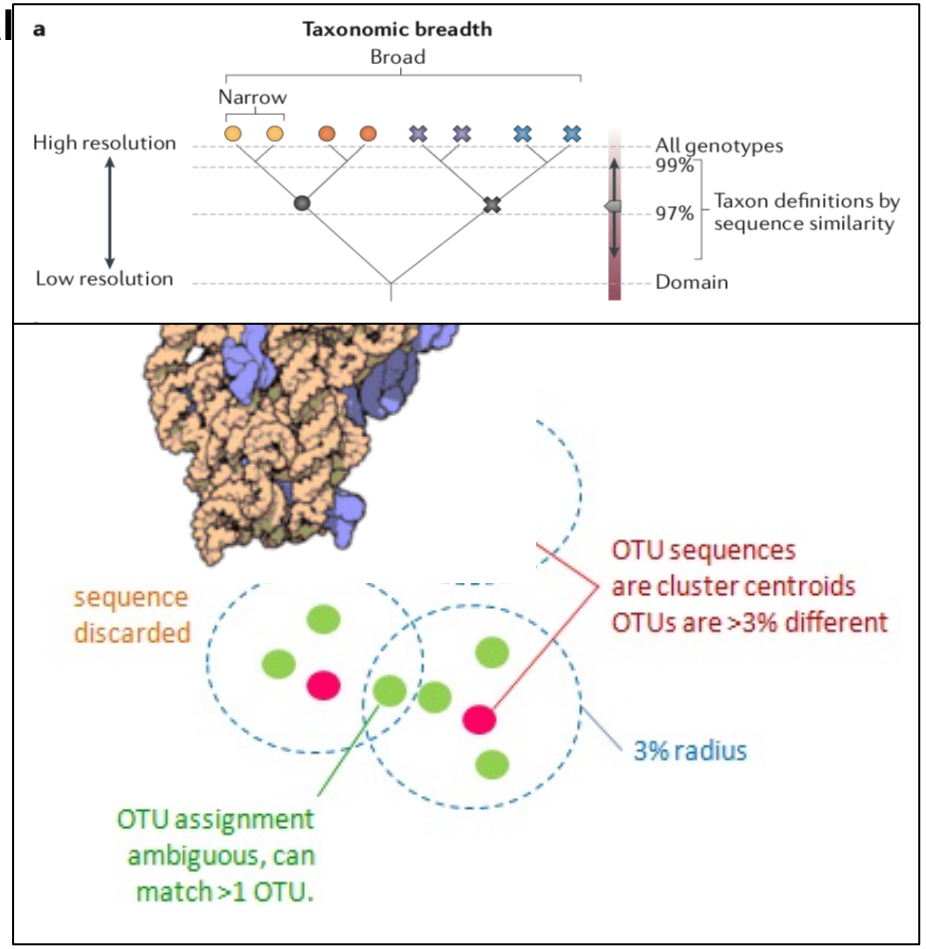


*WGS: whole genome shotgun sequencing

Review: 16S rRNA-based Operational Taxonomic Units (OTUs)

- 16S rRNA
  - present in all prokaryotes
  - conserved function as integral part of the protein synthesis machinery
  - gene is rarely horizontally transferred between different microorganisms
  - similar mutation rate: → molecular clock
- Proxy for phylogenetic relatedness of organisms
- Owing to lack of prokaryotic species definition, 97% sequence similarity is often used to define ‘species’-like:
  “**Operational Taxonomic Units” (OTUs)**

30S small



Hanson et al., Nat Rev Microbiol, 2012

Microbial community compositions

- Goal: determine ‘who’ is there at what abundance in one or more samples

OTU count table

	OTU1	OTU2	OTU3	OTU4	OTU5	…	Sum
S1	68	38	84	60	60		
S2	9	92	24	0	93		
S3	14	0	21	90	80		
S4	41	34	78	65	29		
S5	3	70	74	63	0		
…							

Rows: S1 to Sn = samples
Columns: OTUs

Step 1: Generation of 16S rRNA amplicon reads by PCR

Community DNA extract

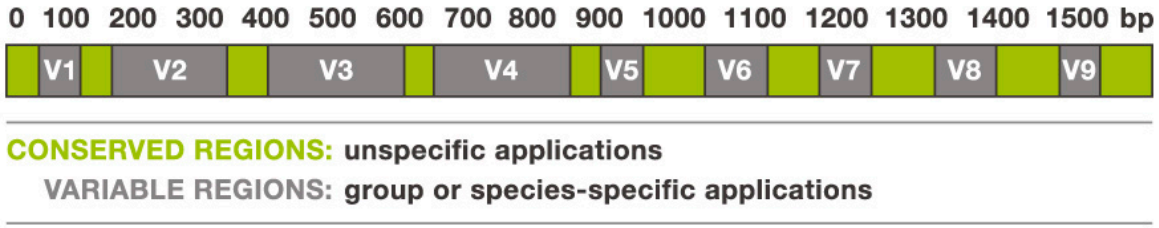
PCR

agtctcgctatgacgtcgtcgtcagactac

gtcgtacgtcgatatttctcgcgccggagc
gtcgtacgtcgatattctcgcggccggagc

agcctacgtcgtcgatagtgcgctagtgtc

- **Primers bind to conserved regions of constant regions.**
- **Variable regions are amplified by PCR**



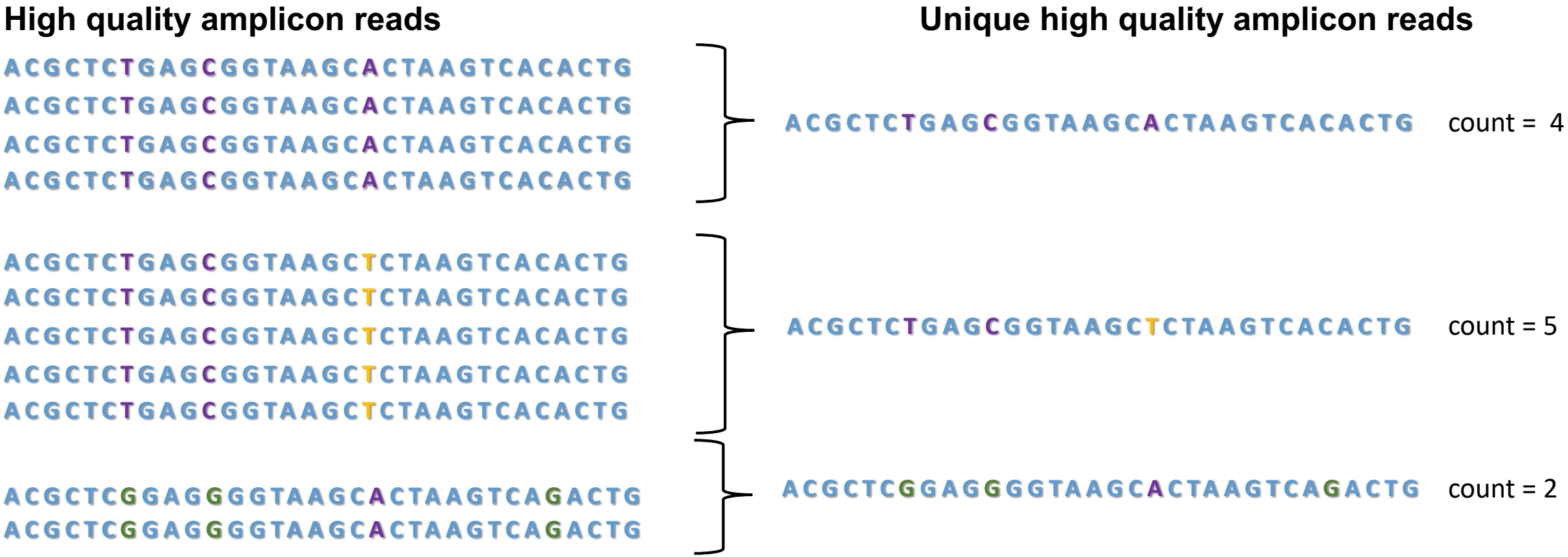
Example

- "V4 primers" yield ca. 250 bp long amplicon reads

- **After sequencing, amplicon reads are quality controlled, yielding high quality amplicon reads**

→ Number of reads are proportional to number of gene copies in the community

Step 2: De-replication of identical sequences



- All reads are aligned to each other to identify identical sequences
- Unique sequences are kept and the number of identical sequences is counted
- Output are unique sequences with records of identical sequences

Step 3: Heuristic clustering of sequences into OTUs

Deterministic approach: calculate all pairwise similarities

→ too "expensive" (resource and time consuming)

Heuristic approach:

1) Unique high quality reads are sorted by counts (high to low)
2) Read with highest count is centroid of a new OTU (N=1)
3) Next read is compared to all OTU centroids

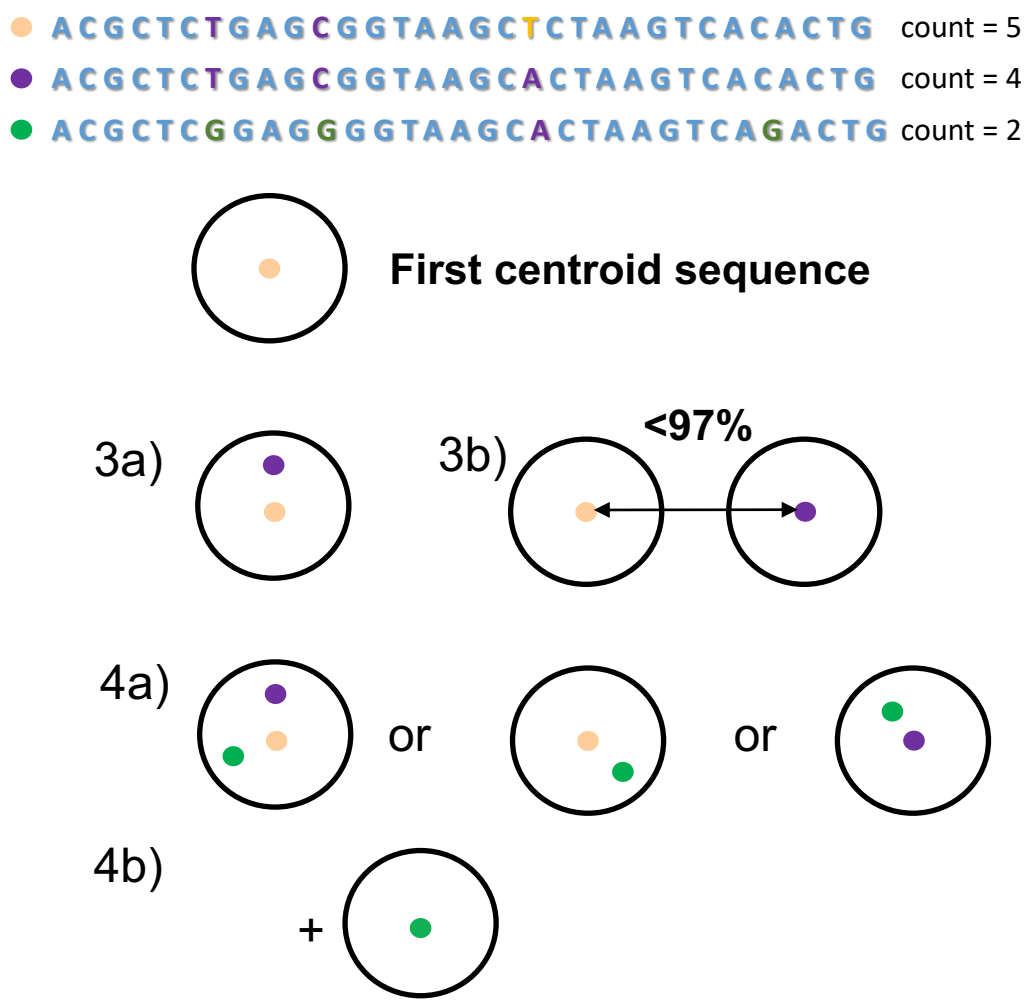
   2 different possibilities:

   a) Centroid sequence and new read are >= 97% identical
      - read becomes new member of the OTU (N=1)
   b) Centroid sequence and new read are < 97% identical
      - read becomes centroid of a new OTU (N=2)

