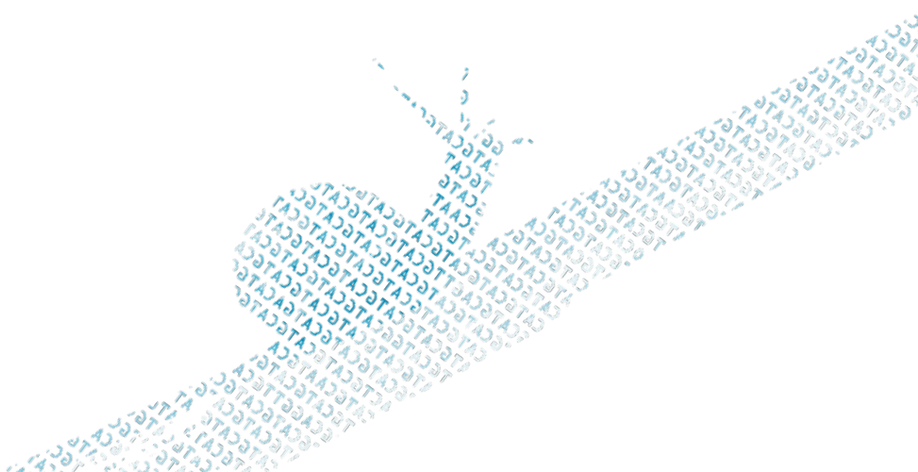
GDC Genetic Diversity Centre Zurich

Genetic Diversity: Analysis

Amp-Seq / Meta-Seq

Thursday, 1. July 2021



GDC Genetic Diversity Centre Zurich

Overview

DNA
eDNA
RNA

Amplicon Sequencing

Marker Gene Analysis

- PCR
- Sequencing
- OTU clustering / amplicon sequence variant
- Database comparison
- Taxonomic / phylogenetic classification

Metagenomics Metatranscriptomics

Binning

- Composition clustering / classification
- Database comparison
- Genome comparison

Assembly

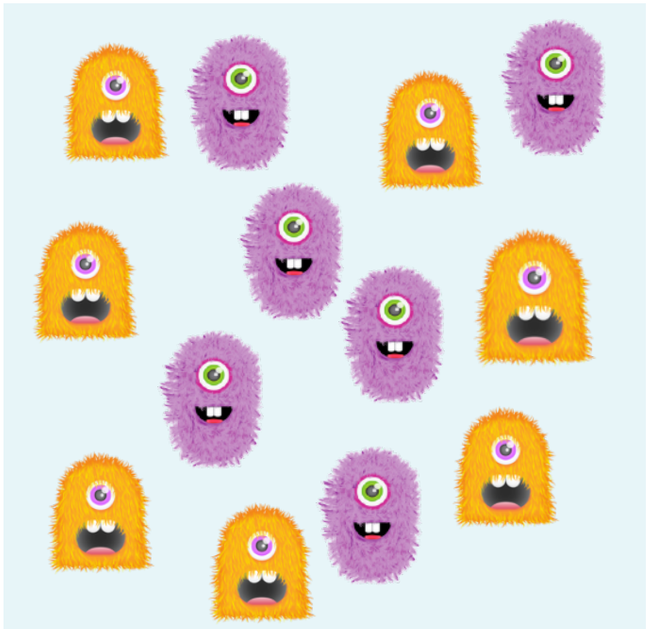
- Metagenome assembly
- Metatranscritpome assembly

! Please think! Cleaning and/or filtering your raw data might safe you some toubles.

