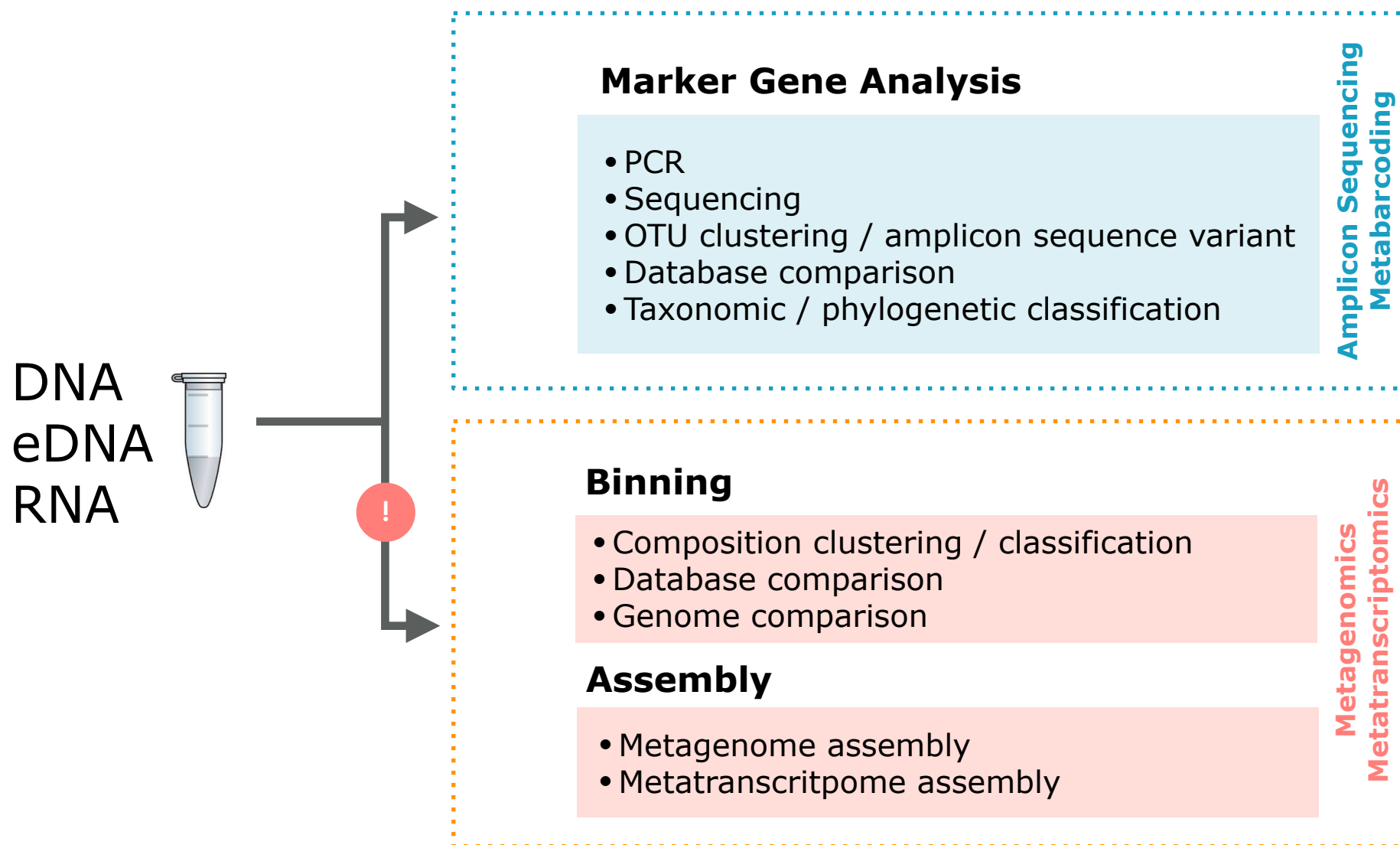


Genetic Diversity: Analysis
Amp-Seq (Metabarcoding) / MetaSeq
Thursday, 26. June 2025

Overview

Amplicon Sequencing (Metabarcoding) - A targeted DNA sequencing method that amplifies specific genomic regions (e.g., 16S rRNA or ITS) using PCR, commonly used to profile microbial communities.

Metagenomics - A shotgun sequencing approach that analyses all genetic material in a sample, allowing comprehensive study of entire microbial communities without prior targeting.

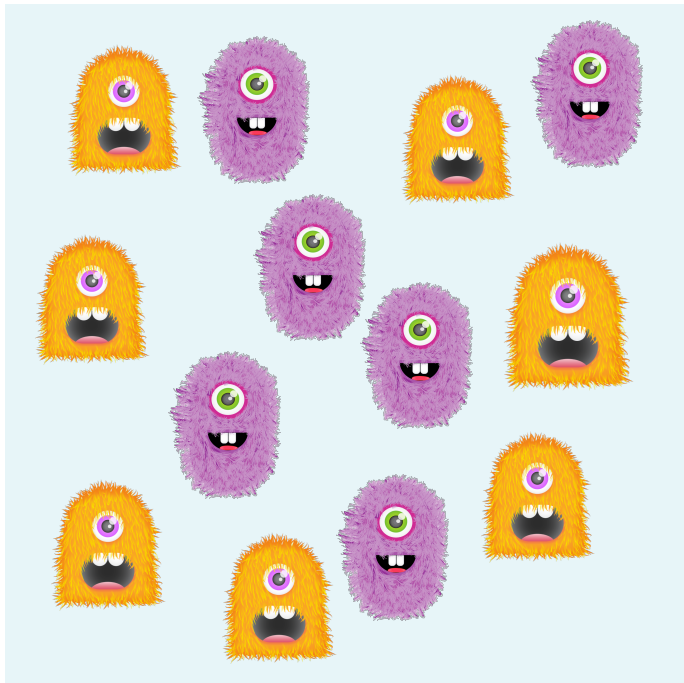


! Please think! Cleaning and/or filtering your raw data may save you some trouble.

