BIO390 - Introduction to Bioinformatics: Metagenomics

October 1st, 2019

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

Overview of the Metagenomics lecture

Part I - Microbial community structure

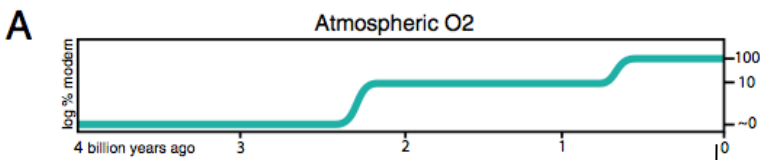
- definition of genotype, taxonomic resolution and operational taxonomic units (OTUs)
- analysis of microbial diversity, richness, evenness
- comparison of microbial community compositions

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

Background: why metagenomics?

Origin of life



- originated some 3.8 billion years ago
- drive biogeochemical cycles of the biosphere
- harbor large portion of genetic diversity on Earth

Impact (examples):

- biogeochemistry: e.g., photosynthesis by microbes, carbon fixation/export, nitrogen fixation
- health: human microbiome

- 2 billion years for eukaryotes and microbes to evolve the most diverse forms of symbioses

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

Background: why metagenomics?

- 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*

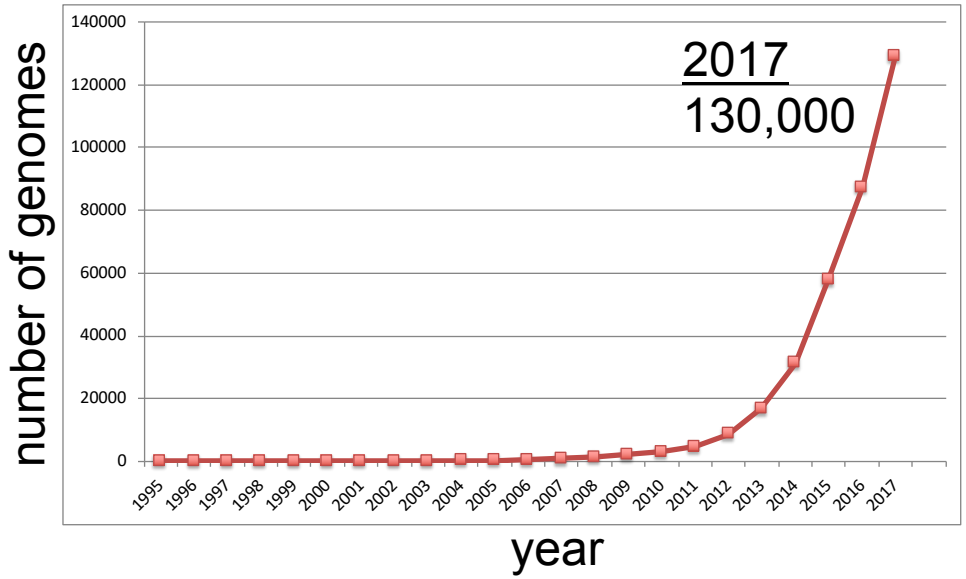
- Most bacteria and archaea cannot be isolated, as they live in communities and depend on other organisms
- Metagenomics provides access, in principle, to all genomic resources of a microbial community. This allows us to ask “who is there?”, and “what can they do?”

Background: why metagenomics?

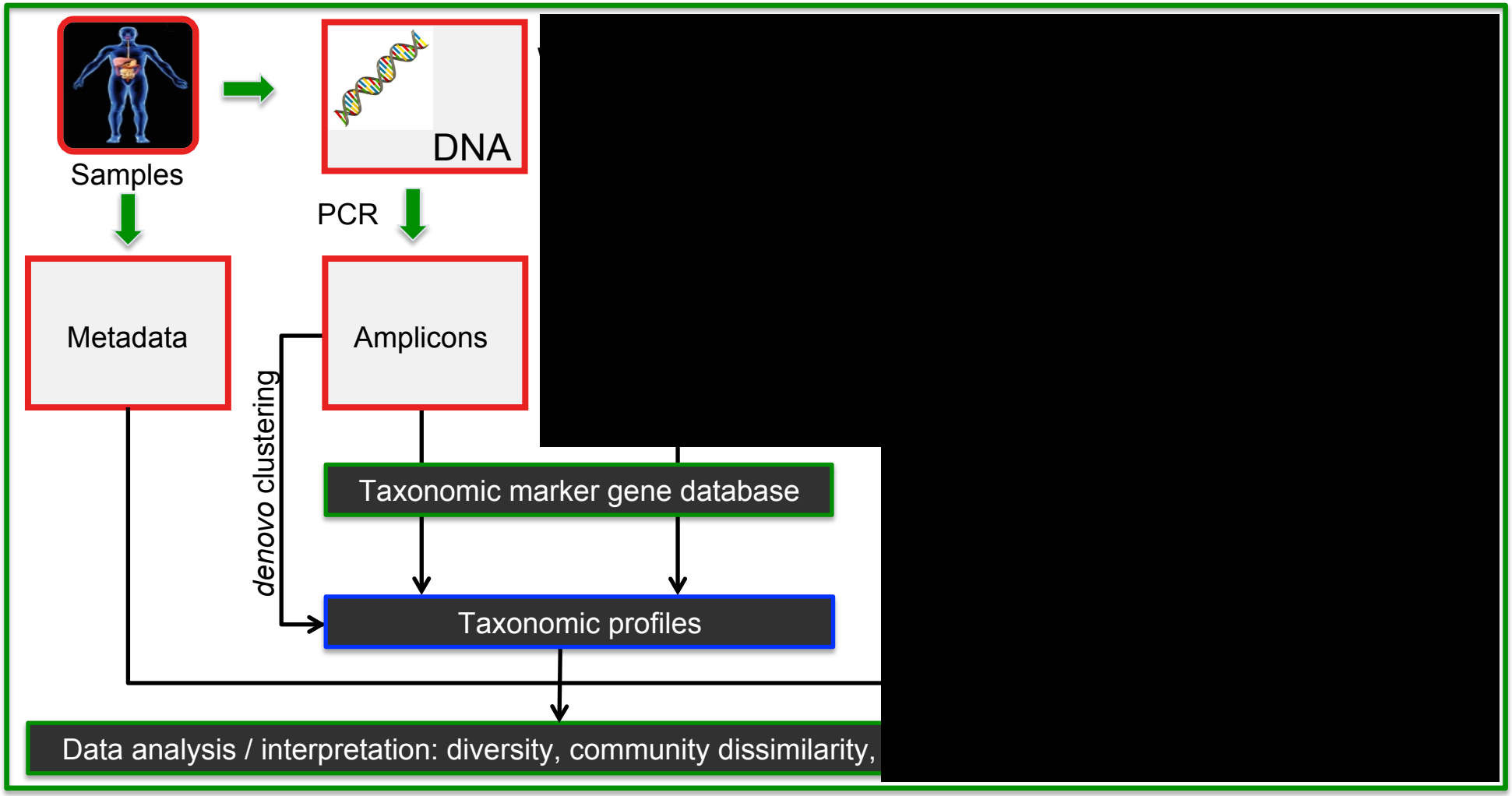
- Bacteria and archaea have ca. 500–10,000 genes, usually arrayed on circular DNA molecules (e.g., chromosomes and plasmids)
- Protein coding genes are ca. 1,000 base pairs long
- Their genomes are ca. 600,000–12 million bp in size (human 2 x 3 billion bp)

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

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



Overview – Part 1



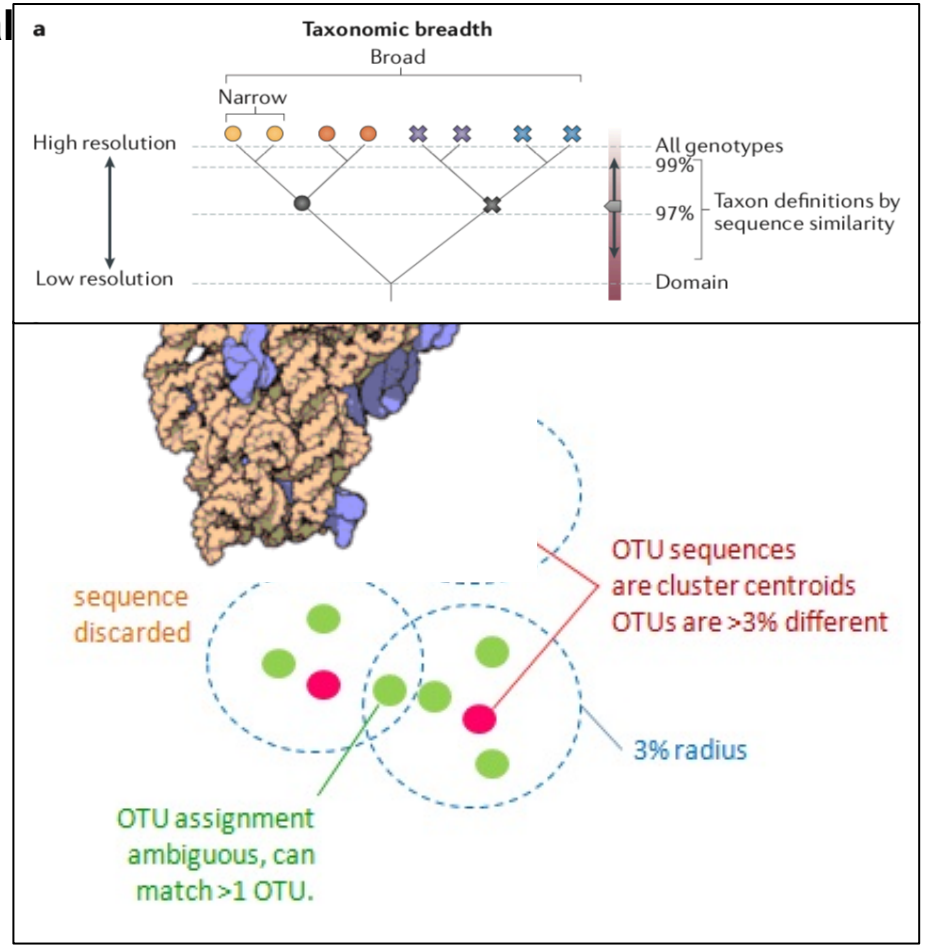
*WGS: whole genome shotgun sequencing

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

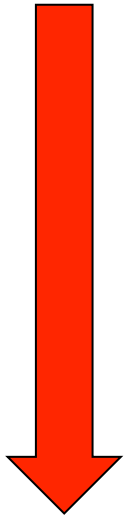
- Goal: determine ‘who’ is there at what abundance in one or more samples

OTU	S1	S2	S3	S4	S5	S6	S7	S8	…
OTU1	234	87	166	240	131	249	0	244	
OTU2	23	0	93	0	146	122	92	5	
OTU3	2	137	191	299	285	0	0	0	
OTU4	455	0	112	0	114	0	289	0	
OTU5	23	229	66	247	0	216	127	116	
OTU6	34	8	206	54	276	214	158	81	
OTU7	0	249	0	76	132	200	0	0	
OTU8	0	6	0	127	0	254	125	272	
OTU9	34	174	207	91	184	0	3	44	
OTU10	356	186	25	134	98	162	0	0	
SUM									

OTU count table
Columns: S1 to Sn = samples
Rows: OTUs

Step 1: PCR-amplification of 16S rRNA amplicons (“tags”)