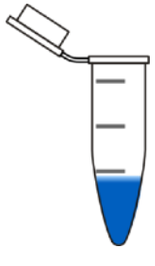
GDC Genetic Diversity Centre Zurich

Amplicon-Sequencing

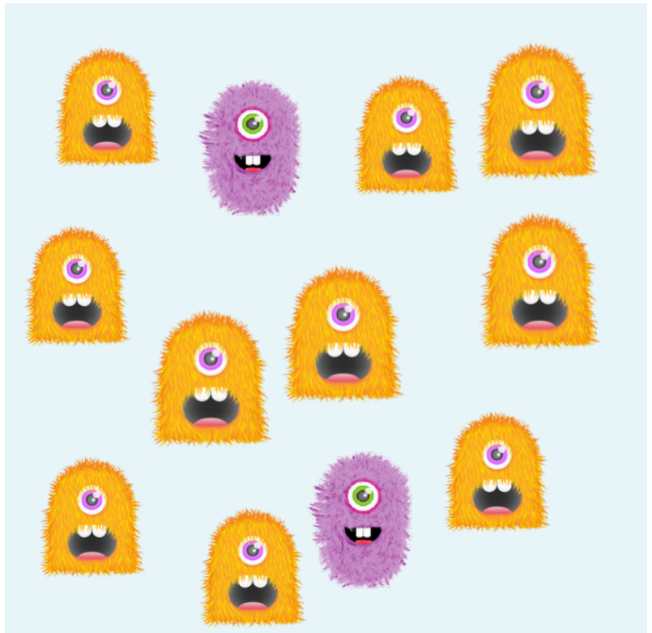
SampleA



DNA

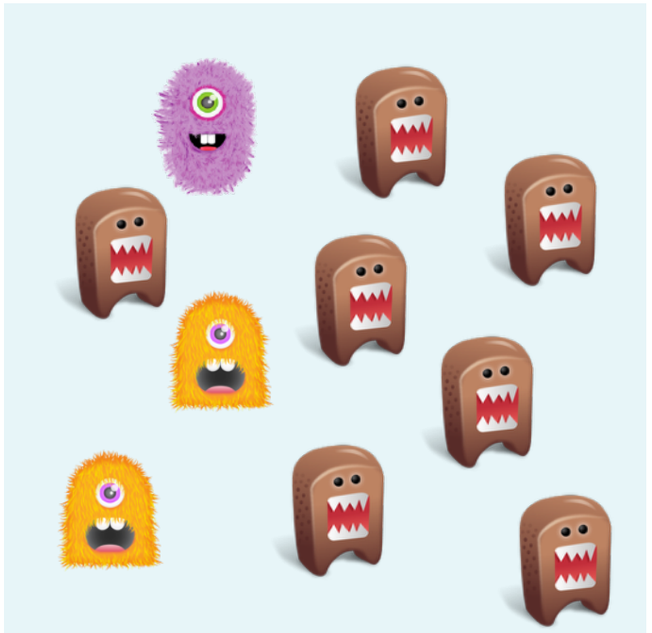


SampleB



DNA

SampleC



DNA

species composition

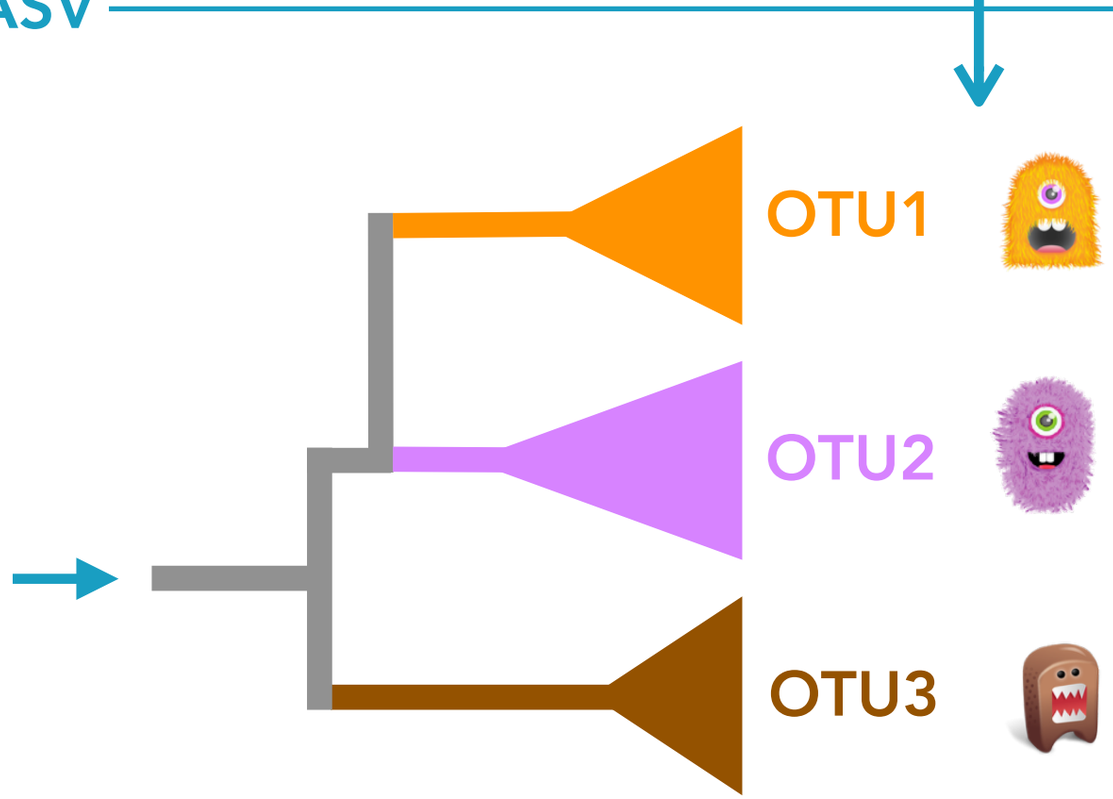
GDC Genetic Diversity Centre Zurich

targeted PCR

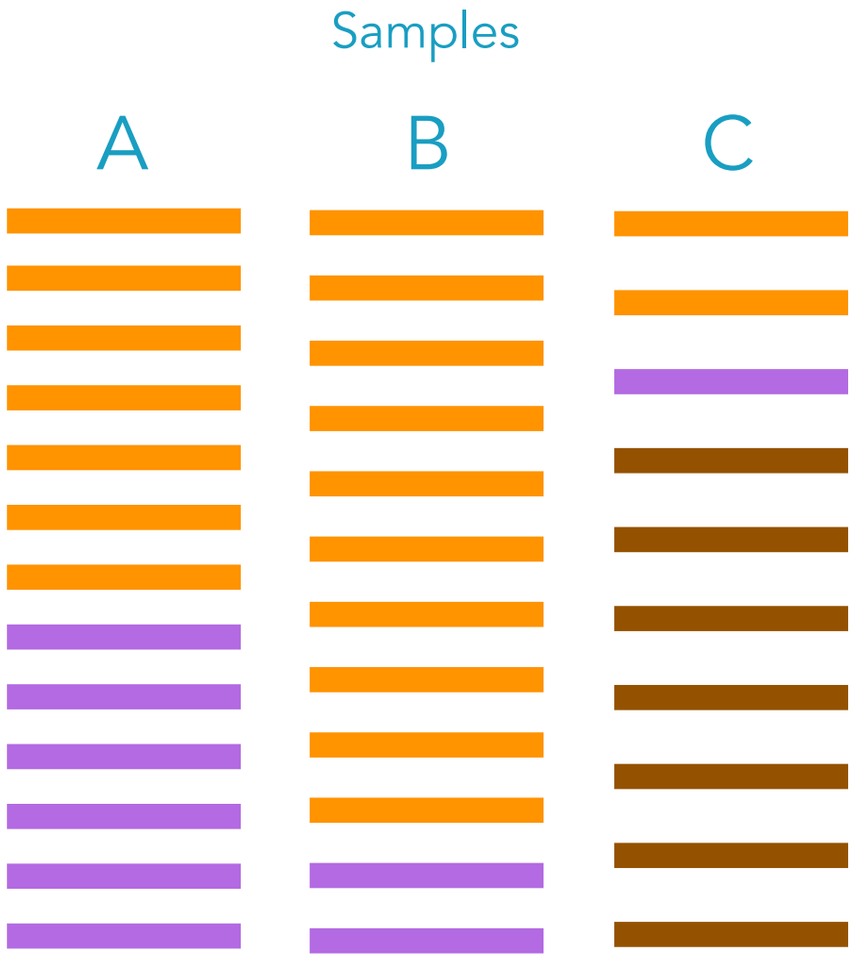
Classification

Reference

CLUSTERING / ASV



GDC Genetic Diversity Centre Zurich



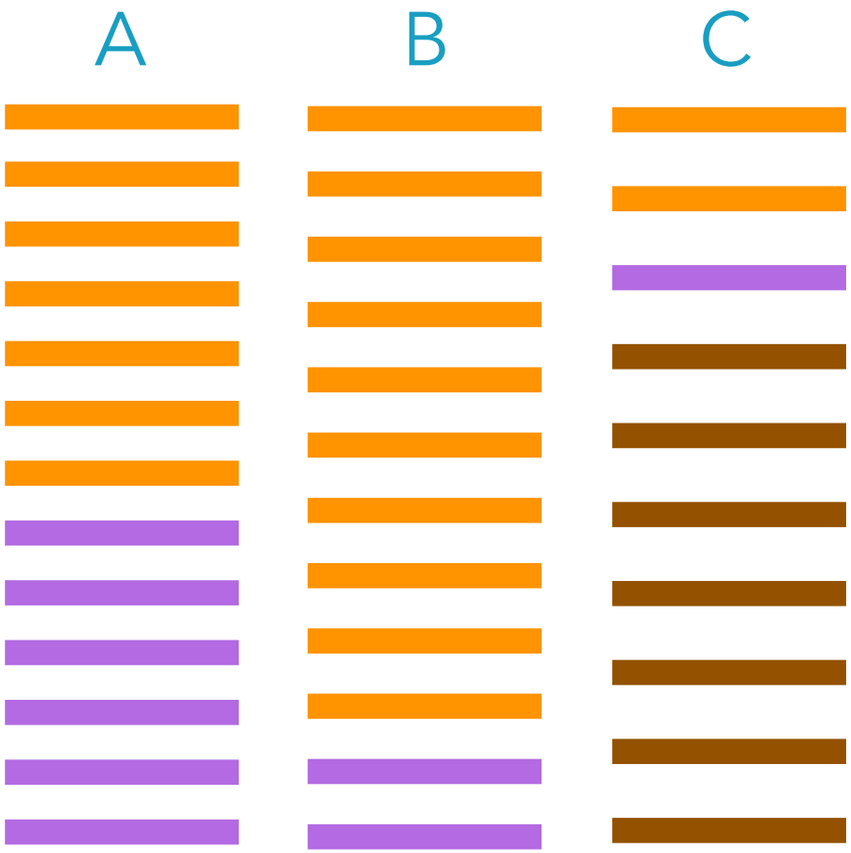
OTU/Count-Table

	A	B	C
OTU1	7	10	2
OTU2	6	2	1
OTU3	0	0	7
Total	13	12	10

Sequening Depth

GDC Genetic Diversity Centre Zurich

Samples

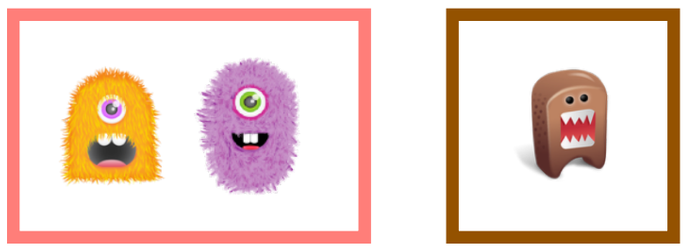


OTU/Count-Table

	A	B	C
OTU1	7	10	2
OTU2	6	2	1
OTU3	0	0	7
Total	13	12	10

Sequening Depth

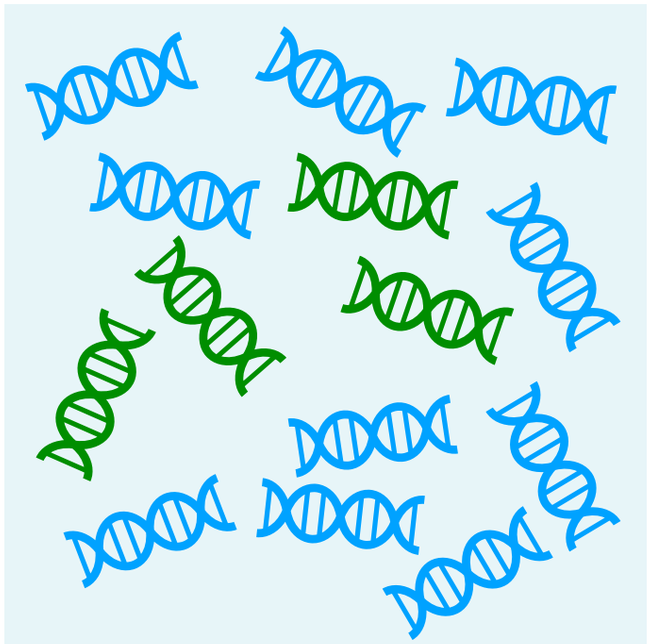
Associations



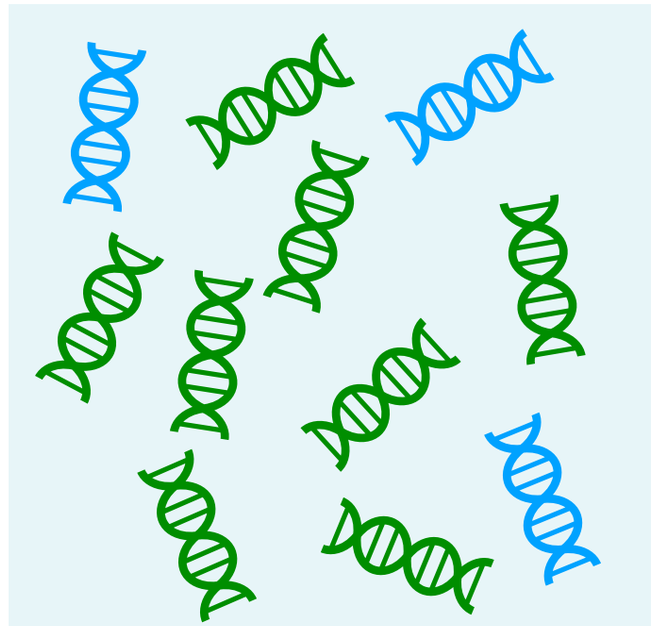
GDC Genetic Diversity Centre Zurich

Metagenome/Metatranscriptome-Sequencing

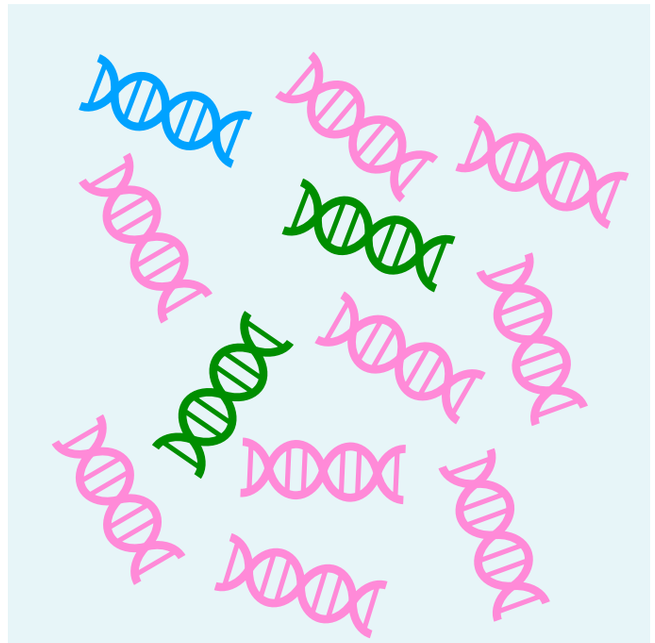
SampleA



SampleB



SampleC



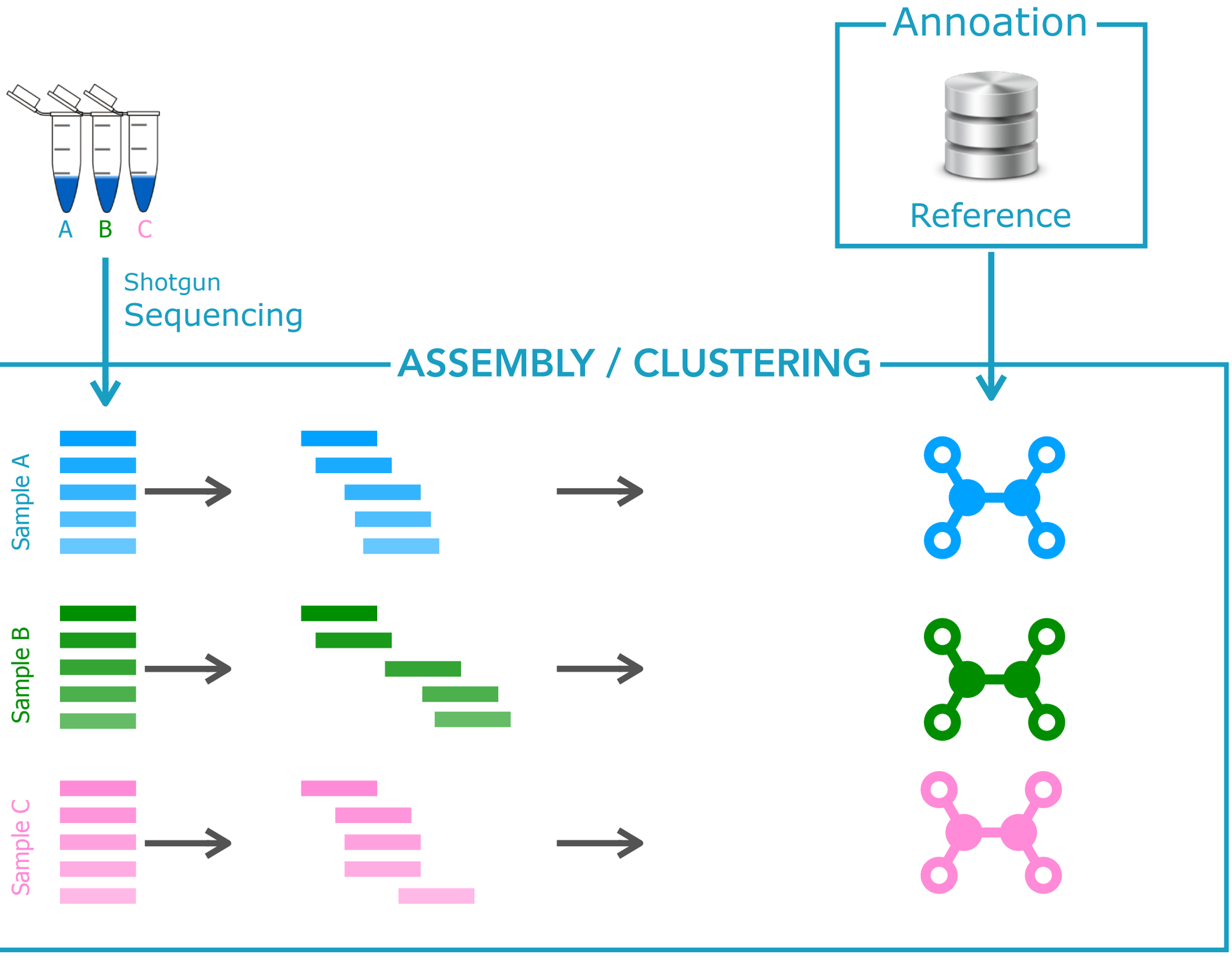
functional composition

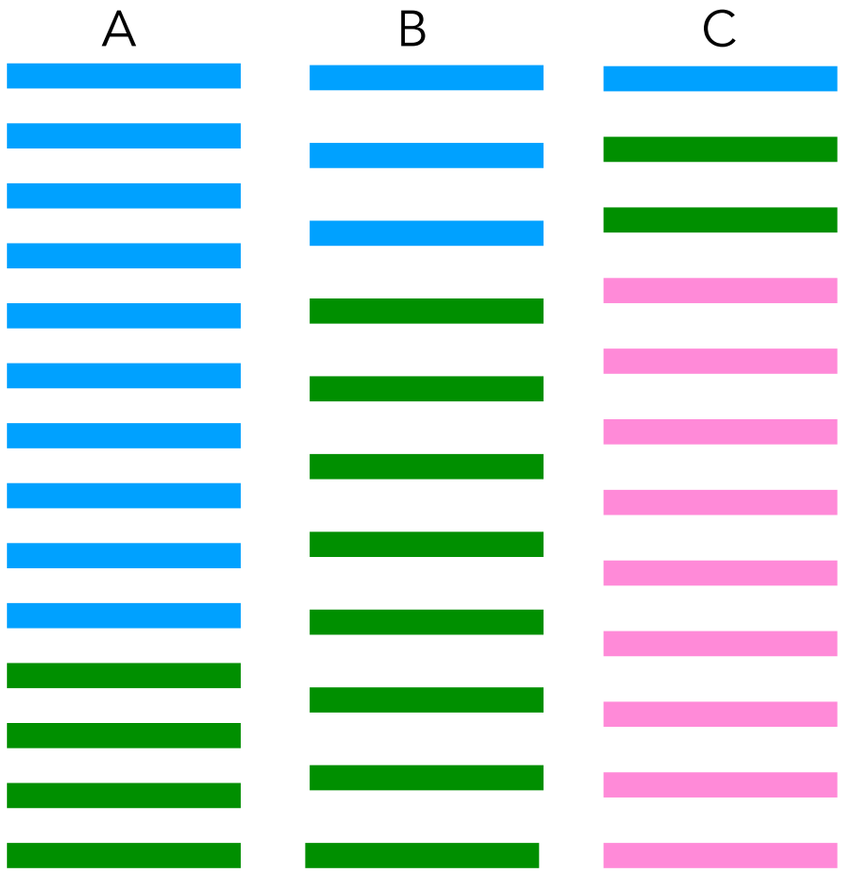
species composition

DNA
RNA

DNA
RNA

DNA
RNA


Annoation
Reference
A
B
C
Shotgun
Sequencing
ASSEMBLY / CLUSTERING