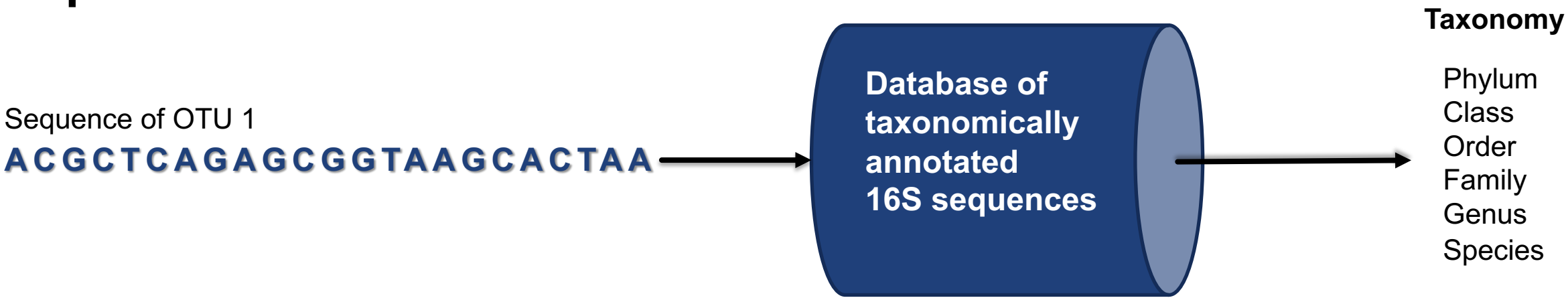
4) Next read is compared to all OTU centroids

   2 different possibilities:

   a) Any centroid sequence and new read are >= 97% identical
      - read becomes new member of the OTU (N=N)
   b) Any centroid sequence and new read are < 97% identical
      - read becomes centroid of a new OTU (N=N+1)



Step 4: Taxonomic annotation of OTUs

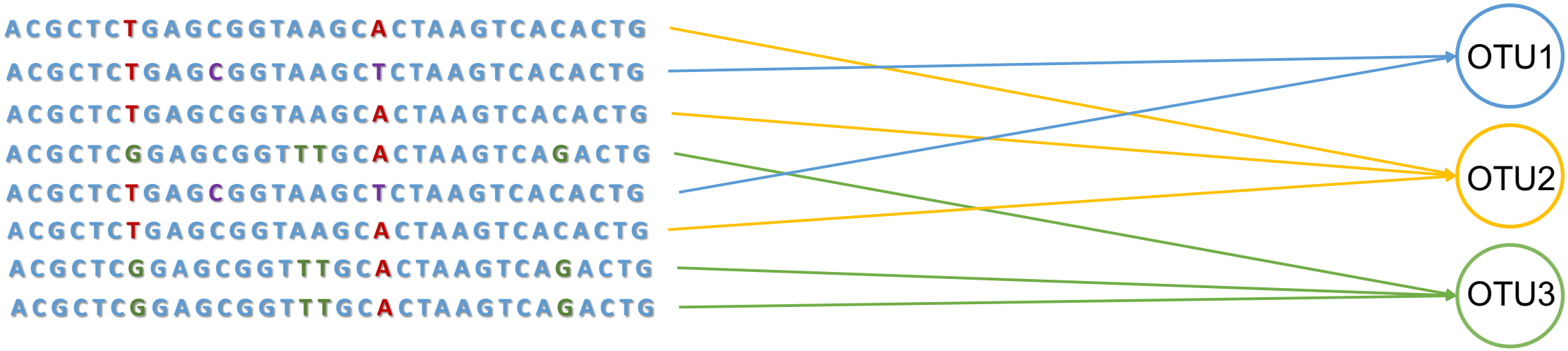


- Identification of taxon to which an OTU belongs
  - The centroid sequence of each OTU is compared to a database of annotated 16S rRNA gene sequences

→ sequences are assigned to taxonomic ranks: phylum, class, family etc.

Step 5: Quantification of OTU abundances

All reads are aligned to best matching OTU centroid sequence (and counted)



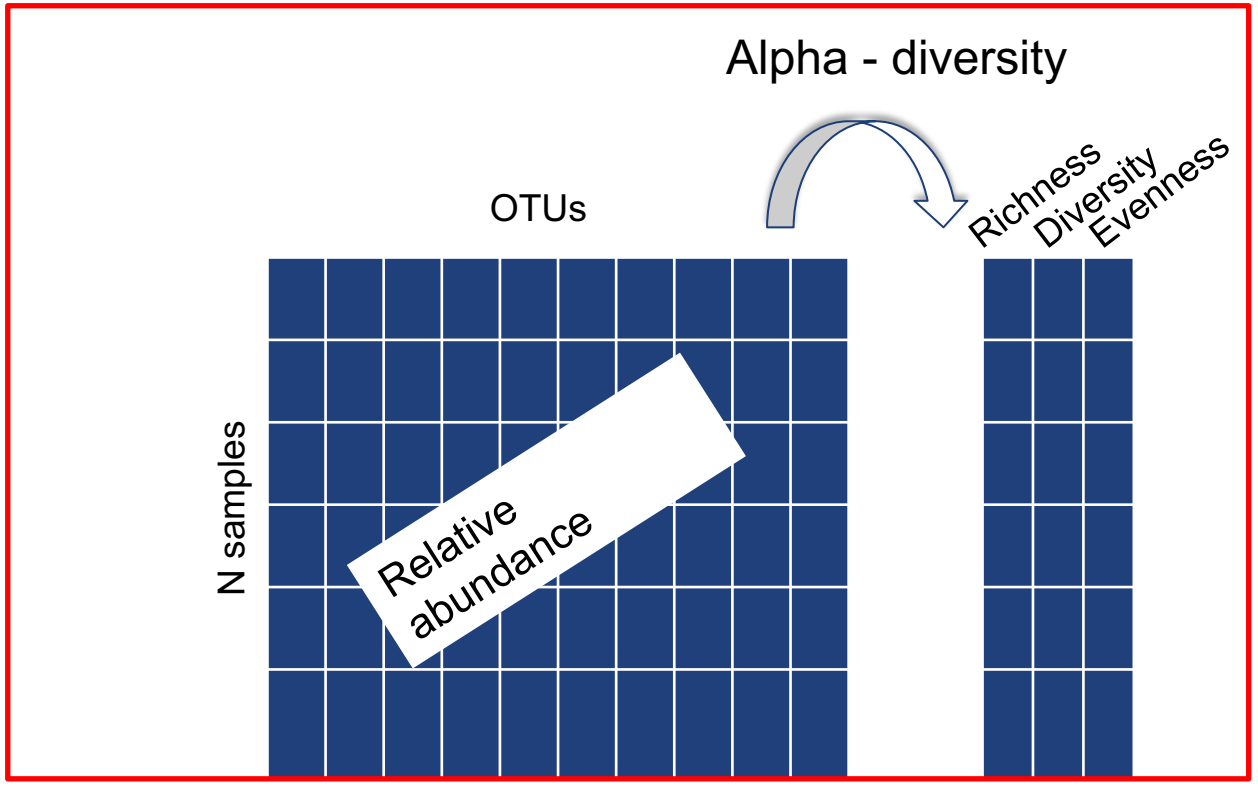
The result is an OTU count table, summarizing read counts for each OTU for each sample:

OTU	S1	S2	S3
OTU1	234	87	166
OTU2	23	0	93
OTU3	2	137	191
OTU4	455	0	112
OTU5	23	229	66

Data analysis / interpretation: diversity, community dissimilarity, sample classification

Concept of diversity

- How can we use an OTU count table to formally describe the diversity of a microbial community?



In-class task 1: alpha diversity

- In groups of 2, discuss how the diversity of one sample could be formally described (i.e., measured in quantitative terms)?
- Assume an example of 5 OTUs and 100 individuals (or 16S amplicons)

OTUs	Sample A	Sample B	Sample C	Sample D
1	20	1	25	0
2	20	10	25	0
3	20	20	0	0
4	20	30	25	0
5	20	39	25	100
Sum	100	100	100	100

- Consider distributing the 100 individuals in different ways among the 5 OTUs
- Also note that not all OTUs need to be present in a given sample

→ **What are the factors that influence the differences between samples?**

In-class task 1: alpha diversity

Commonly used indices are:

Shannon's diversity index (*H'*)

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

R = richness
p*i* = the proportion of the *i*-th OTU
n*i* = the number individuals of the *i*-th OTU
n = total number of individuals

Pielou's evenness (*J'*):

$$J' = H' / \log R$$

R = richness

Test your newly acquired knowledge

OTUs	Sample A
1	1
2	1
3	1
4	0
5	0

OTUs	Sample B
1	1
2	1
3	1
4	1
5	1

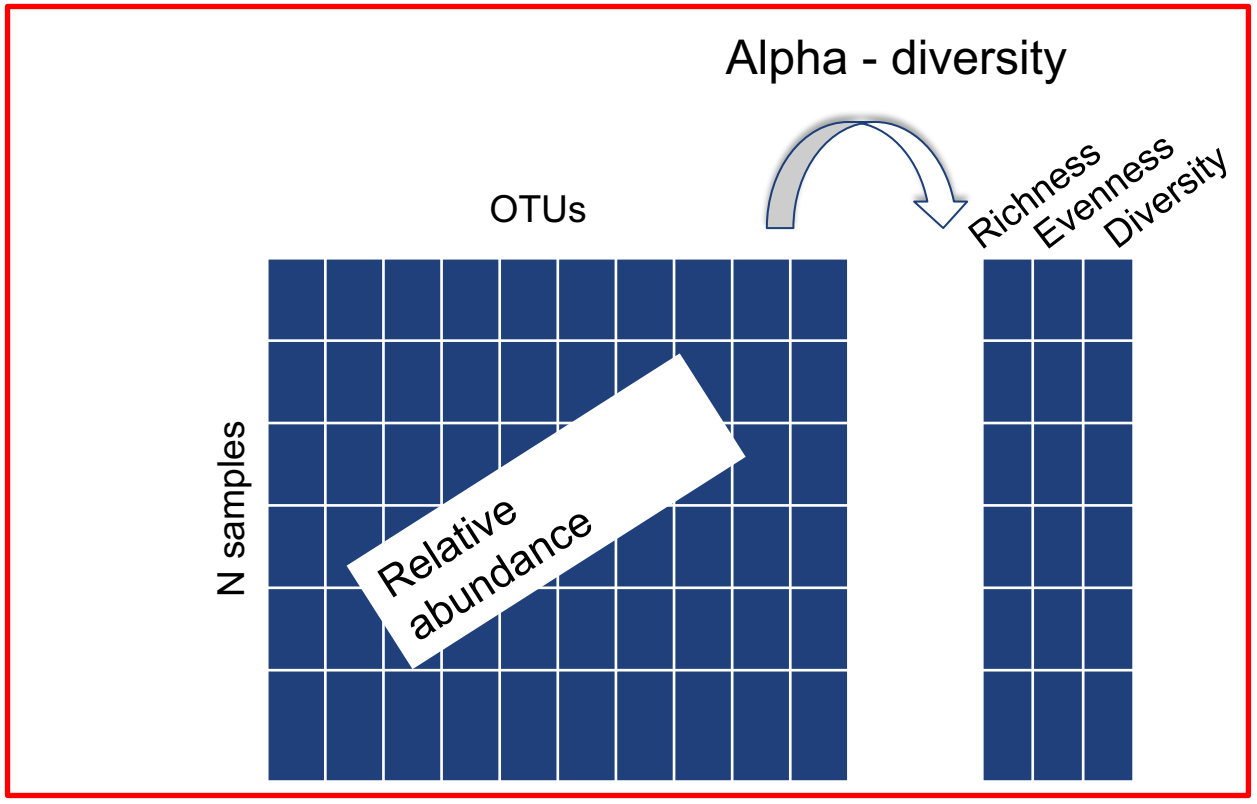
OTUs	Sample C
1	4
2	1
3	1
4	0
5	0

OTUs	Sample D
1	2
2	2
3	2
4	0
5	0

Which of the samples on the left (A-D) is:

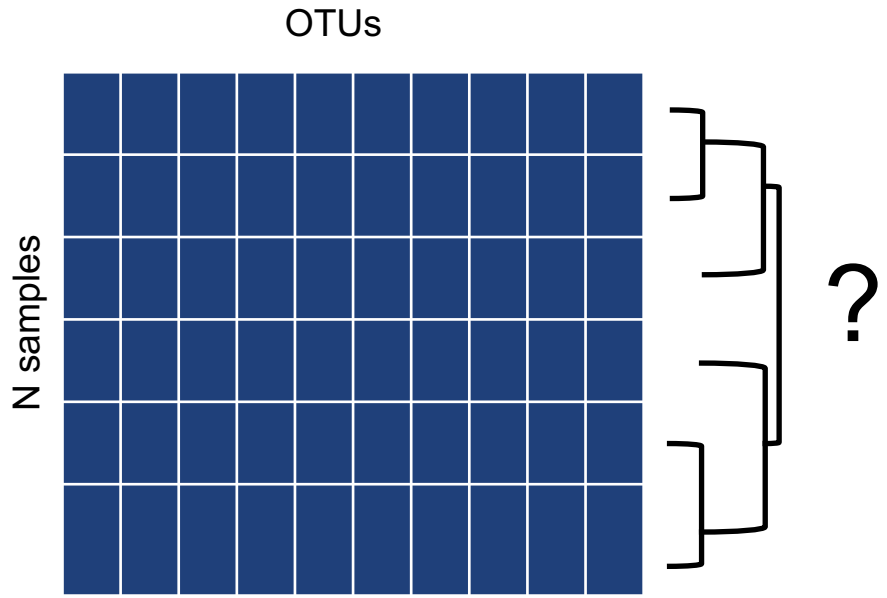
- **the richest?**
- **the most even?**
- **the most diverse?**

Concept of diversity - summary



Beta diversity: between sample dissimilarity

- Now that we learned how to describe the diversity of an individual sample, how do we compare different communities to each other?



In-class task 2: beta diversity

OTUs	Sample A
1	1
2	1
3	1
4	0
5	0

OTUs	Sample B
1	1
2	1
3	1
4	1
5	1

OTUs	Sample C
1	4
2	1
3	1
4	0
5	0

OTUs	Sample D
1	2
2	2
3	2
4	0
5	0

→ In pairs, please discuss how pairwise similarities of samples A, B, C, and D could be quantified?

→ Both qualitative differences vs quantitative differences can be taken into account.

In-class task 2: beta diversity

OTUs	Sample A
1	1
2	1
3	1
4	0
5	0

OTUs	Sample B
1	1
2	1
3	1
4	1
5	1

OTUs	Sample C
1	4
2	1
3	1
4	0
5	0

OTUs	Sample D
1	2
2	2
3	2
4	0
5	0

Example: Jaccard index/dissimilarity

Jaccard index: **J = a / (a + b + c)**

where
a = # of species shared
b= # of species unique to sample 1
c= # of species unique to sample 2

Jaccard distance / dissimilarity: **D = 1 - J**

→ **Note: For Jaccard distance, only presence/absence of species are considered!**

Other distance (dissimilarity) measures

The formulae for calculating the ecological distances are:

Bray-Curtis: $$D = 1 - 2\frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S}(a_i + c_i)}$$

Kulczynski: $$D = 1 - \frac{1}{2}\left(\frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S} a_i} + \frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S} c_i}\right)$$

Euclidean: $$D = \sqrt{\sum_{i=1}^{S}(a_i - c_i)^2}$$

Chi-square: $$D = \sqrt{\sum_{i=1}^{S}\frac{(a_+ + c_+)}{(a_i + c_i)}\left(\frac{a_i}{a_+} - \frac{c_i}{c_+}\right)^2} \text{ with } a_+ = \sum_{i=1}^{S} a_i$$

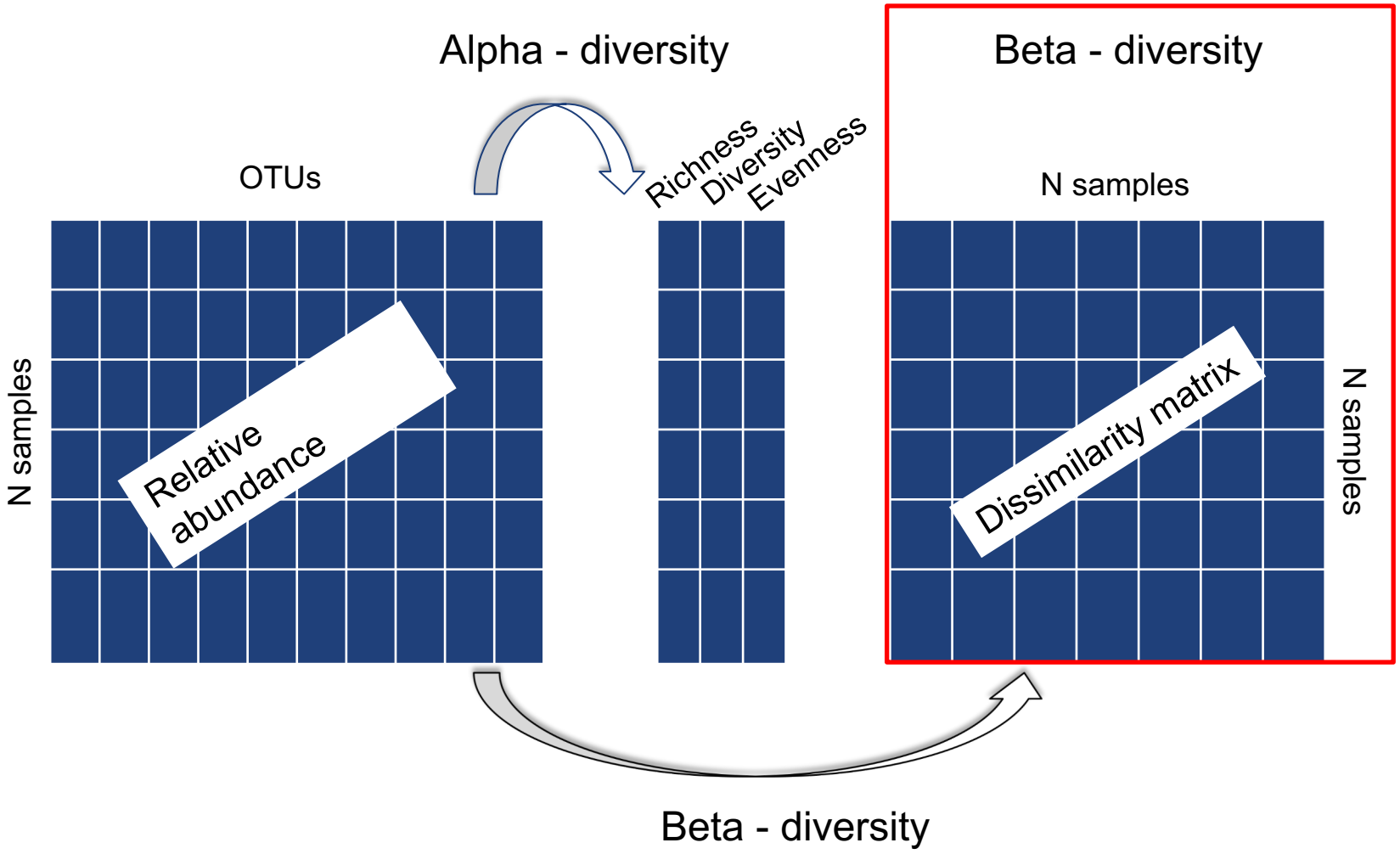
Hellinger: $$D = \sqrt{\sum_{i=1}^{S}\left(\sqrt{\frac{a_i}{a_+}} - \sqrt{\frac{c_i}{c_+}}\right)^2} \text{ with } a_+ = \sum_{i=1}^{S} a_i$$

UniFrac distance = phylogenetically weighted distance

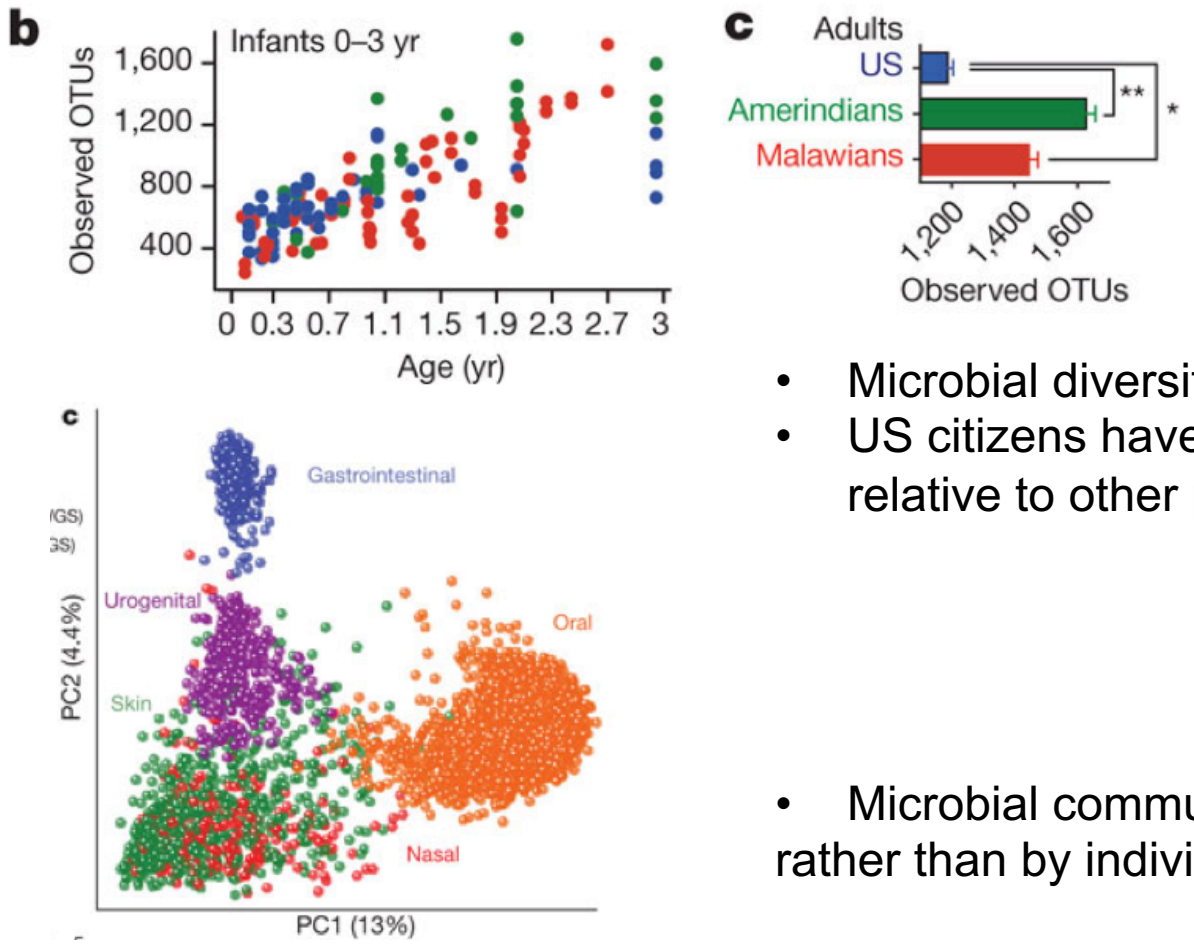
Lozupone et al., 2005

a_i = abundance of taxon *i* in sample *a*, and
c_i = abundance of taxon *i* in sample *c*