Community DNA extract



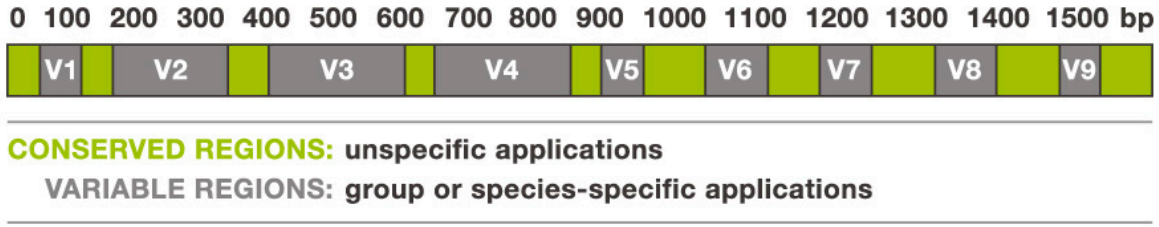
PCR

agtctcgctatgacgtcgtcgtcagactac

gtcgtacgtcgatatttctcgcgccggagc
gtcgtacgtcgatattctcgcggccggagc

agcctacgtcgtcgatagtgcgctagtgtc

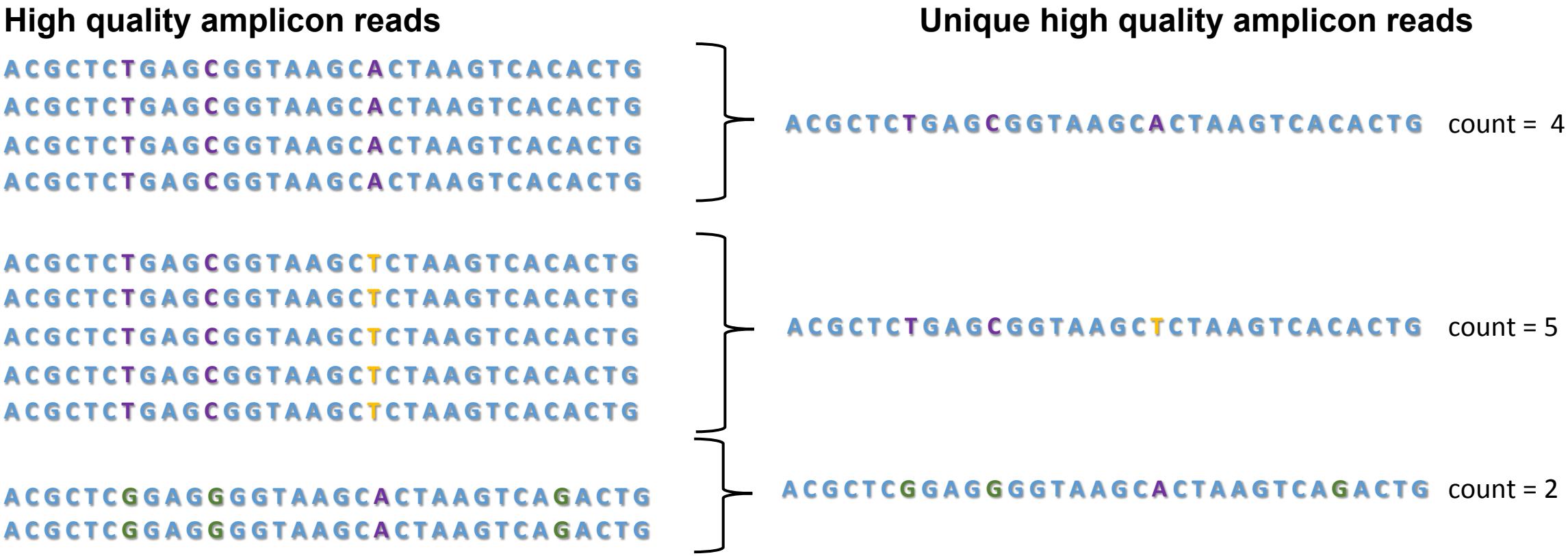
Primers bind to conserved regions of constant regions. Variable regions are amplified by PCR



Example

- “V4 primers” yield ca. 250 bp long sequences

Step 2: de-replication



- 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

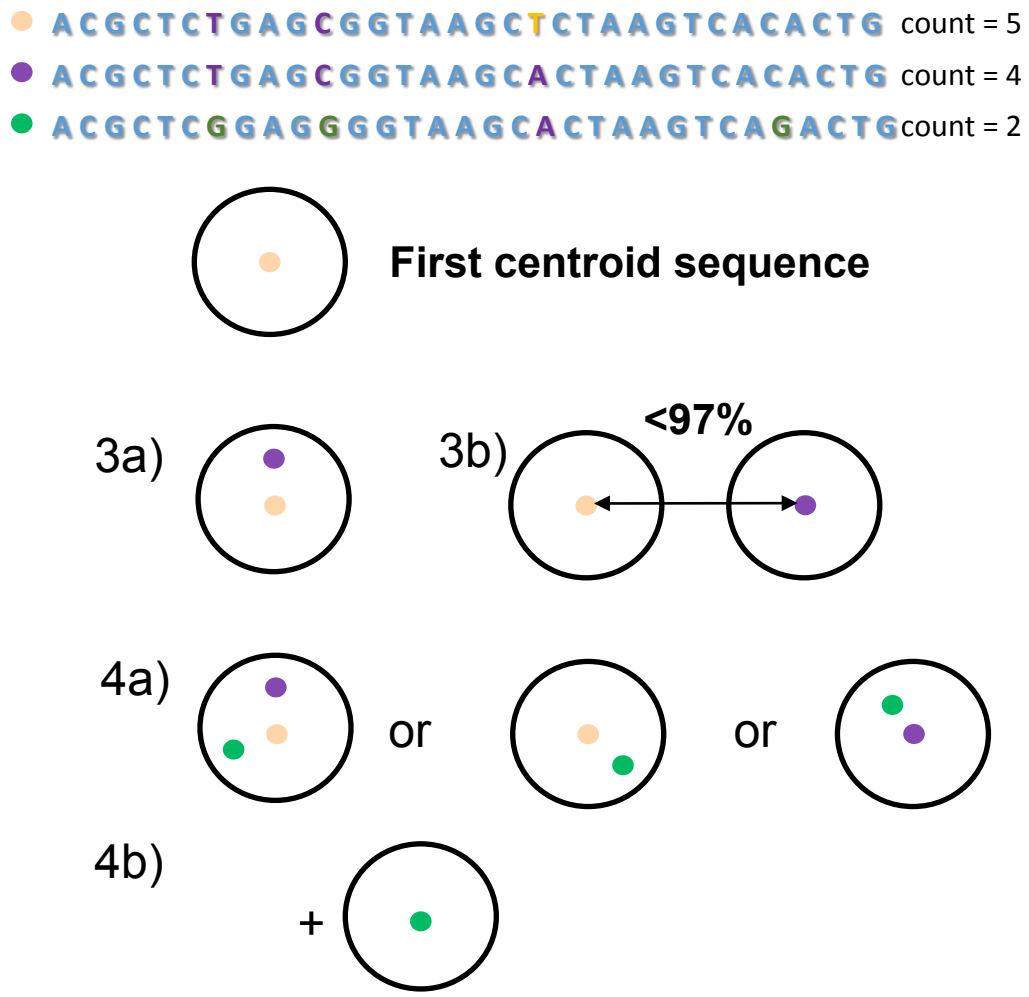
   2 different possibilities:

   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)

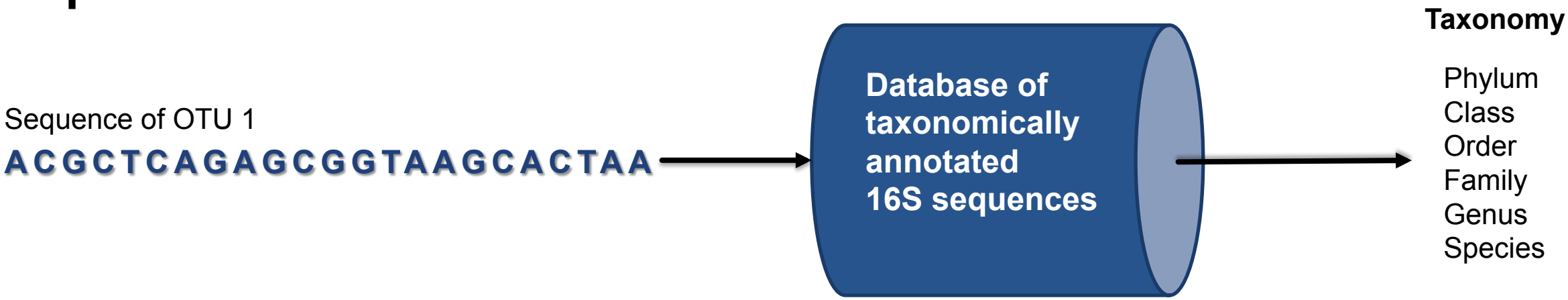
4) Next read is compared to all OTU centroids

   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

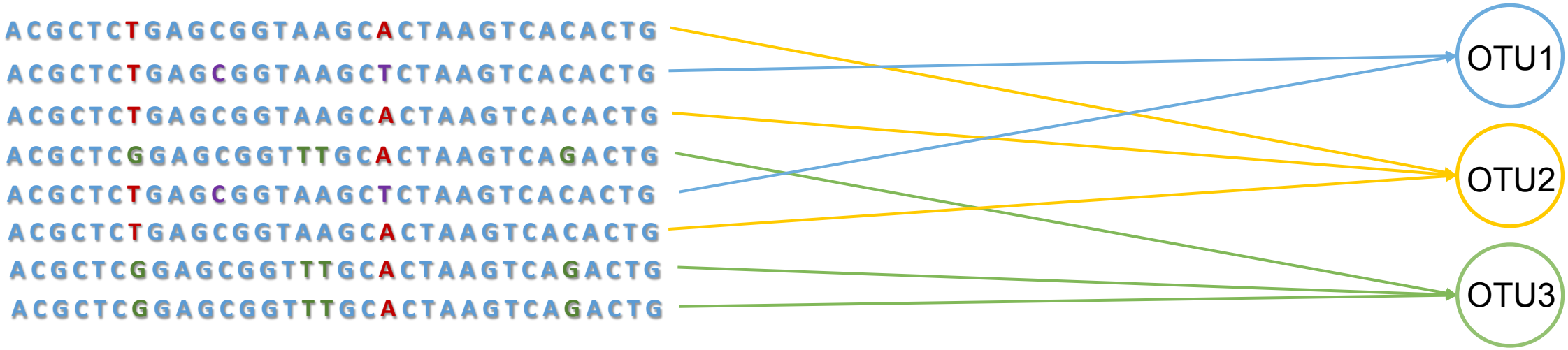


- Identification of taxon to which an OTU belongs
  - The centroid sequence of each OTU is compared to a database of annotated 16S rRNA gene sequences

→ 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)



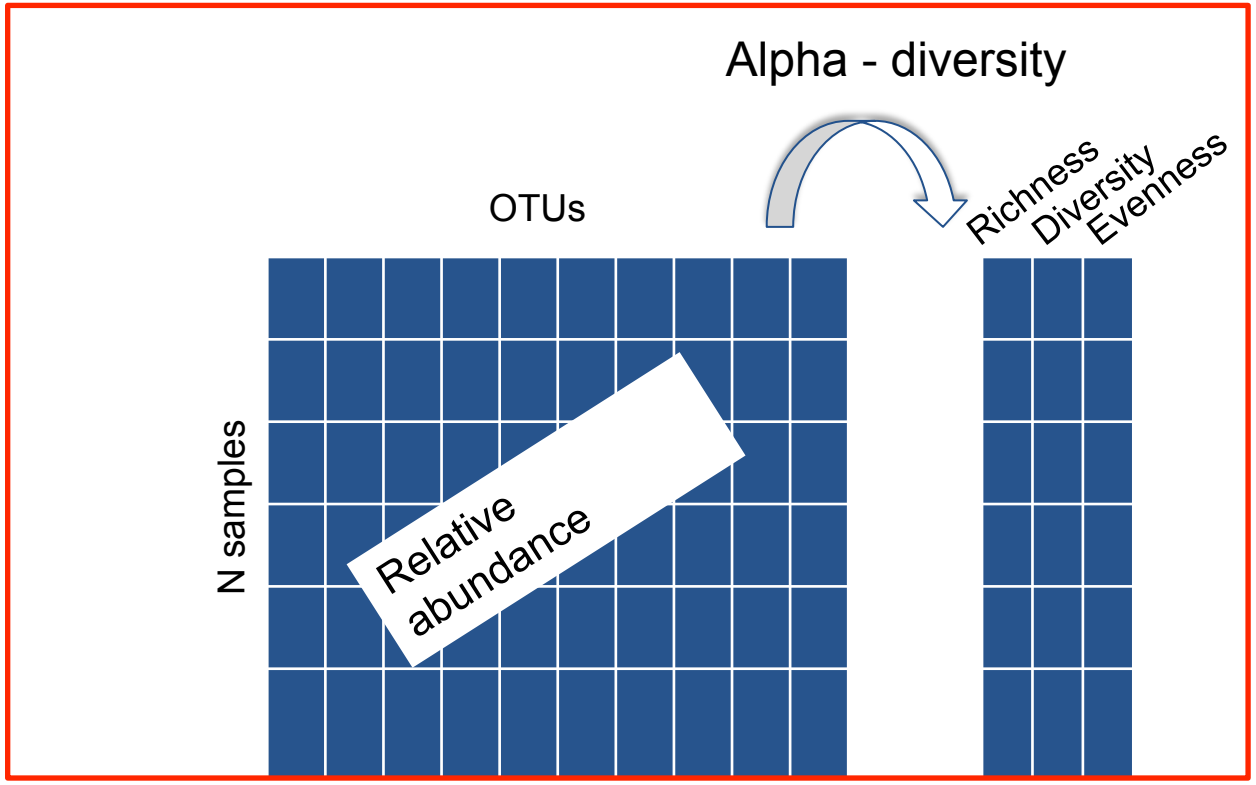
The result is an OTU (feature) count table, summarizing read counts for each OTU for each sample:

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

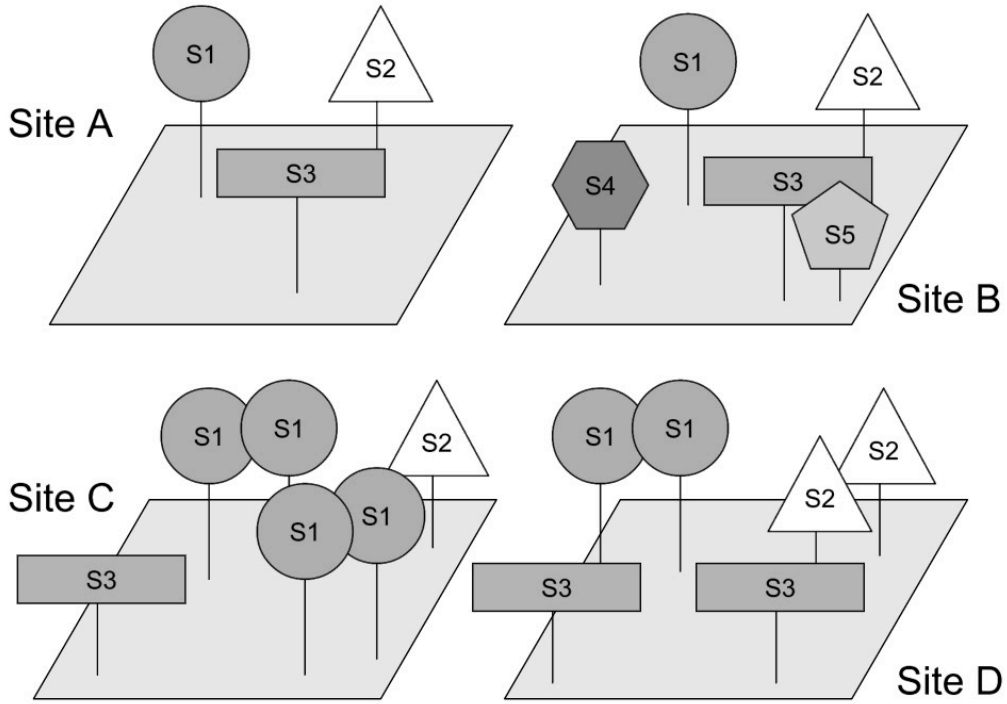
Concept of diversity

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



Alpha diversity: a function of richness and evenness

richness = 3
high evenness



richness = 5
high evenness

Note that in this figure
Site = sample
S1-5 = species 1-5

richness = 3
low evenness

richness = 3
high evenness

Alpha - diversity = f(richness, evenness)

Shannon's diversity index:

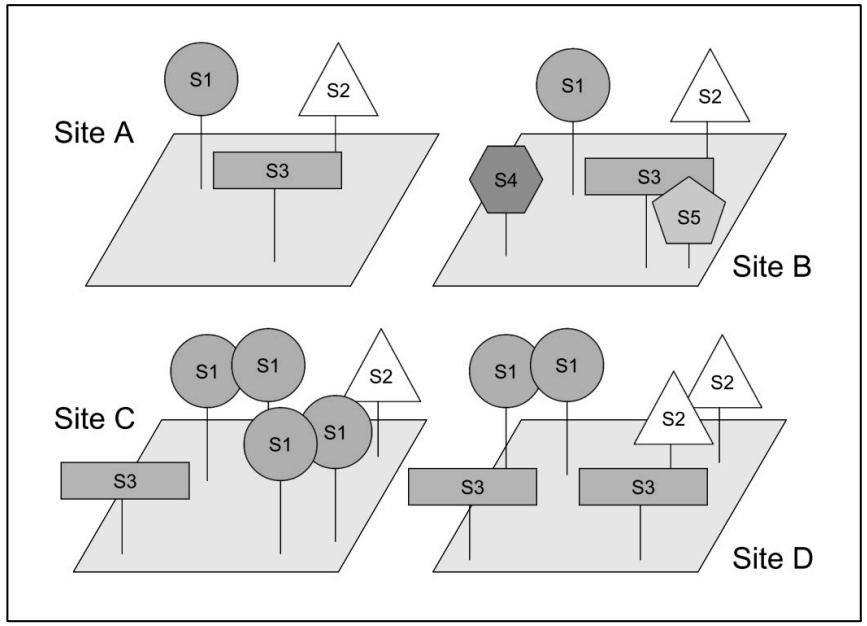
$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

where p_i is the relative abundance of species *i,* and *R* the richness

Pielou's evenness:

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

where *S* is the observed richness in sample X

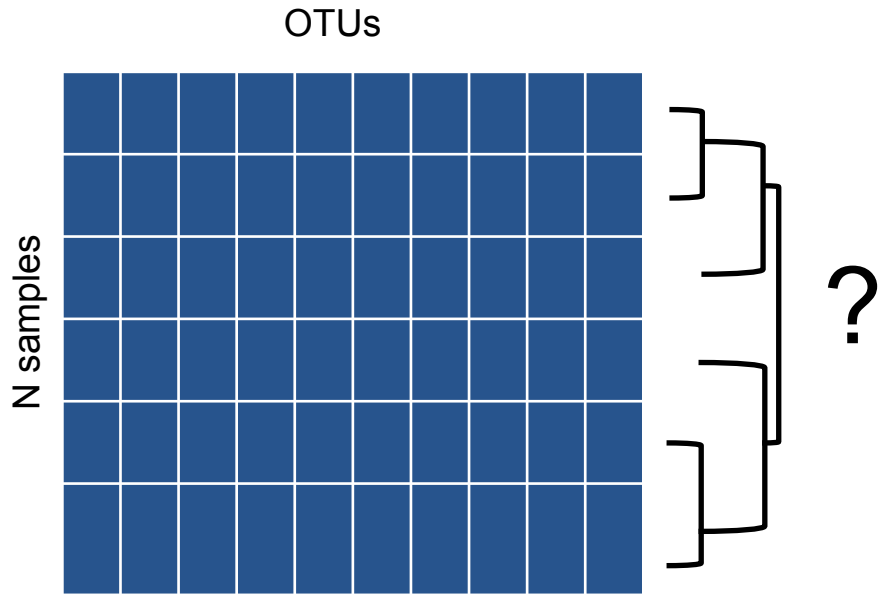


Note that in this figure
Site = sample
S1-5 = species 1-5

Site A:
H' = -(1/3*ln(1/3) + 1/3*ln(1/3) + 1/3*ln(1/3)) = **1.0986**
Site B:
H' = -(1/5*ln(1/5) + 1/5*ln(1/5) + 1/5*ln(1/5) + 1/5*ln(1/5) + 1/5*ln(1/5)) = **1.6094**
Site C:
H' = -(4/6*ln(4/6) + 1/6*ln(1/6) + 1/6*ln(1/6) = **0.8676**

Beta diversity: between sample dissimilarity

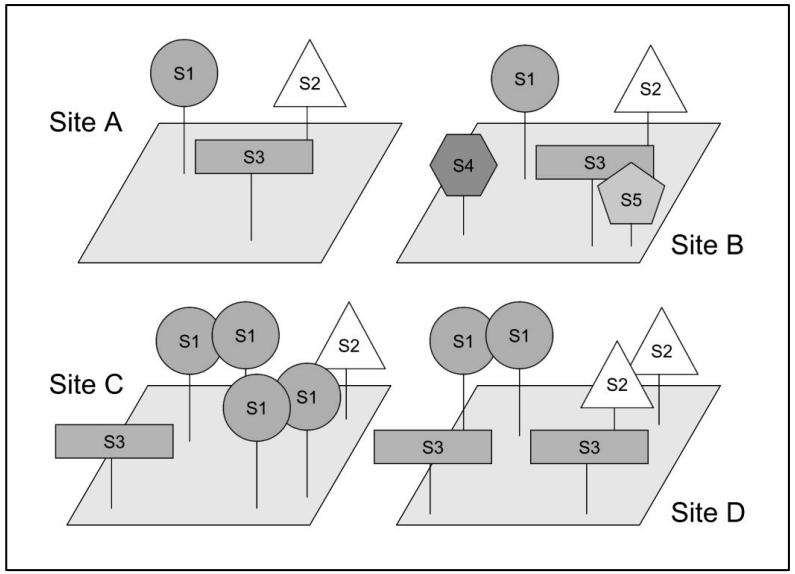
- Now that we learned how to describe the structure of a microbial community, how do we compare different communities to each other?



Beta diversity: between sample dissimilarity

Index	Equation	Properties
Jaccard	$S_7 = \frac{a}{a+b+c}$	Compares the number of shared species to the number of species in the combined assemblages placing more emphasis on taxa not shared between sites
Sørensen	$S_8 = \frac{2a}{(2a+b+c)}$	Compares the number of shared species to the mean number of species in a single assemblage placing more emphasis on similarity of samples owing to shared species

In the above table, a = the number of species shared between assemblages, b = the number of unique species in the first assemblage, and c = the number of unique species in the second assemblage.