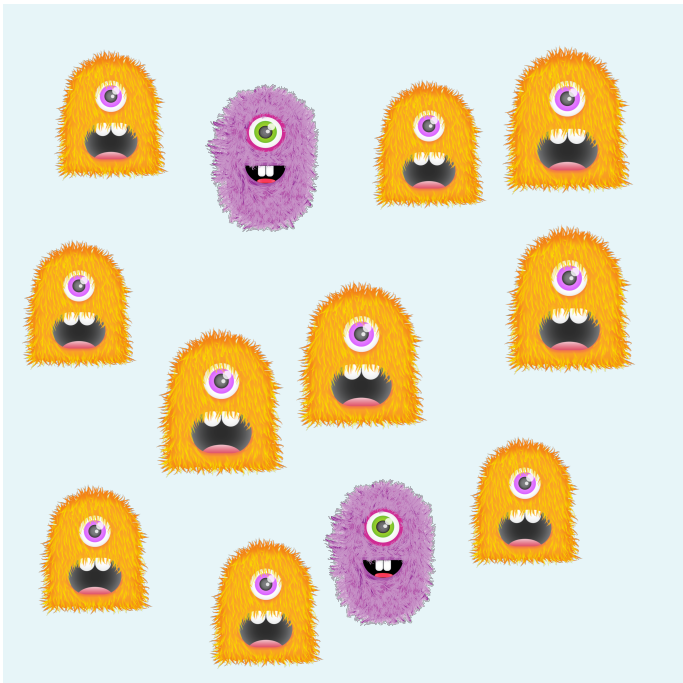
Amplicon-Sequencing

Metabarcoding

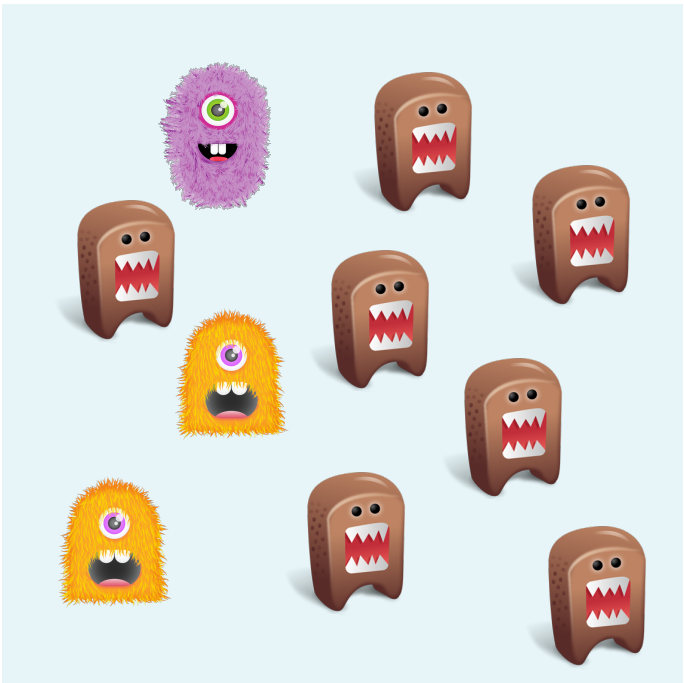
SampleA



SampleB

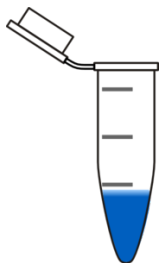


SampleC

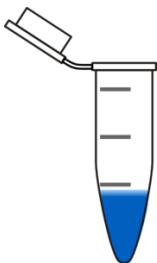


Species
Composition

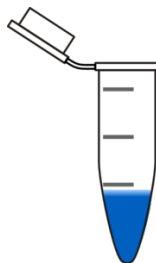
DNA

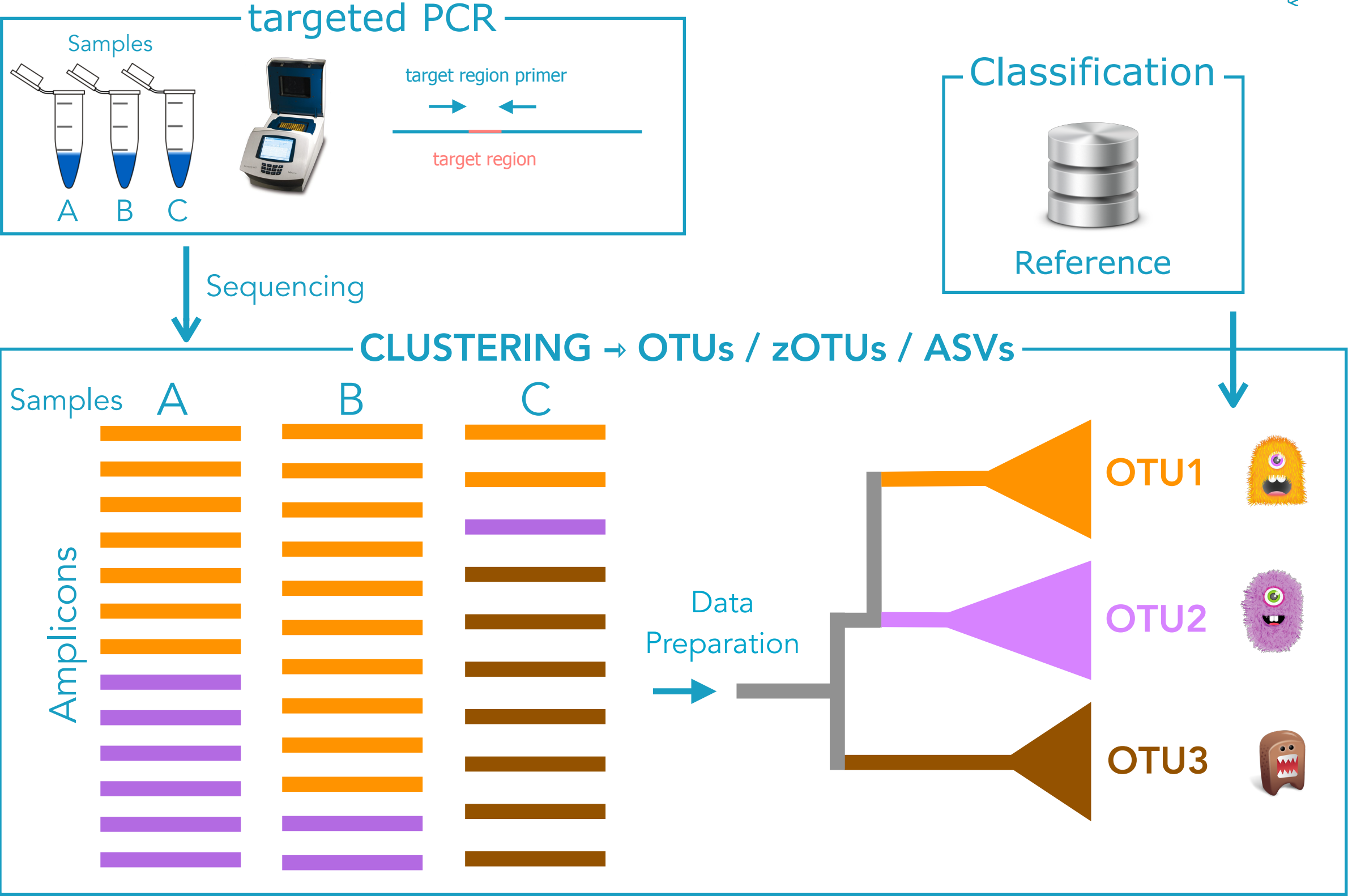


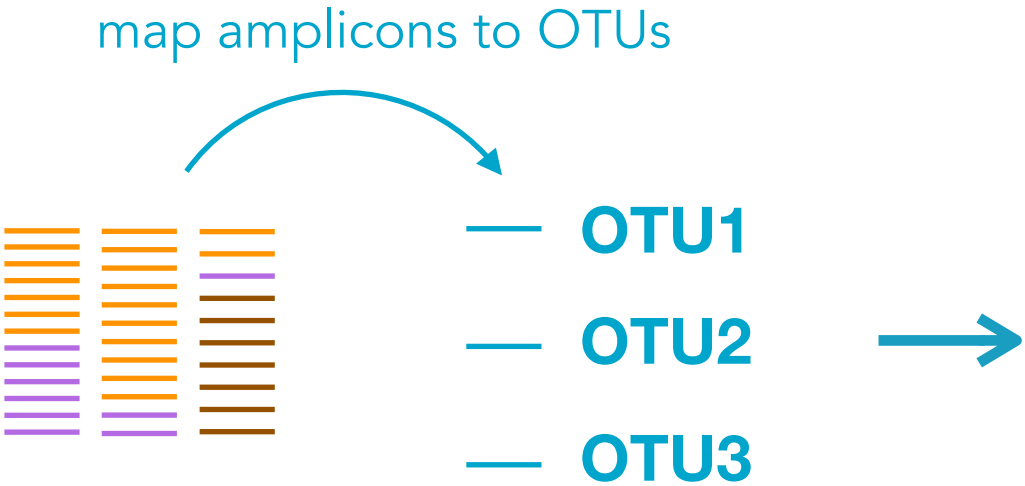
DNA



DNA

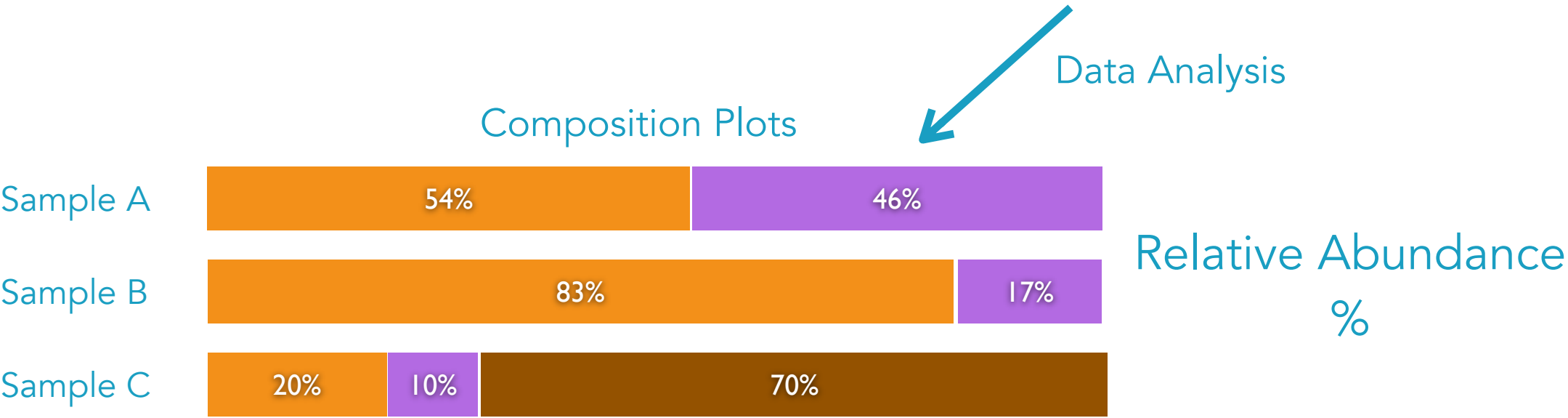






OTU/Count-Table			
	A	B	C
OTU1	7	10	2
OTU2	6	2	1
OTU3	0	0	7
Total	13	12	10

Sequencing Depth



ARTICLE

Received 6 Nov 2015 | Accepted 12 Jul 2016 | Published 30 Aug 2016

DOI: 10.1038/ncomms12544

OPEN

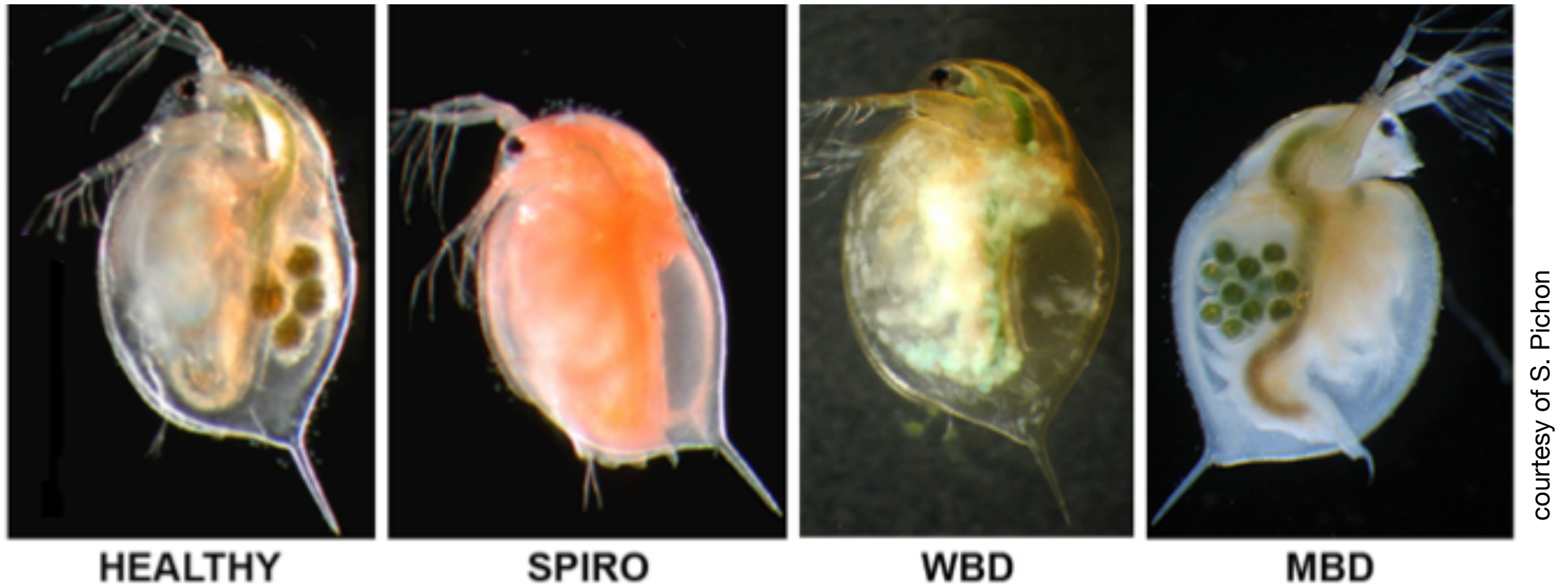
Environmental DNA reveals that rivers are conveyer belts of biodiversity information

Kristy Deiner^{1,2}, Emanuel A. Fronhofer^{1,3}, Elvira Mächler^{1,3}, Jean-Claude Walser⁴ & Florian Altermatt^{1,3}

DNA sampled from the environment (eDNA) is a useful way to uncover biodiversity patterns. By combining a conceptual model and empirical data, we test whether eDNA transported in river networks can be used as an integrative way to assess eukaryotic biodiversity for broad spatial scales and across the land–water interface. Using an eDNA metabarcode approach, we detect 296 families of eukaryotes, spanning 19 phyla across the catchment of a river. We show for a subset of these families that eDNA samples overcome spatial autocorrelation biases associated with the classical community assessments by integrating biodiversity information over space. In addition, we demonstrate that many terrestrial species are detected; thus suggesting eDNA in river water also incorporates biodiversity information across terrestrial and aquatic biomes. Environmental DNA transported in river networks offers a novel and spatially integrated way to assess the total biodiversity for whole landscapes and will transform biodiversity data acquisition in ecology.

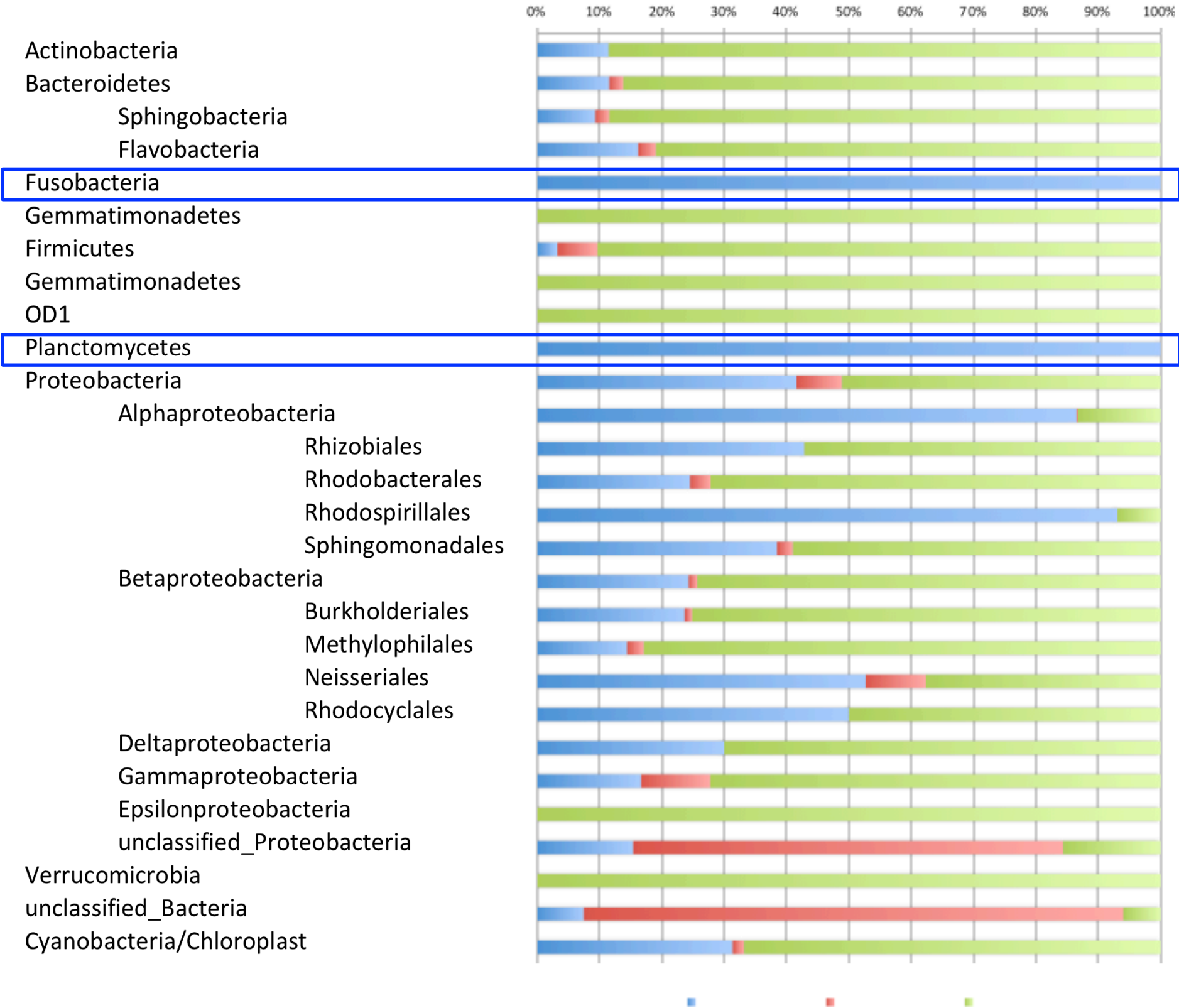
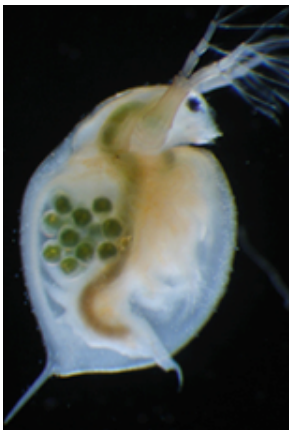
NATURE COMMUNICATIONS | 7:12544 | DOI: 10.1038/ncomms12544 | www.nature.com/naturecommunications

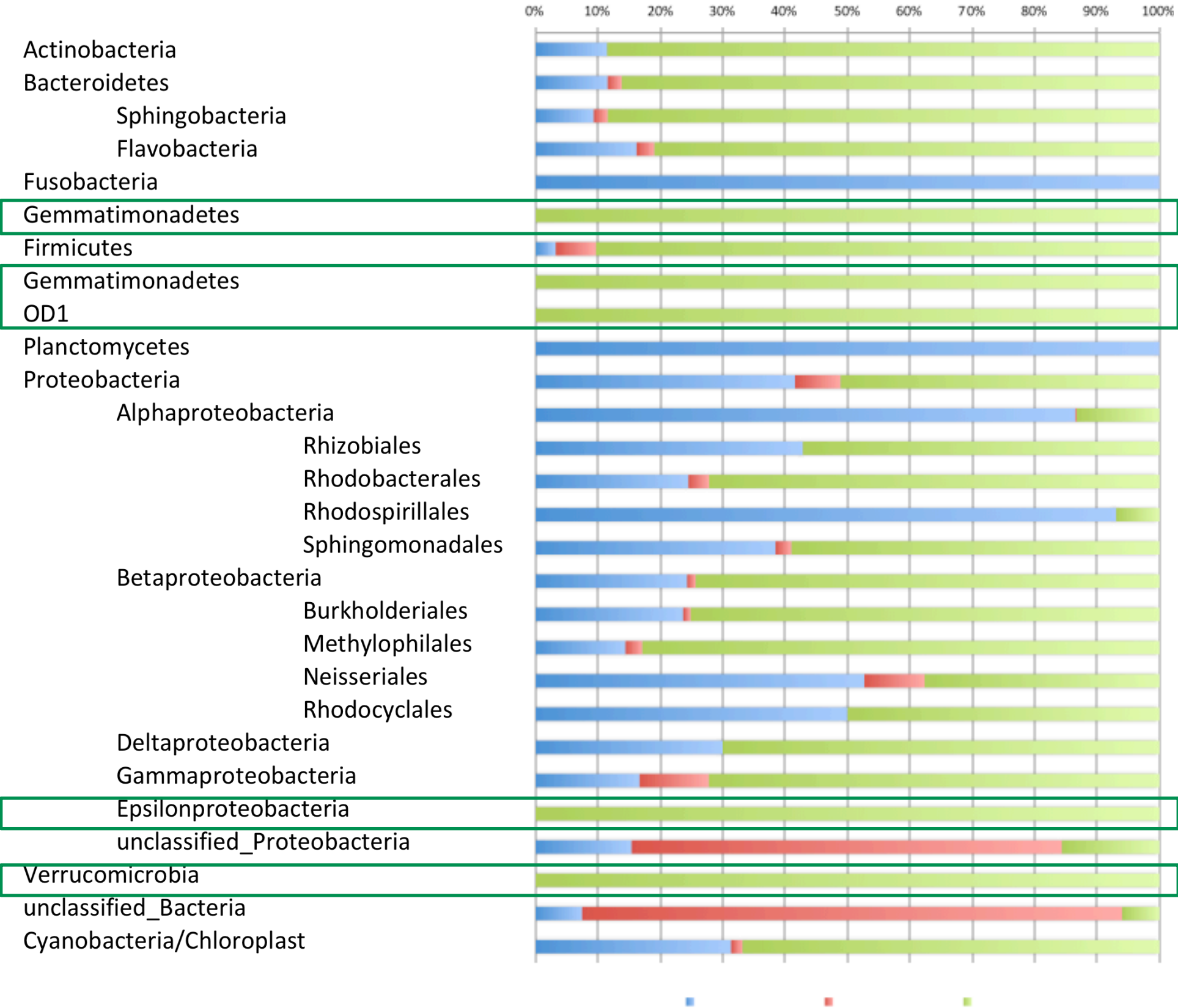
daphnia disease phenotypes



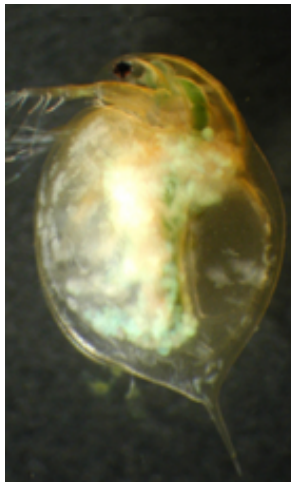
Daphnia magna disease phenotype. From the left, a healthy female *D. magna* and three disease phenotypes: infected with *Spirobacillus* sp. (SPIRO), white bacterial disease (WBD), and milky blood disease (MBD).