Within sample descriptions → between sample comparisons



Applied examples I

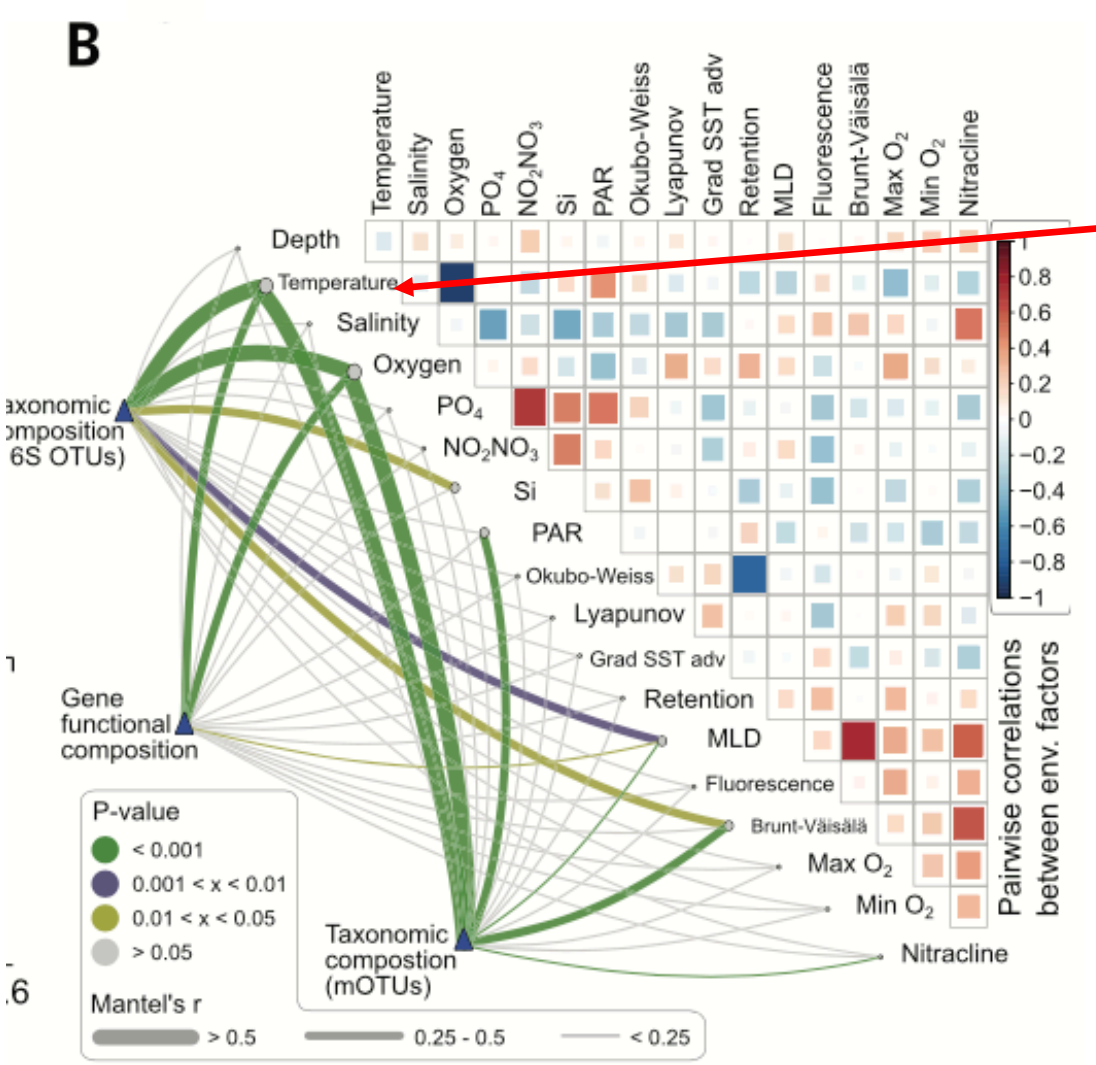


- Microbial diversity in human gut increases with age
- US citizens have harbor less diverse gut microbiota relative to other populations

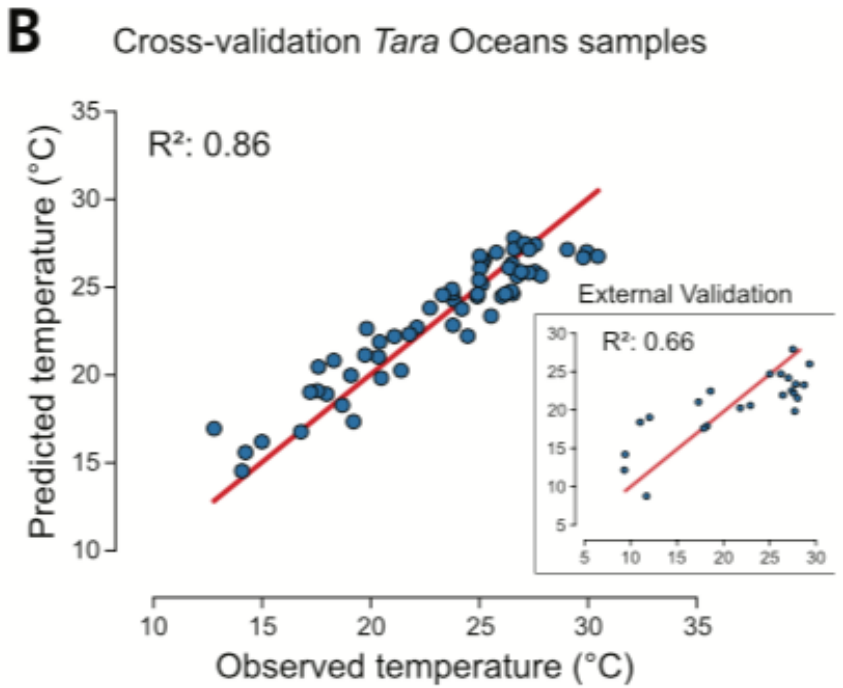
- Microbial communities cluster by human body site rather than by individual

Top: Yatsunenko *et al. Nature* (2012); Bottom: C Huttenhower *et al. Nature* (2012)

Applied examples II



Temperature has highest correlation with surface microbial community composition in the open ocean



Sunagawa et al. *Science* (2015)

Summary – Part I

- Metagenomics provides information about microorganisms that are often difficult or impossible to be cultivated in their natural environment
- Due to the lack of species concept for prokaryotes, researchers use sequence identity cutoffs of marker genes to define operational taxonomic units
- Microbial community structure describes the richness (number of species) of taxa and their evenness (the distribution of their abundances)
  - Alpha diversity (within sample diversity) is a function of richness and evenness
  - Alpha diversity can be quantified by diversity indices (e.g., Shannon, Simpson)
- Beta diversity describes differences in microbial community structures
  - Differences are quantified by dissimilarity indices (e.g., Jaccard, Bray-Curtis)

Overview of the Metagenomics lecture

Part II – Reconstruction and annotation of microbial community genomes

- assembly of individual genomes and metagenomes
- binning of metagenomic assemblies into metagenome-assembled genomes
- taxonomic and functional annotation of metagenomes

Background: from taxonomic composition to genomic content

- First microbial genome (1995): *Haemophilus influenzae* *Fleischmann et al. 1995*
- Followed by many isolated pathogens of diseases (plague, anthrax, tuberculosis, Lyme disease, malaria, and sleeping sickness)
- Many isolates of important non-pathogenic species: e.g., *Prochlorococcus*, *Lactobacillus*, *Bradyrhizobium*

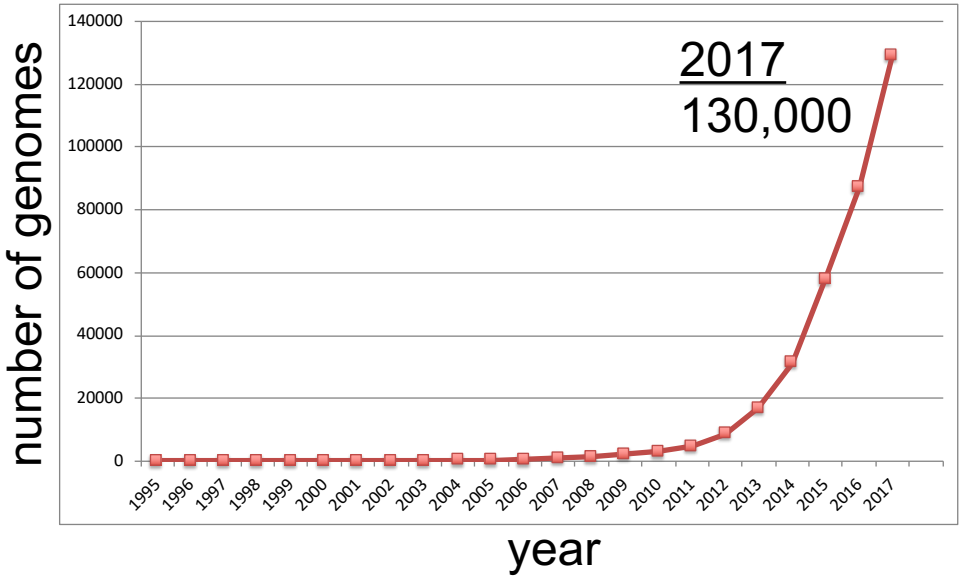
- Bacteria and archaea have ca. 500–10,000 genes, usually arrayed on circular DNA molecules (e.g., chromosomes and plasmids)
- Protein coding genes are on average ca. 1,000 base pairs long
- Their genomes are ca. 600,000–12 million bp in size (human 2 x 3 billion bp)

Background: added value of metagenomics

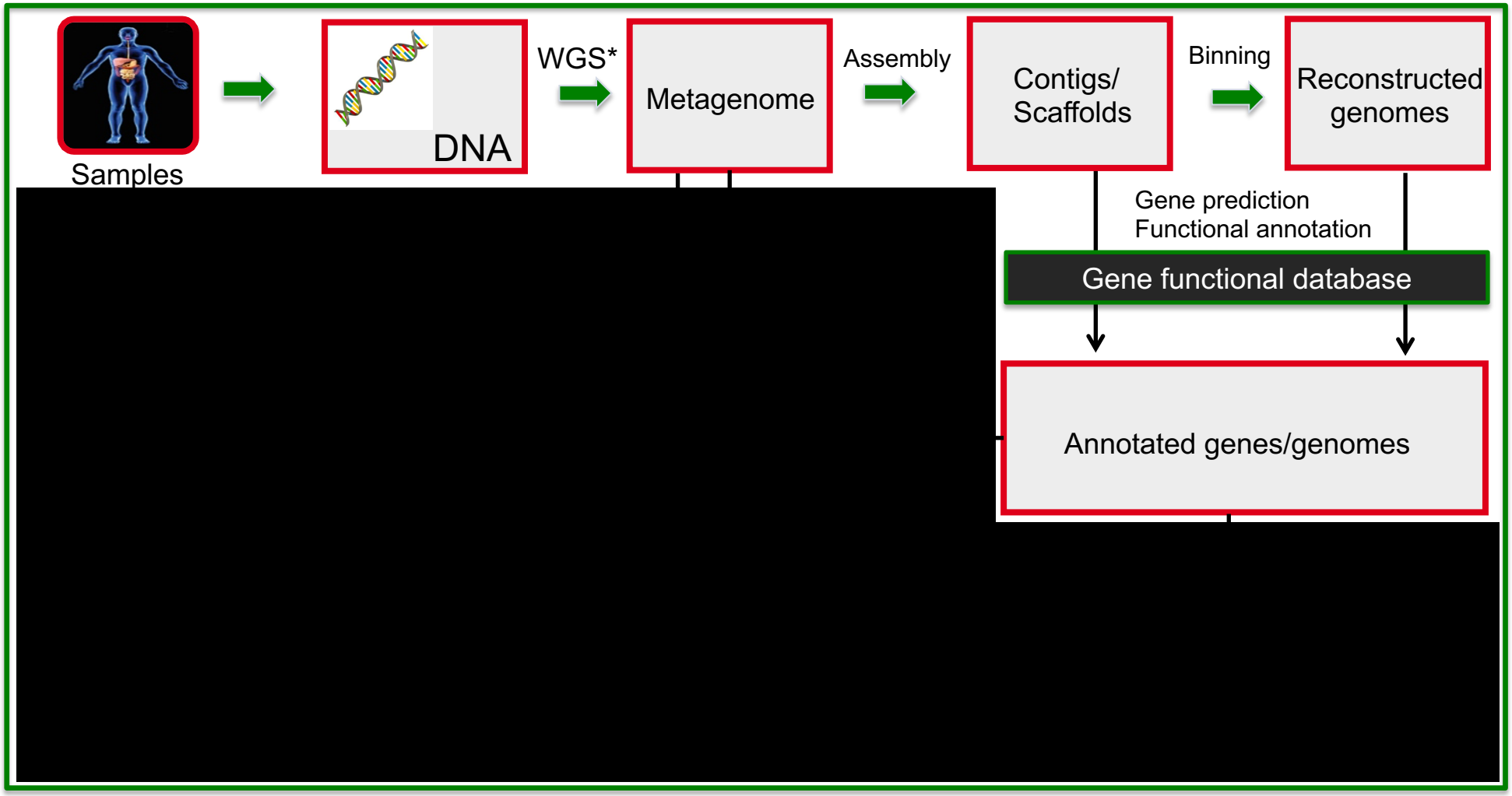
- Most bacteria and archaea cannot be isolated, as they live in communities and depend on other organisms
- Metagenomics provides access, in principle, to all genomic resources of a microbial community. This allows us to ask “what can they do?” (in addition to “who is there?”).

