Genetic Diversity Centre Zurich

701-1425-00L - Genetic Diversity: Analysis

NGS: Amp-Seq

Thursday, June 25, 2020

Jean-Claude Walser

jean-claude.walser@env.ethz.ch

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

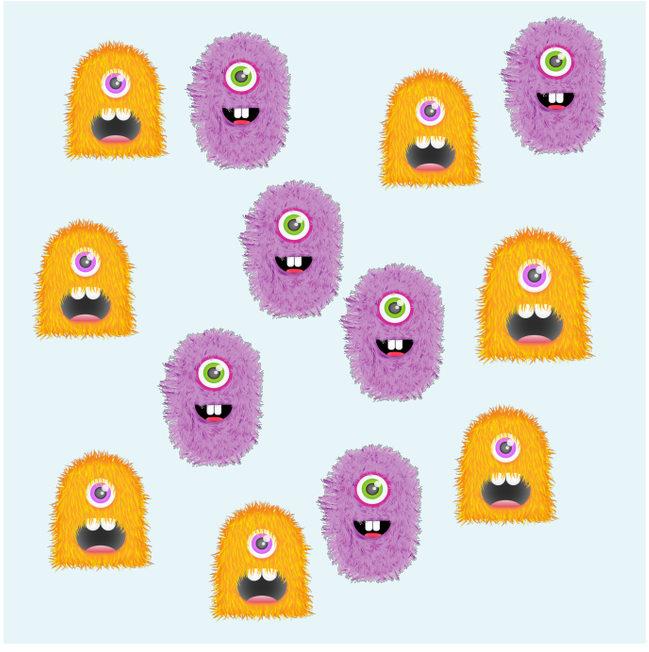
Assembly

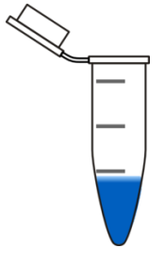
- Metagenome assembly
- Metatranscritpome assembly