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

Site A-B: J = 3/(3+0+2) = **0.6**
Site A-C: J = 3/(3+0+0) = **1**

Distance / Dissimilarity

D = 1-J = **0.4**
D = 1-J = **0**

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

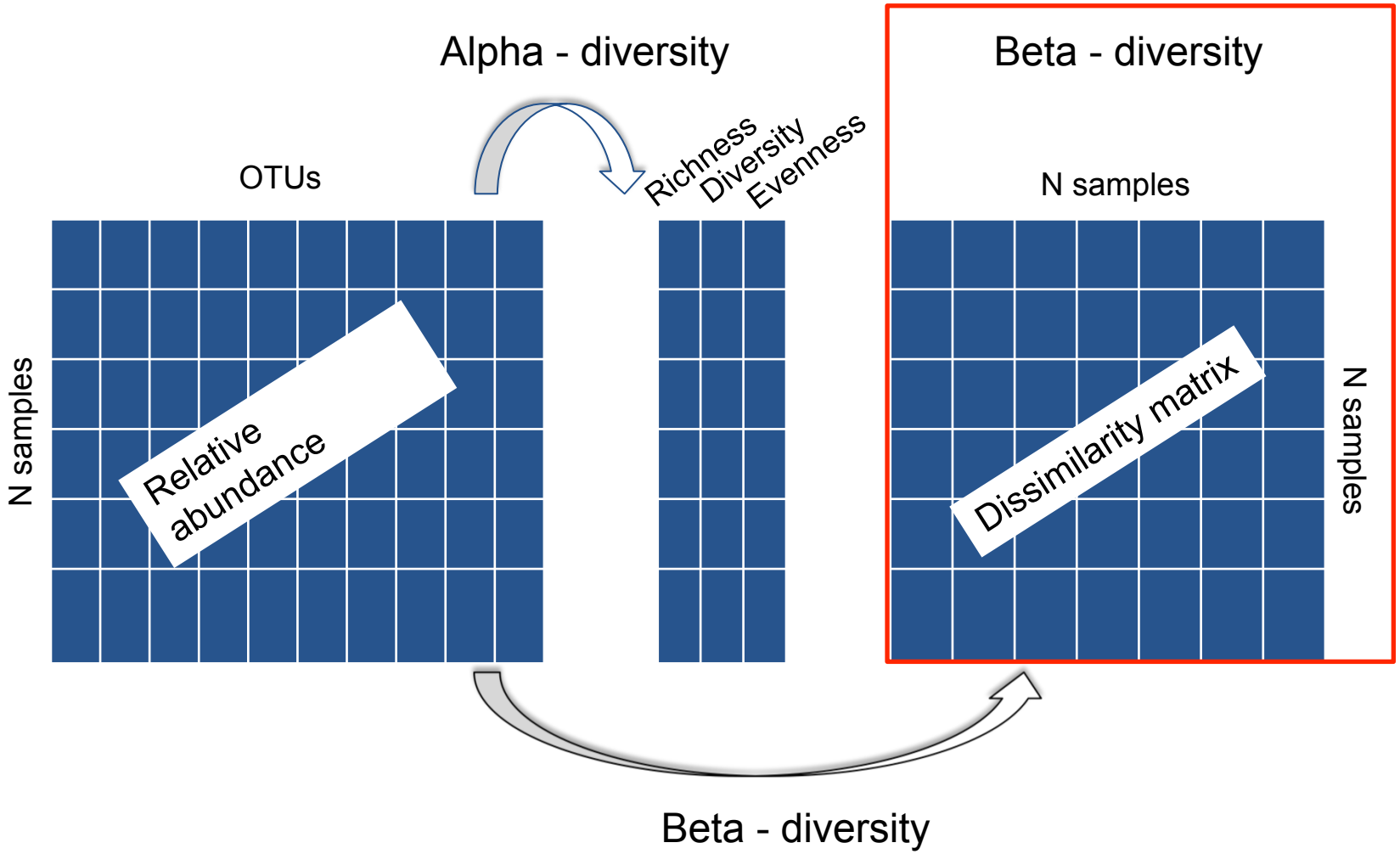
Hellinger: $$D = \sqrt{\sum_{i=1}^{S}\left(\sqrt{\frac{a_i}{a_+}} - \sqrt{\frac{c_i}{c_+}}\right)^2} \quad \text{with } a_+ = \sum_{i=1}^{S} a_i$$

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*

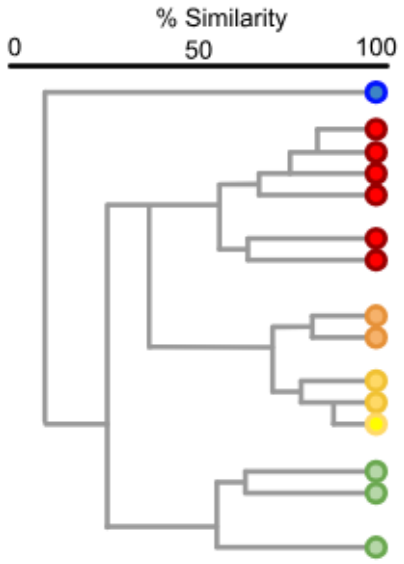
Within sample descriptions → between sample comparisons



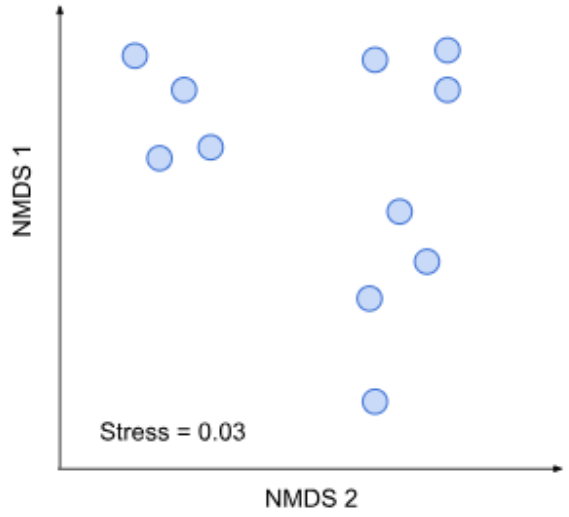
Visualize dissimilarities between microbial communities

- For 2 (xy) or 3 (xyz) variables, data can be easily visualized
- For multi (n>3) dimensional data, calculate distances and ‘project’ into lower dimensional space

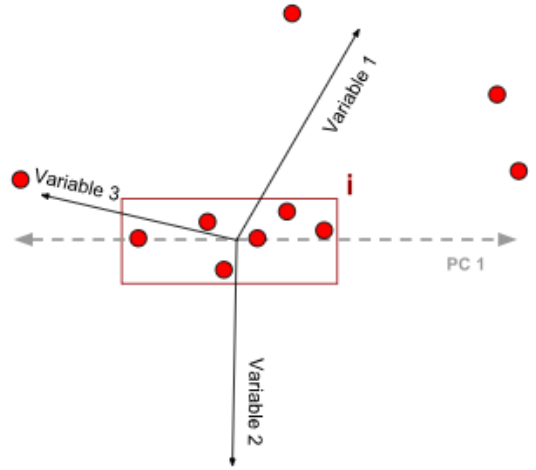
Hierarchical clustering



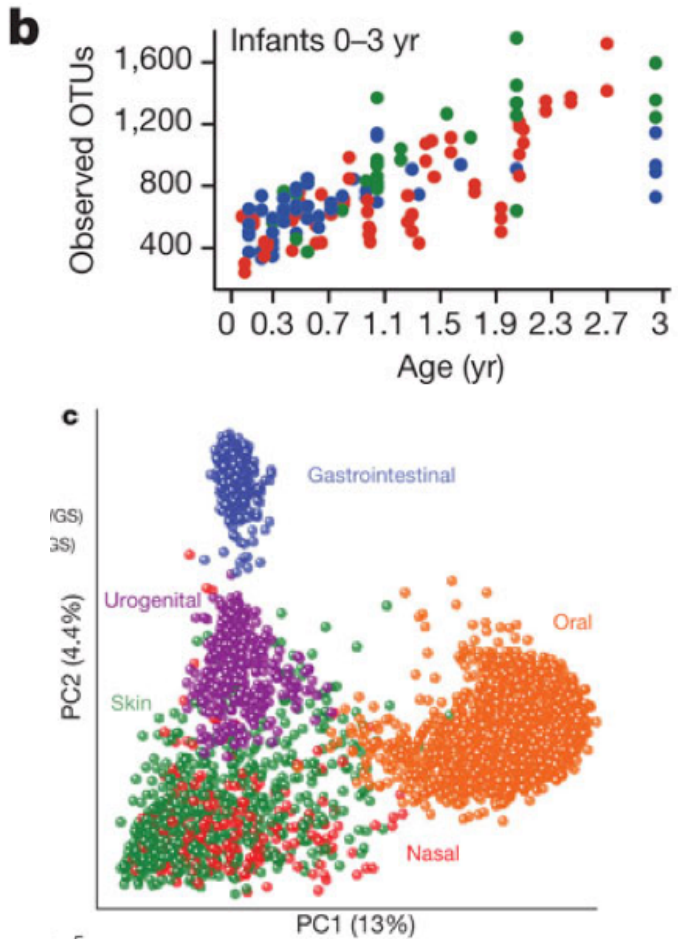
Non-metric dimensional scaling (NMDS)

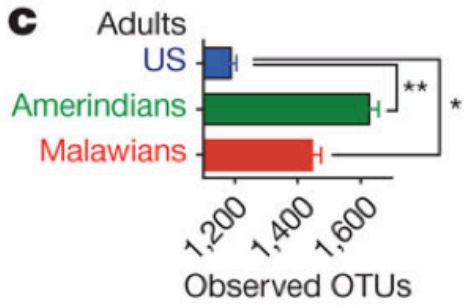


Principal component or coordinate analysis (PCA or PCoA)



Applied examples I



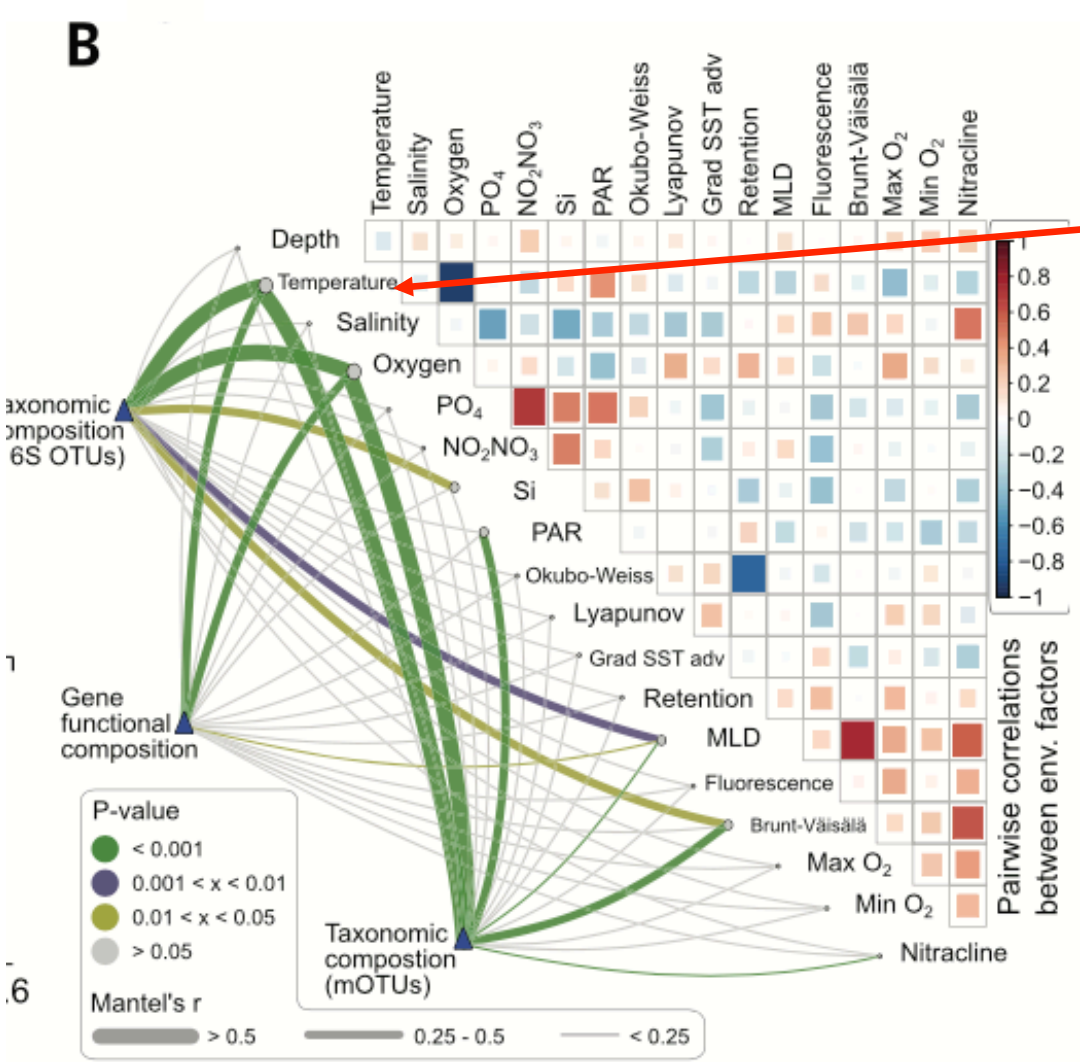


- Microbial diversity in human gut increases with age
- US citizens have harbor less diverse gut microbiota relative to other populations