- Sequencing costs human genome: $1,000
- Microbial genomes: ~$1

→ Costs for data analysis are now much higher than for data generation

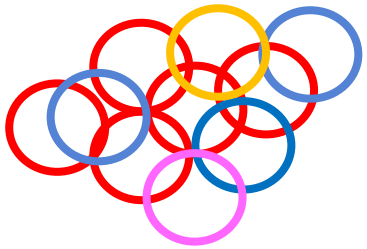


Overview – Part 2



*WGS: whole genome shotgun sequencing

Reconstruction of (community) genomes: overview



Metagenome: many genomes with different abundances

High throughput sequencing →

Raw reads

Data quality control →

HQ reads

Assembly →

contig 1

contig 3

contig 2

Scaffolding ↓

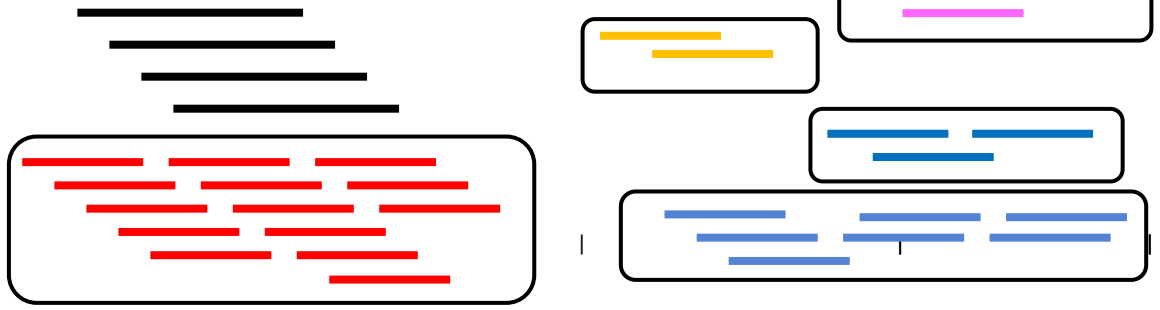
scaffold 1

scaffold 2

Binning ↓

Bins or metagenome assembled genomes (MAGs):
Contigs / scaffolds that originate from the same organism

(many) unbinned contigs

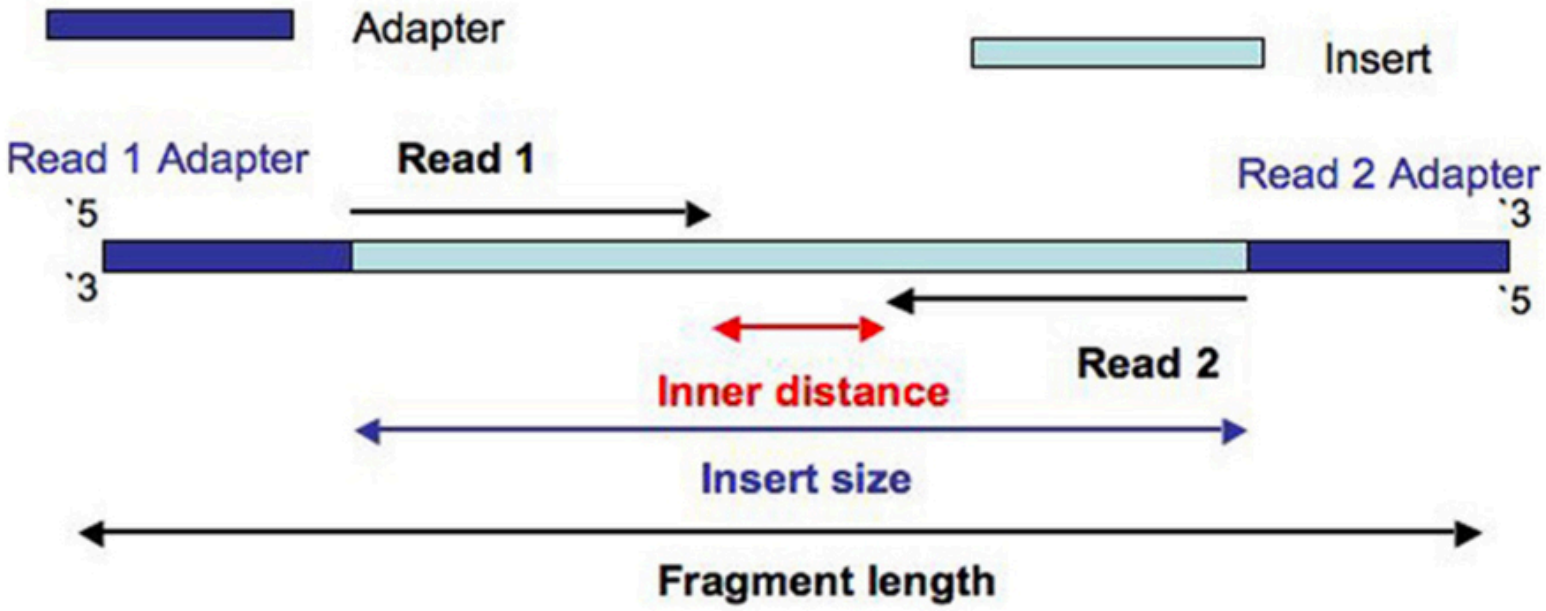


Gene prediction and annotation ←

Analysis

Background: DNA sequencing libraries

- DNA extracted from a metagenomic sample is randomly sheared into inserts of known size distribution (i.e., min, max, mean)
- Adapters are added to facilitate the sequencing of these inserts



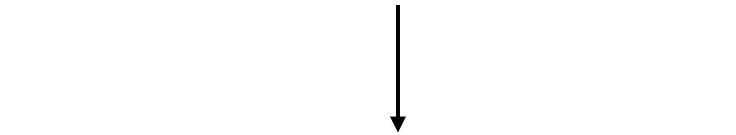
Note: Paired end reads may be overlapping providing the possibility to “merge” reads into one or not. In the latter case, the sequence between the paired reads remains unknown, while the length can be estimated due to the known insert size distribution

Step 1: Data quality control - sources of errors

1. Low base calling quality scores

original sequence

A C T G A A C T A A G T A



sequenced read

A C T G A A C T C A A N A T T T A G C T G C A

40 40 40 40 30 20 40 40 10 30 20 5 30

adapter

1. Same base
2. Different base
3. Additional base
4. Undefined base

Other sources of error

2. Residual adapter sequences
3. Residual control DNA sequences (e.g., “PhiX spike-ins”)
4. Contamination from non-target organisms

Base calling quality (*phred*) scores

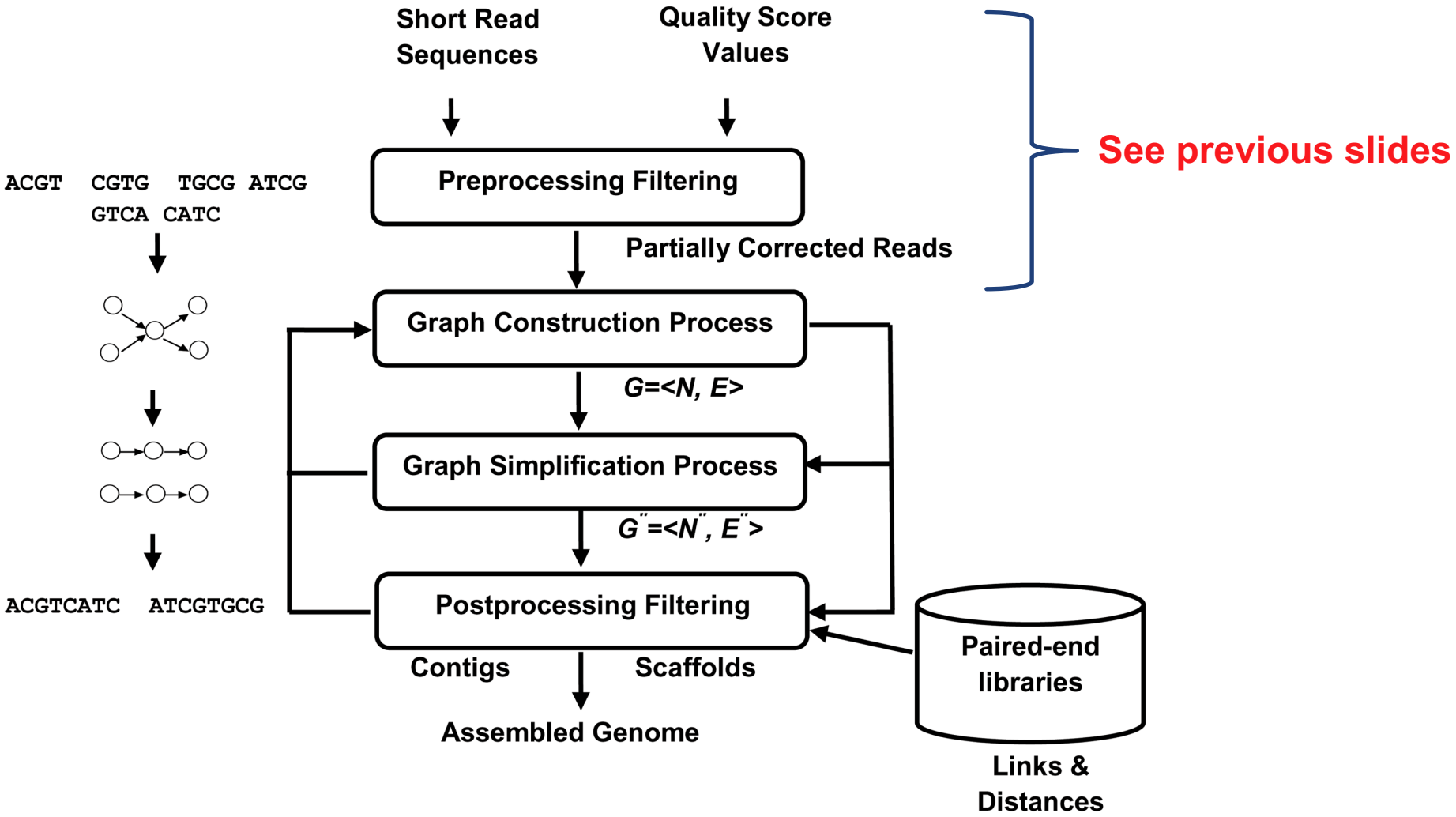
$Q = -10 \log_{10} P$

Probability of error: $P = 10^{-Q/10}$

Probability of truth: $1 - P$

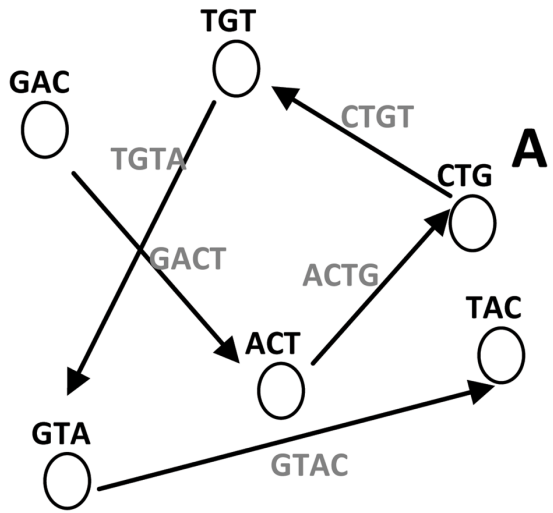
Quality score	% Correct Base
40	99.99
30	99.9
20	99
10	90