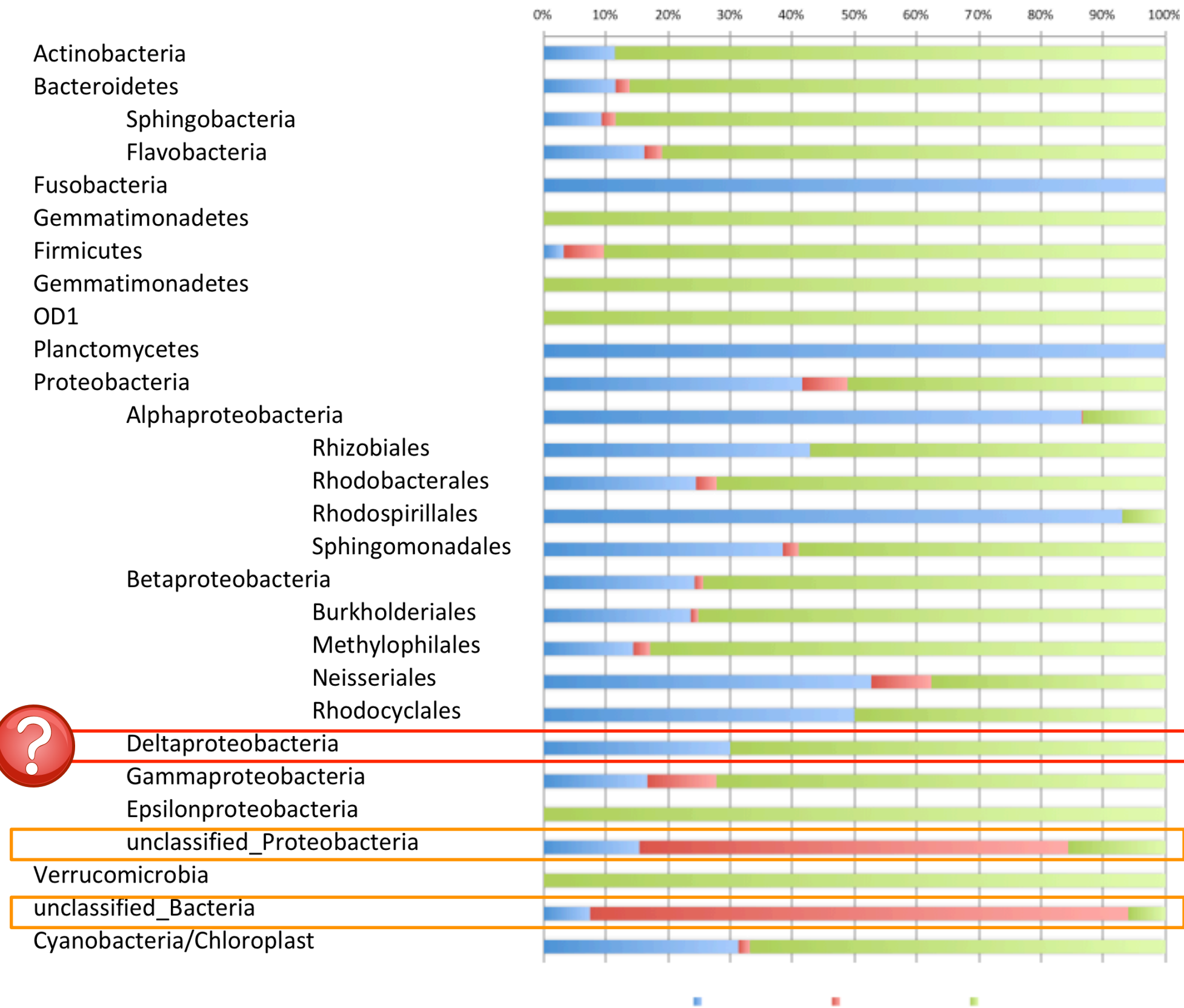
MBD





WBD





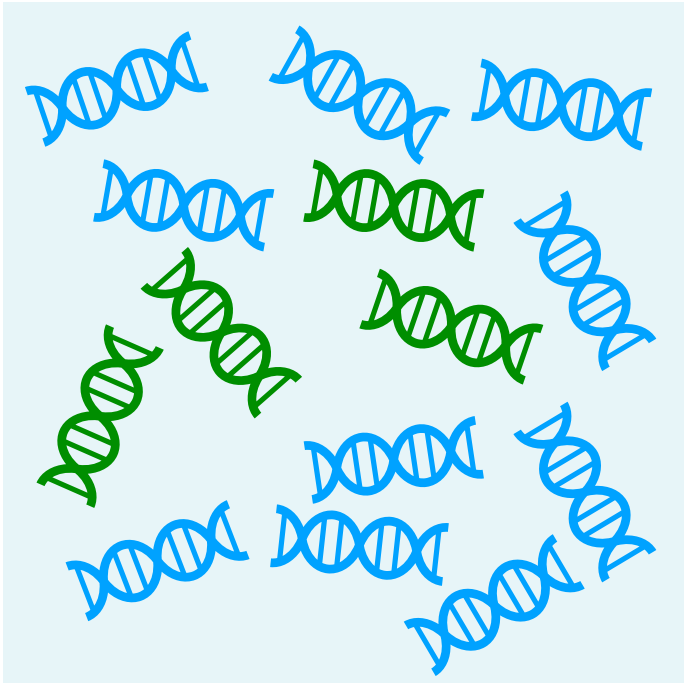
SPIRO



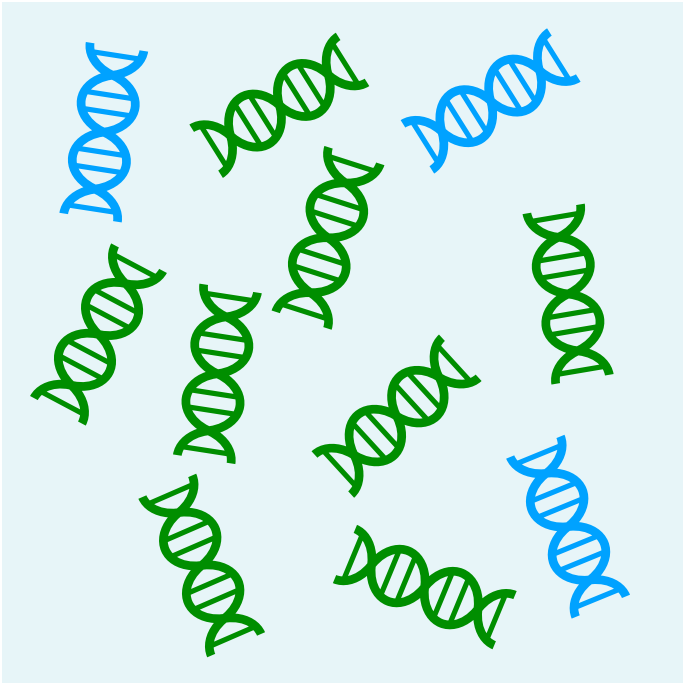
d: Bacteria
p: Proteobacteria
c: Deltaproteobacteria
o: Oligoflexales
f: Oligoflexaceae
g: Spirobacillus
s: Spirobacillus_cienkowskii

Metagenome/Metatranscriptome-Sequencing

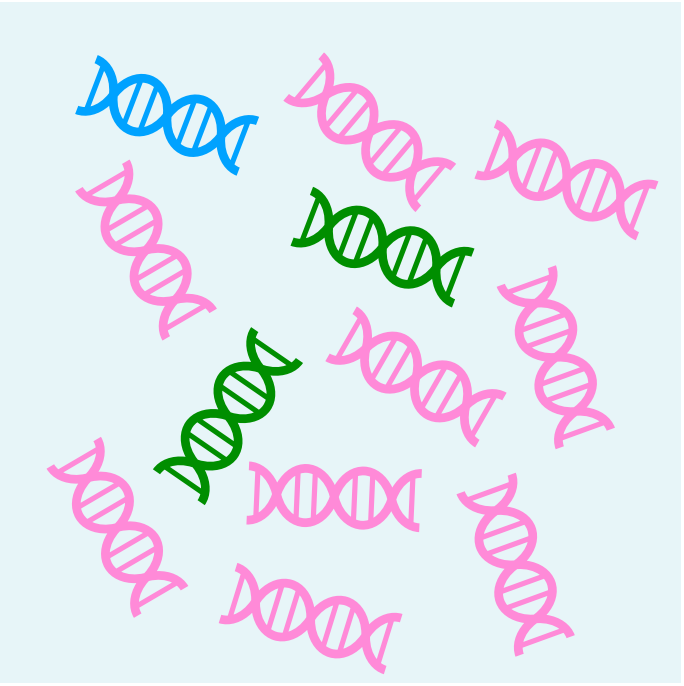
SampleA



SampleB



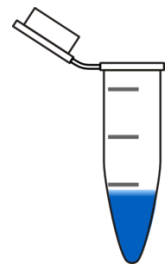
SampleC



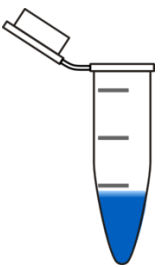
species
composition

functional
composition

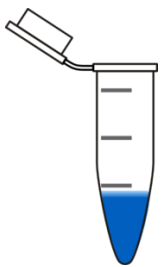
DNA
RNA

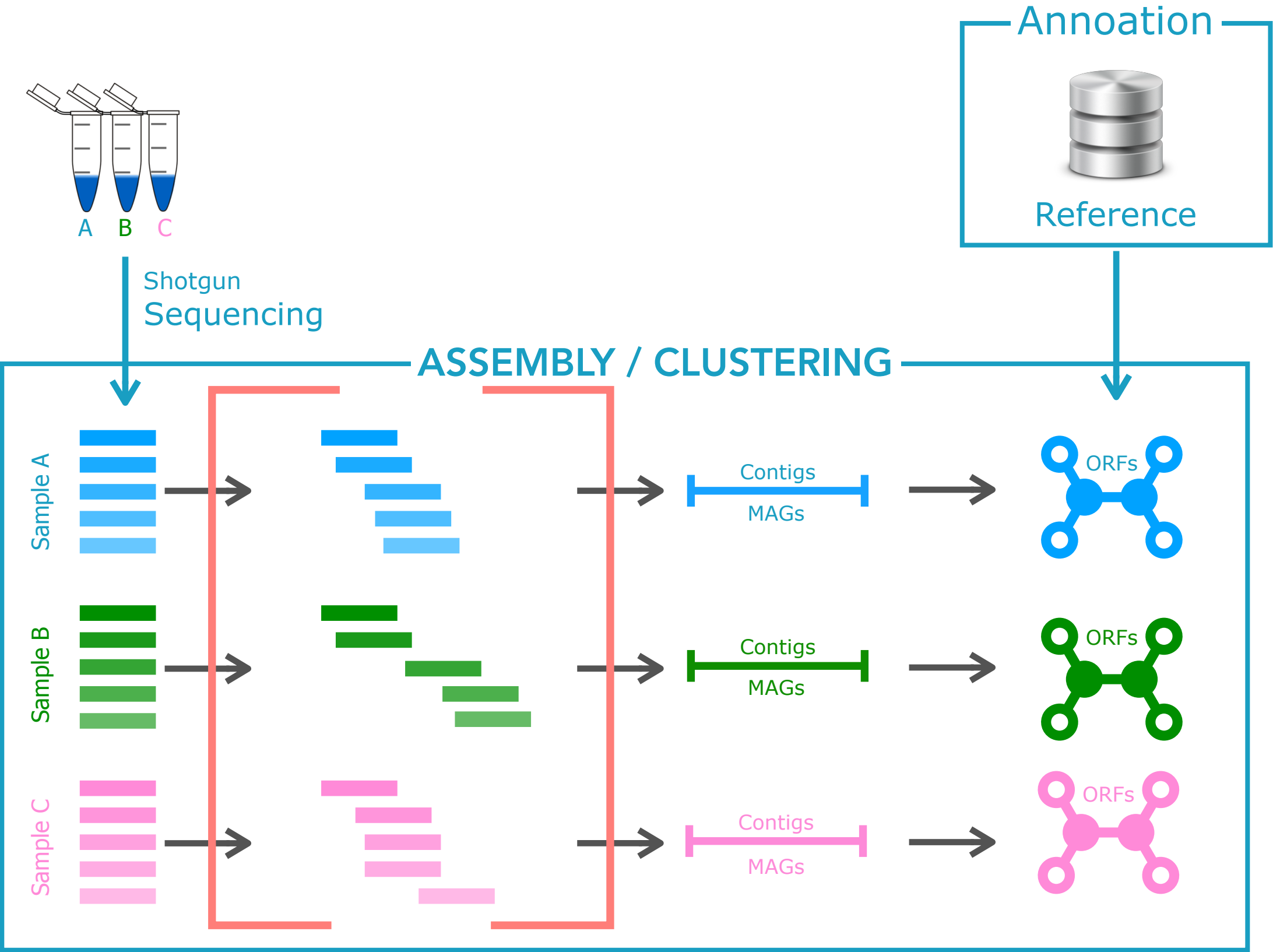


DNA
RNA

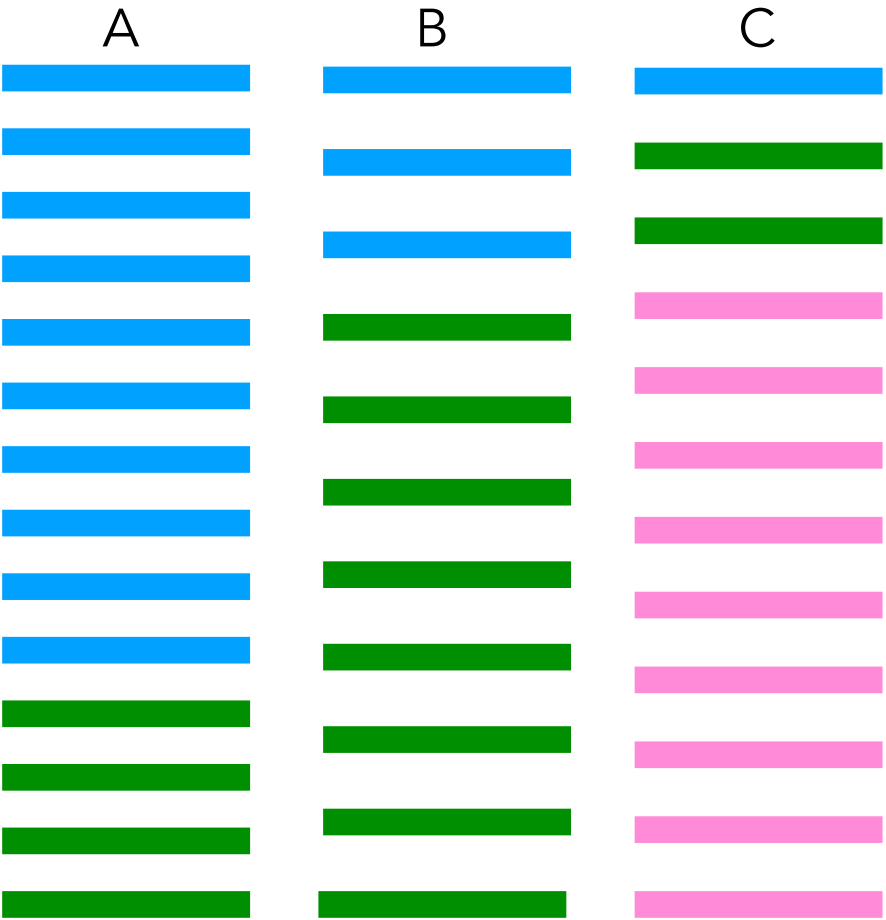


DNA
RNA





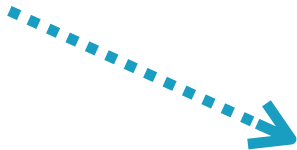
MAG - Meta-Assembled Genomes



Functional Annoation






Species Annoation

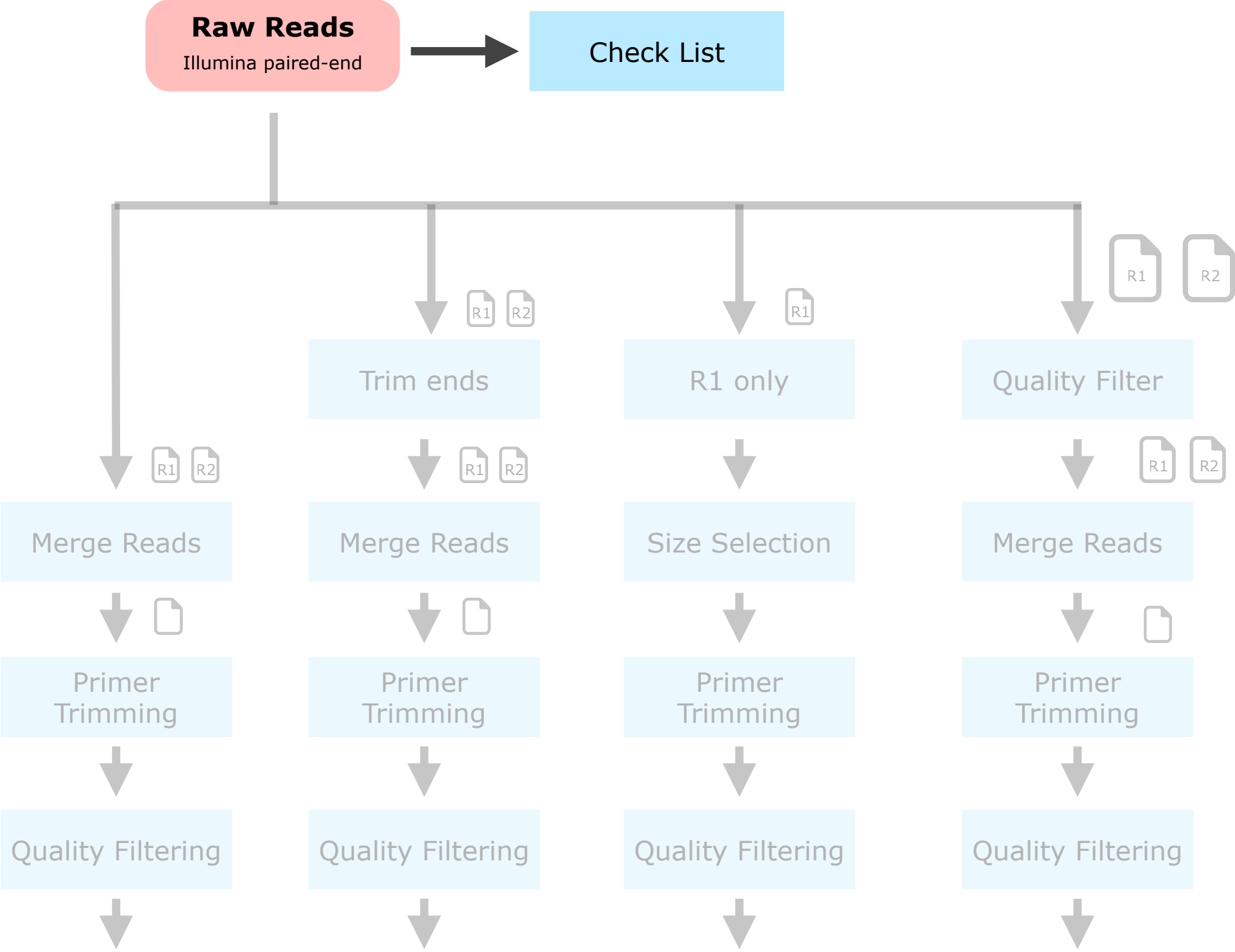


	A	B	C
EC	10	3	1
GO	4	8	2
MP	0	0	9
Total	14	11	12

Enzyme Commission Number (EC)
Gene Ontology (GO)
Metabolic Pathway (MP)