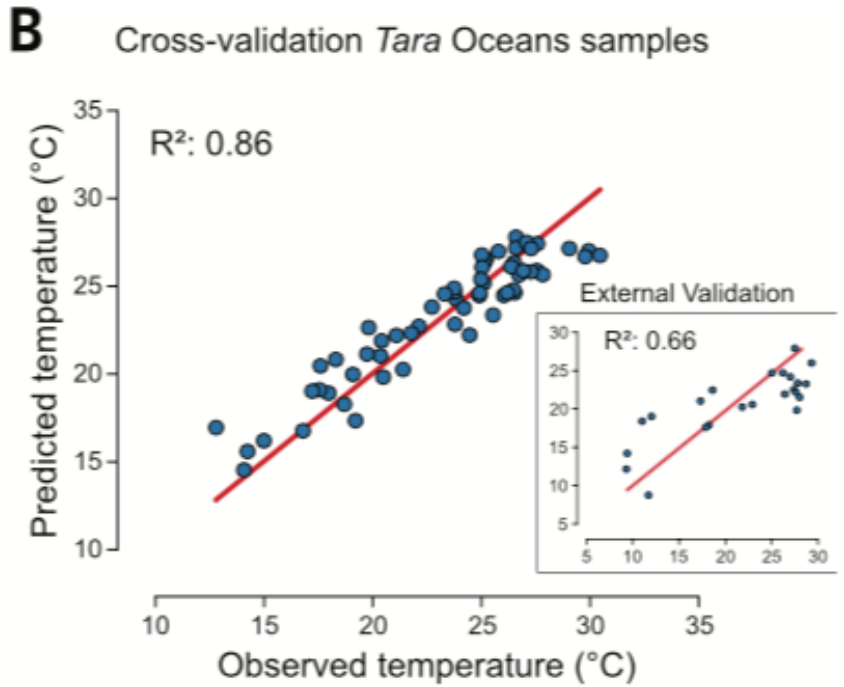
- Microbial communities cluster by human body site rather than by individual

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

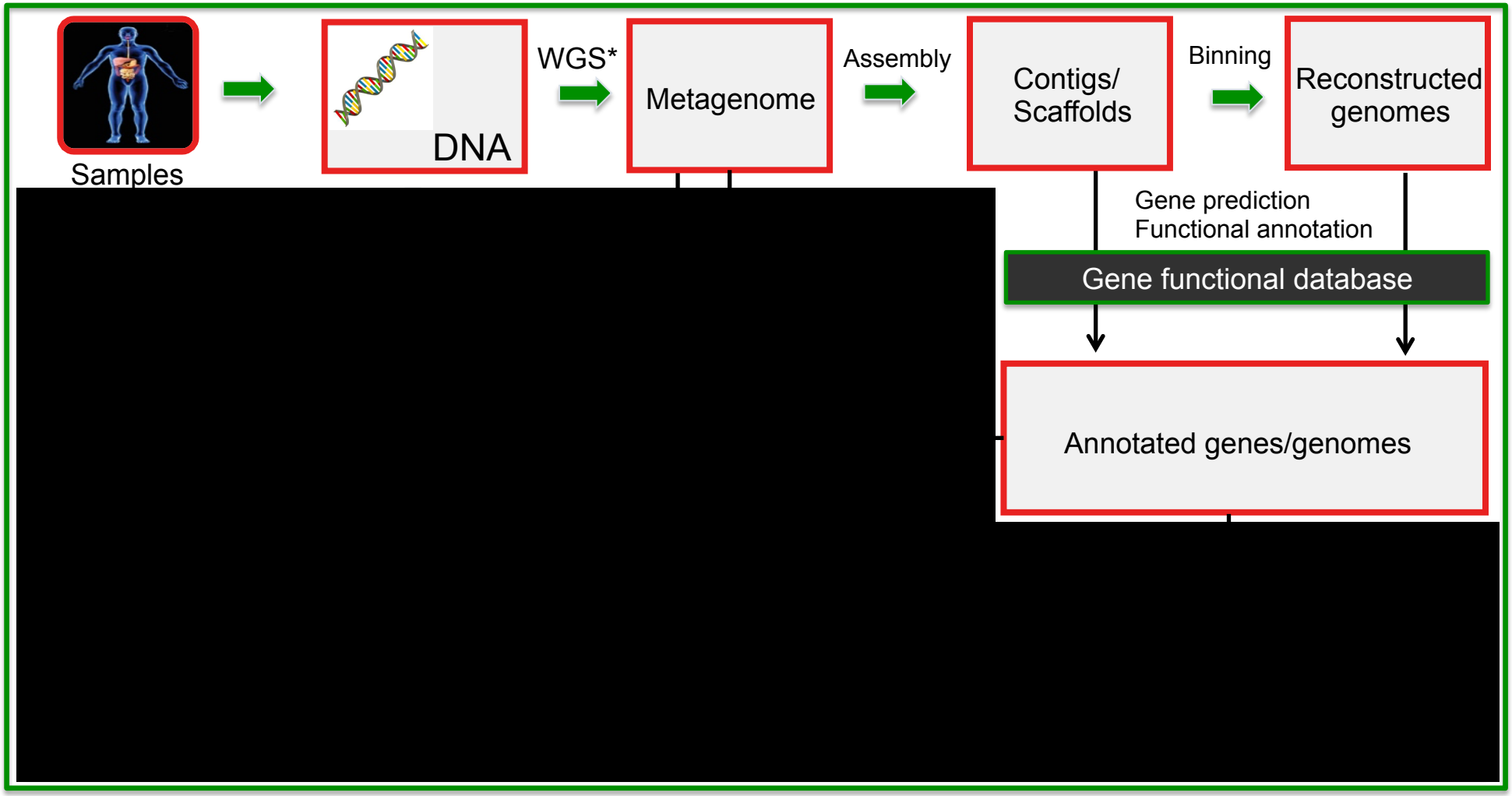
- Metagenomics provides information about microorganisms that are often difficult or impossible to be cultivated in their natural environment
- Due to the lack of species concept for prokaryotes, researchers use sequence identity cutoffs of marker genes to define operational taxonomic units
- Microbial 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

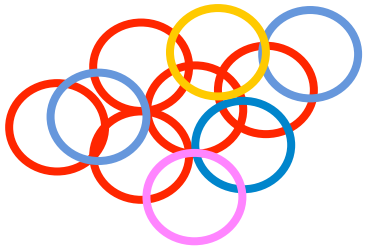
- assembly of individual genomes and metagenomes
- binning of metagenomic assemblies into metagenome-assembled genomes
- taxonomic and functional annotation of metagenomes

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 ↓

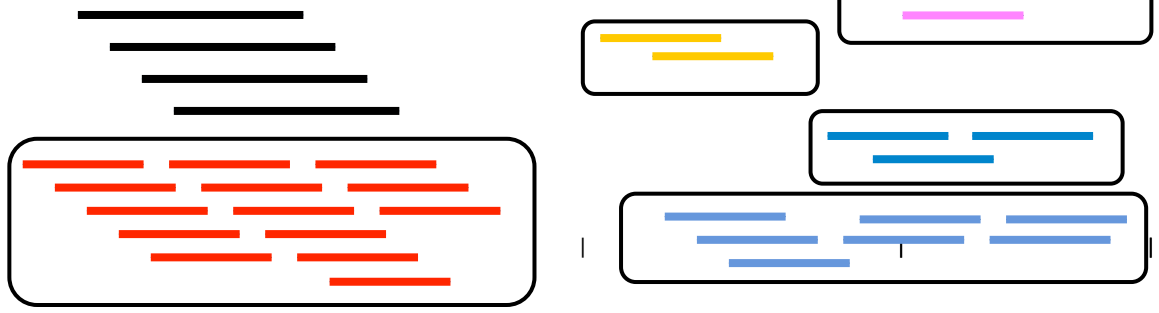
scaffold 1

scaffold 2

Binning ↓

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

(many) unbinned contigs

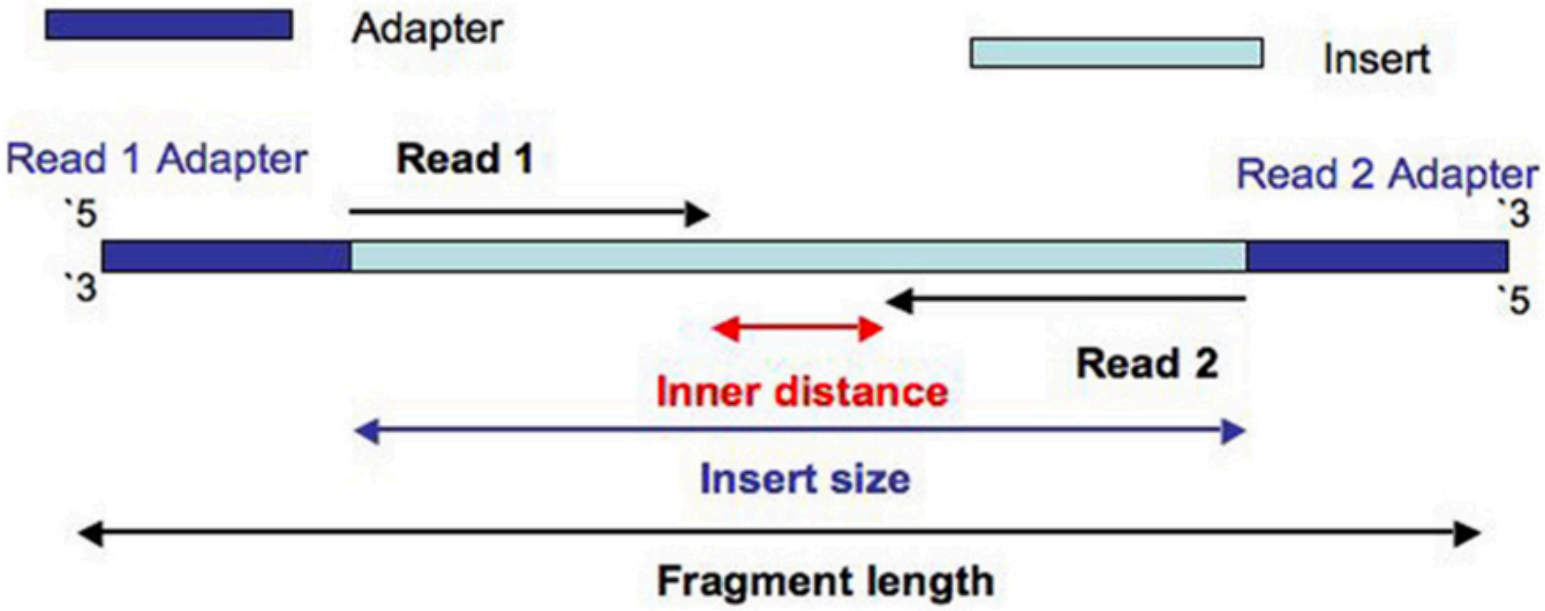


Gene prediction and annotation ←

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



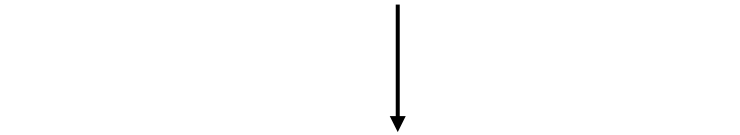
Note: Paired end reads may be overlapping providing the possibility to “merge” reads into one or not. In the latter case, the sequence between the paired reads remains unknown, while the length can be estimated due to the known insert size distribution