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

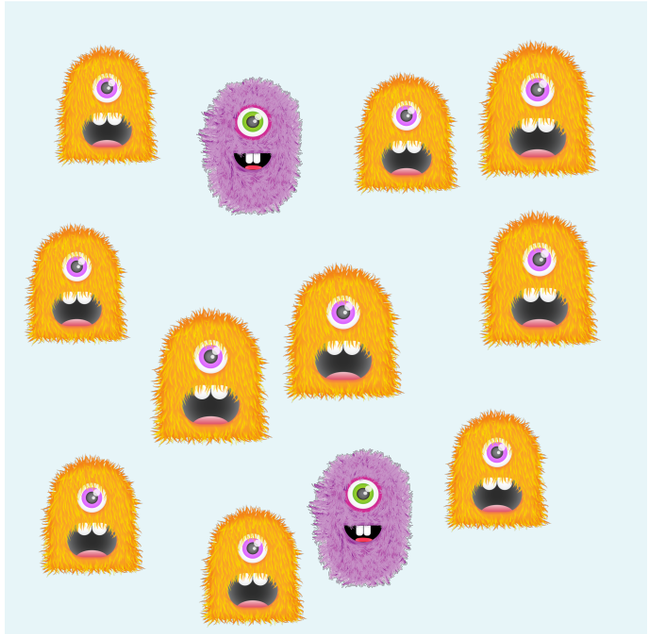
Amplicon-Sequencing

SampleA

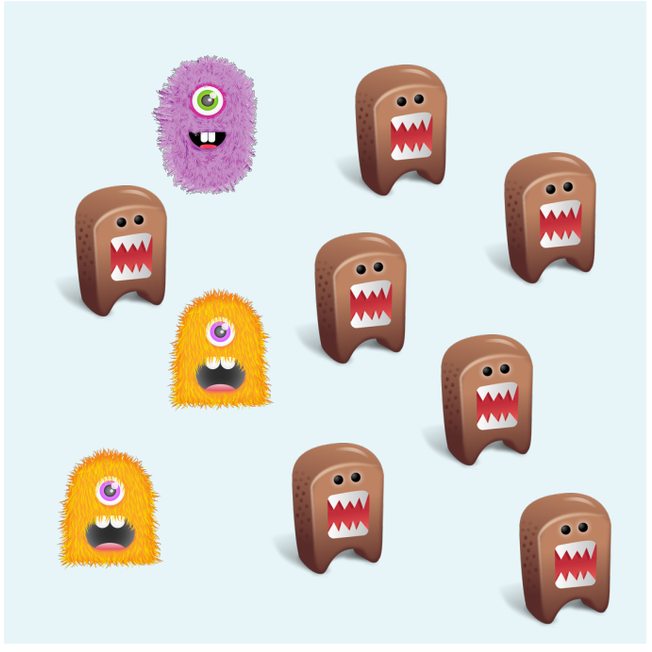


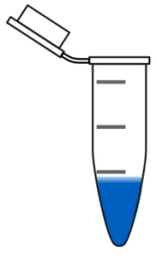