Sample A
Sample B
Sample C



	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

GDC Genetic Diversity Centre Zurich

Raw Data ▶ Data Preparation ▶ Data Analysis

USEARCH

QIIME

R

mothur

DADA2

PLOS ONE

RESEARCH ARTICLE

Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing

Andrei Prodan[1]*, Valentina Tremaroli[2], Harald Brolin[2], Aeilko H. Zwinderman[3], Max Nieuwdorp[1], 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-uclust	QIIME-uclust (default)	19 (0)[a]	134 (27)	412 (236)
	QIIME-uclust (e30.ee1)	19 (0)[a]	133 (31)	341 (198)
	QIIME-uclust (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-uclust 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).

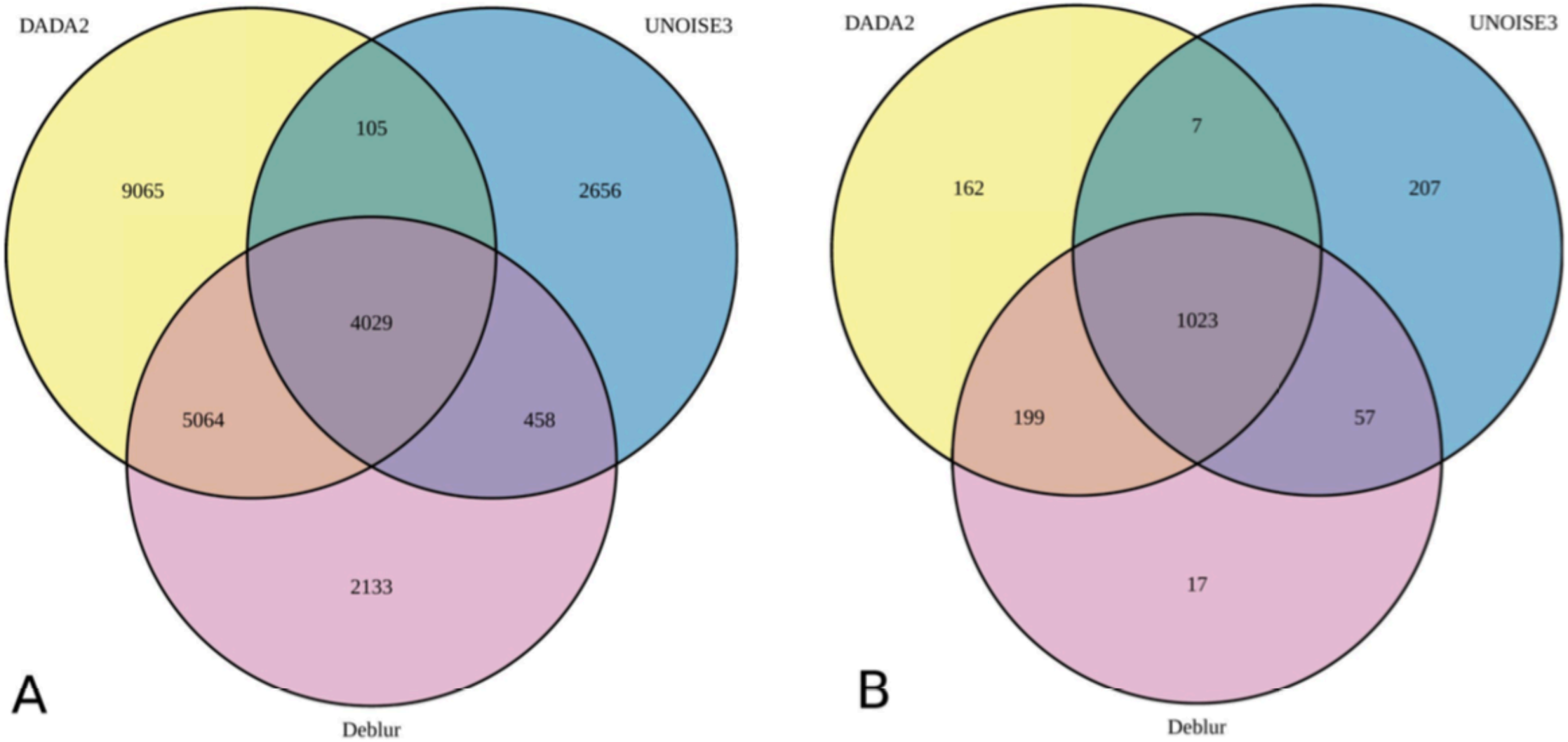


Fig 7. Venn diagram showing the overlap between the ASVs produced by three denoising pipelines from the HELIUS fecal sample data (N = 2170). Workflows shown are DADA2 (no filter), Qiime2-Deblur (e30.ee1), and USEARCH-UNOISE3. A) ASVs remaining after rarefaction to 10 000 counts. B) Filtered ASVs (mean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

PLOS ONE

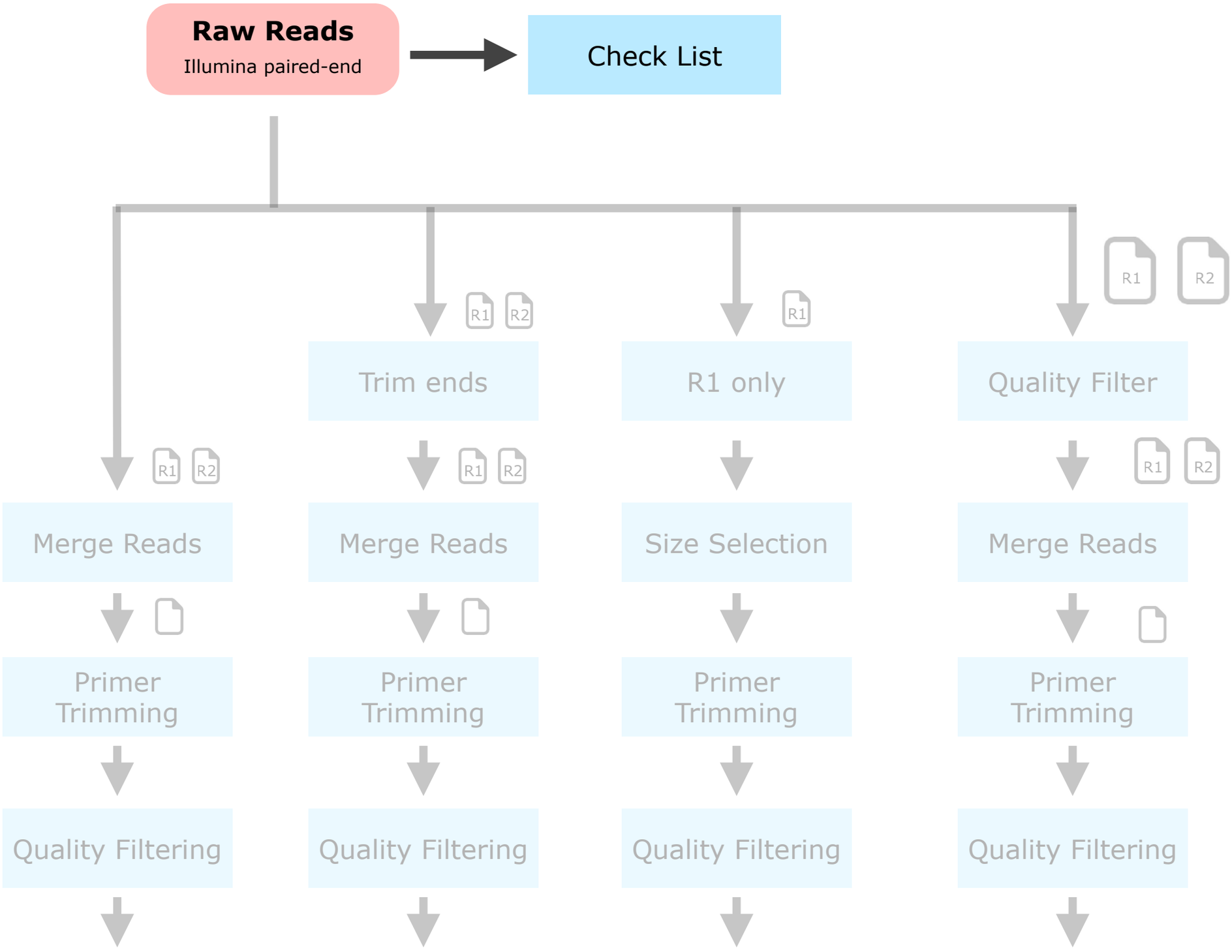
Conclusion

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 sensi- tivity with excellent specificity.