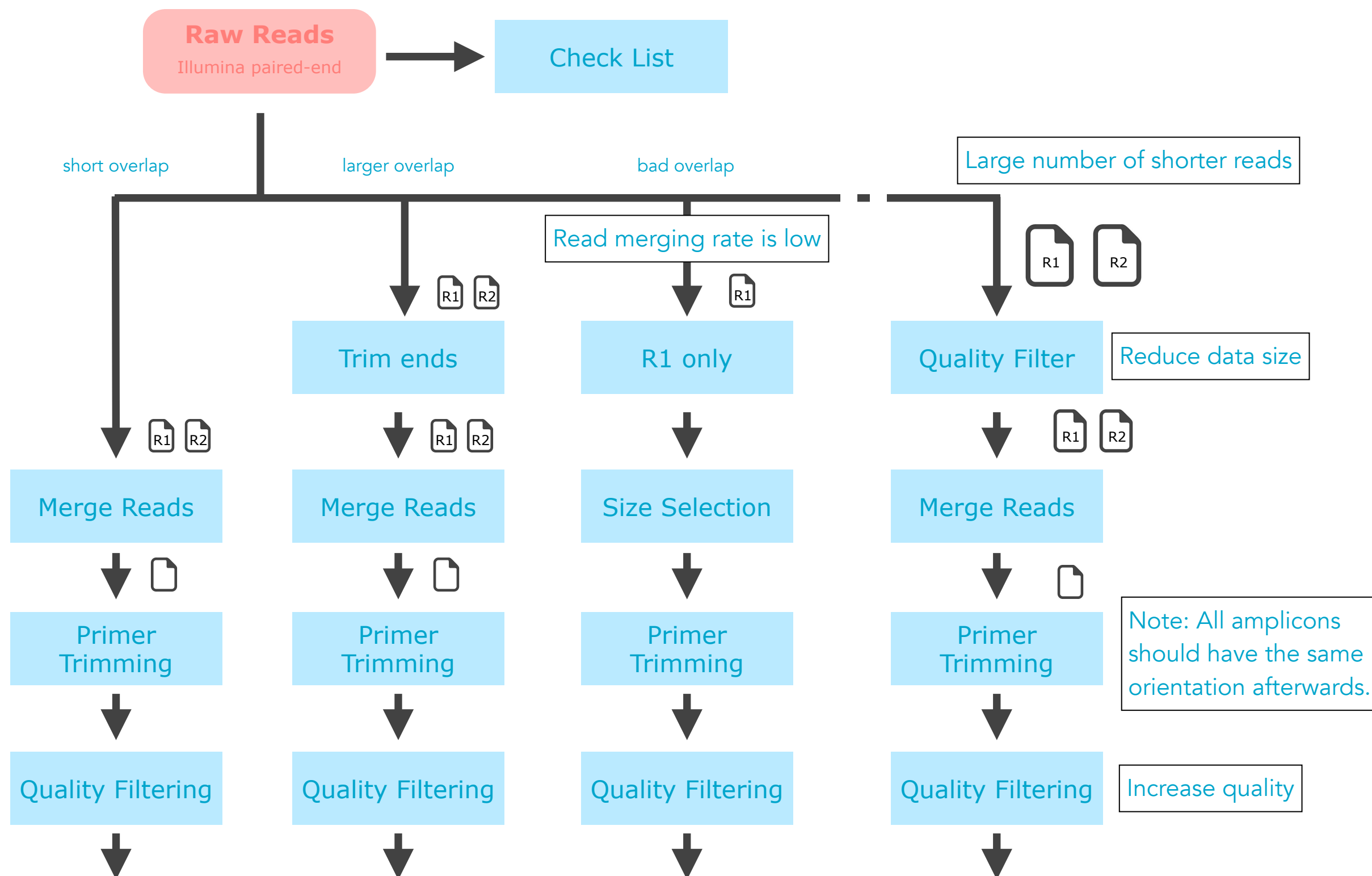
	A	B	C
	7	4	2
	0	0	9
	7	7	1
	14	11	11

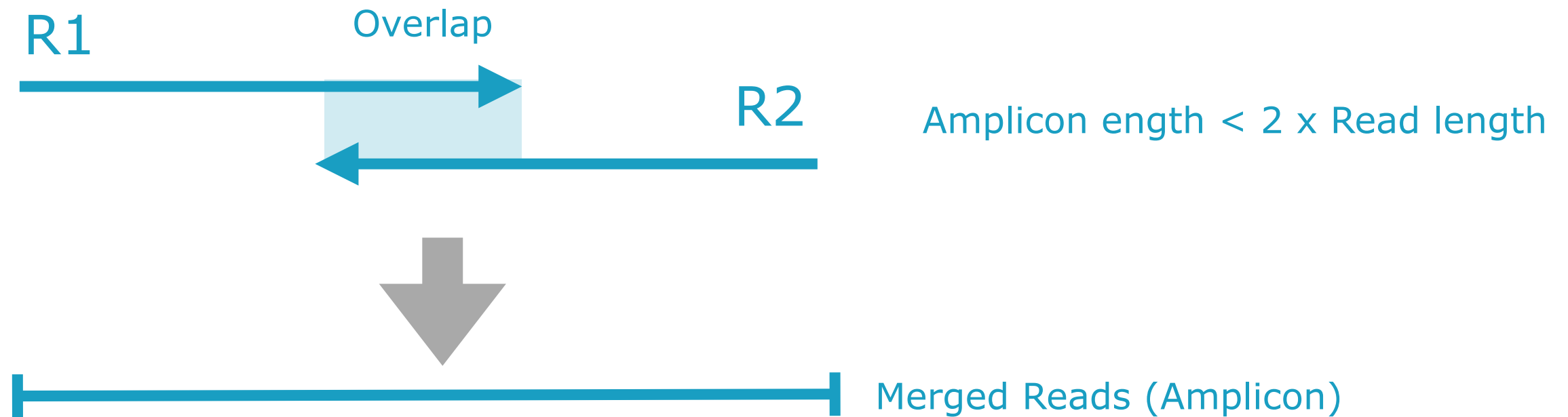
Data preparation



Check-List

1. Download data (if possible via terminal e.g. sftp, wget)
2. Verify file integrity (md5sum)
3. Check data: $N_{\text{samples}} = N_{R1} = N_{R2} / N_{(R1)=N_{(R2)}}$
4. Blast a few random reads
5. Run a quality control (e.g. FastQC, FastScreen)
6. Look at the read size distribution
7. Check fastq header - how many runs?
8. Check for PhiX "contamination"
9. Have a closer look at your control (negative) samples
10. Archive a copy of the raw data
11. Submit the raw data (e.g. ENA)





Paired-end Read Merging:

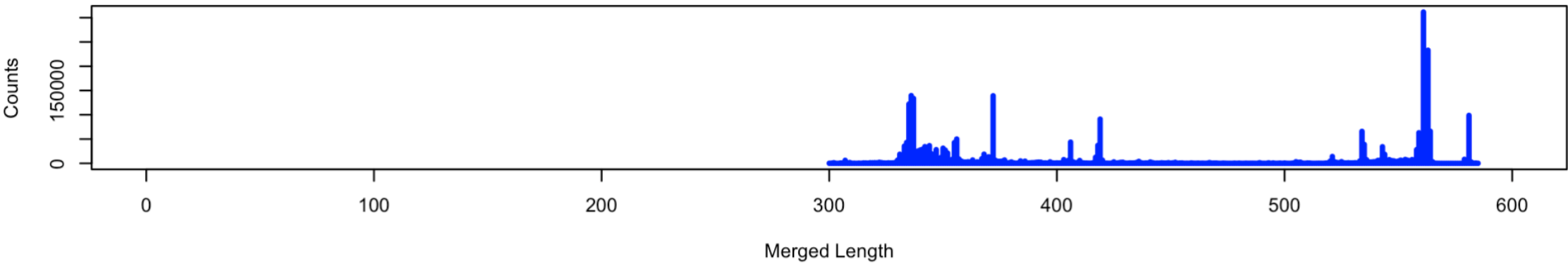
flash: ***2,854,051*** ***(95.1%)***

bbmerge: ***2,490,347*** ***(83.0%)***

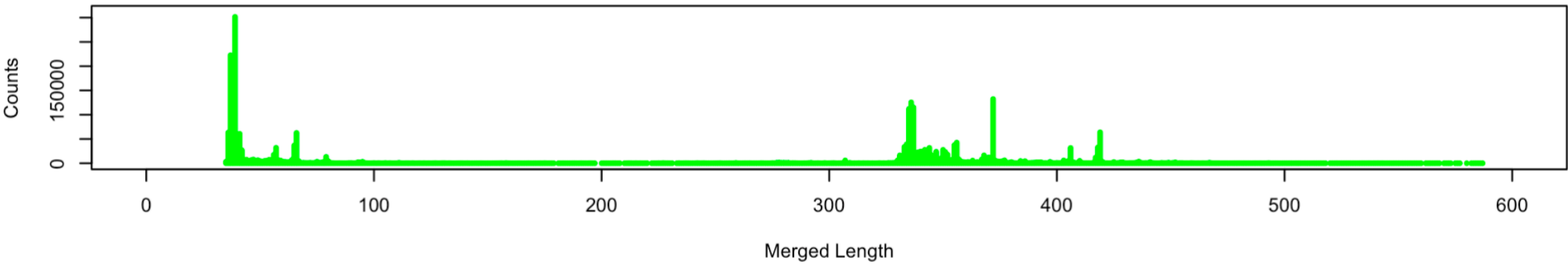
usearch: ***1,857,602*** ***(61.9%)***

?