SampleB



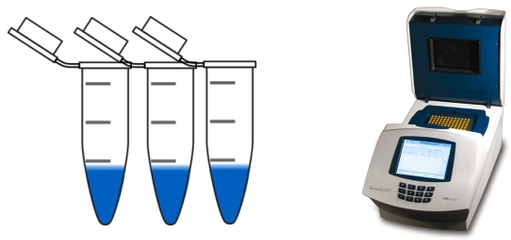
SampleC





Genetic Diversity Centre Zurich

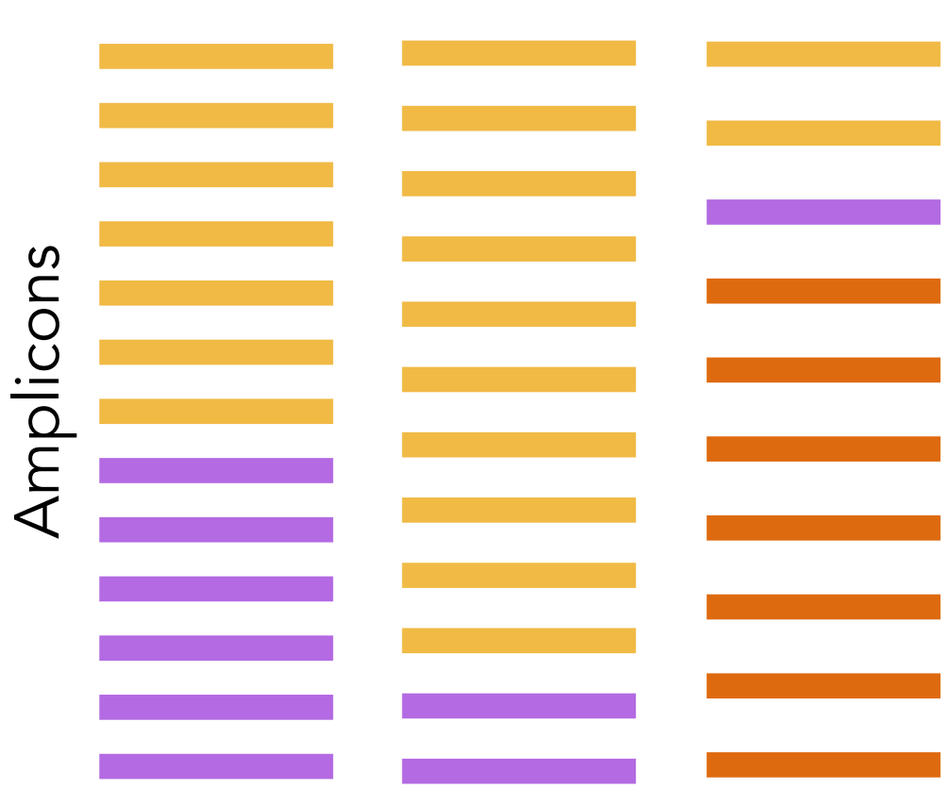
targeted PCR

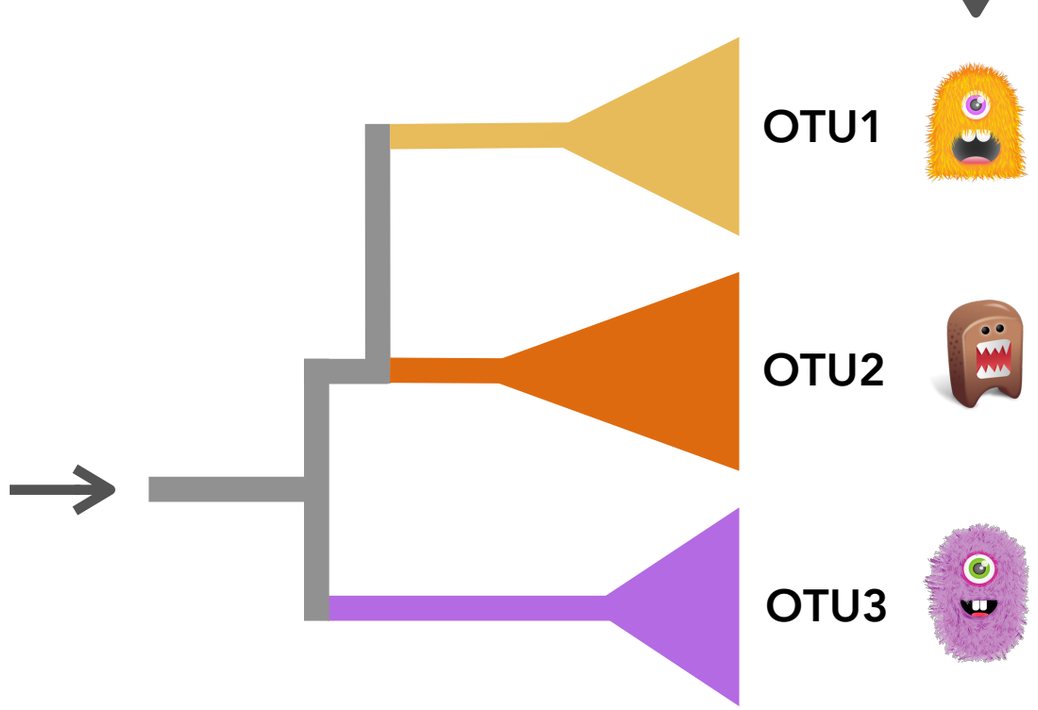