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

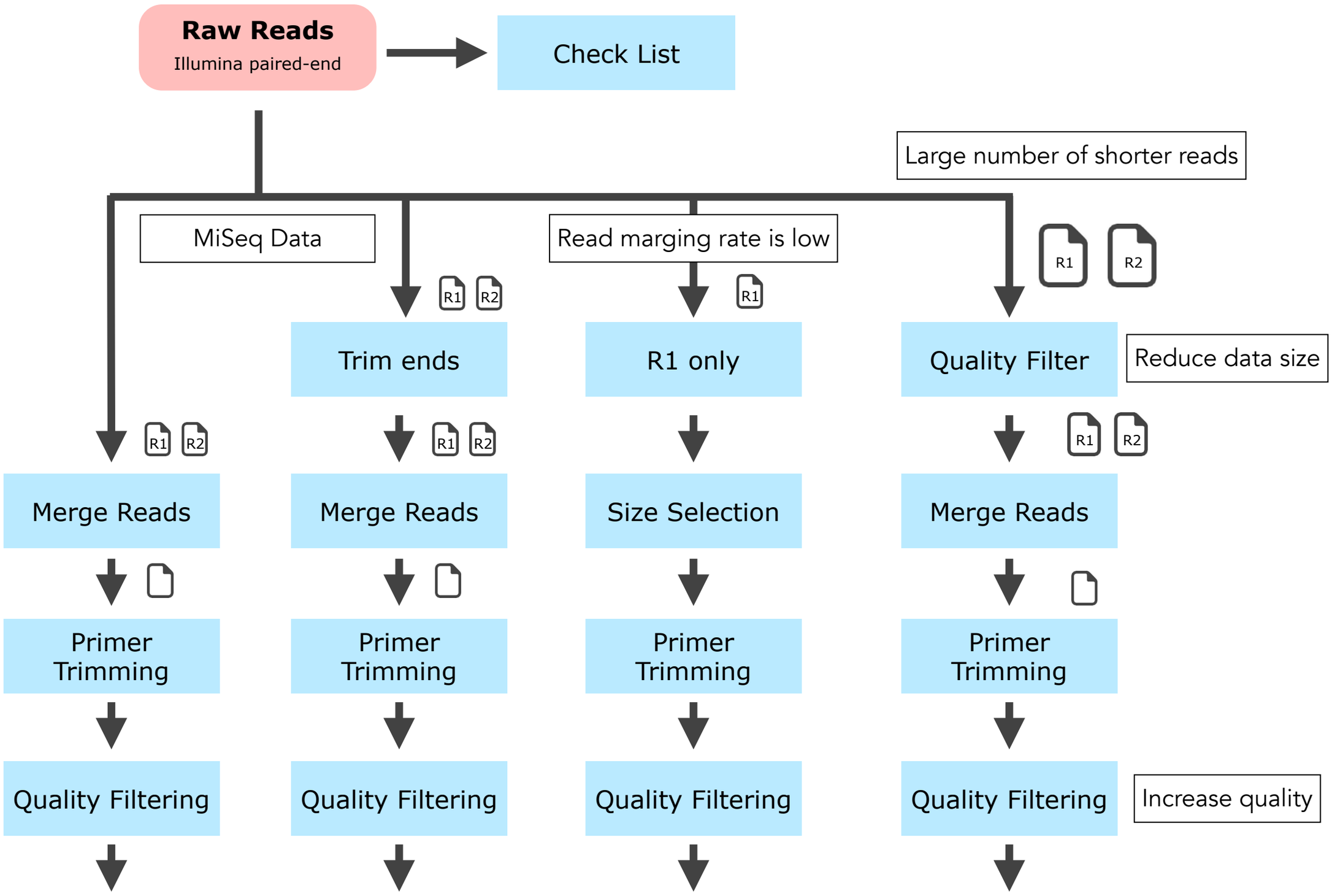
PLOS ONE


Raw Reads
Illumina paired-end
Check List
R1
R2
Trim ends
R1
R1 only
R1
R2
Quality Filter
R1
R2
Merge Reads
R1
R2
Merge Reads
Size Selection
R1
R2
Merge Reads
Primer Trimming
Primer Trimming
Primer Trimming
Primer Trimming
Quality Filtering
Quality Filtering
Quality Filtering
Quality Filtering

GDC Genetic Diversity Centre Zurich

Check-List

1. Download data (if possible via terminal e.g. sftp, wget)
2. Verify file integrity (md5sum)
3. Check data: $N_{samples} = N_{R1} = N_{R2}$
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)


Raw Reads
Illumina paired-end
Check List
Large number of shorter reads
MiSeq Data
Read marging rate is low
R1
R2
Trim ends
R1 only
Quality Filter
Reduce data size
Merge Reads
Merge Reads
Size Selection
Merge Reads
Primer Trimming
Primer Trimming
Primer Trimming
Primer Trimming
Quality Filtering
Quality Filtering
Quality Filtering
Quality Filtering
Increase quality

GDC Genetic Diversity Centre Zurich

Paired-end Read Merging:

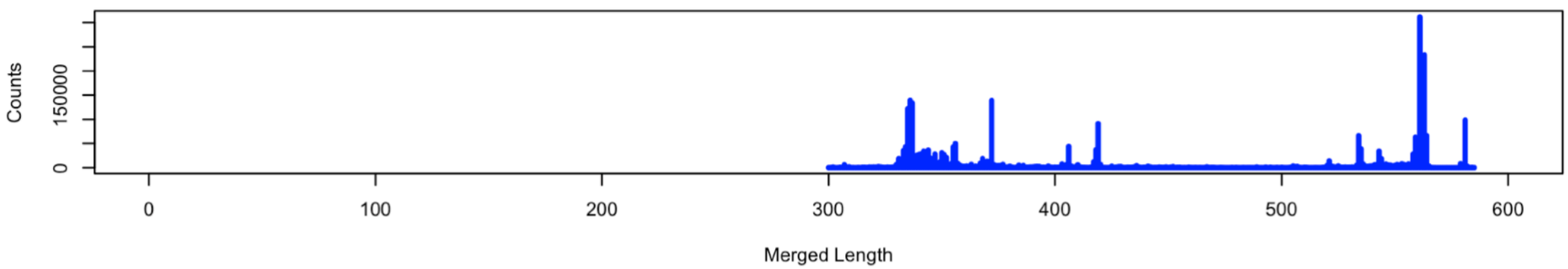
flash: ***2,854,051 (95.1%)***

bbmerge: 2,490,347 (83.0%)

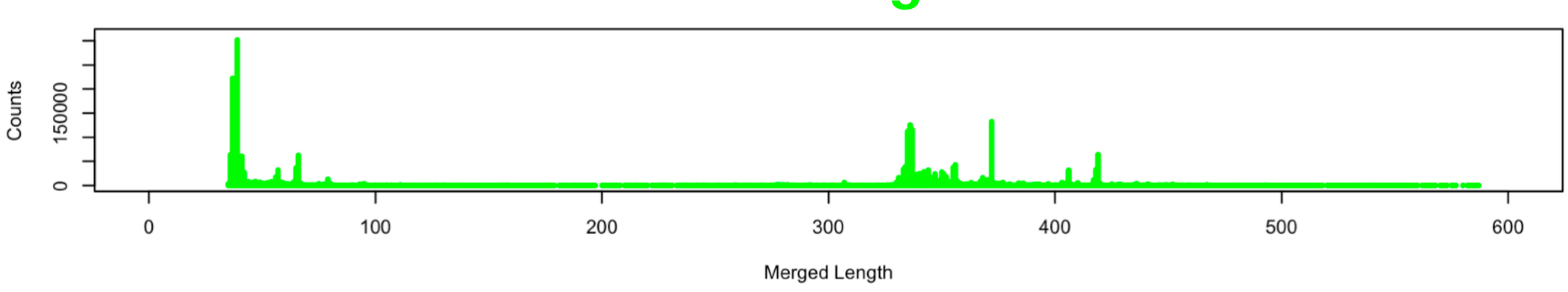
usearch: 1,857,602 (61.9%)