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

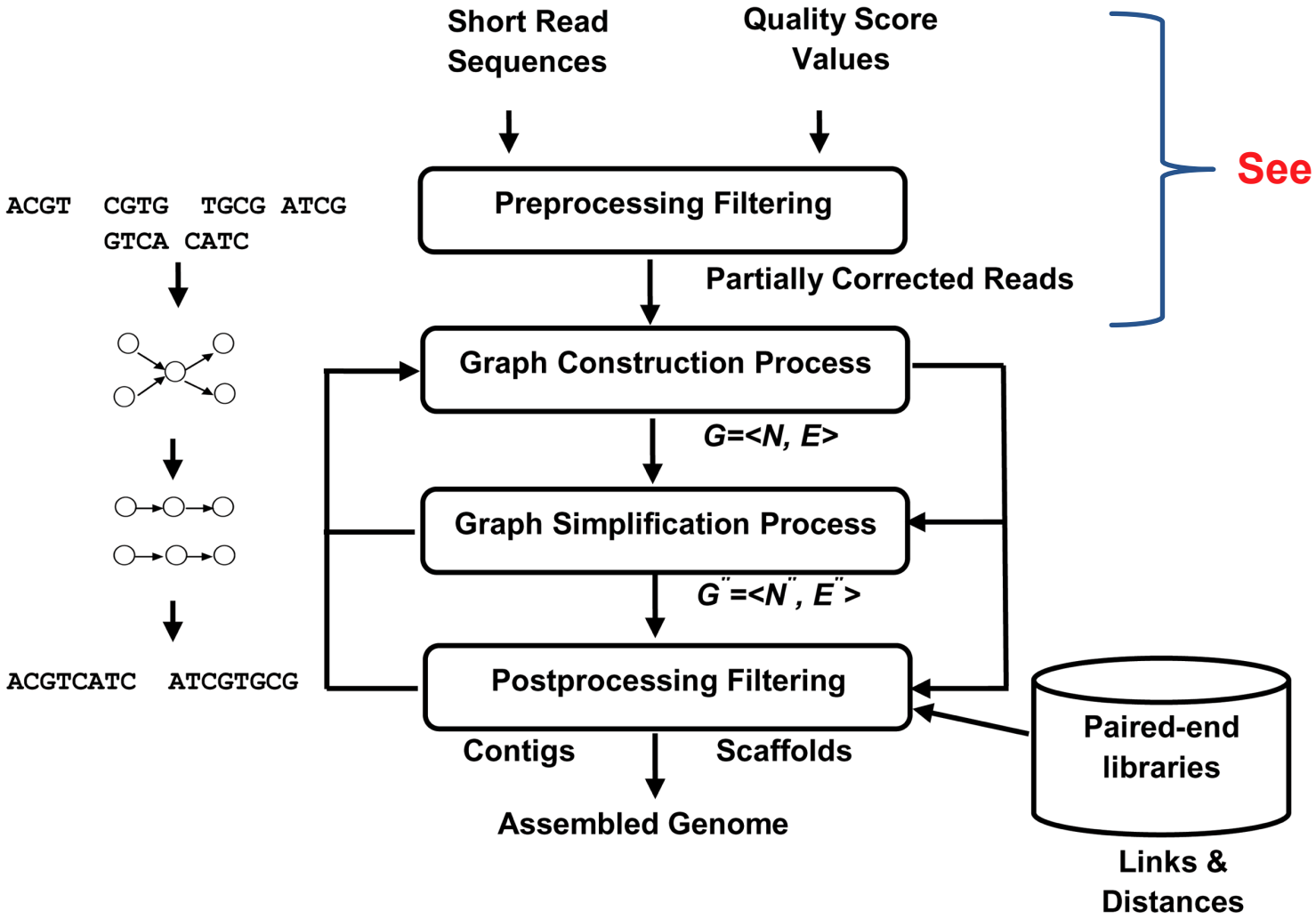
$Q = -10 \log_{10} P$

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

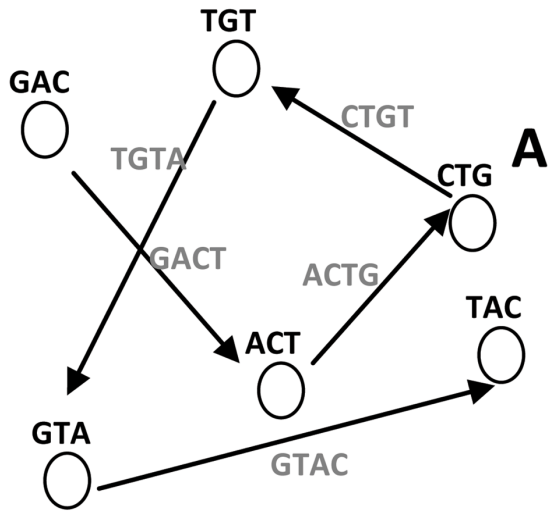
Step 2: Assembly and scaffolding: overview



See previous slides

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

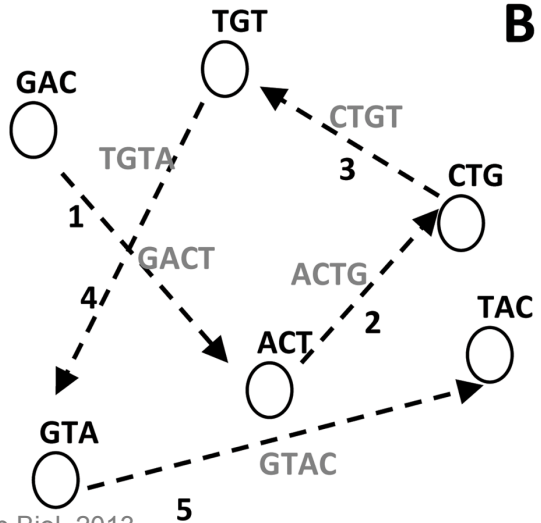
Graph construction: k-mer based assembly



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) k-mer-based graph

Nodes = k-mers

Edges = k-1 overlaps

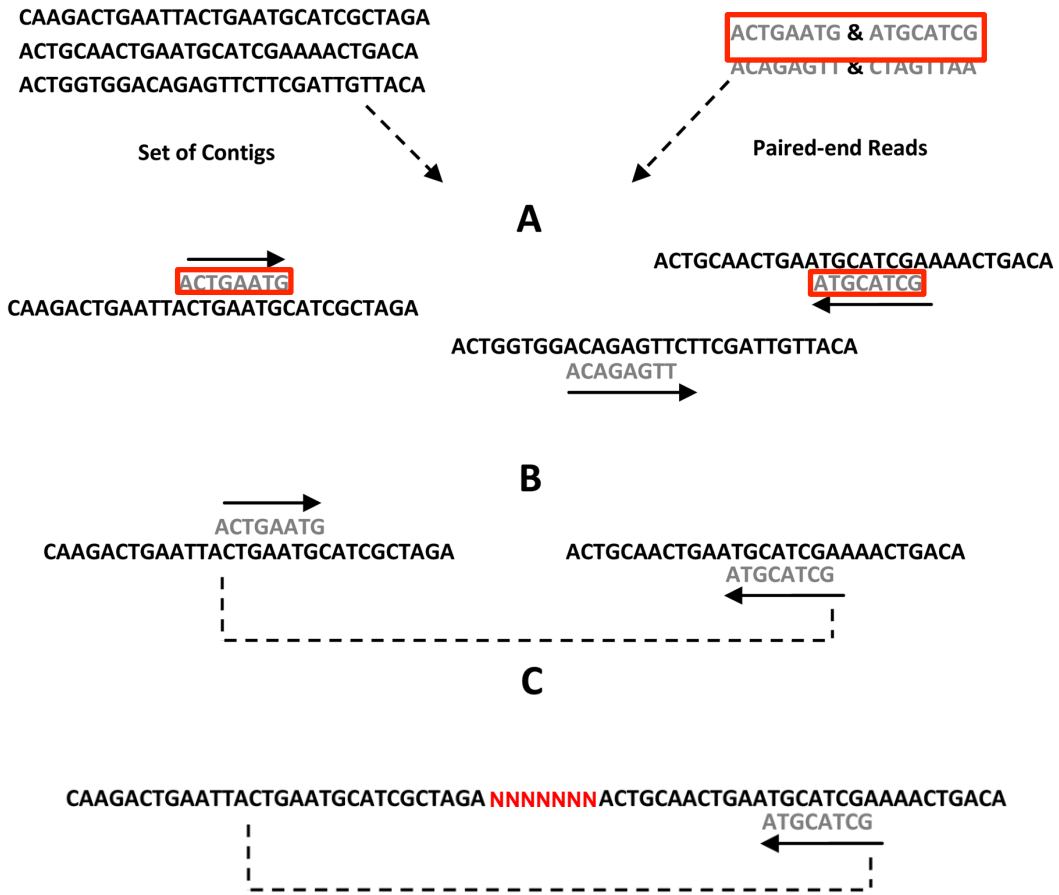
B) Layout shortest Eulerian path

Visit each edge once

C) Combine into consensus

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

Post processing: scaffolding



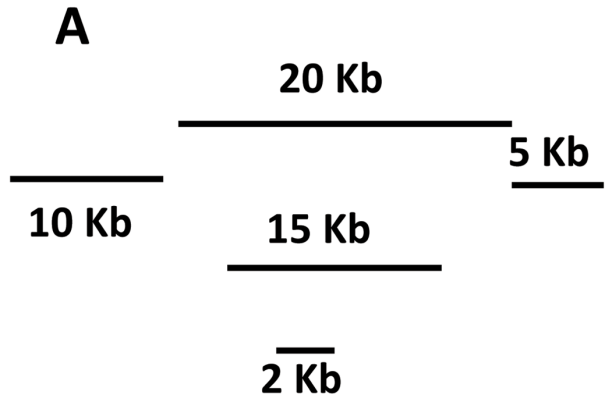
A) Align paired-end reads

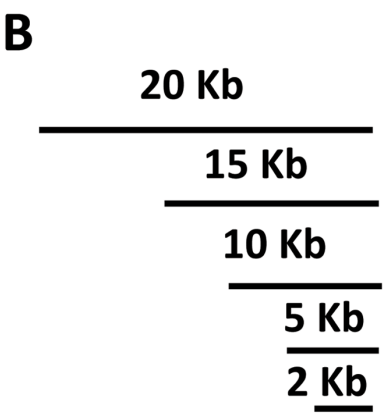
B) Orientation of contigs
→ according to read orientation

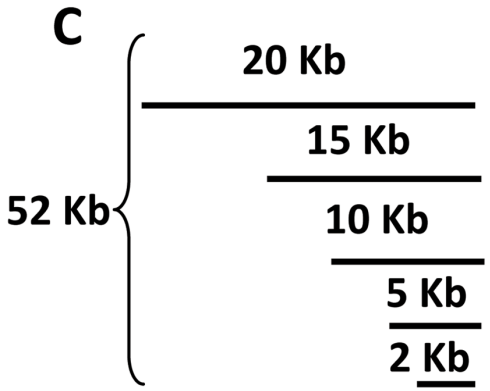
C) Scaffolding
→ use information on insert size distribution and fill 'gaps' with 'Ns'

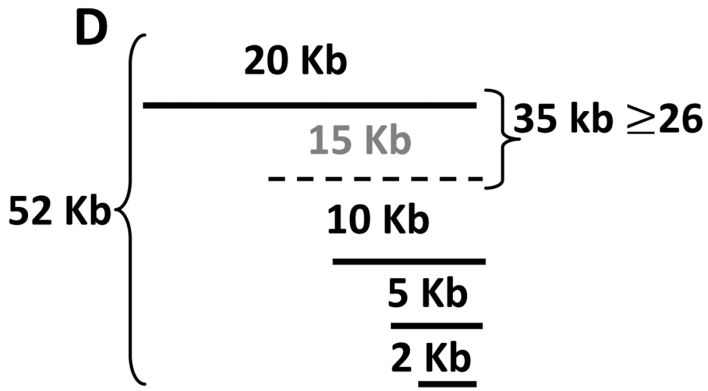
El-Metwally et al., PLoS Comp Biol, 2013

Assembly quality: N50 and L50









A) Set of contigs

B) Sort contigs by size in descending order

C) Calculate total length of all contigs

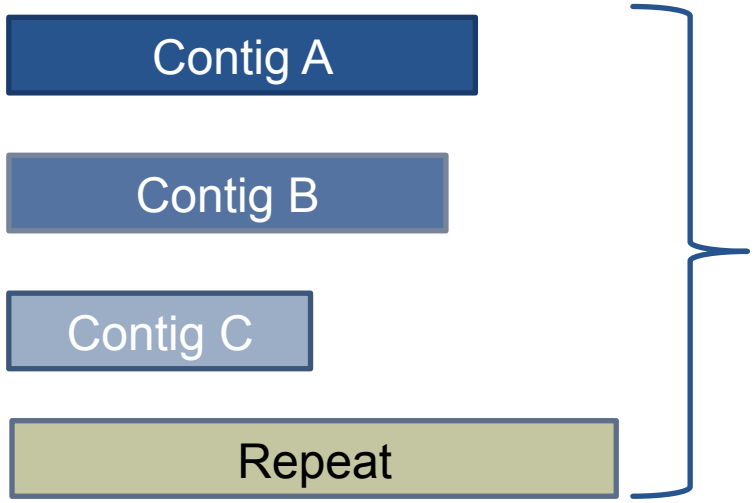
D) Add length in descending order until sum equals or exceeds 50% of the total length.