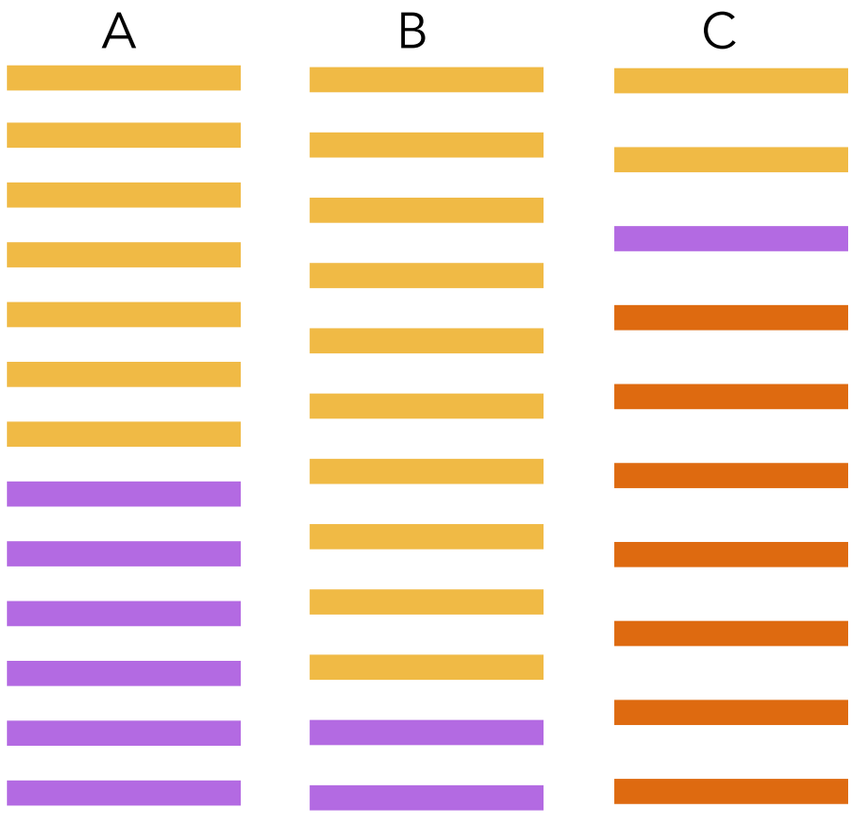
Classification

Reference

CLUSTERING / ASV







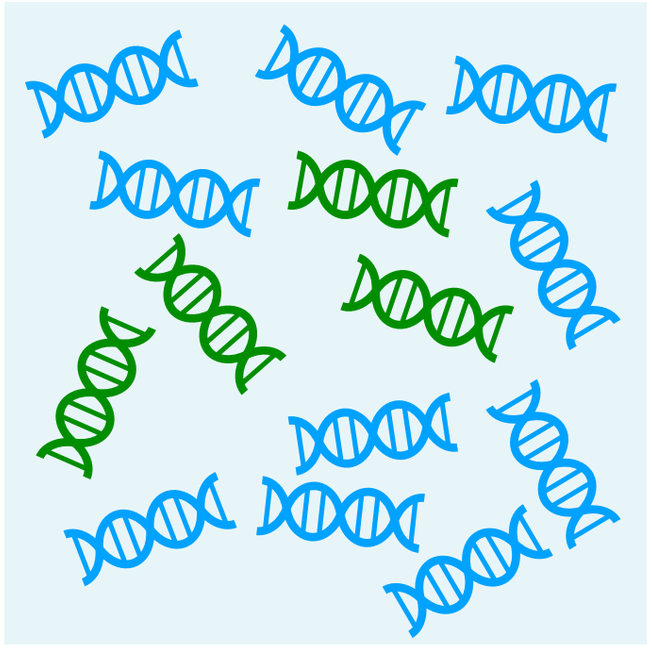
Count Table

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