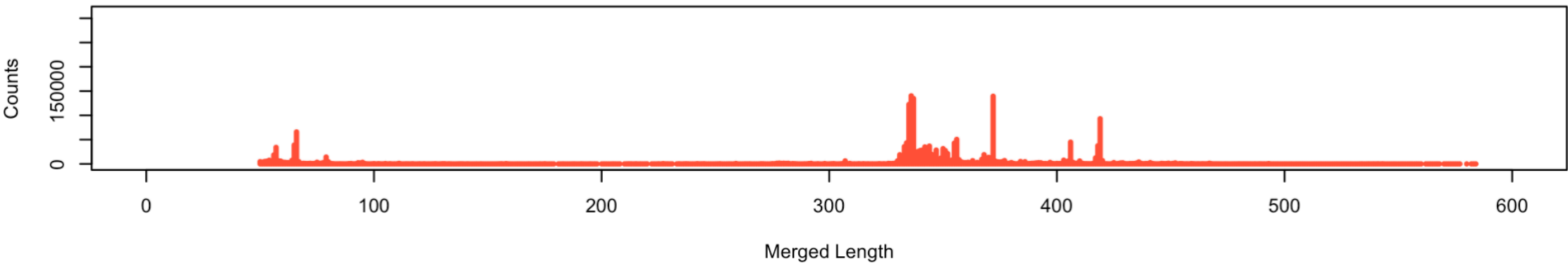
flash

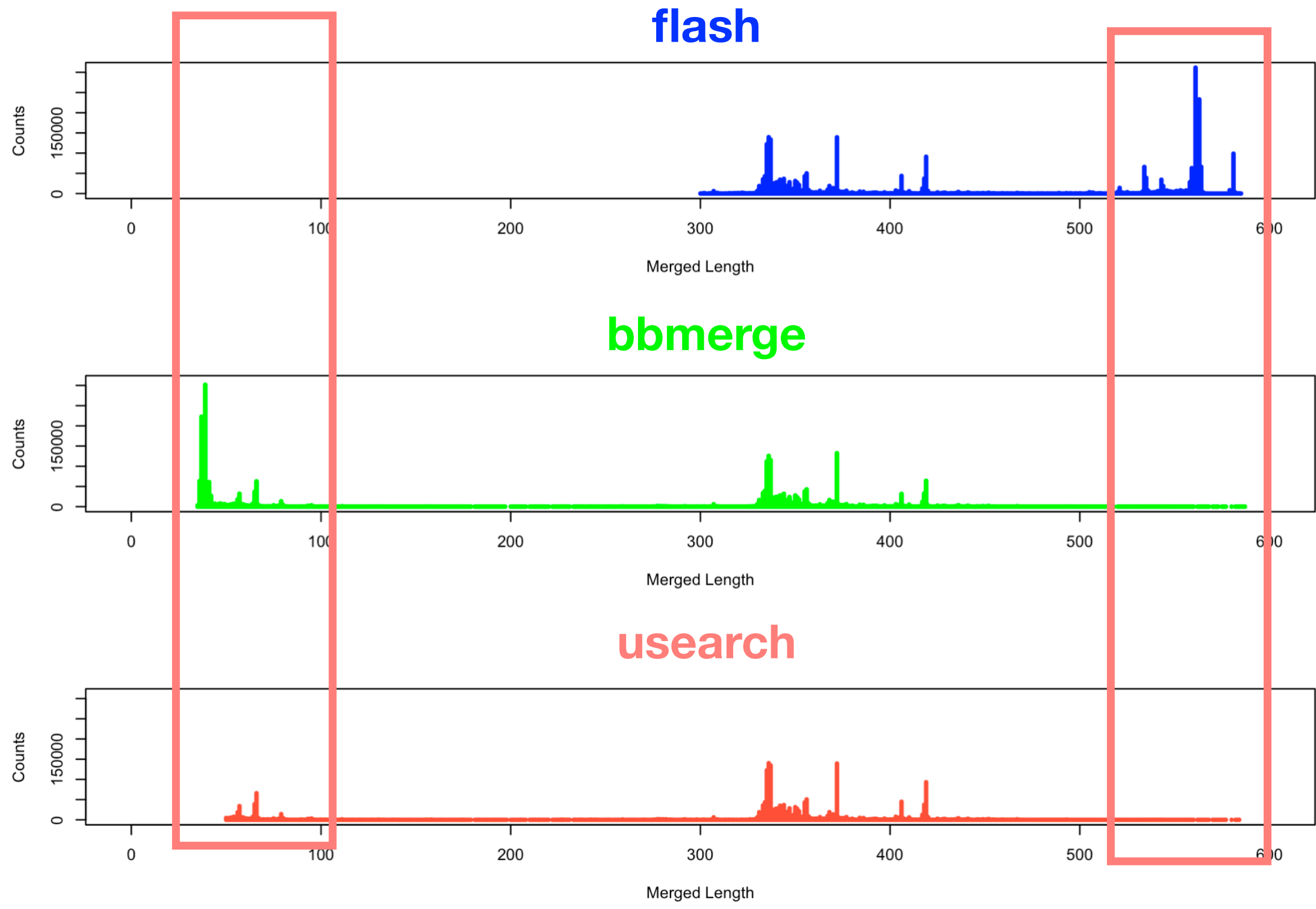


bbmerge

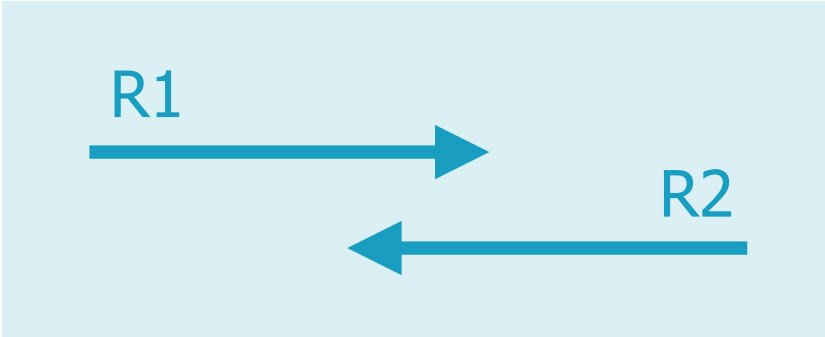


usearch

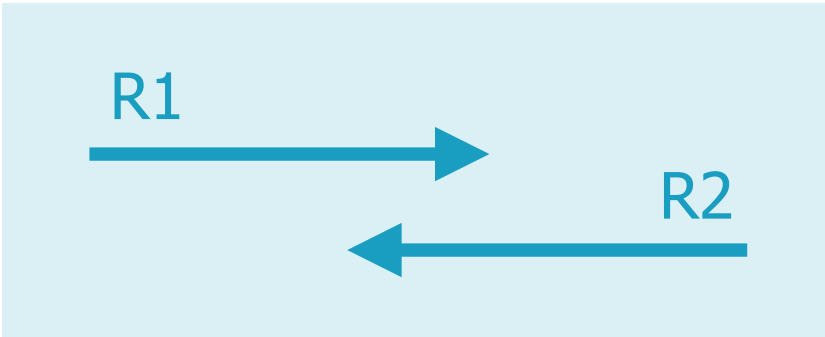




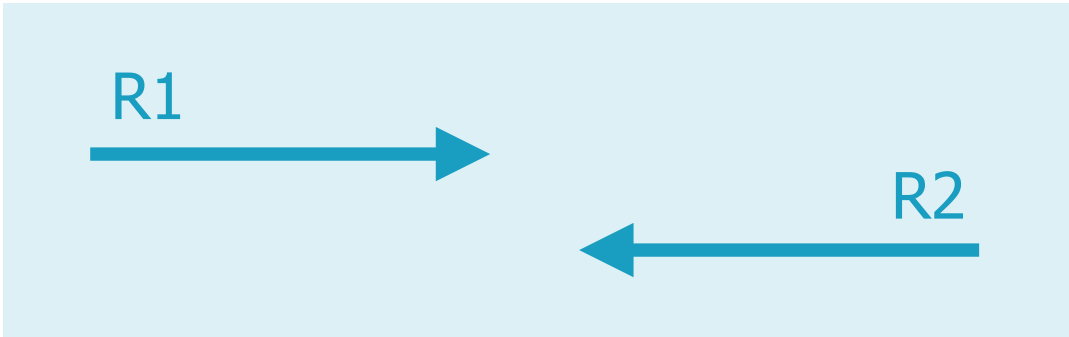
Overlapping Reads



Overlapping Reads

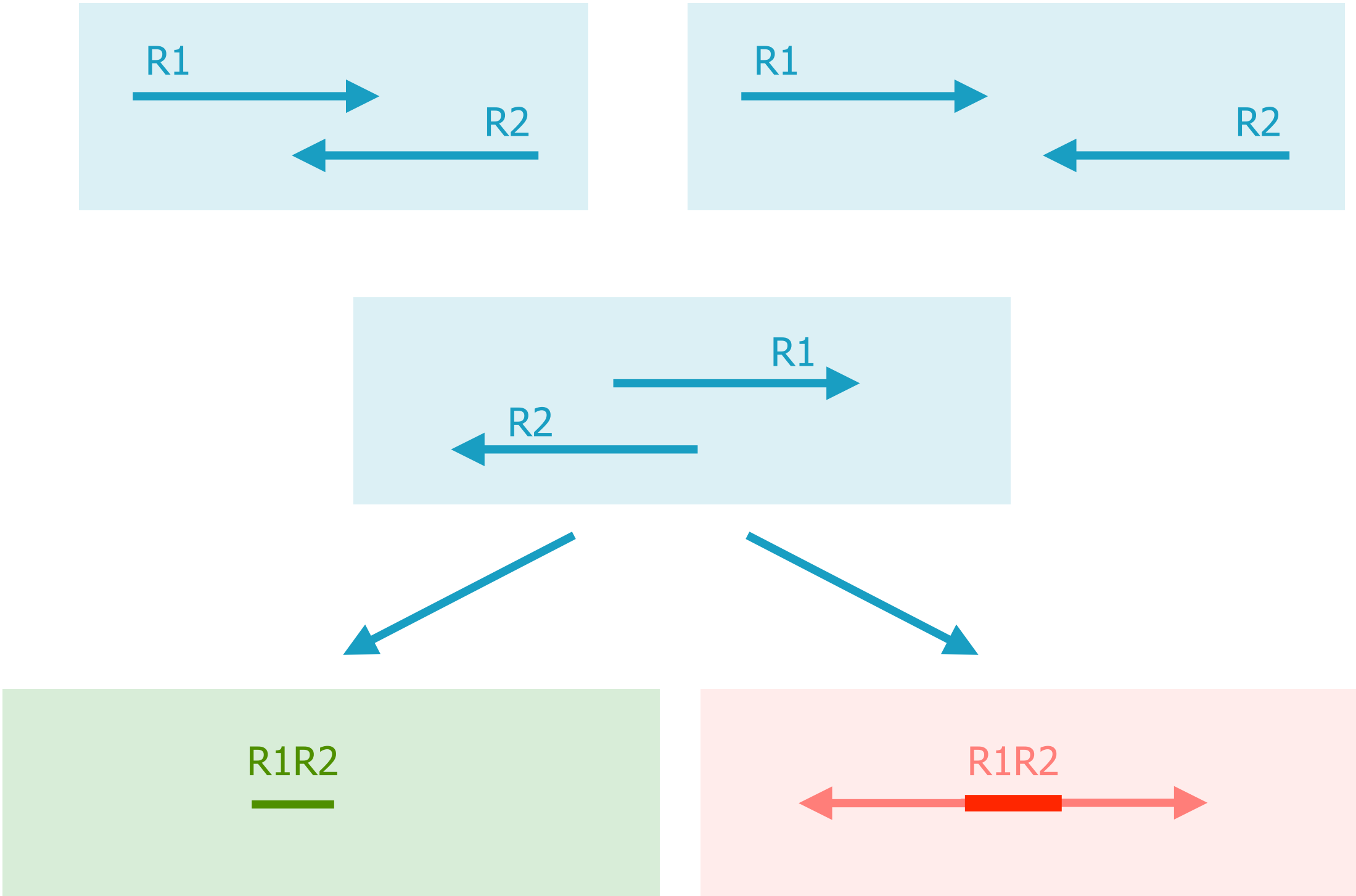


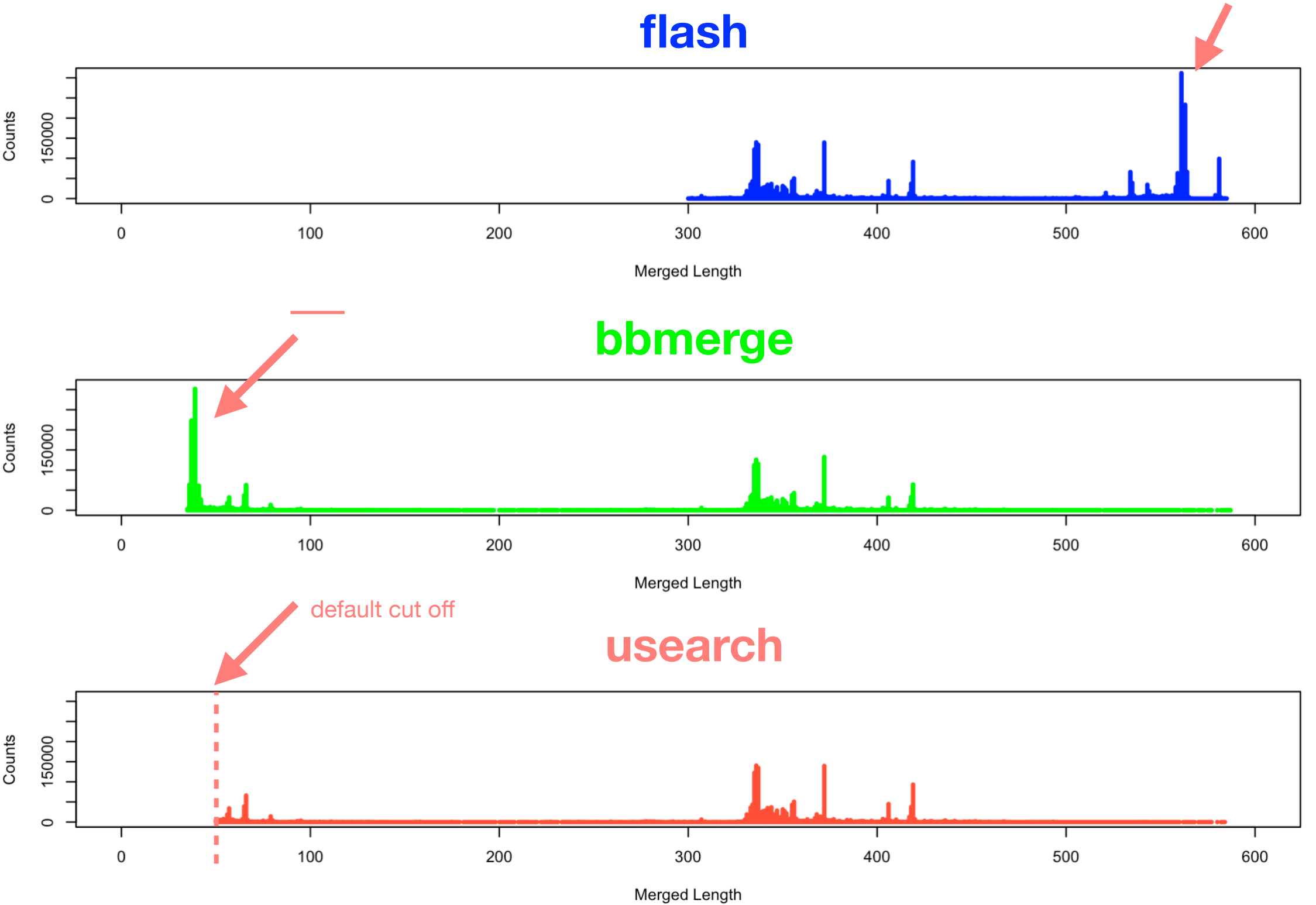
Non-Overlapping Reads



Staggered Reads



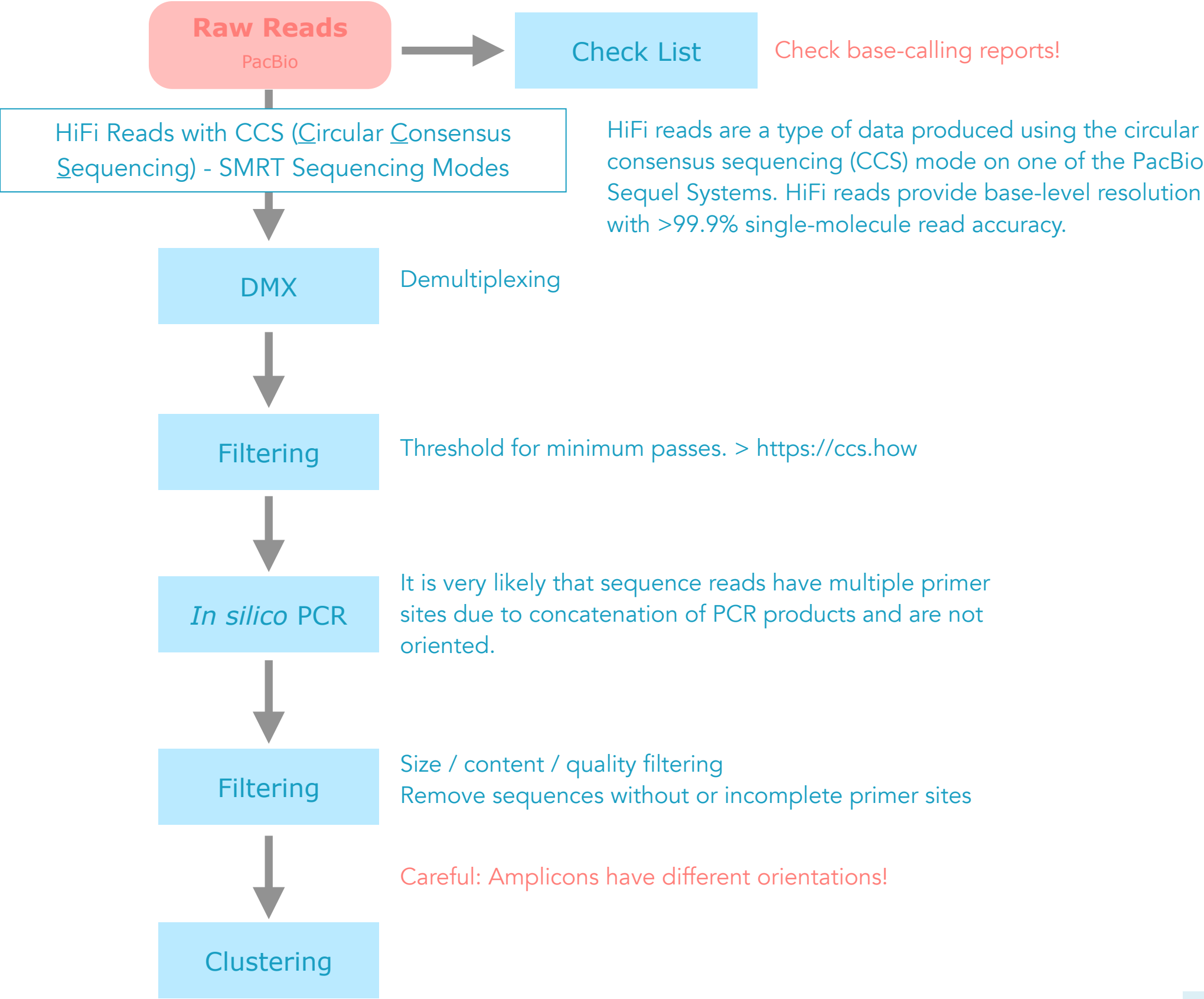


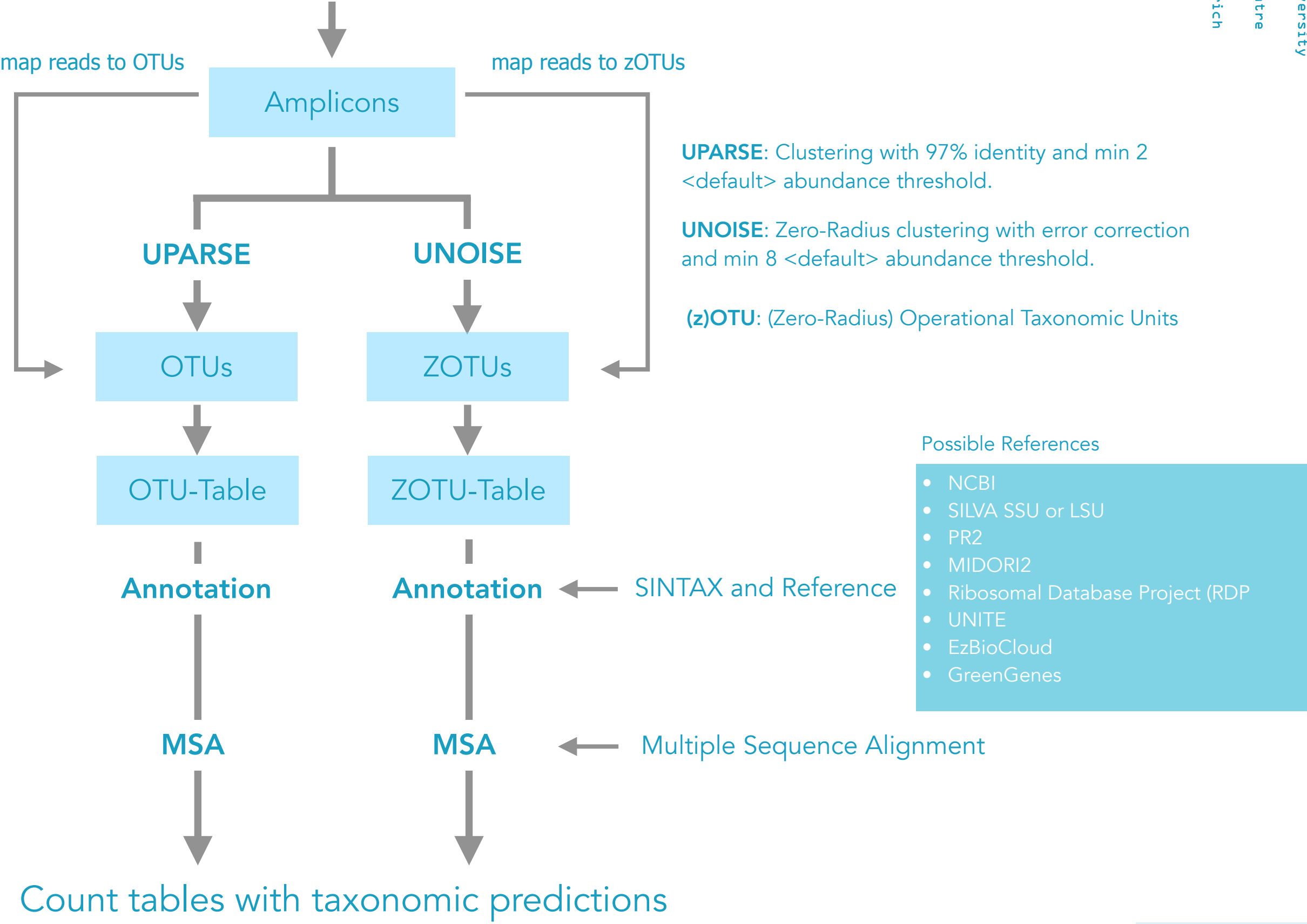


Detailed results of USEARCH::mergepairs

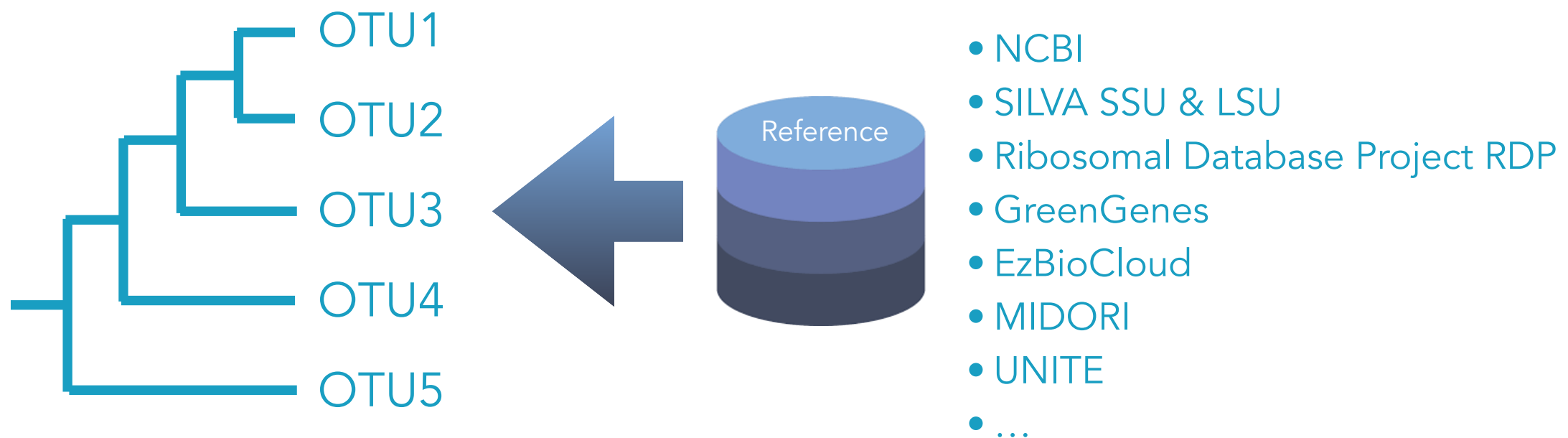
```
► Aa-007 - Merging Summary
  Merging Rate: 116101 / 122531 (94.8%)
  Median Merged Length: 465

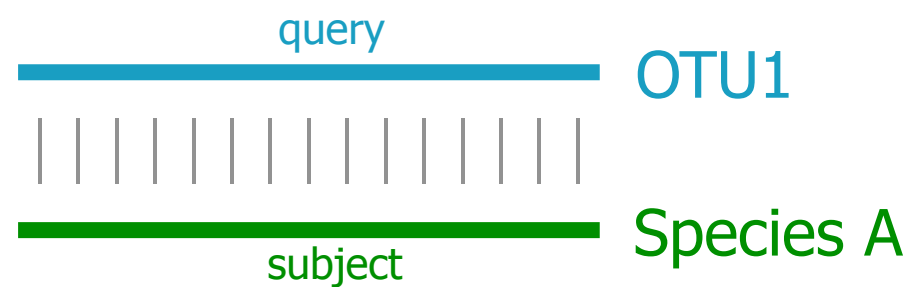
122531 Pairs (122.5k)
116101 Merged (116.1k, 94.75%)
 64263 Alignments with zero diffs (52.45%)
  3037 Too many diffs (> 20) (2.48%)
    48 Fwd too short (< 64) after tail trimming (0.04%)
    12 Rev too short (< 64) after tail trimming (0.01%)
 3330 No alignment found (2.72%)
    0 Alignment too short (< 30) (0.00%)
    3 Merged too short (< 100)
    0 Min Q too low (<0) (0.00%)
    65 Staggered pairs (0.05%) merged & trimmed
 89.40 Mean alignment length
459.45 Mean merged length
  0.71 Mean fwd expected errors
  1.61 Mean rev expected errors
  0.87 Mean merged expected errors
```



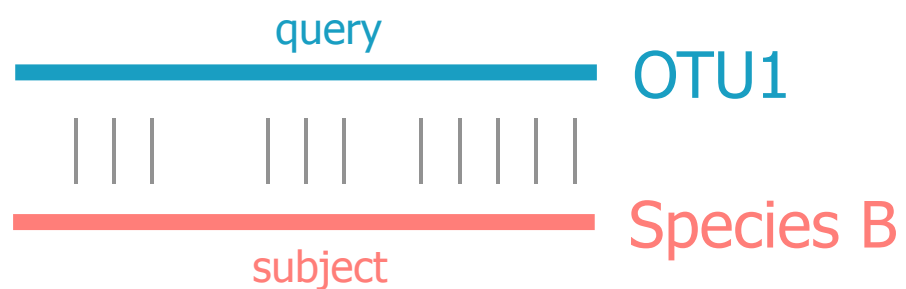


OTU - Annotation (-Prediction)