Step 2: Assembly and scaffolding: overview



El-Metwally et al., PLoS Comp Biol, 2013

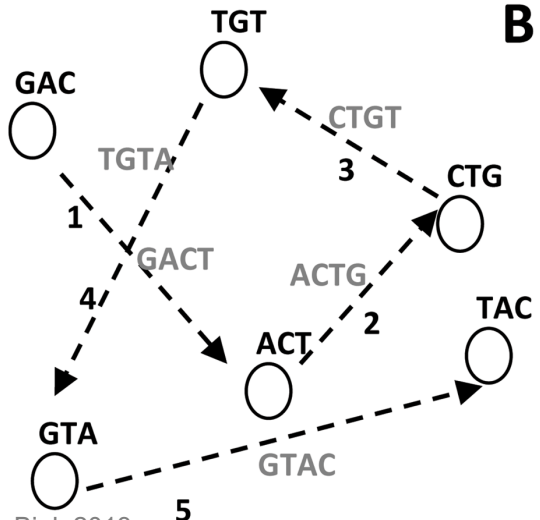
Graph construction: k-mer based assembly



R_1= GACTGTA R_2= ACTGTAC

Set of 3-Kmers of R_1= GAC, ACT,CTG,TGT,GTA

Set of 3-Kmers of R_2= ACT,CTG,TGT,GTA ,TAC



Example of an Eulerian path :

GACT

ACTG

CTGT

TGTA

GTAC

C Assembled Reads :

GACTGTAC

A) <u>k-mer-based</u> graph

Nodes = k-mers

Edges = k-1 overlaps

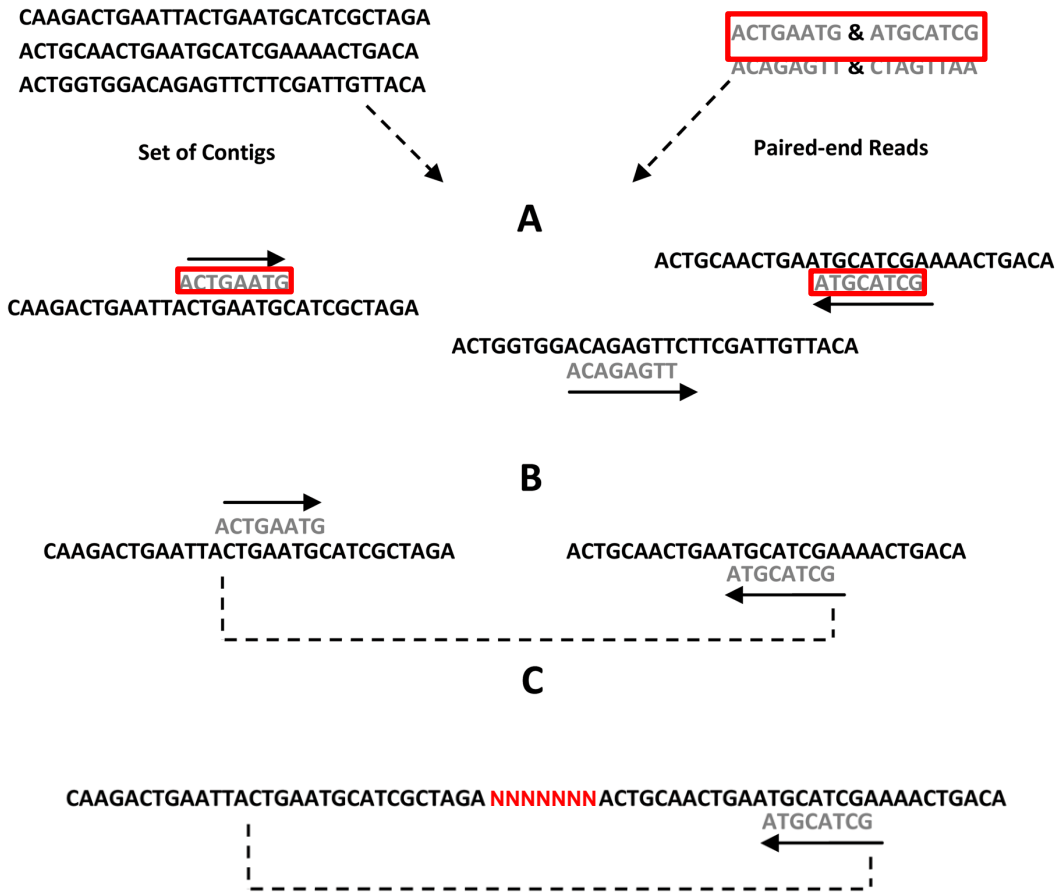
B) <u>Layout</u> shortest Eulerian path

Visit each <u>edge</u> once

C) Combine into <u>consensus</u>

El-Metwally et al., PLoS Comp Biol, 2013

Post processing: scaffolding



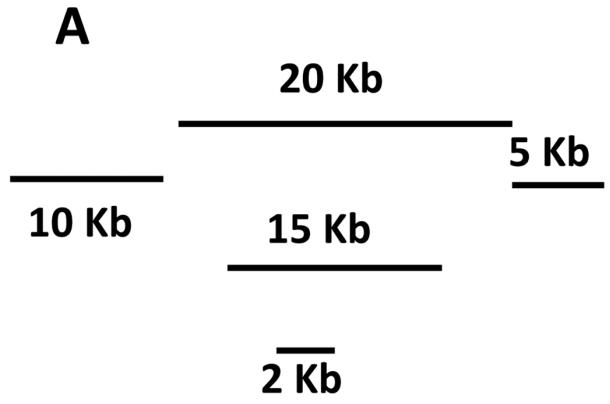
A) Align paired-end reads

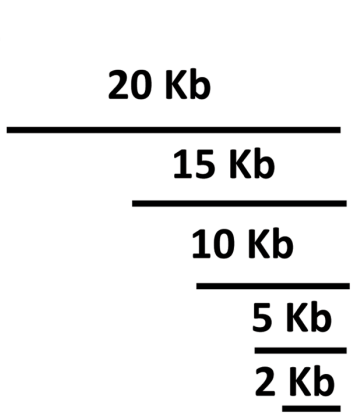
B) Orientation of contigs
→ according to read orientation

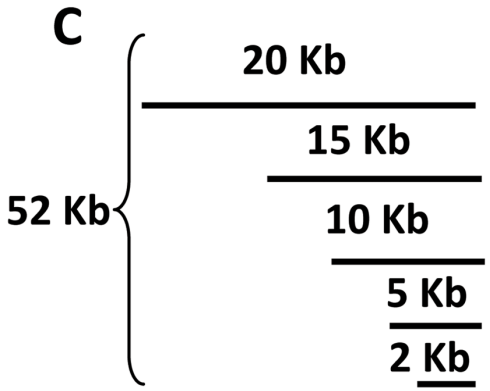
C) Scaffolding
→ use information on insert size distribution and fill 'gaps' with 'Ns'

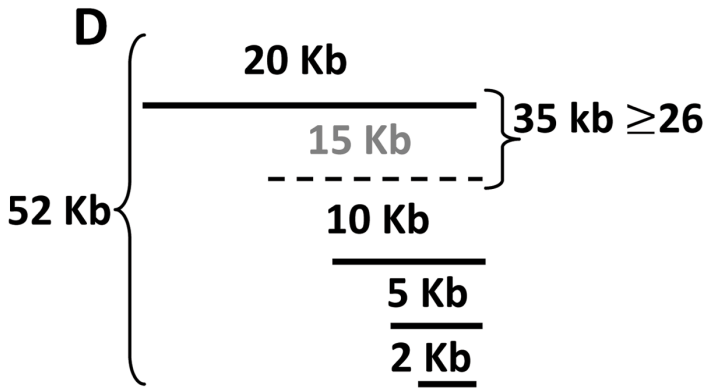
El-Metwally et al., PLoS Comp Biol, 2013

Assembly quality: N50 and L50









A) Set of contigs

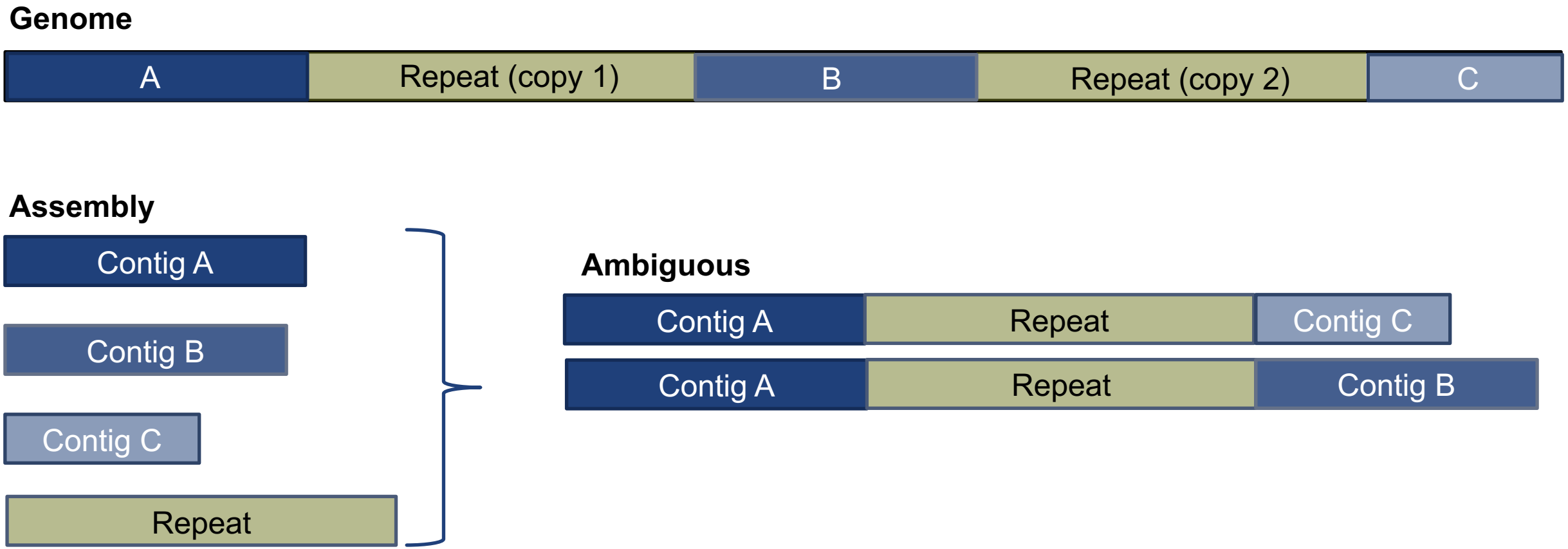
B) Sort contigs by size in descending order