GDC Genetic Diversity Centre Zurich

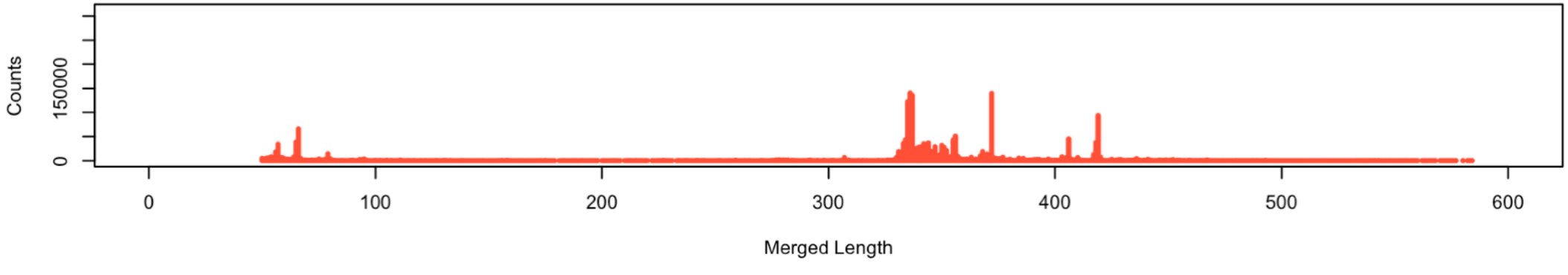
flash

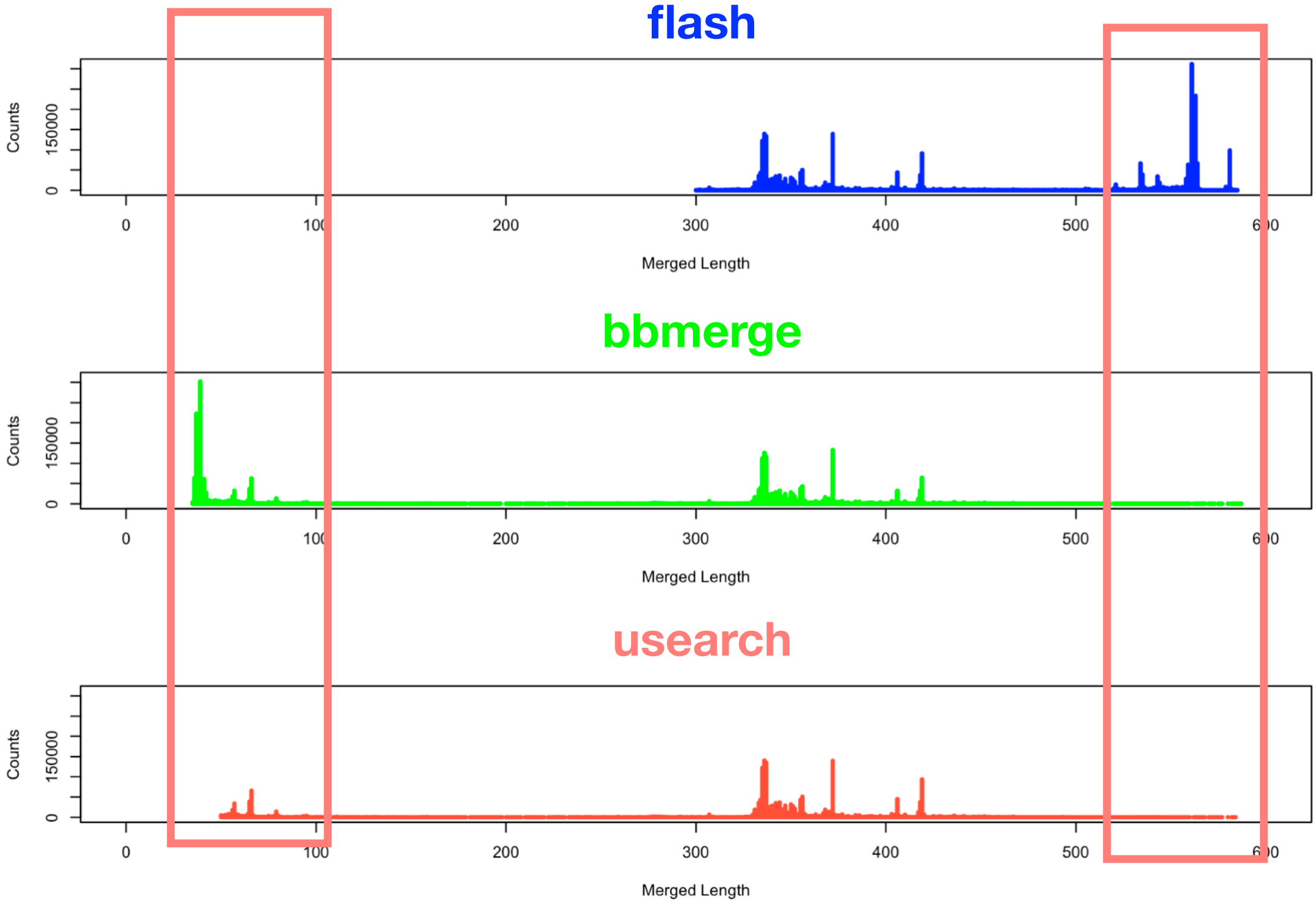


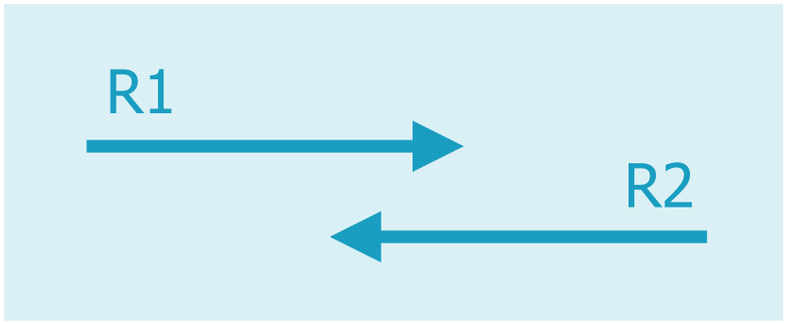
bbmerge



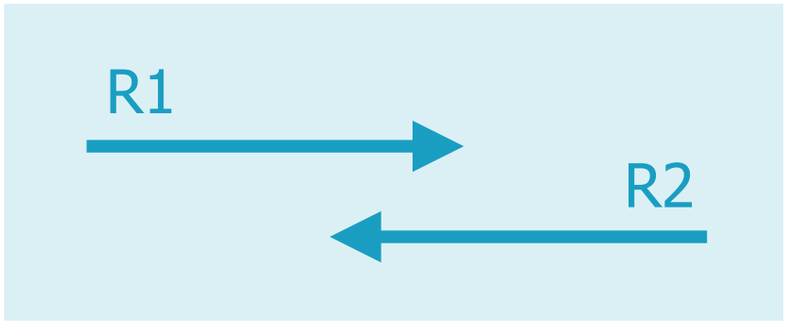
usearch

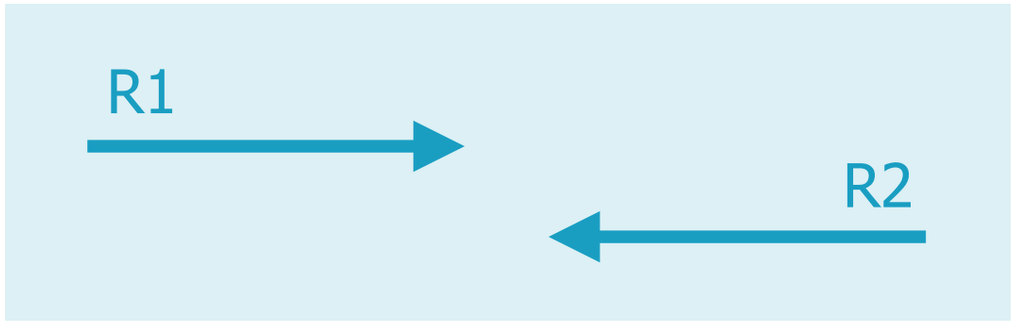


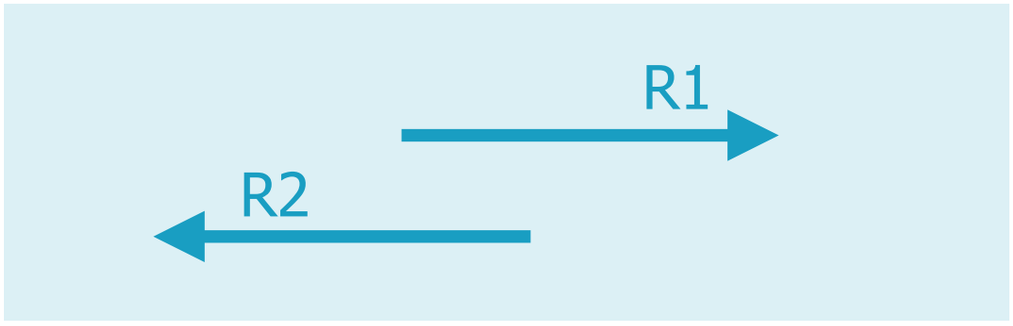

flash
Counts
0
150000
0
100
200
300
400
500
600
Merged Length
bbmerge
Counts
0
150000
0
100
200
300
400
500
600
Merged Length
usearch
Counts
0
150000
0
100
200
300
400
500
600
Merged Length