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

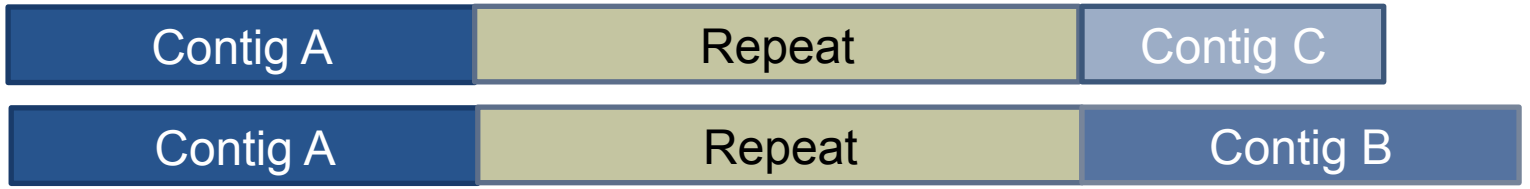
Repetitive sequences in genomes prevent full assembly

Genome

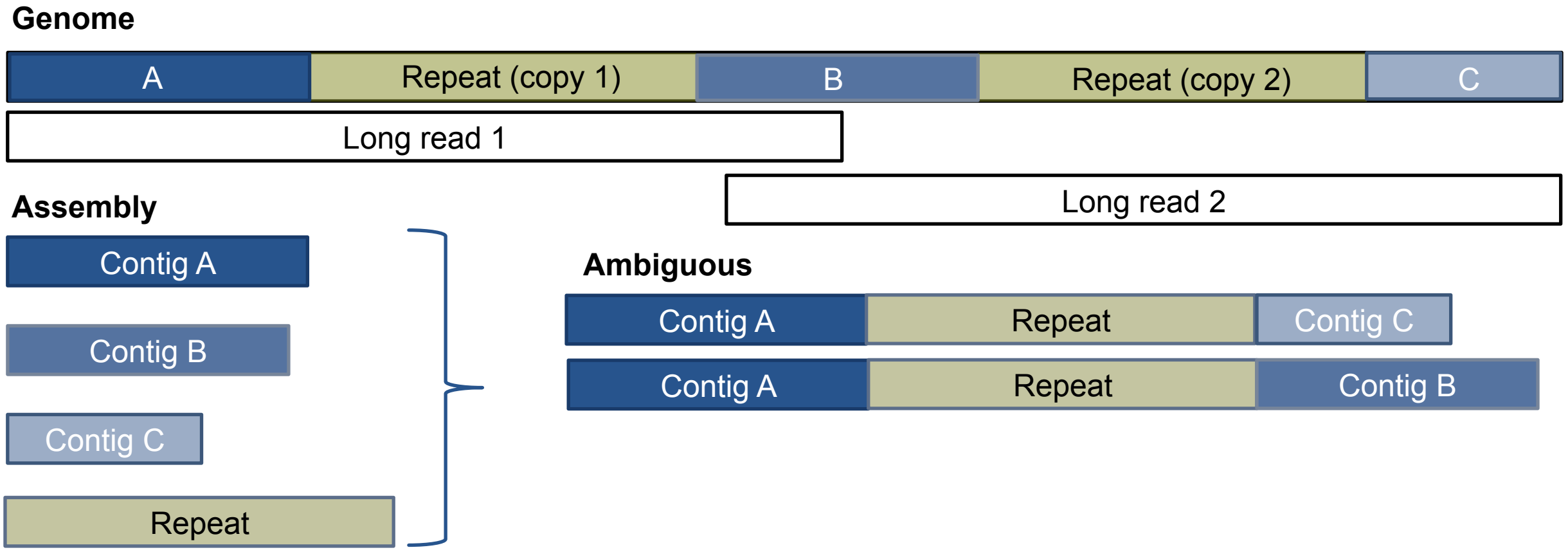
Assembly



Ambiguous



Repetitive sequences in genomes prevent full assembly

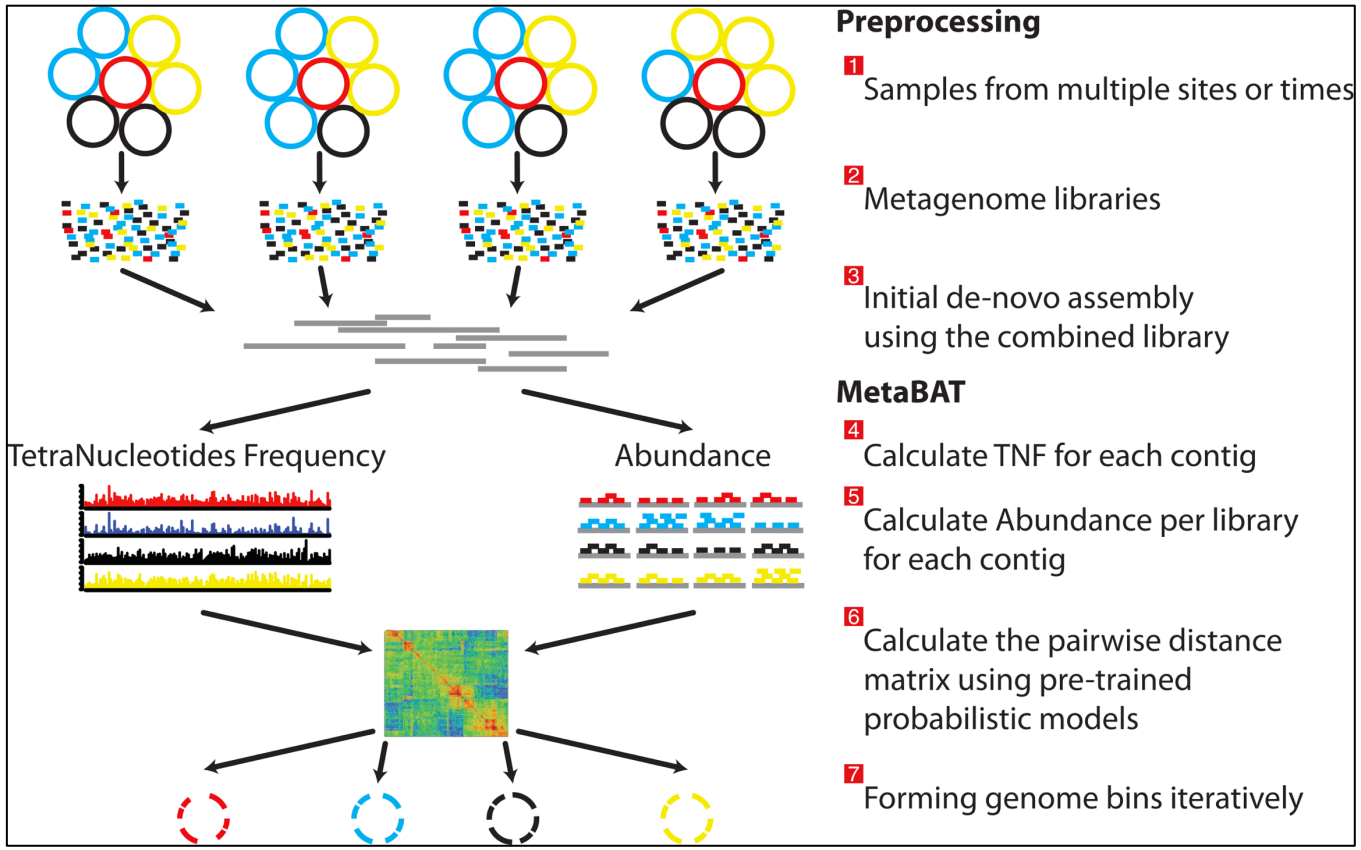


→ **Long sequencing reads can be used to resolve repeats**

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

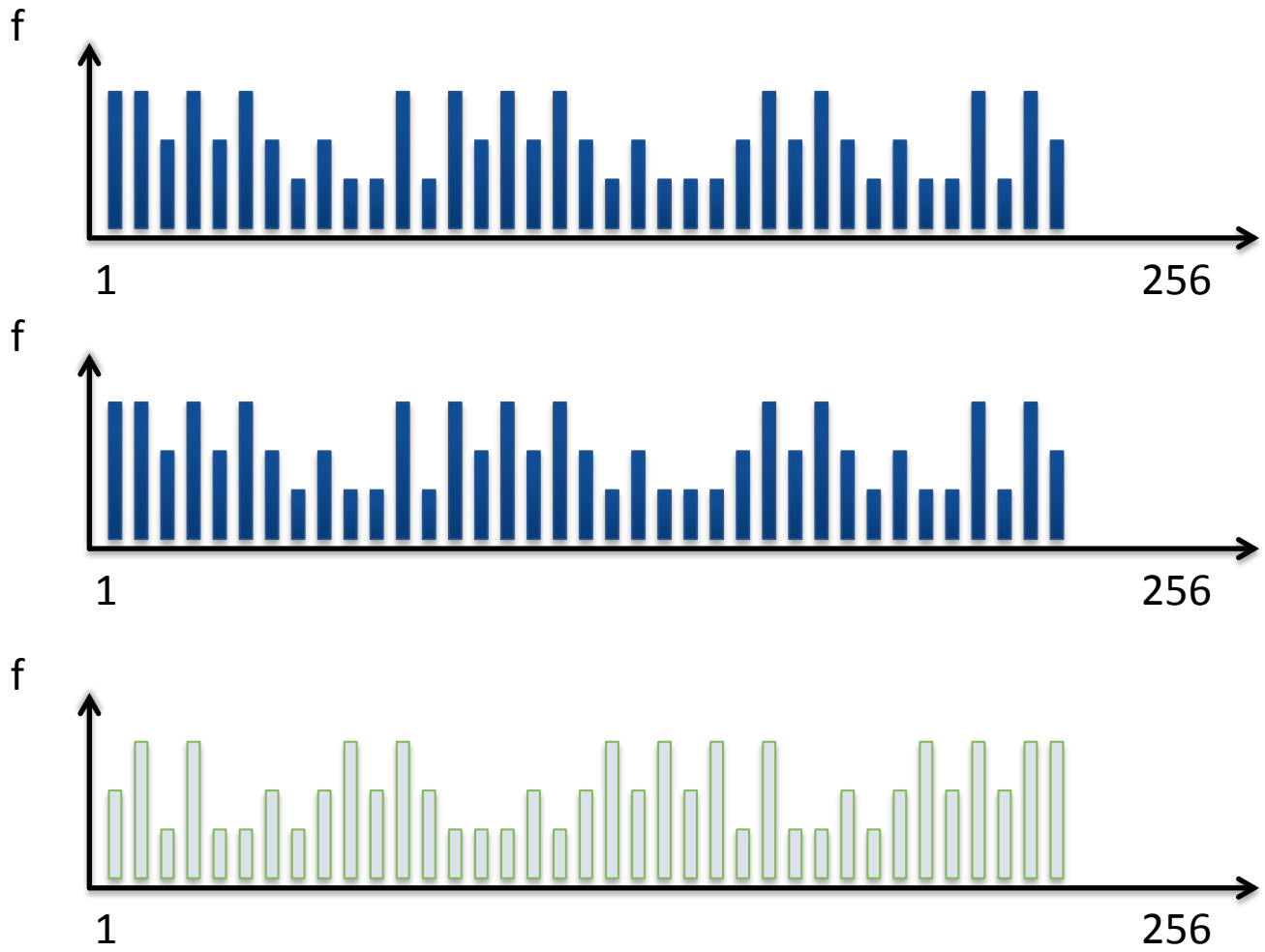


- For each contig:
  a) calculate k-mer frequencies
  b) calculate read abundance
- Combine a) and b) into a distance matrix
- Resolve distance matrix into clusters of highly correlated contigs

Kang et al. PeerJ (2015)