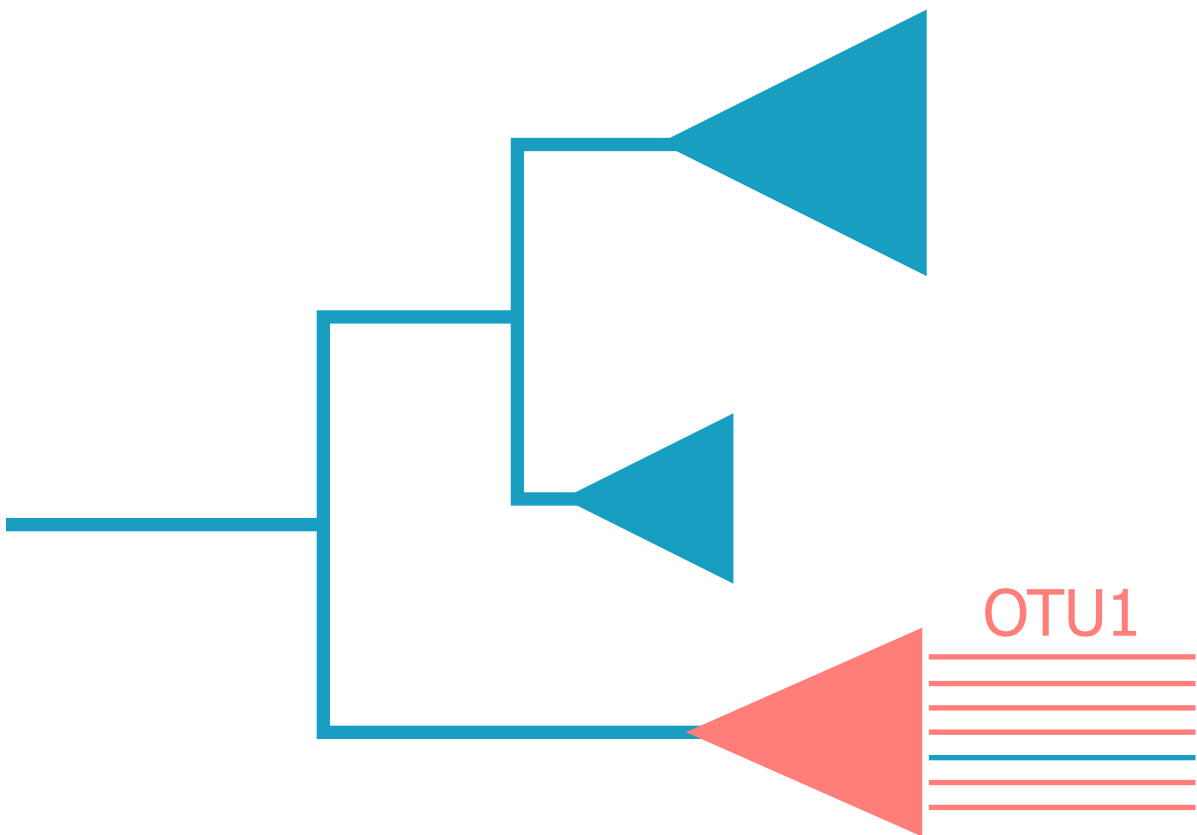


OTU1 \approx Species A

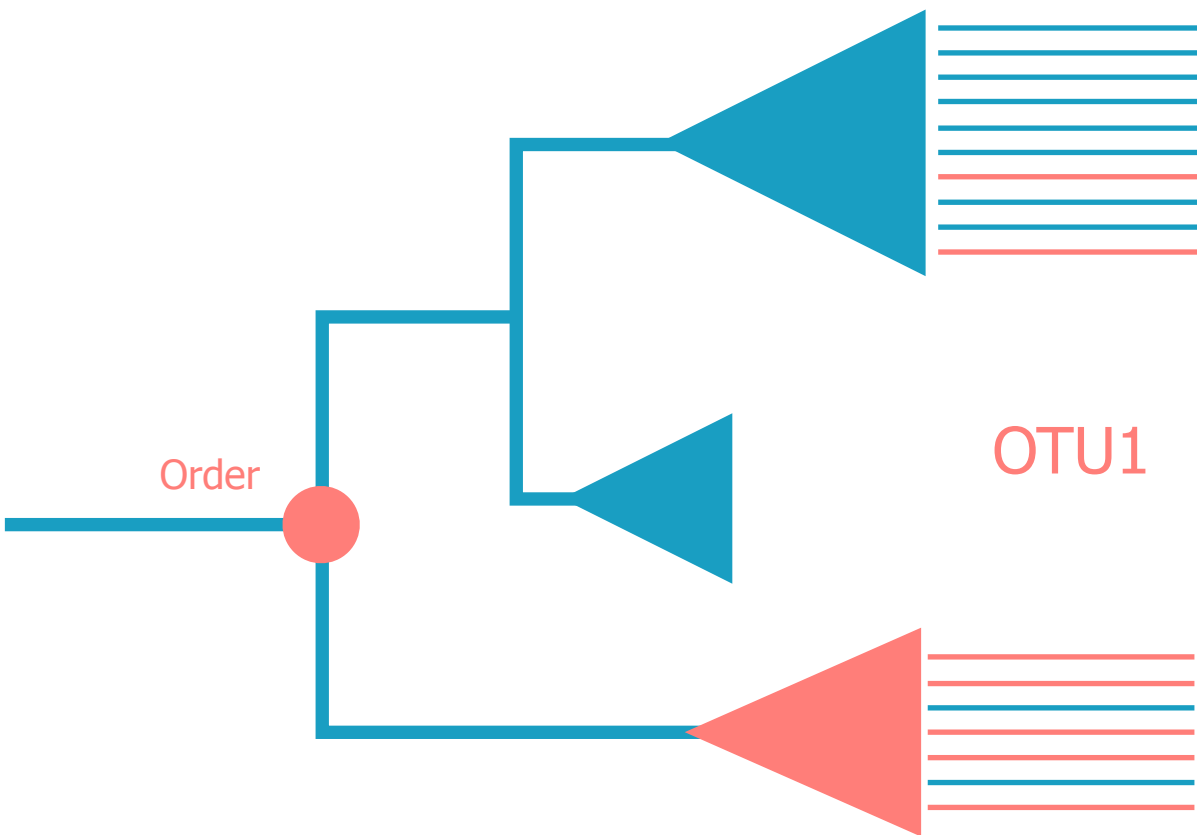


This approach might work if ...

- ... there is only one clear best-hit
- ... there is no other possible hit
- ... the reference is complete
- ... the query is correct
- ... the subject is correct



d:Bacteria (1.000)
p:Tenericutes (1.000)
c:Mollicutes(1.000)
o:Mycoplasmatales(1.000)
f:Mycoplasmataceae(1.000)
s:Echinogammarus_veneris(0.980)



d:Bacteria (1.000)
p:Proteobacteria (1.000)
c:Gammaproteobacteria (1.000)
o:Aeromonadales (0.830)
f:Aeromonadaceae (0.689)
g:Aeromonas (0.572)

Raw Data ► Data Preparation ► Data Analysis

USEARCH

QIIME

R

Mothur

DADA2

...



RESEARCH ARTICLE

Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing

Andrei Prodan^{1*}, Valentina Tremaroli², Harald Brolin², Aeilko H. Zwinderman³, Max Nieuwdorp¹, Evgeni Levin^{1,4}

¹ Department of Experimental Vascular Medicine, Amsterdam University Medical Centers, Amsterdam, The Netherlands, ² Wallenberg Laboratory for Cardiovascular and Metabolic Research, Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, ³ Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam University Medical Centers, Amsterdam, The Netherlands, ⁴ Horaizon BV, Delft, the Netherlands

Table 1. Sensitivity and specificity over three mock sequencing runs. Values are reported as mean (standard deviation).