R1
R2

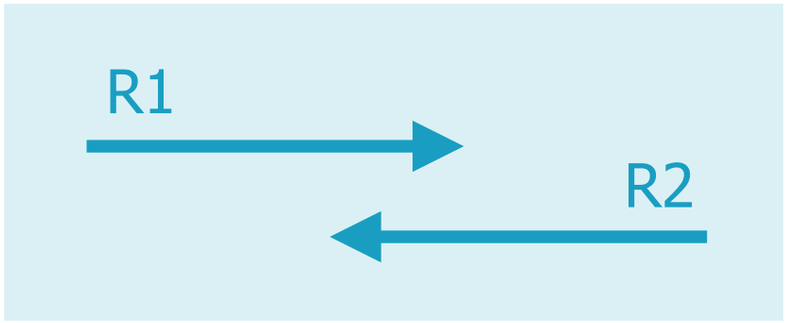
Zurich Centre Diversity Genetic

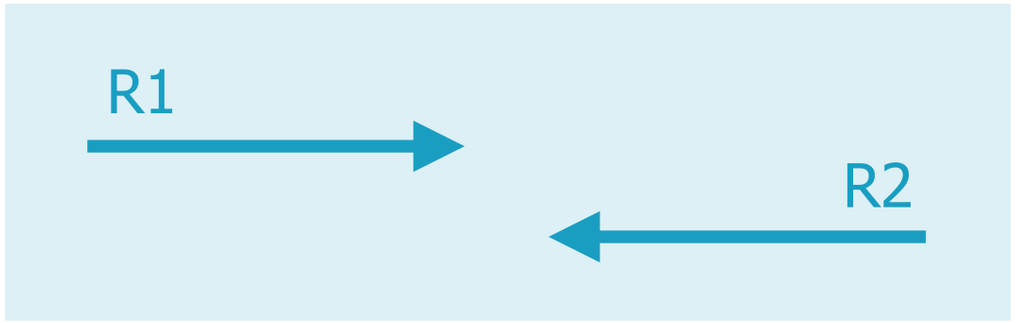




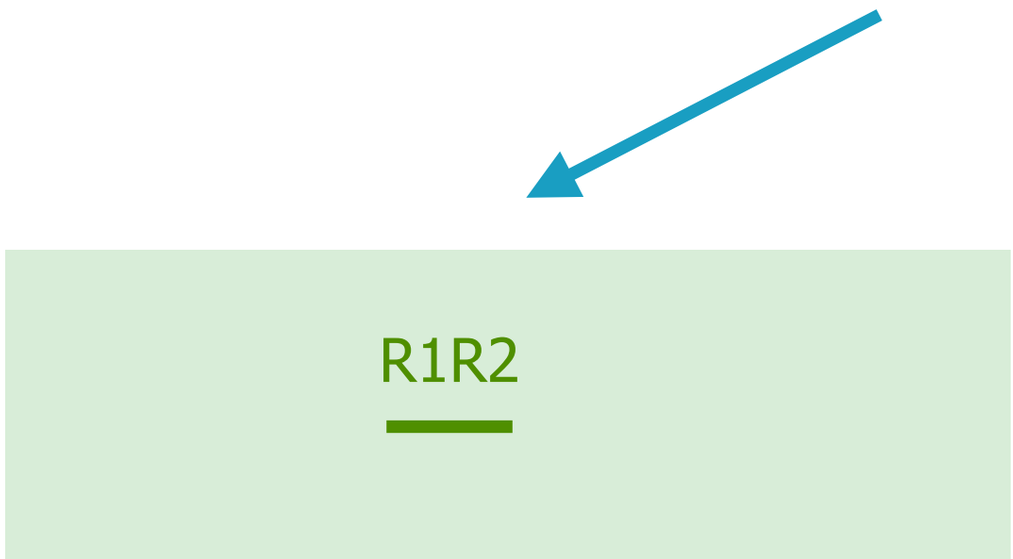


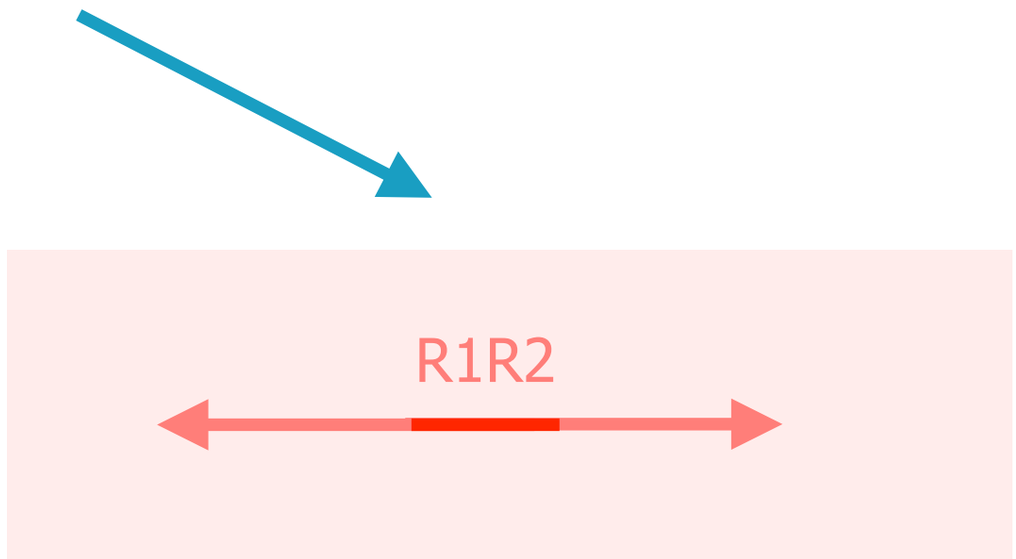
Staggered Reads


R1
R2


R1
R2

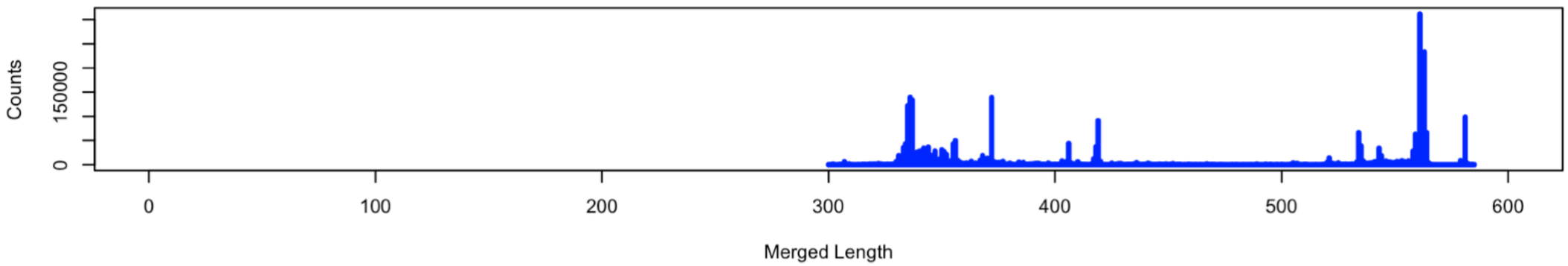

R1
R2


R1R2

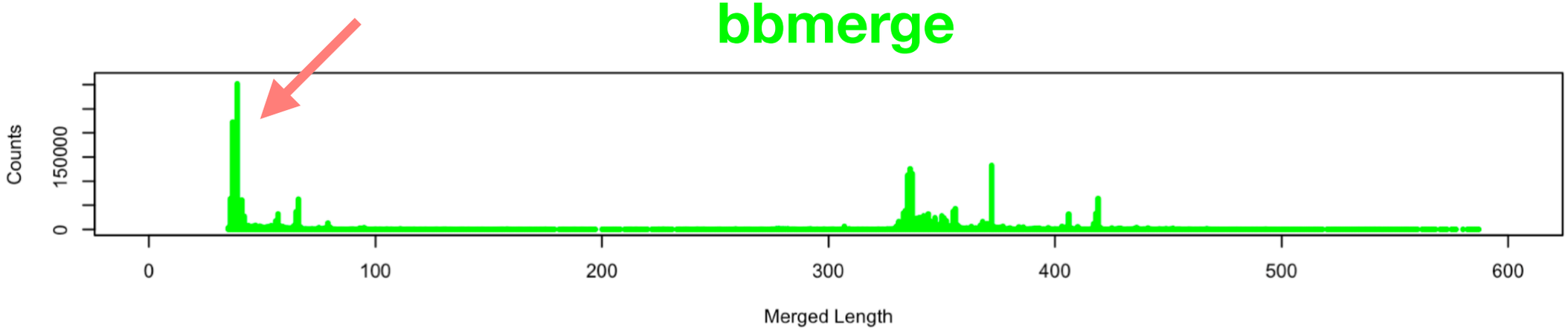

R1R2

GDC Genetic Diversity Centre Zurich

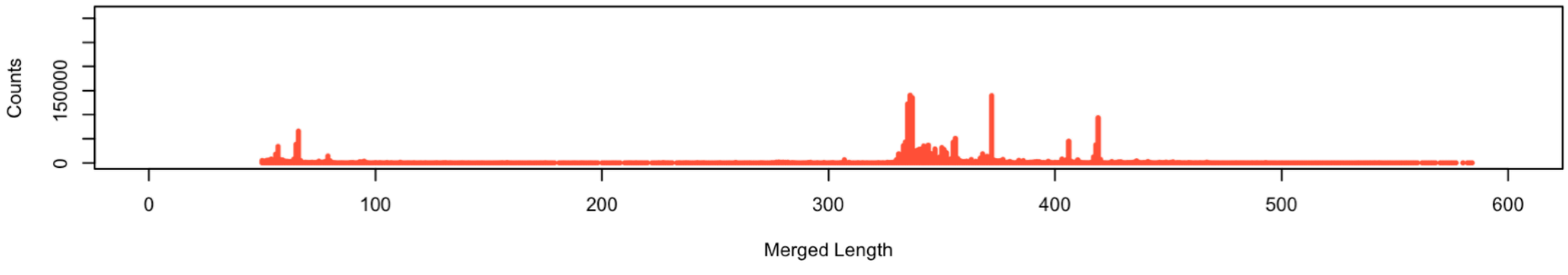
flash



bbmerge

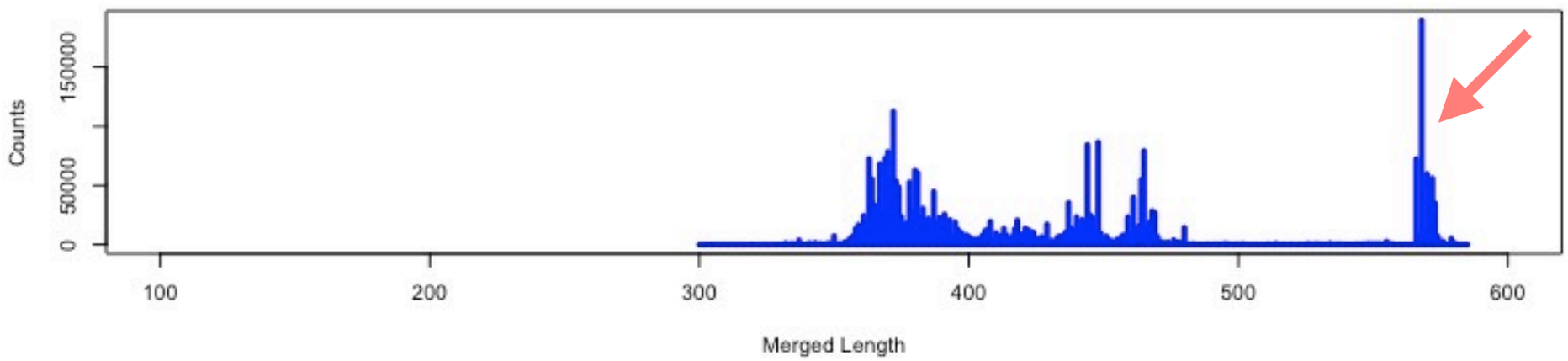


usearch

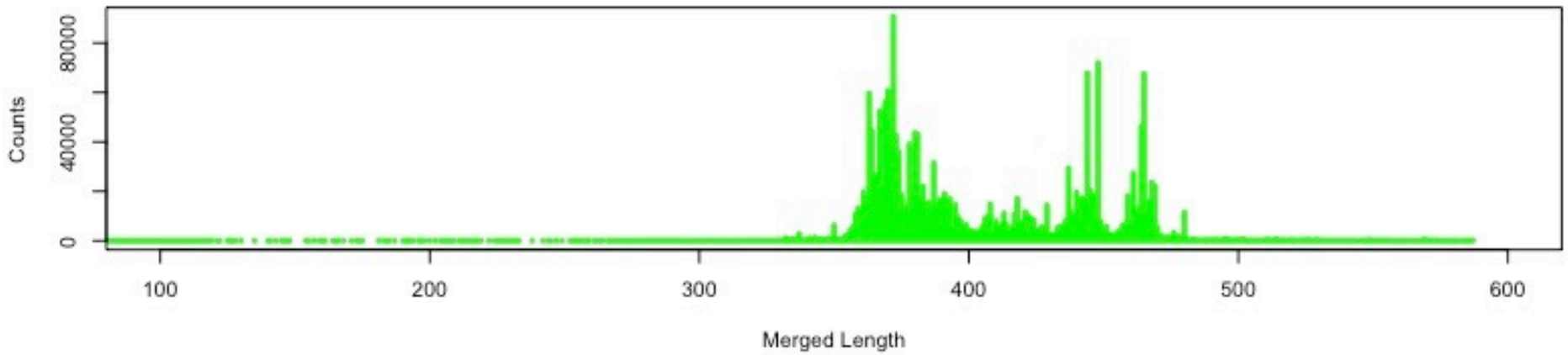


GDC Genetic Diversity Centre Zurich

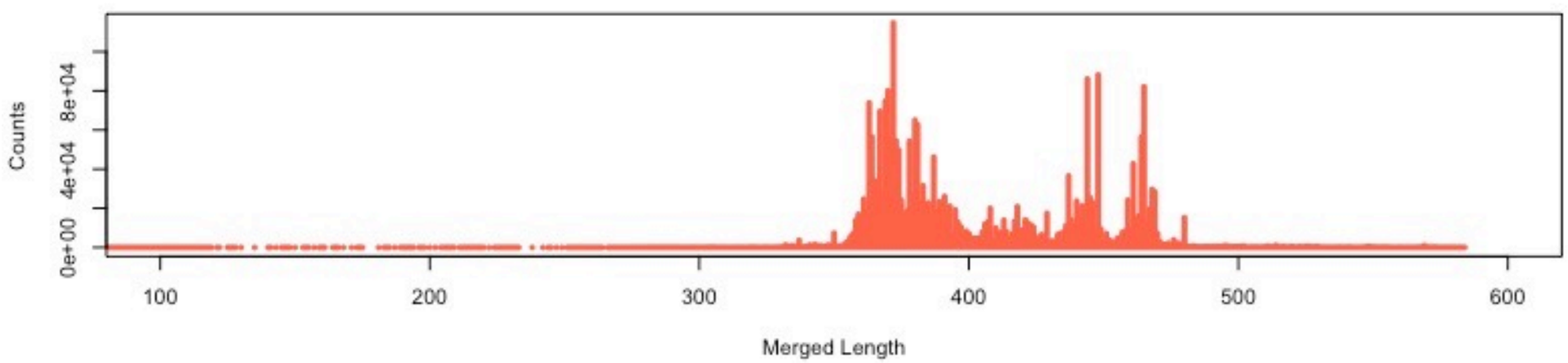
flash



bbmerge



usearch

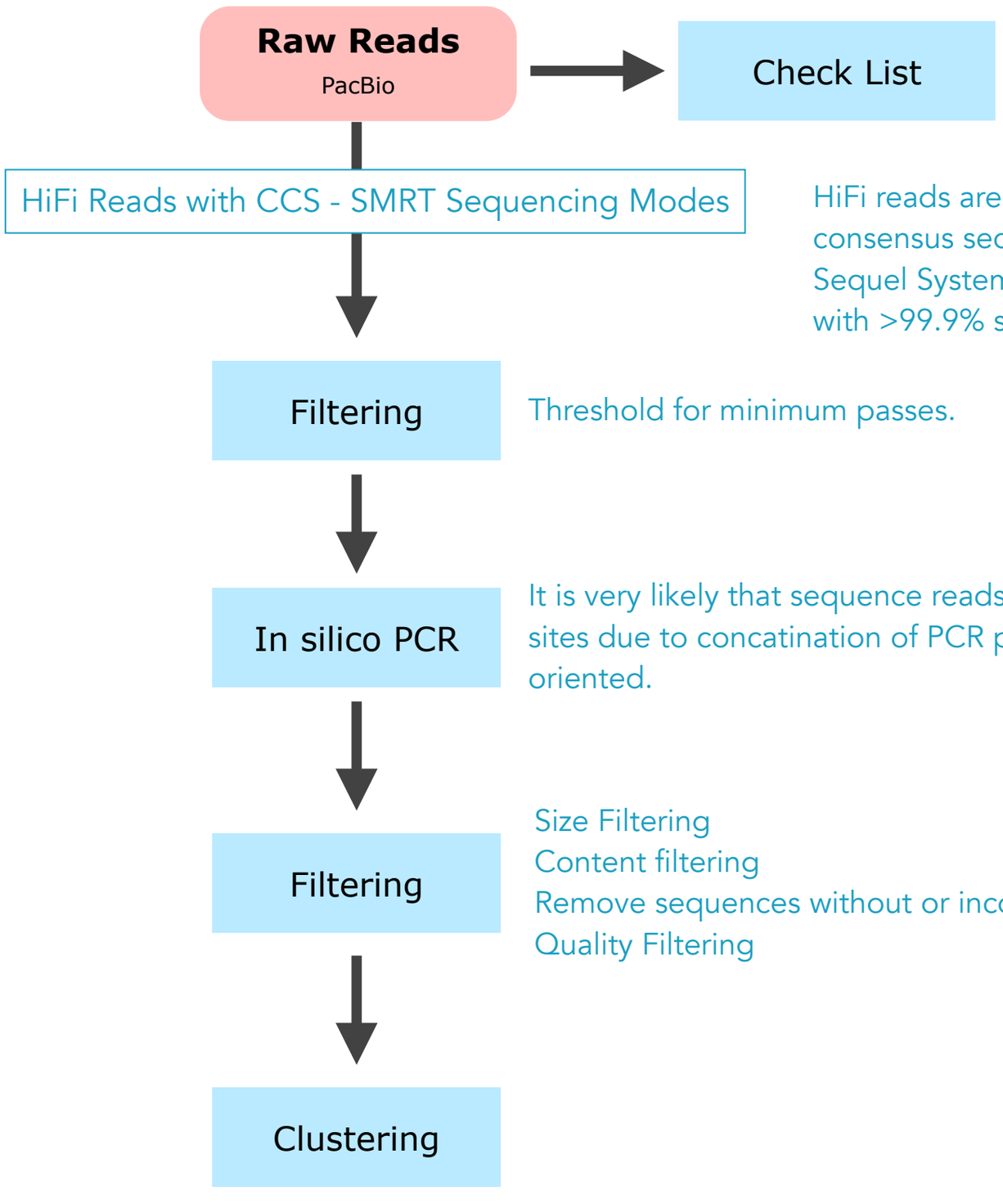


GDC Genetic Diversity Centre Zurich

- Aa-007 - Merging Summary

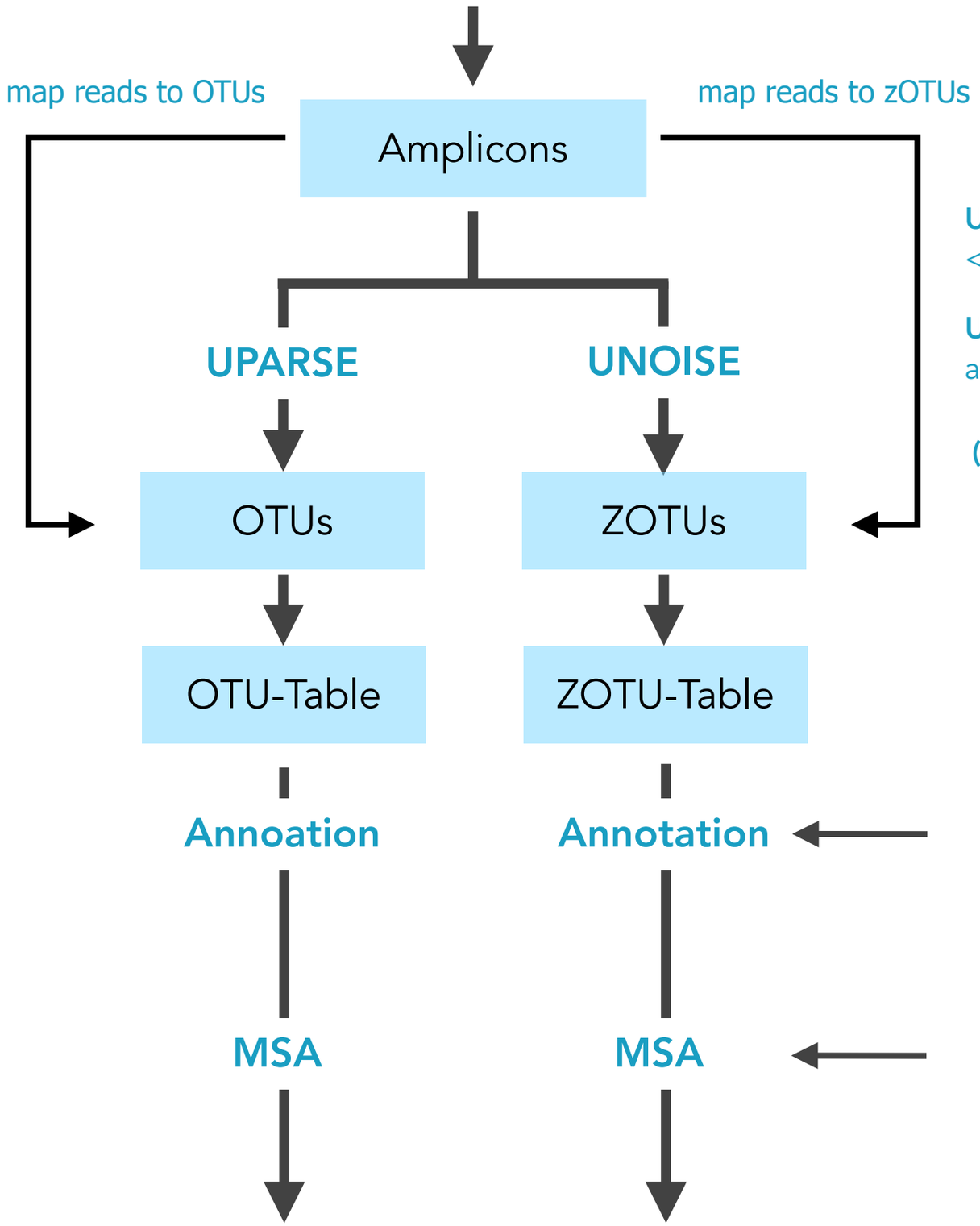
```
  Merging Rate: 116101 / 122531 (94.8%)
  Median Merged Length: 465

  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
```