C) Calculate total length of all contigs

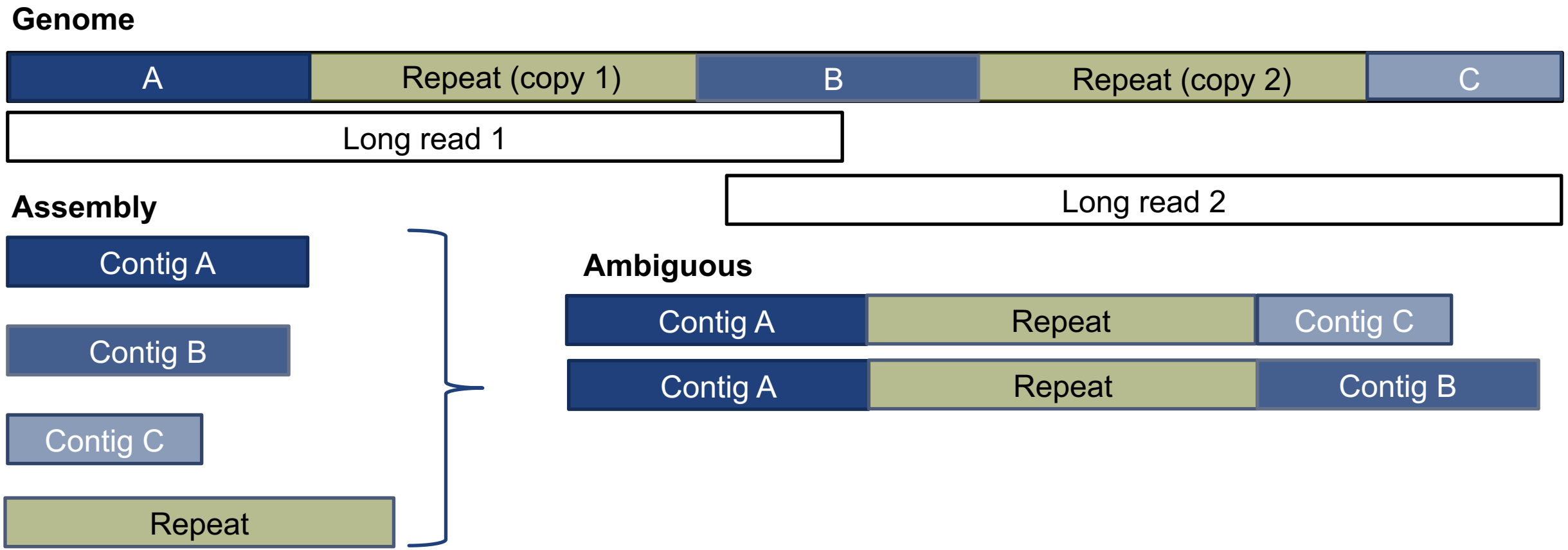
D) Add length in descending order until sum equals or exceeds 50% of the total length.

→ N50 = size of last contig added (15kb)
→ L50 is the number of contigs required to equal or exceed N50

Repetitive sequences in genomes prevent full assembly



Repetitive sequences in genomes prevent full assembly

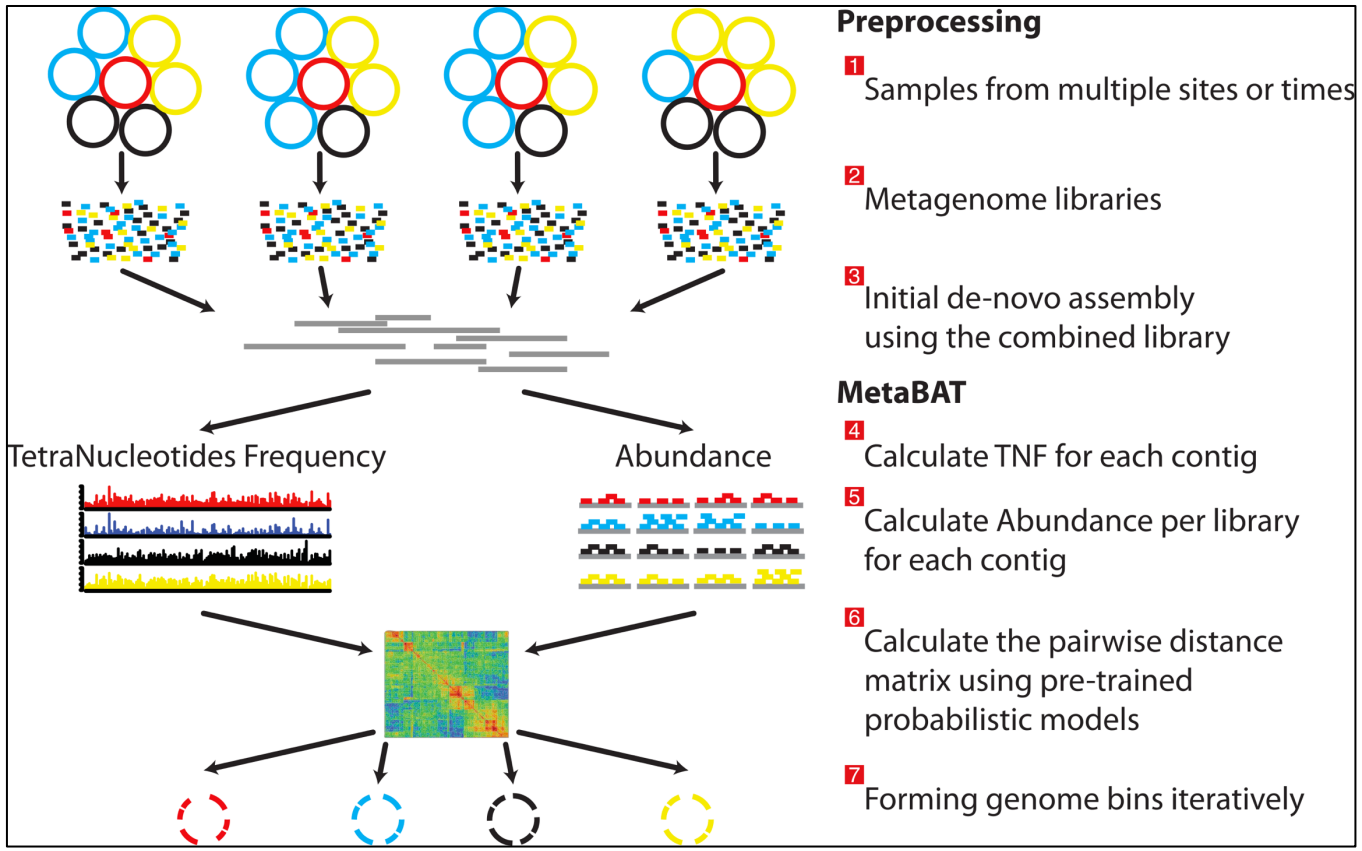


→ **Long sequencing reads can be used to resolve repeats**

Step 3: Binning contigs/scaffolds – composition guided

Composition-guided binning (independent from external information):

Tetranucleotide (or other k-mer) frequencies (e.g., GGAG vs. GGAC) + contig abundance

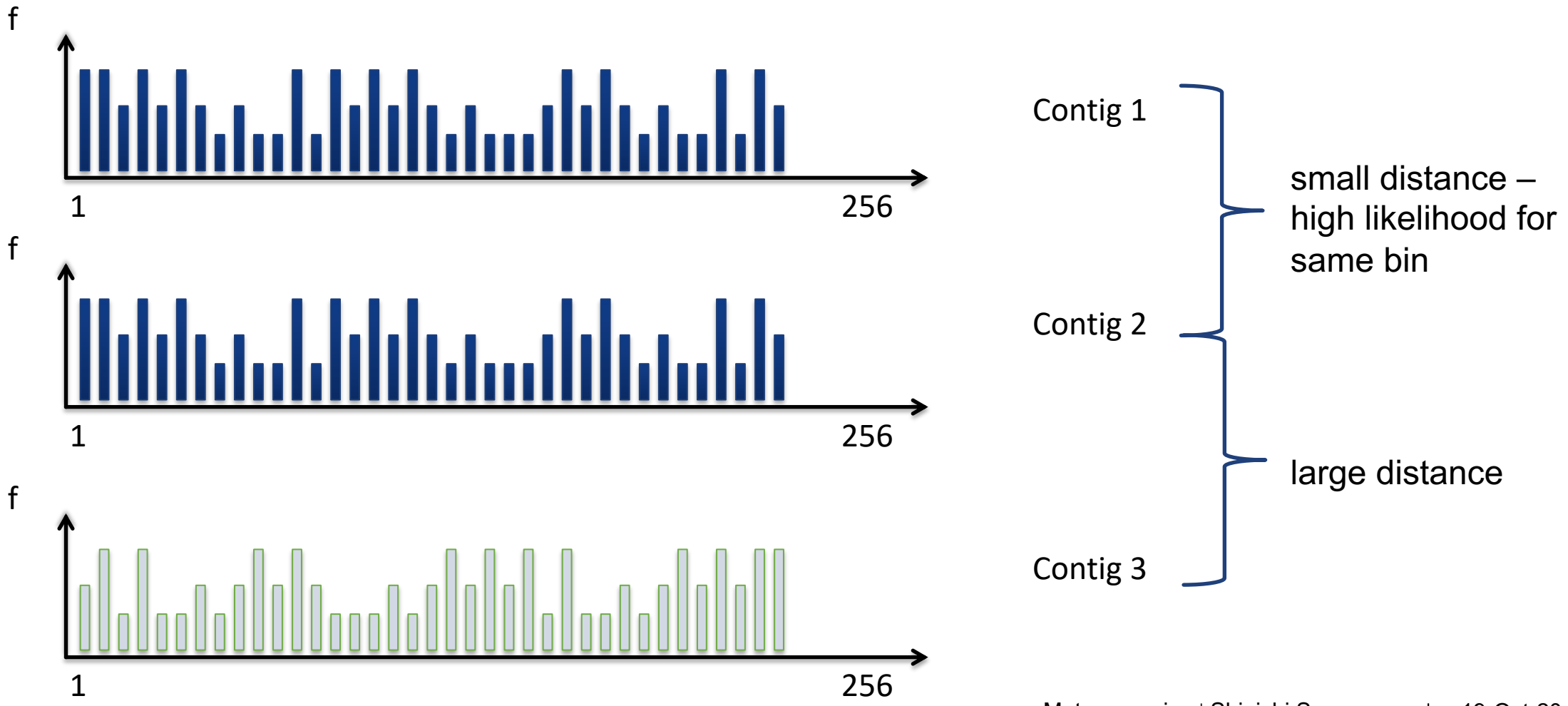


- For each contig:
  - a) calculate k-mer frequencies
  - b) calculate read abundance
- Combine a) and b) into a distance matrix
- Resolve distance matrix into clusters of highly correlated contigs

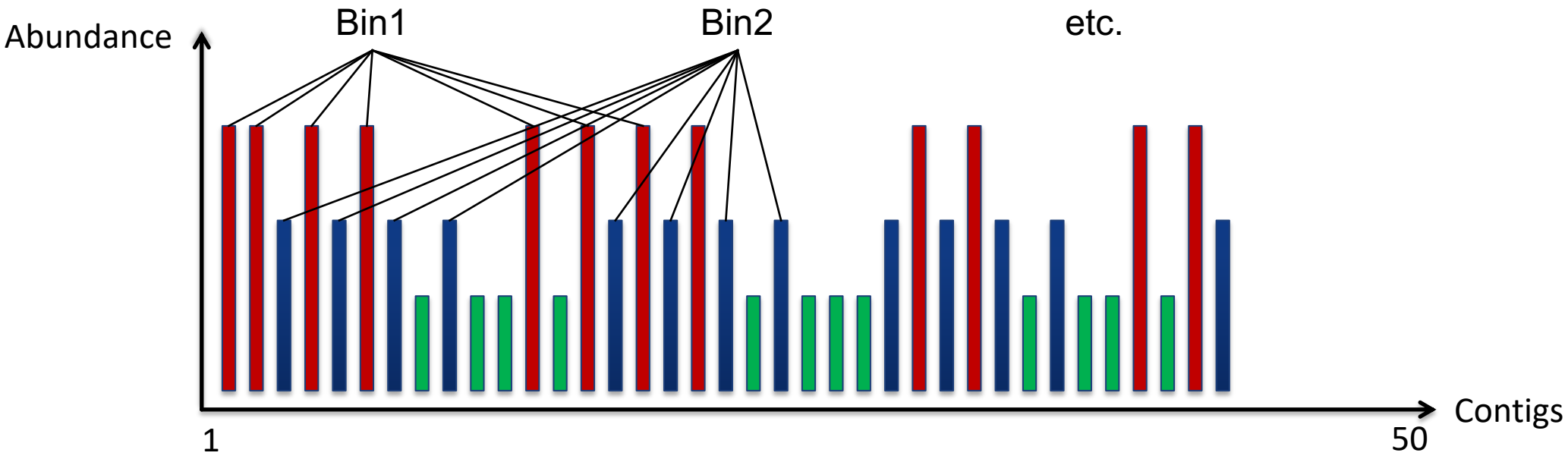
Kang et al. PeerJ (2015)