	Pipeline workflow	Exact	One-Off	Spurious
OTU-level				
QIIME-ucrust	QIIME-ucrust (default)	19 (0) ^a	134 (27)	412 (236)
	QIIME-ucrust (e30.ee1)	19 (0) ^a	133 (31)	341 (198)
	QIIME-ucrust (Q20)	19 (0) ^a	132 (26)	400 (232)
MOTHUR	MOTHUR (DGC.0)	19 (0)	none	48 (14)
	MOTHUR (DGC.1)	19 (0)	none	24 (8)
	MOTHUR (DGC.3)	19 (0)	none	5 (1)
	MOTHUR (Opticlust.3)	19 (0)	none	9 (4)
UPARSE	USEARCH-UPARSE	19 (0)	none	13 (7)
ASV-level				
DADA2	DADA2 (ee2)	21.7 (0.6) ^b	none	6 (4)
	DADA2 (no filter)	21.7 (0.6) ^b	none	5 (4)
Qiime2-Deblur	Qiime2-Deblur (default)	19 (0)	none	none
	Qiime2-Deblur (e30.ee1)	19 (0)	none	none
	Qiime2-Deblur (Q20)	19 (0)	none	none
UNOISE3	USEARCH-UNOISE3	21 (0) ^c	none	none

^a QIIME-ucrust erroneously produced separate OTUs for the two *C. beijerinckii* sequence variants, even though they have only 1 bp difference. It did not detect *P. acnes* in one of the three mock runs.
^b DADA2 did not find the lower copy number *C. beijerinckii* variant in one of the three mock runs.
^c USEARCH-UNOISE3 could not differentiate the two *C. beijerinckii* variants (13:1 copy number ratio).

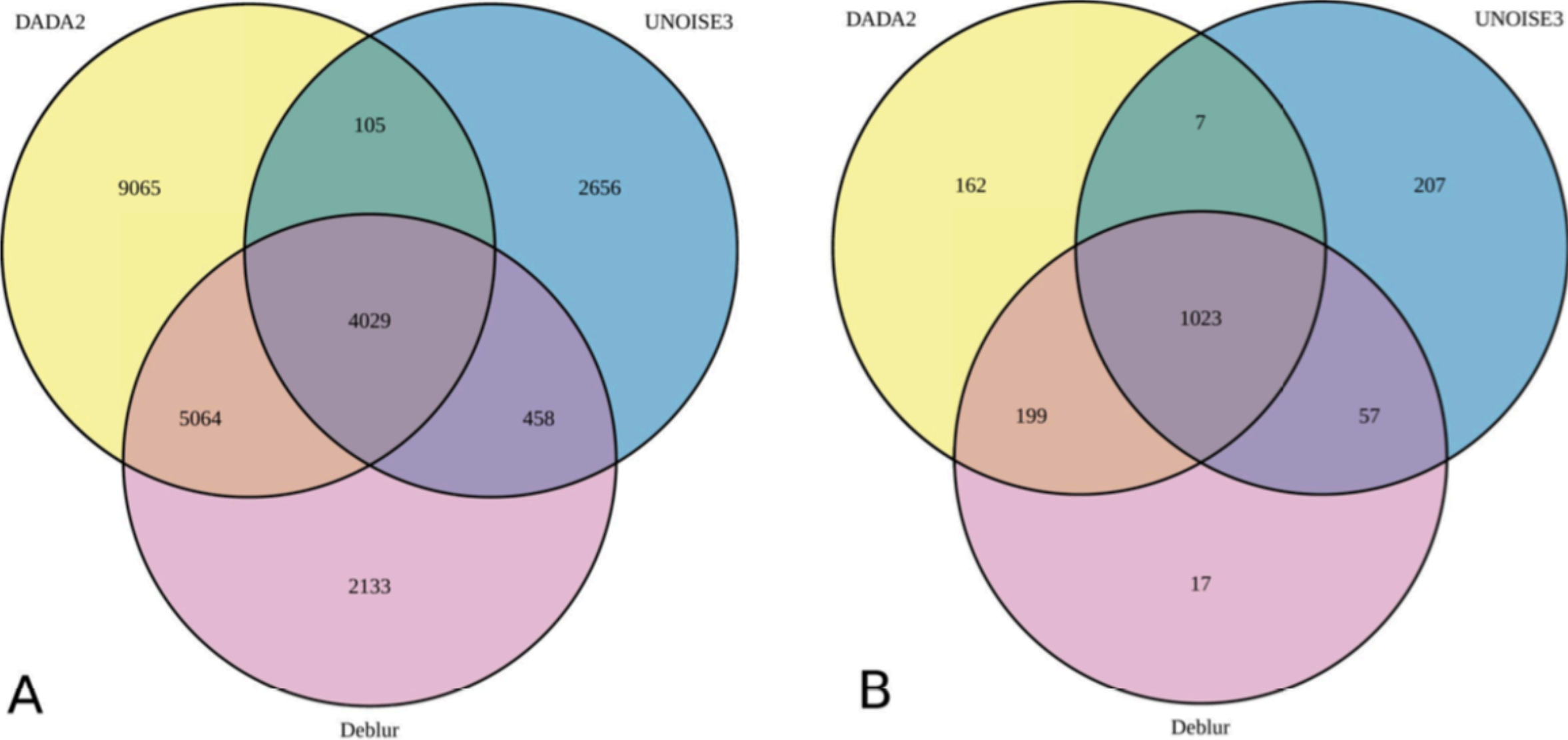


Fig 7. Venn diagram showing the overlap between the ASVs produced by three denoising pipelines from the HELIUS fecal sample data (N = 2170). Workflows shown are DADA2 (no filter), Qiime2-Deblur (e30.ee1), and USEARCH-UNOISE3. A) ASVs remaining after rarefaction to 10 000 counts. B) Filtered ASVs (mean relative abundance of at least 0.002% of rarefied counts).

Prodan et al (2020) Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS ONE 15(1): e0227434.

Conclusion

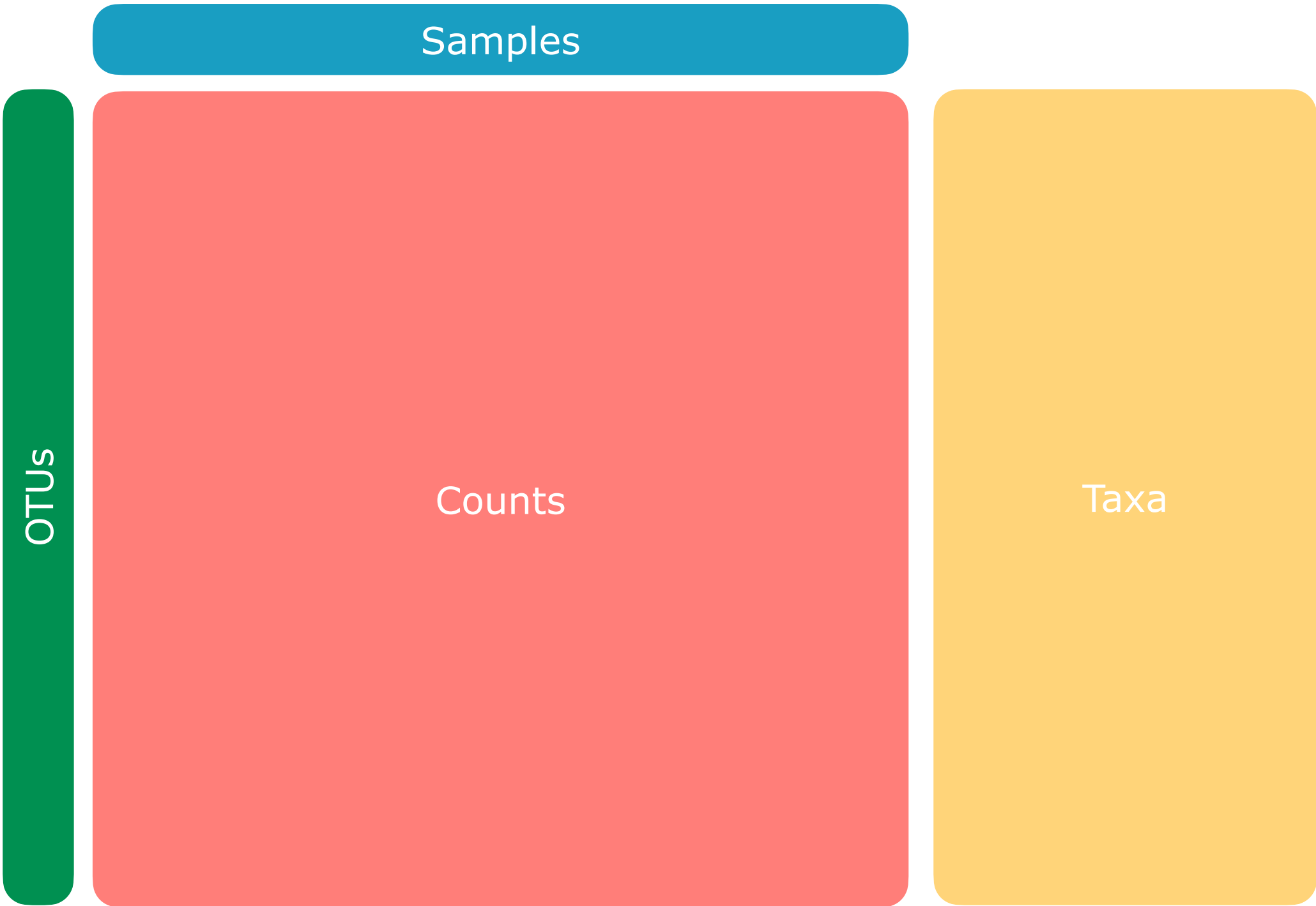
Large differences in sensitivity and specificity were observed between different pipelines. **DADA2 showed the best sensitivity and resolution (followed by USEARCH-UNOISE3) at the cost of producing higher number of spurious ASVs compared to USEARCH-UNOISE3 and Qiime2-Deblur.** USEARCH-UPARSE and MOTHUR produced similar numbers of OTUs, especially when a cutoff value was used in MOTHUR to remove singletons or extremely low abundance sequences before clustering. QIIME-uclust workflows produced huge numbers of spurious OTUs as well as inflated alpha-diversity measures, regardless of quality filtering parameters. Current QIIME users may consider switching to other pipelines. Indeed, the authors of QIIME have stopped supporting the platform since 1st January 2018 and are encouraging users to switch over to Qiime2. Biological conclusions based on alpha-diversity measures obtained from QIIME-uclust pipelines may warrant revisiting or confirmation other pipelines. ASV-level workflows offer superior resolution compared to OTU-level, and in this study showed better specificity and lower spurious sequence rates. Moreover, ASV-level pipelines allow for easier inter-study integration of biological features, as ASVs have intrinsic biological meaning, independent of reference database or study context.

We found DADA2 to be the best choice for studies requiring the highest possible biological resolution (e.g. studies focused on differentiating closely related strains). However, USEARCH-UNOISE3 showed arguably the best overall performance, combining high sensitivity with excellent specificity.

Prodan et al (2020) Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS ONE 15(1): e0227434.

**MORE
THINGS
CONSIDERED**

Data Analysis





OPEN ACCESS Freely available online



phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data

Paul J. McMurdie, Susan Holmes*

Abstract

Background: The analysis of microbial communities through DNA sequencing brings many challenges: the integration of different types of data with methods from ecology, genetics, phylogenetics, multivariate statistics, visualization and testing. With the increased breadth of experimental designs now being pursued, project-specific statistical analyses are often needed, and these analyses are often difficult (or impossible) for peer researchers to independently reproduce. The vast majority of the requisite tools for performing these analyses reproducibly are already implemented in R and its extensions (packages), but with limited support for high throughput microbiome census data.

Results: Here we describe a software project, phyloseq, dedicated to the object-oriented representation and analysis of microbiome census data in R. It supports importing data from a variety of common formats, as well as many analysis techniques. These include calibration, filtering, subsetting, agglomeration, multi-table comparisons, diversity analysis, parallelized Fast UniFrac, ordination methods, and production of publication-quality graphics; all in a manner that is easy to document, share, and modify. We show how to apply functions from other R packages to phyloseq-represented data, illustrating the availability of a large number of open source analysis techniques. We discuss the use of phyloseq with tools for reproducible research, a practice common in other fields but still rare in the analysis of highly parallel microbiome census data. We have made available all of the materials necessary to completely reproduce the analysis and figures included in this article, an example of best practices for reproducible research.

Conclusions: The phyloseq project for R is a new open-source software package, freely available on the web from both GitHub and Bioconductor.

$$\beta = \frac{\gamma}{\alpha}$$

α local diversity
mean diversity per treatment

β total diversity

γ between sample/treatment diversity

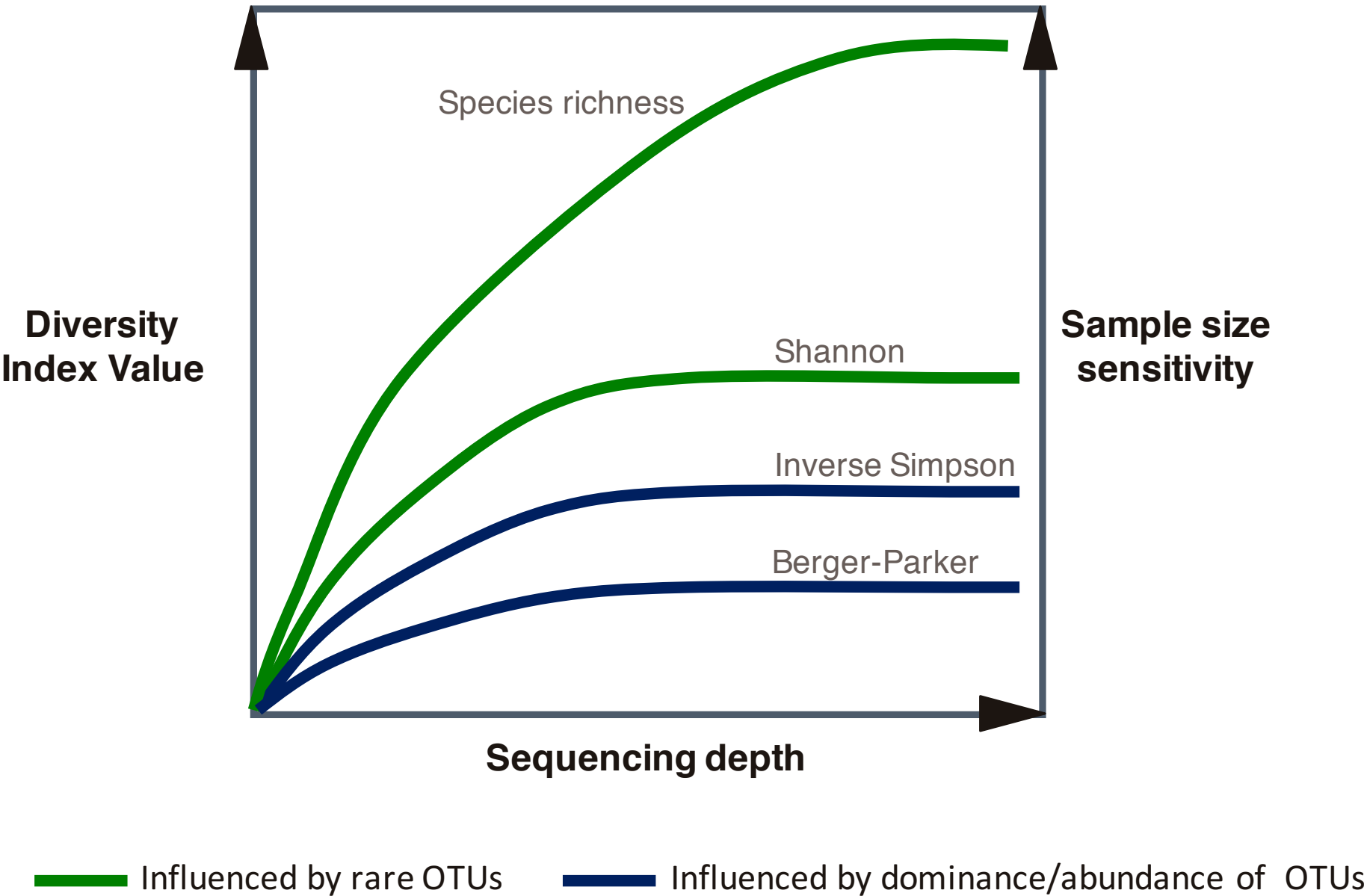
Minimum Differentiation



Maximum Differentiation



α -Diversity



β -Diversity

ECOLOGY LETTERS

Ecology Letters, (2011) 14: 19–28

doi: 10.1111/j.1461-0248.2010.01552.x

IDEA AND
PERSPECTIVENavigating the multiple meanings of β diversity: a roadmap for the practicing ecologist