HiFi reads are a type of data produced using the circular consensus sequencing (CCS) mode on one of the PacBio Sequel Systems. HiFi reads provide base-level resolution with >99.9% single-molecule read accuracy.

GDA ▷ Amp-Seq / MG-Seq / MT-Seq / (SG-Seq)



UPARSE: Clustering with 97% identiy and min 2 <default> abundance threshold.

UNOISE: Zero-Radius clustering with error correction and min 8 <default> abundance threshold.

(Z)OTU: (Zero-Radius) Operational Taxonomic Units

Annotation with SINTAX

Possible References

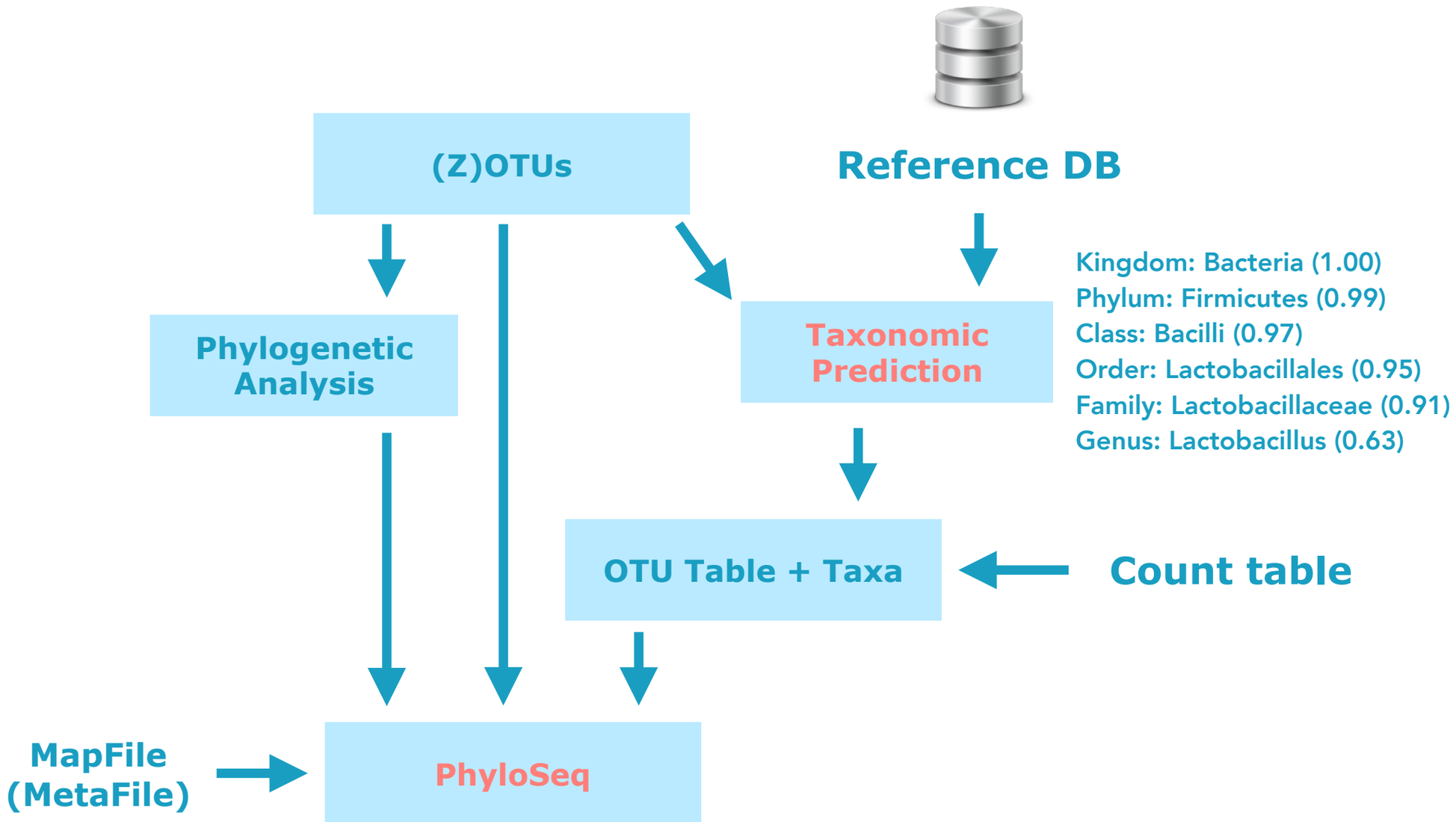
- NCBI
- SILVA SSU & LSU
- Ribosomal Database Project (RDP)
- GreenGenes
- EzBioCloud
- MIDORI
- UNITE ITS

Multiple Sequence Alignment

Count tables with taxonomic predictions

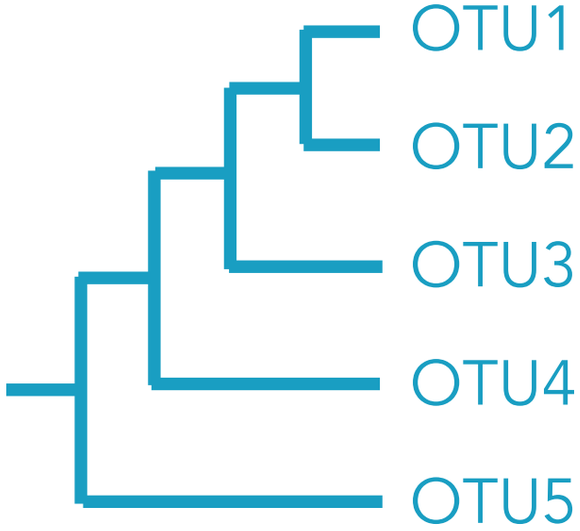
GDC Genetic Diversity Centre Zurich

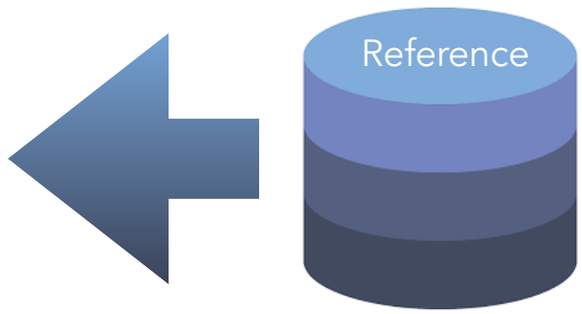
OTU - Annoation (-Prediction)



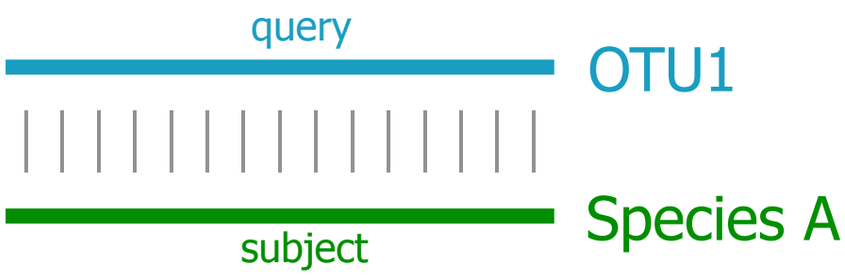
Genetic Diversity Centre Zurich

OTU - Annoation (-Prediction)

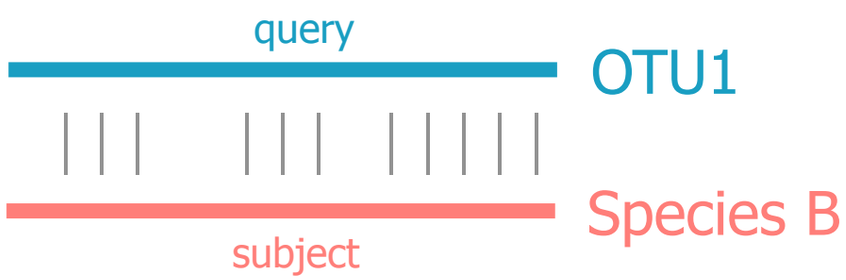




- NCBI 16S
- SILVA SSU & LSU
- Ribosomal Database Project RDP
- GreenGenes
- EzBioCloud
- …