Example for tetranucleotide frequency (TNF) distances

$[ATGC]^4$ = 256 possible combinations



Contig 1

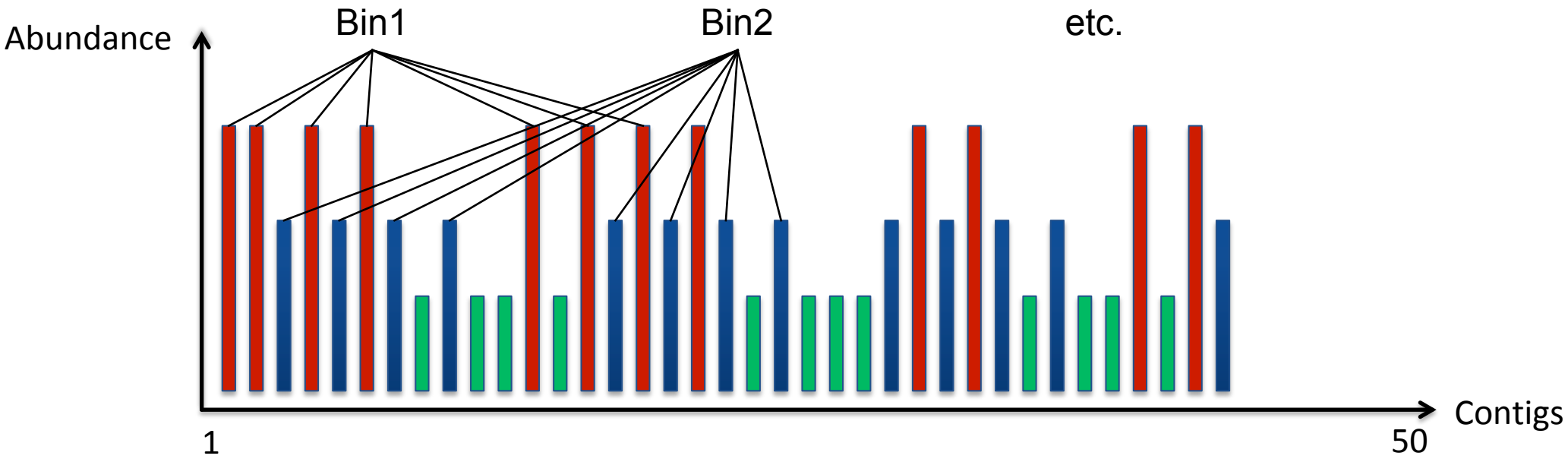
Contig 2

Contig 3

small distance – high likelihood for same bin

large distance

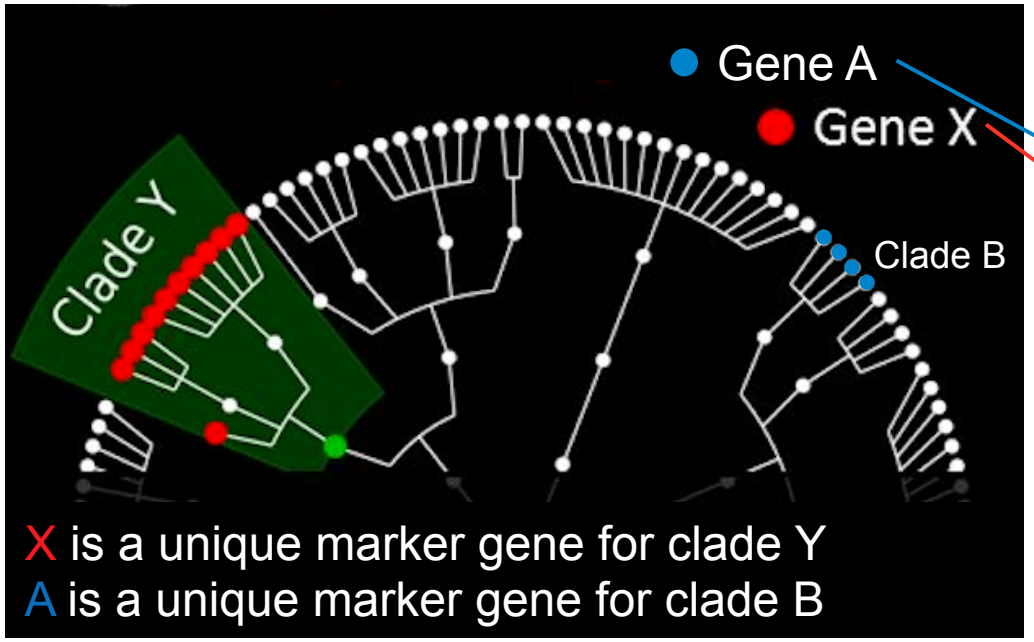
Example for abundance-based distances

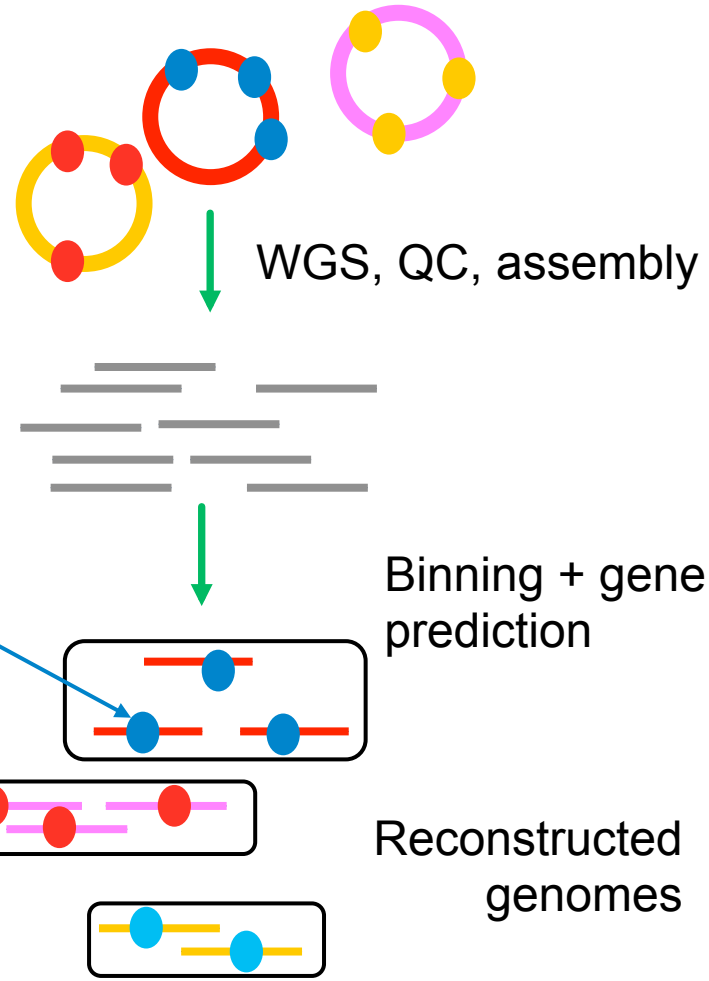


→ TNF and abundance-based distances can be combined and used for iterative binning

Step 3: Binning contigs/scaffolds – taxonomy guided

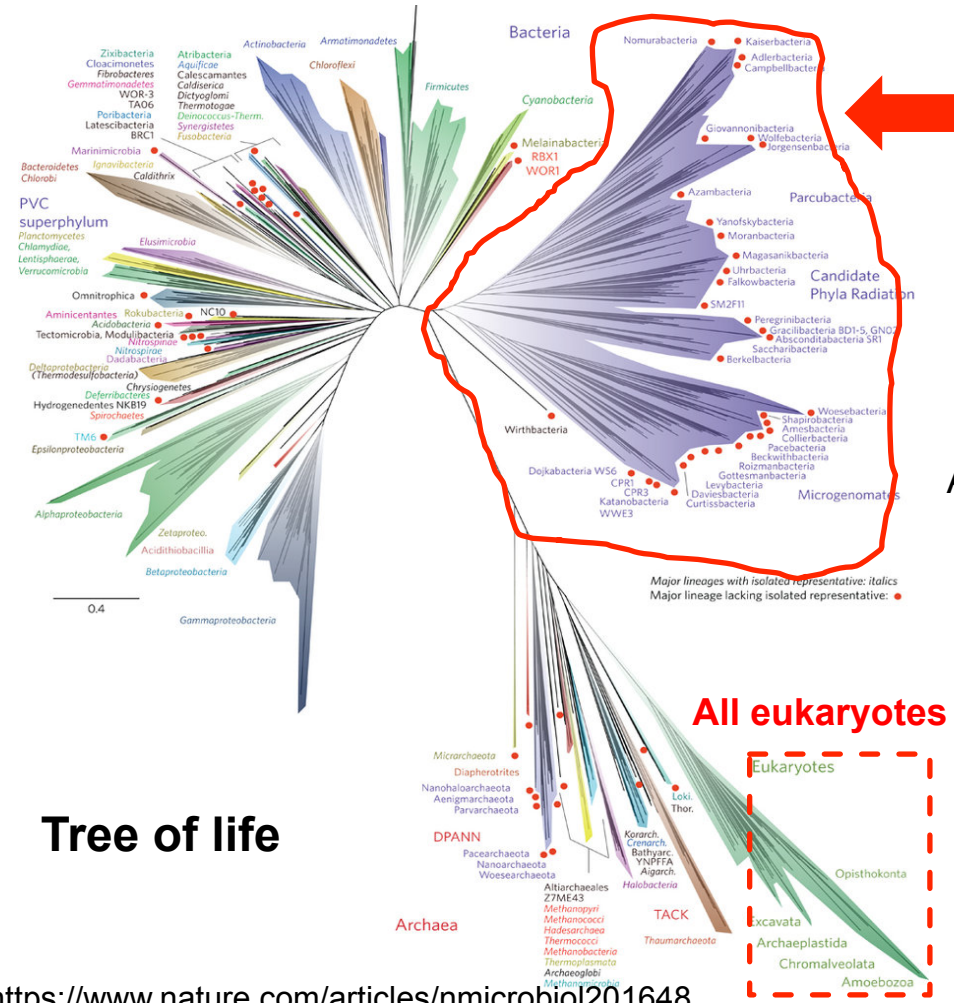
Taxonomy-guided (based on reference genomes) use of clade specific marker genes for binning metagenomic assemblies into draft genomes





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

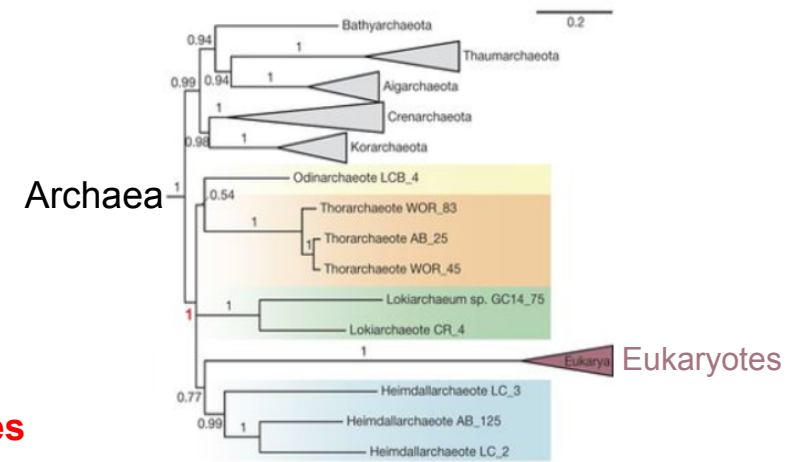
Applied examples III



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

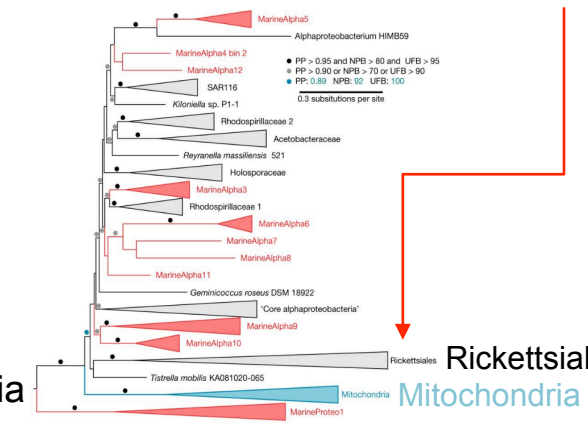
Candidate phyla radiation discovered by metagenomics

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
- Strain level differences between genomes of the same species result in core and pan-genomes
- MAGs (metagenome assembled genomes) have revealed unexpected diversity and new implications on evolutionary origin of eukaryotes and mitochondria