Marti J. Anderson,^{1*} Thomas O. Crist,²
Jonathan M. Chase,³ Mark Vellend,⁴ Brian D.
Inouye,⁵ Amy L. Freestone,⁶ Nathan J. Sanders,⁷
Howard V. Cornell,⁸ Liza S. Comita,⁹ Kendi F.
Davies,¹⁰ Susan P. Harrison,⁸ Nathan J. B.
Kraft,¹¹ James C. Stegen¹² and Nathan G.
Swenson¹³

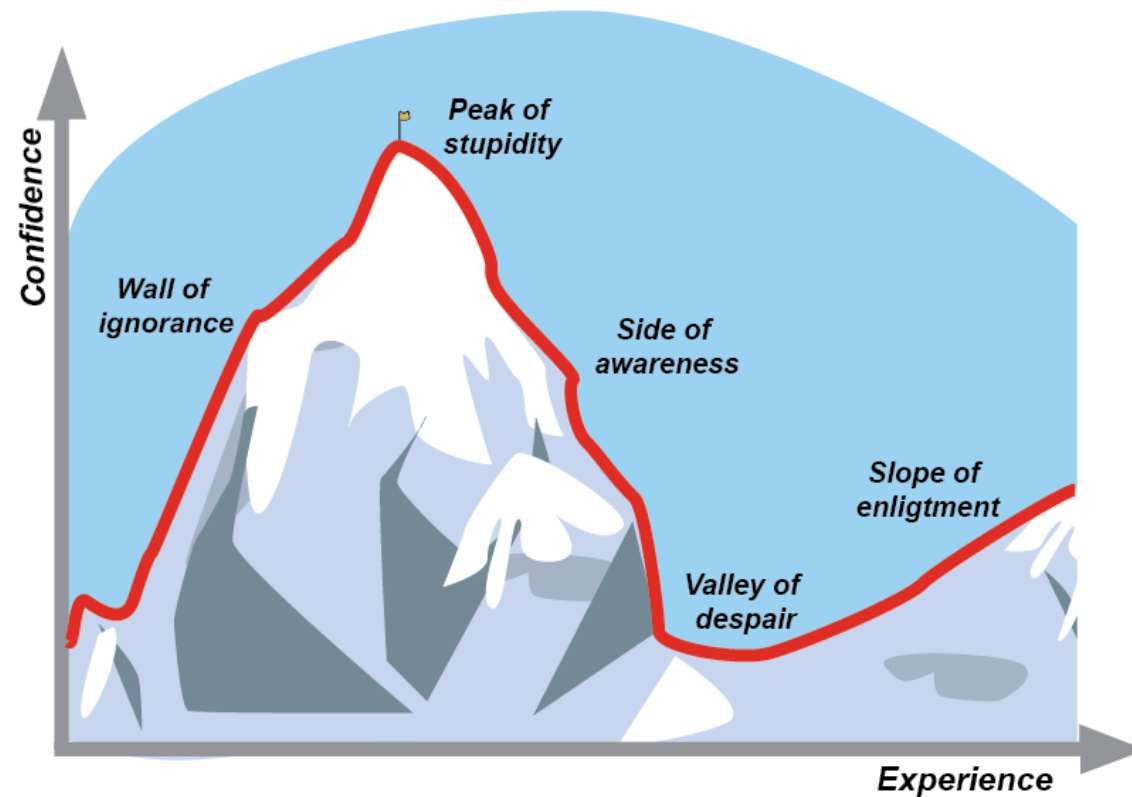
Abstract

A recent increase in studies of β diversity has yielded a confusing array of concepts, measures and methods. Here, we provide a roadmap of the most widely used and ecologically relevant approaches for analysis through a series of mission statements. We distinguish two types of β diversity: directional turnover along a gradient vs. non-directional variation. Different measures emphasize different properties of ecological data. Such properties include the degree of emphasis on presence/absence vs. relative abundance information and the inclusion vs. exclusion of joint absences. Judicious use of multiple measures in concert can uncover the underlying nature of patterns in β diversity for a given dataset. A case study of Indonesian coral assemblages shows the utility of a multi-faceted approach. We advocate careful consideration of relevant questions, matched by appropriate analyses. The rigorous application of null models will also help to reveal potential processes driving observed patterns in β diversity.

Food for thought

Dunning-Kruger Effect Curve

Dunning-Kruger Effect Curve



Suddenly everyone is a microbiota specialist

Boers et al (2016) Suddenly everyone is a microbiota specialist. *Clinical Microbiology and Infection* 22: 581-582.

Check List for Metabarcoding / AmpSeq Projects

- Caution is required when comparing microbiota studies (**meta-analysis**). Factors such as extraction protocols, polymerase, primer patch, etc.
- **Negative controls** should be included and analysed in the experimental protocol.
- Be aware of the potential for amplification **bias** in PCR amplification reactions.
- PCR and sequencing **artefacts** (e.g. chimera formation, false priming, index hopping, carry-overs)
- Accurate **taxonomic assignments** depend on the quality and completeness of reference databases and the method used.
- Limited **discriminatory power** of the amplified DNA region.
- **Multi-copy genes** (e.g. 16S) do not provide accurate information for quantification of bacterial species.
- Microbiota **profiling** is difficult due to lack of knowledge.
- DNA-based studies do not allow accurate differentiation between **viable, non-viable or dead bacterial cells**.
- **Small** (cohort) **size** studies with one or a few time points.
- Potential **conflicts of interest** between sponsors of microbiota research or beliefs.

Based on Boers et al. (2016) Suddenly everyone is a microbiota specialist. Volume 22, Issue 7, p581-582

Sample Diversity

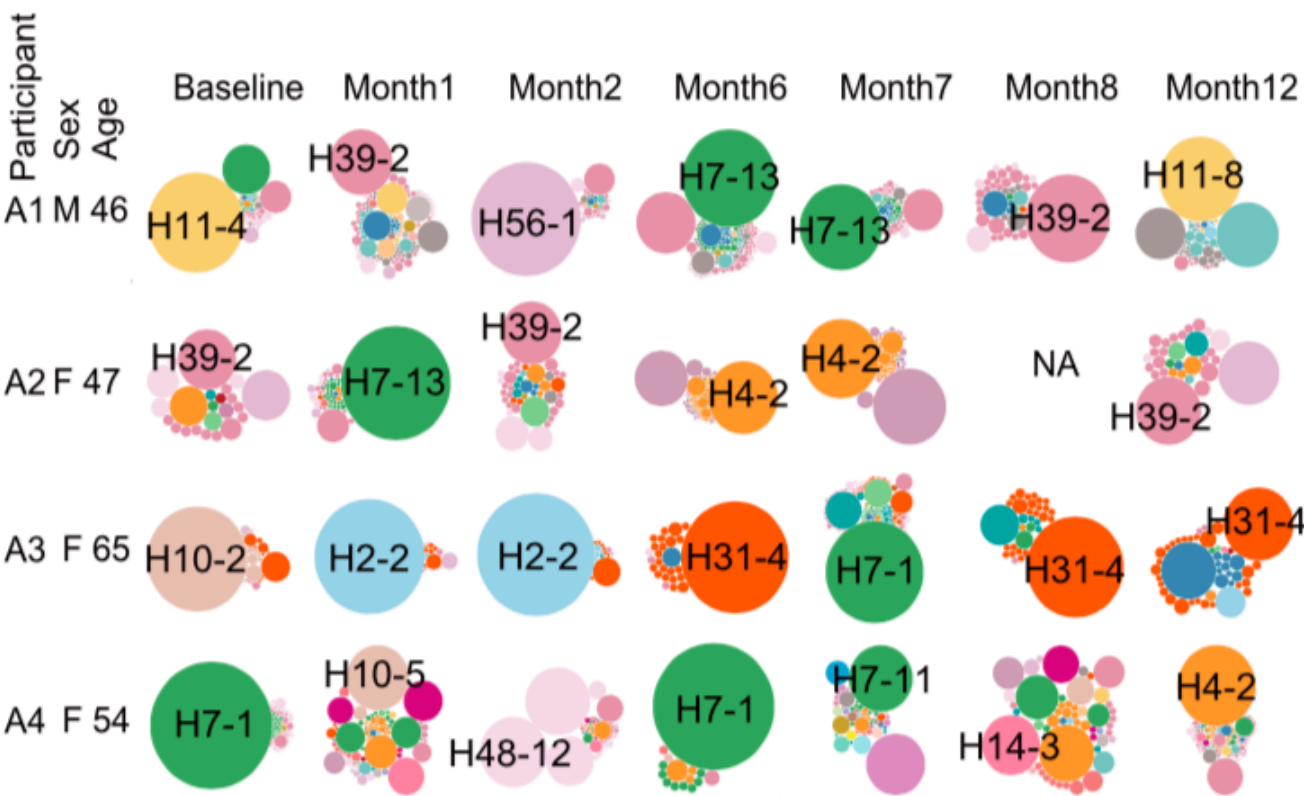
Cell Reports



Resource

Single-gene long-read sequencing
illuminates *Escherichia coli* strain
dynamics in the human intestinal microbiome

Dalong Hu,¹ Nicholas R. Fuller,^{2,3} Ian D. Caterson,^{1,2,3} Andrew J. Holmes,¹ and Peter R. Reeves^{1,4,*}

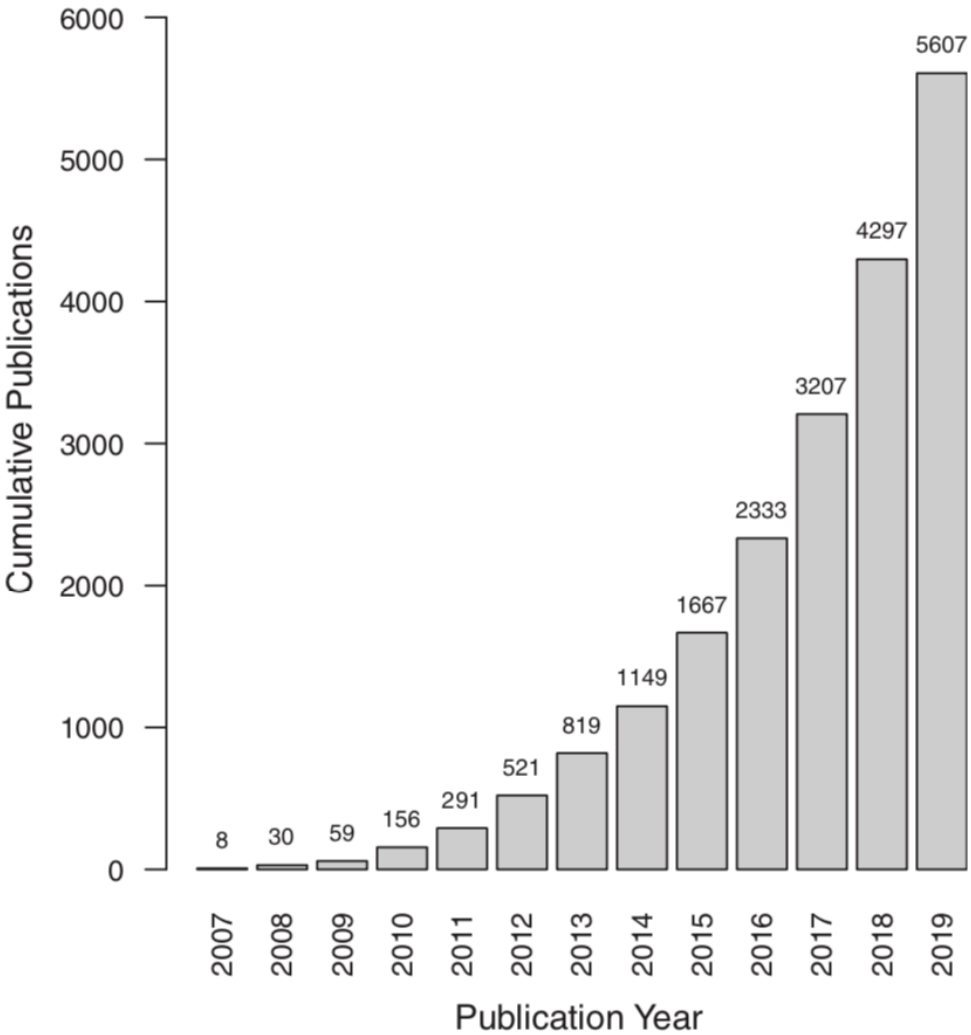


Sampling Design:
16 participants
7 time points

Flagellin diversity was discovered as strain differences in immune responses, which were codified and used for serotyping. The variable flagellin domains are known in serology as the H antigen, and 53 H antigens were distinguished in *E. coli*, of which 52 have the high-level divergence, with essentially no significant sequence alignment.

Core Microbiome

Core Microbiome



Source: Neu et al. (2021) Defining and quantifying the core microbiome: Challenges and prospects. PNAS

Core Microbiome

Metric type	Pros	Cons	Suggestions for improvement
Occurrence only	Computationally simple	No abundance information	Maximizing sequencing depth and replicate sampling
	Commonly used in the literature Includes rare taxa	Arbitrary cutoffs Can be heavily impacted by sampling coverage and sequencing depth	Using multiple occurrence cutoffs Using a range-through approach in conjunction with deeper sequencing
Relative abundance only	Computationally simple	Impacted by sequencing depth and inadequate spatial and temporal sampling coverage	Increasing geographic and temporal sampling, especially for widespread and/or lower abundance taxa
	Incorporates abundance information to identify taxa likely to be of functional importance	Arbitrary cutoffs Affected by rarefaction and related methods for sample standardization	Ensuring uniform sequencing depth across samples
Abundance–occurrence	Based on macroecological theory	Often use arbitrary cutoffs of abundance and occurrence	Need to better establish macroecological relationships for microbial taxa
	New methods (including code) are being developed in this space	Currently assumes that macroecological relationships in microbial taxa are similar to those in plants and animals	Constraining analyses by phylogenetic or functional groups
	Can potentially differentiate the stochastic from the deterministically selected core		Ensuring that scale of spatial and temporal sampling is adequate to reliably capture macroecological relationships

Source: Neu et al. (2021) Defining and quantifying the core microbiome: Challenges and prospects. PNAS

Core Microbiome

Determining the most effective way to quantify the core micro- biome remains challenging, with some arguing that a taxonomic approach is no longer useful and that a core functional micro- biome should be prioritized.

Recommendations to determine a core microbiome:

- Describe the criteria used for determining the core microbiome.
- Adequate sampling: spatial and temporal coverage.
- Provide adequate information about spatial (local, regional, and range-wide) and temporal context.
- Sequence as deeply as possible and ensure adequate number of sequencing replicates.
- Do not rarefy samples.

Source: Neu et al. (2021) Defining and quantifying the core microbiome: Challenges and prospects. PNAS



Metabarcoding is a powerful technique used to identify species within a mixed sample by amplifying and sequencing a specific genetic marker. The choice between short-read sequencing (e.g. Illumina) and long-read sequencing (e.g. PacBio or Oxford Nanopore Technologies, ONT) can significantly affect the results of metabarcoding studies.

Short-Read Sequencing (Illumina)

- + Higher accuracy
- + High throughput
- + Cost-effective
- + Mature bioinformatics tools
- Short read length
- Assembly challenges
- PCR bias

Long-Read Sequencing (PacBio, ONT)

- + Long read length
- + Complex region resolution
- + Reduced assembly complexity
- + Direct RNA sequencing
- Lower accuracy
- Lower throughput
- Higher cost
- Developing bioinformatics tools

Illumina (short reads): Ideal for high-throughput, cost-effective sequencing with high accuracy, but limited by short read lengths that can make assembly and resolution of complex regions difficult.

PacBio and ONT (long reads): Provide comprehensive insights with long reads, excellent for resolving complex regions and full-length barcodes, but have historically been less accurate and more expensive. Recent advances are closing the gap in accuracy and cost.

Choosing the right technology depends on the specific requirements of your metabarcoding study, including sample complexity, need for read length versus throughput, and budget constraints.