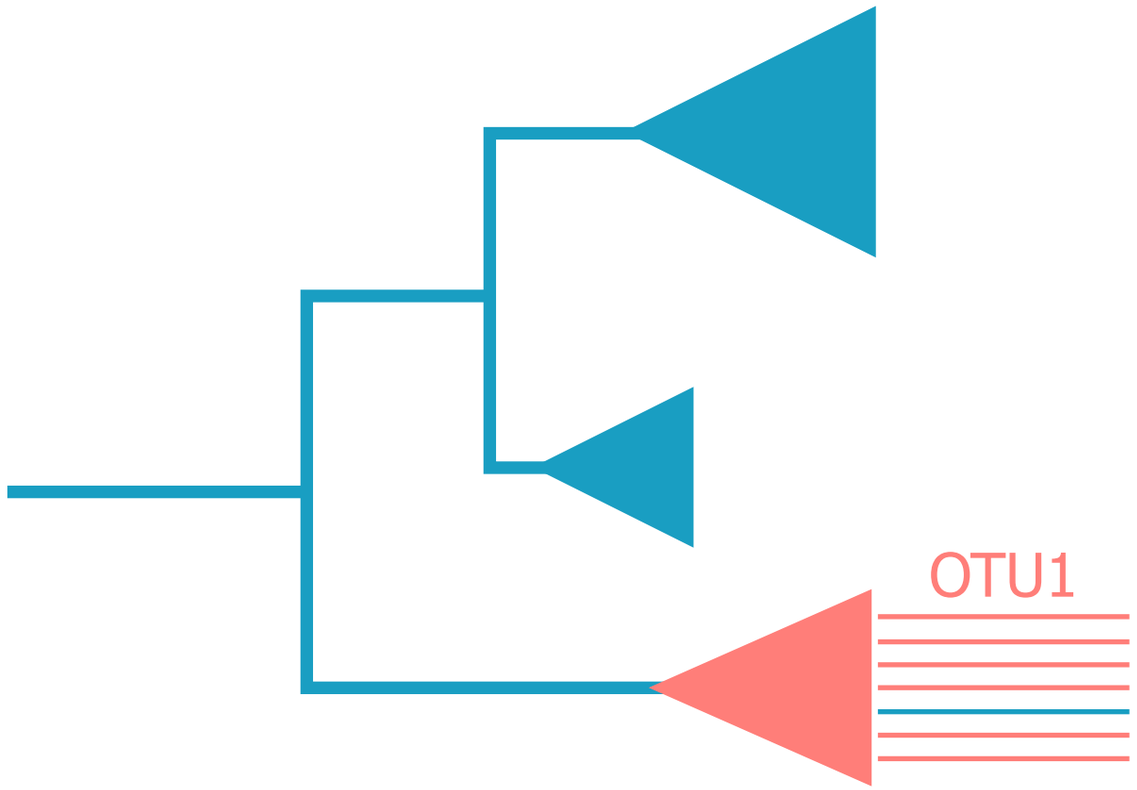


OTU1 ≈ Species A

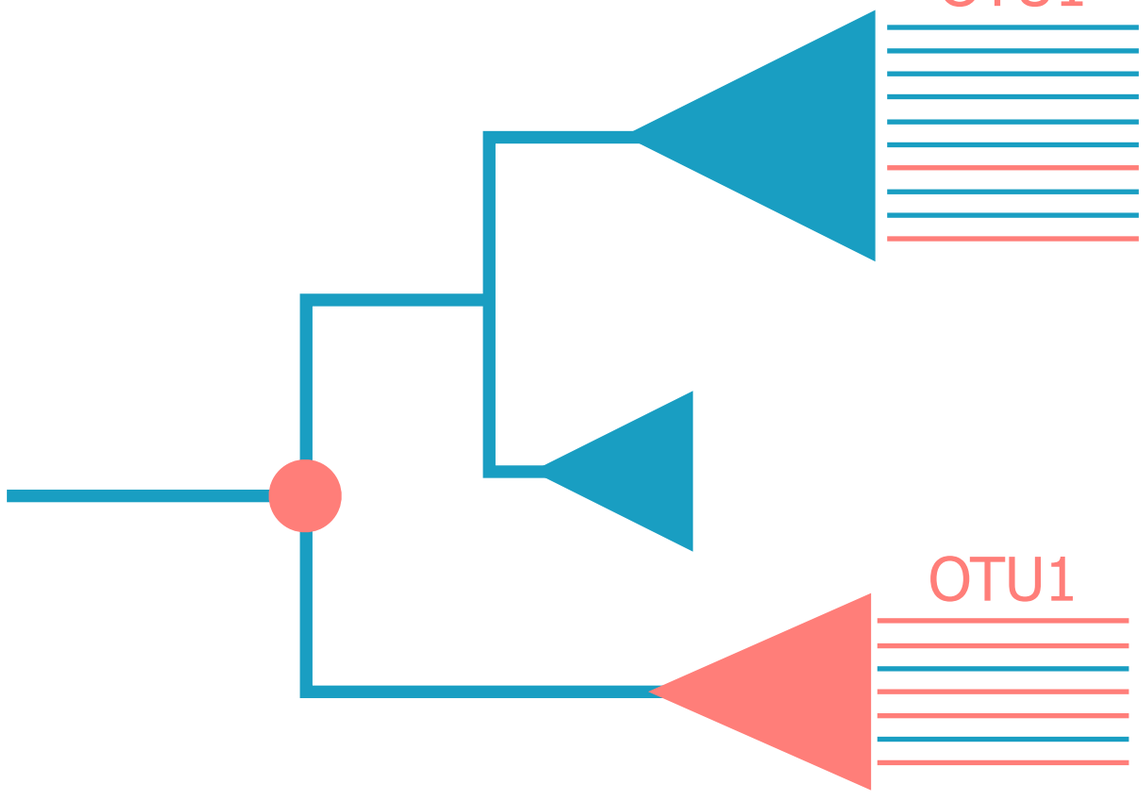


This approach might work if ...

...if there is only one clear best-hit
...if there is no other possible hit
...if the query is correct
...if 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)

Workflows

A1 Scientific Question

A2 Sample Design (e.g. NC, PC, N_{rep}, $N_{samples}$, N_{runs})

A3 Sample Collection (e.g. time, location, contamination)

A4 Sample Storage (e.g. EtOH, ice)

A5 Sample Processing (e.g. DNA/RNA isolation)

B1 Amplicon Design (e.g. size and region)

B2 Primer Design (e.g. two-step PCR)

B3 Library Preparation (e.g. two-step PCR)

B4 Sequencing

B5 Quality Control (e.g. FastQC and FastScreen reports)

Illumina - Paired-end data

I1 Read Merging > Amplicons

I2 Primer Trimming

I3 Quality Filtering

I4 Clustering / Amplicon Sequence Variants

I5 Count Table

Genetic Diversity Centre Zurich

PacBio - CCS Data

P1 De-multiplexing

P2 In-silicon PCR with Size Selection

P3 Quality Filtering

P4 Clustering / Amplicon Sequence Variants

P5 Count Table

GDC Genetic Diversity Centre Zurich

M1 (Quality) Filtering (e.g. complexity filter)

M2 Clustering / Sorting (e.g. rRNA removal)

M3 Meta - Genome/Transcriptome Assembly

M4 Taxonomic Annotation

M5 Functional Annotation

MORE THINGS CONSIDERED

Zurich Centre Genetic Diversity

Preparing NGS reads for OTU and denoising analysis
(7 minutes)

Measuring diversity by 16S sequencing
(14 minutes)

Taxonomy reference databases for 16S
(17 minutes)

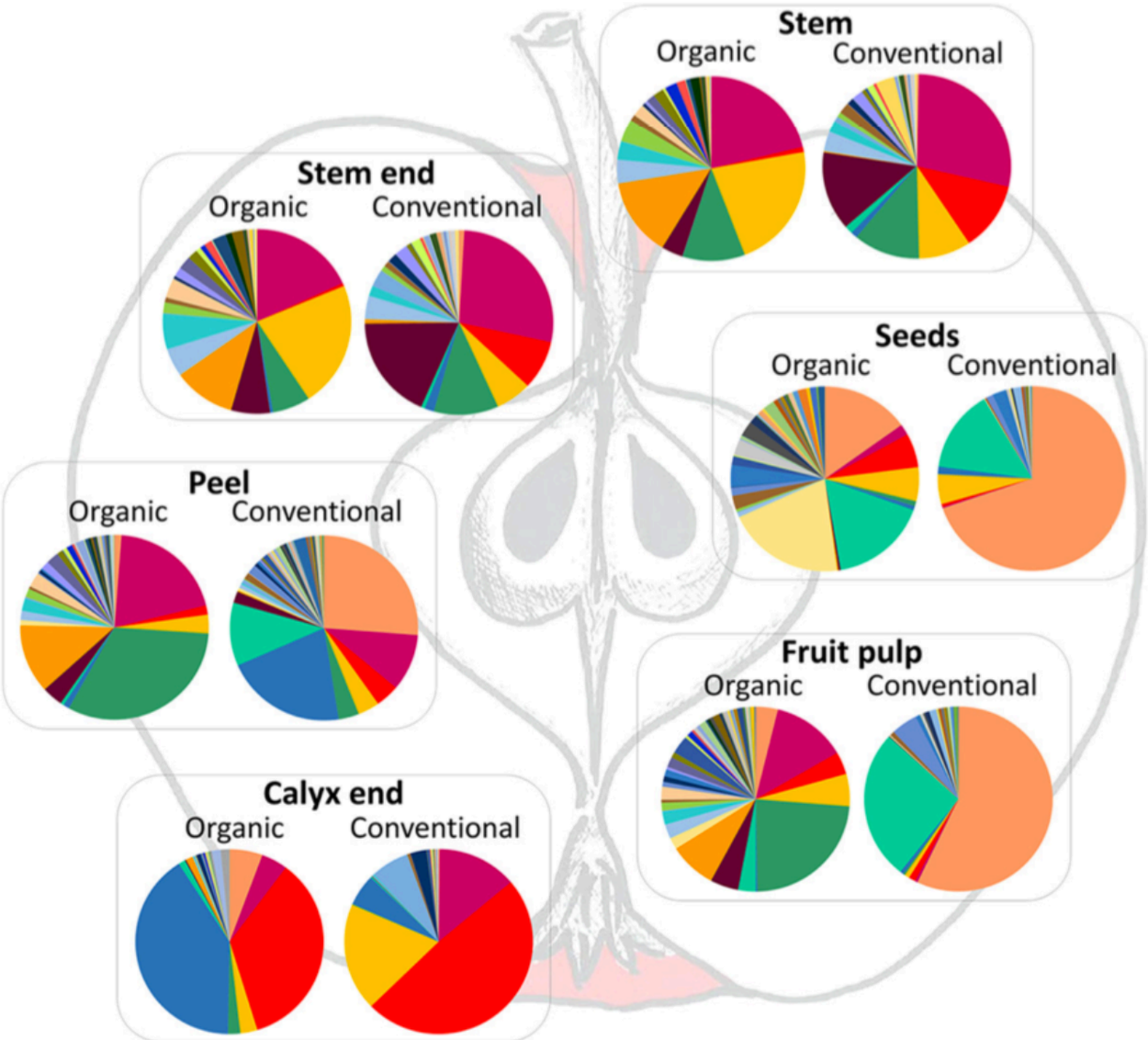
16S taxonomy and sequence identity
(8 minutes)

Taxonomy prediction methods for 16S sequences
(10 minutes)

An Apple a Day: Which Bacteria Do We Eat With Organic and Conventional Apples?

*Birgit Wassermann, Henry Müller and Gabriele Berg**

Institute of Environmental Biotechnology, Graz University of Technology, Graz, Austria



- What do you like about the article?
- It there anything you dislike about the article?
- What are the main findings according to the authors?
- Do you understand the sampling design?
- Would you be able to reproduce the data analysis?
- Where can you find the raw data?
- Do you agree with the statistical tests applied?
- Do you agree with the conclusions?
- Do you understand figures and tables?
- ?