Example for tetranucleotide frequency (TNF) distances

$[ATGC]^4$ = 256 possible combinations



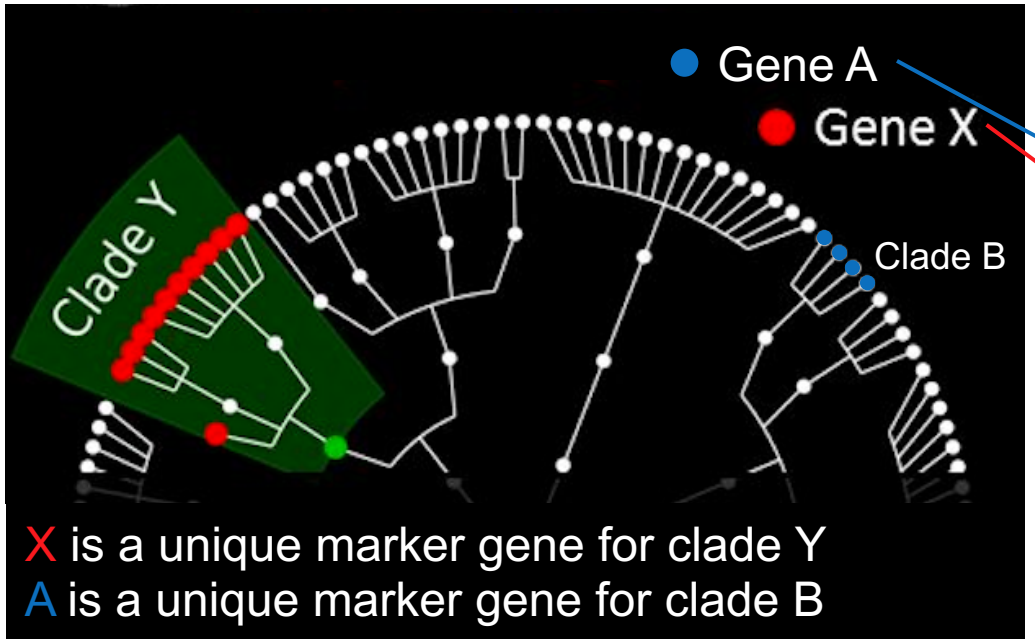
Example for abundance-based distances

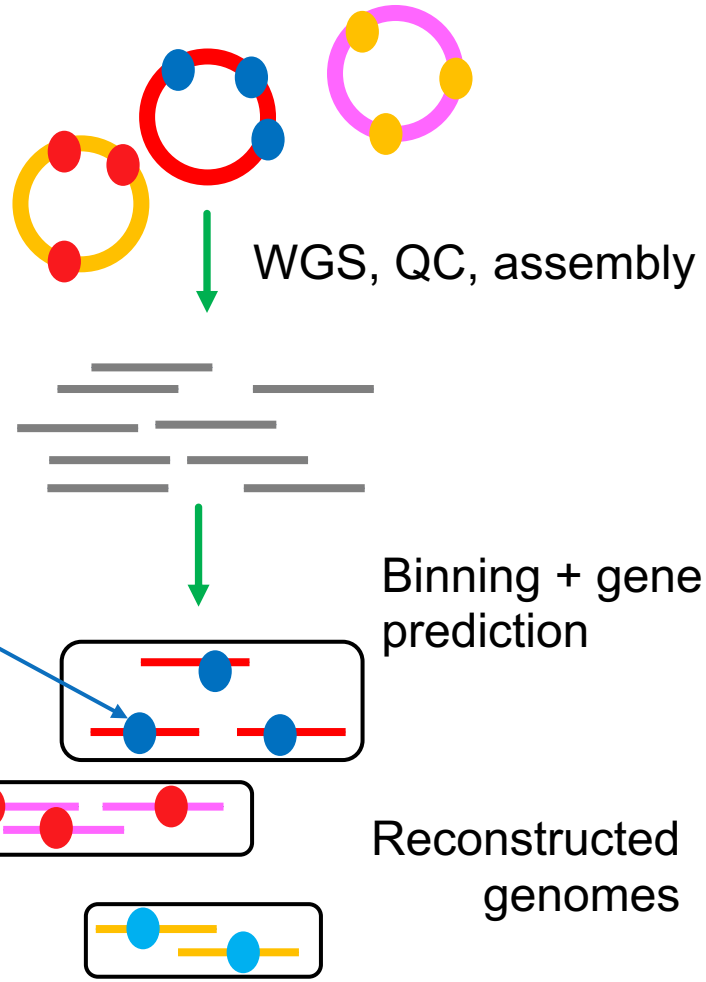


→ TNF and abundance-based distances can be combined and used for iterative binning

Step 3: Binning contigs/scaffolds – taxonomy guided

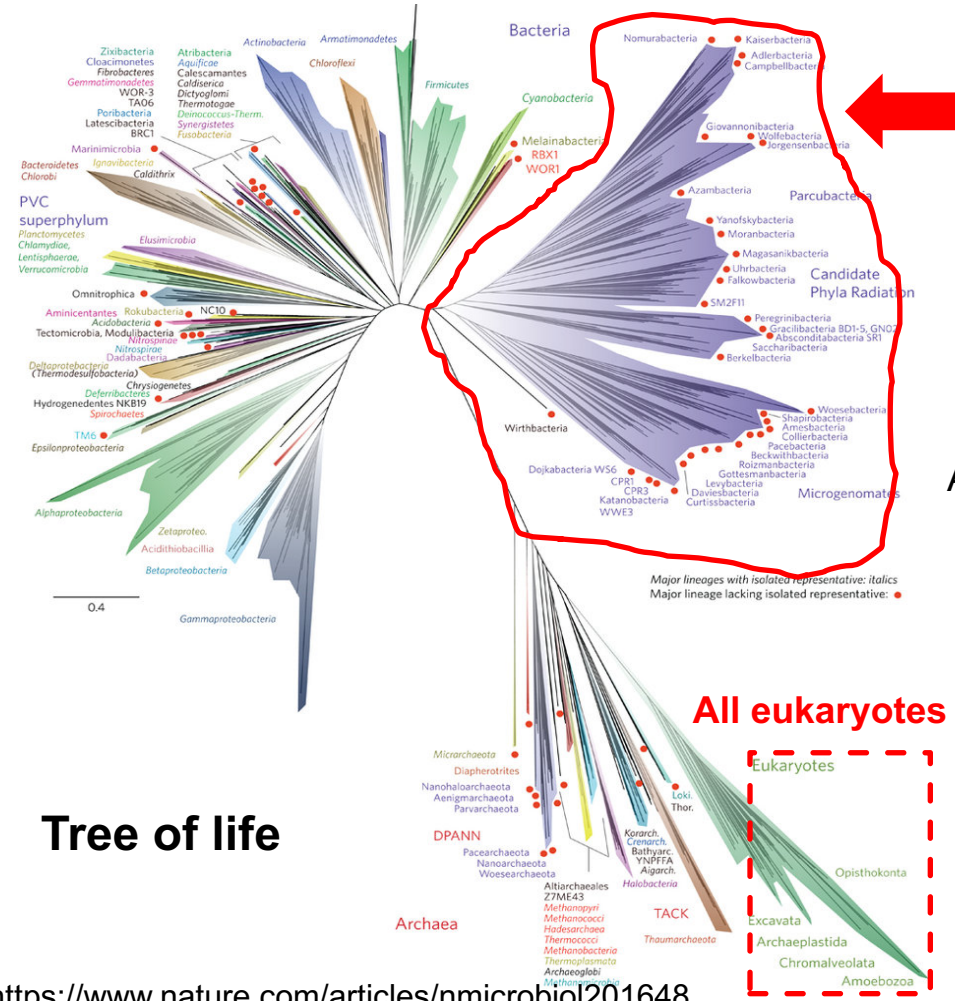
Taxonomy-guided (based on reference genomes) use of clade specific marker genes for binning metagenomic assemblies into draft genomes





→ **Note that clade specific marker genes can also be used to identify contaminated bins**

Applied examples III

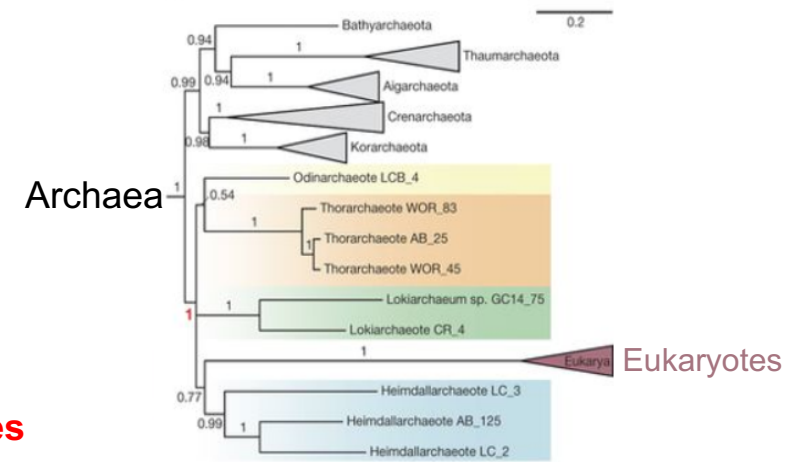


Tree of life

https://www.nature.com/articles/nmicrobiol201648

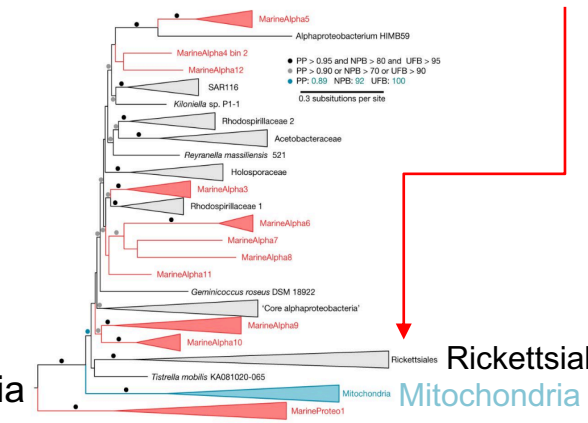
Candidate phyla radiation discovered by metagenomics

Only two domains of life?



https://www.nature.com/articles/nature21031

Mitochondrial origin not within Rickettsiales?



https://www.nature.com/articles/s41586-018-0059-5

Hug et al., Nat. Microbiol., 2016; Zaremba-Niedzwiedzka et al., Nature, 2017; Martijn et al., Nature, 2018

Summary – Part 2

- Reconstruction of microbial genomes from metagenomes is challenging as natural microbial communities are complex (many co-existing strains, uneven distribution of abundance)
- Assembled contigs/scaffolds can be binned by composition and/or taxonomy-based approaches
- Assembly metrics provide means of quality control and lineage-specific genes can reveal contaminations in genomic bins
- Functional annotation of genes/genomes can involve several search strategies against many different databases
- MAGs (metagenome assembled genomes) have revealed unexpected diversity and new implications on evolutionary origin of eukaryotes and mitochondria