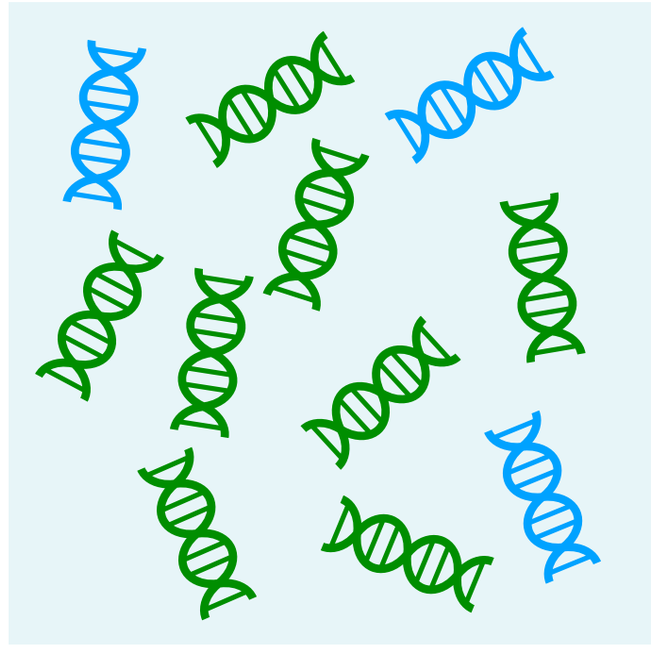
Sequening Depth

Metagenome-Sequencing

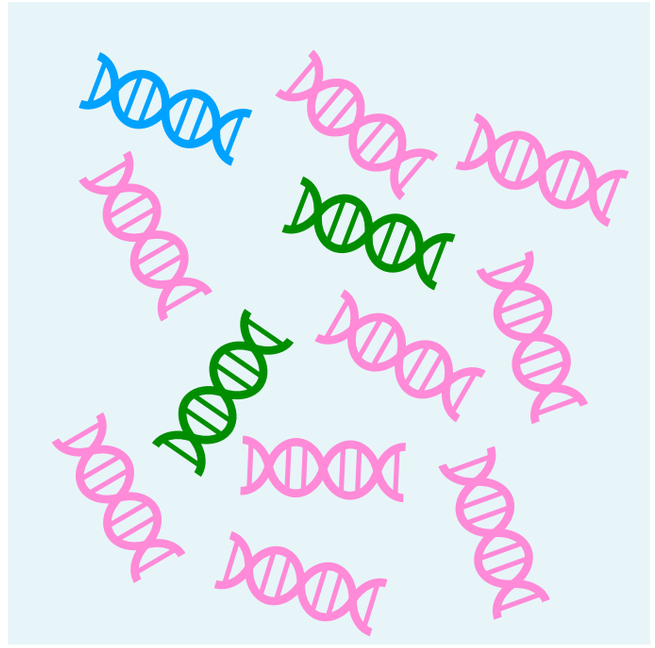
SampleA

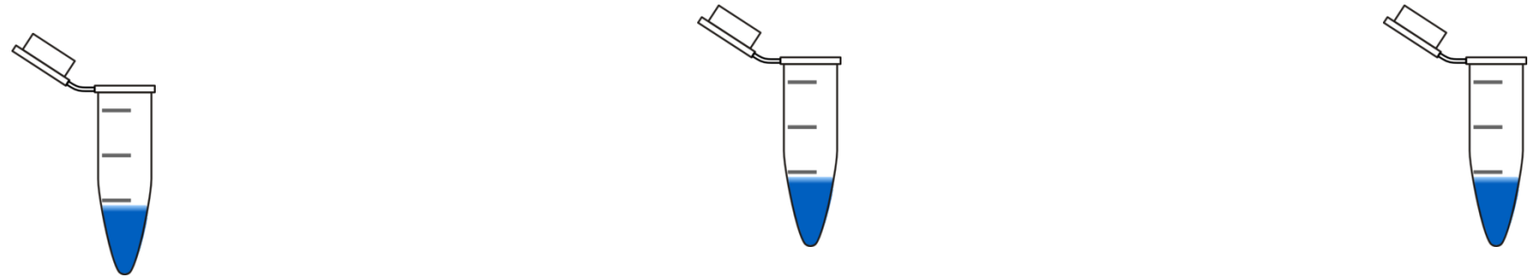


SampleB

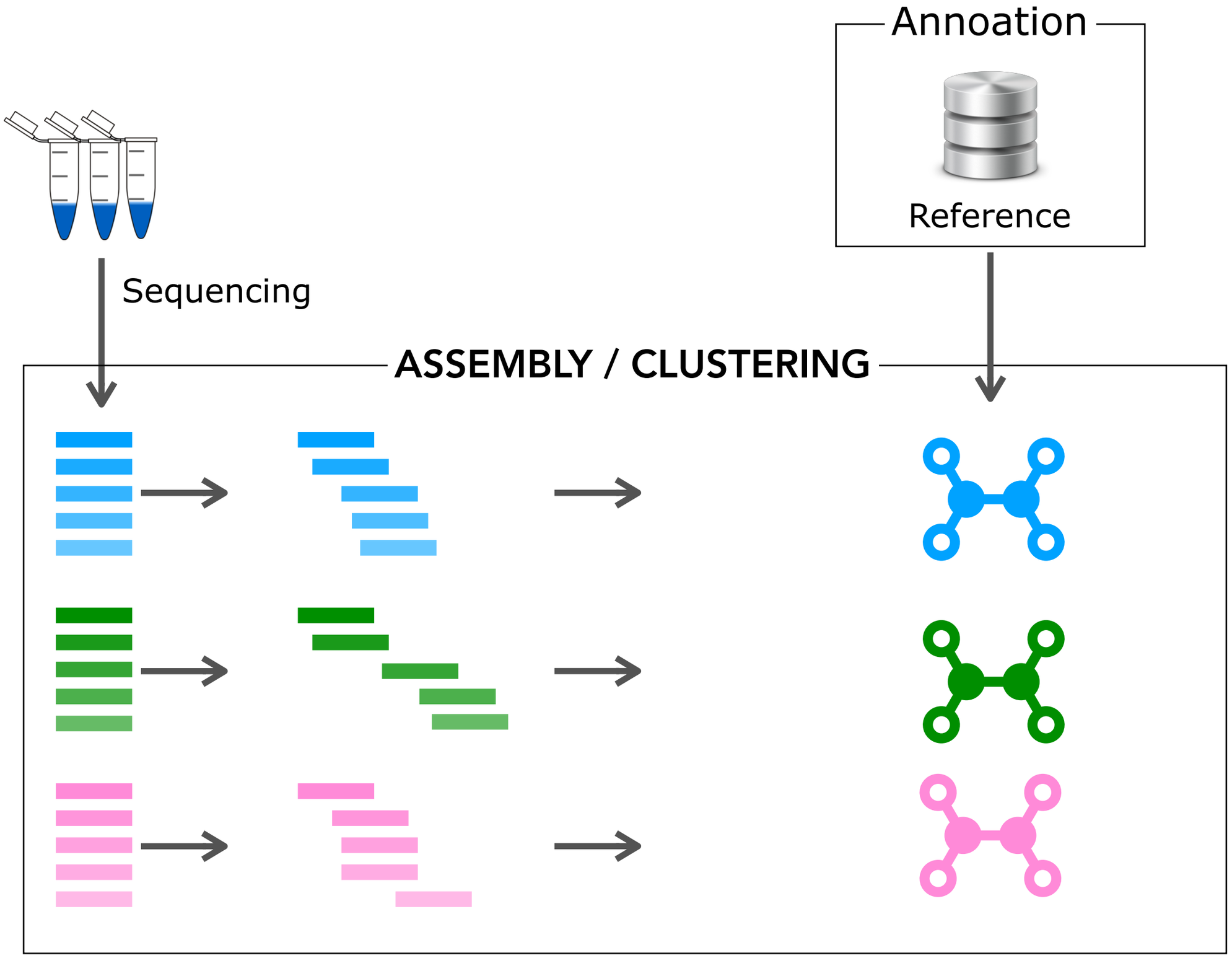


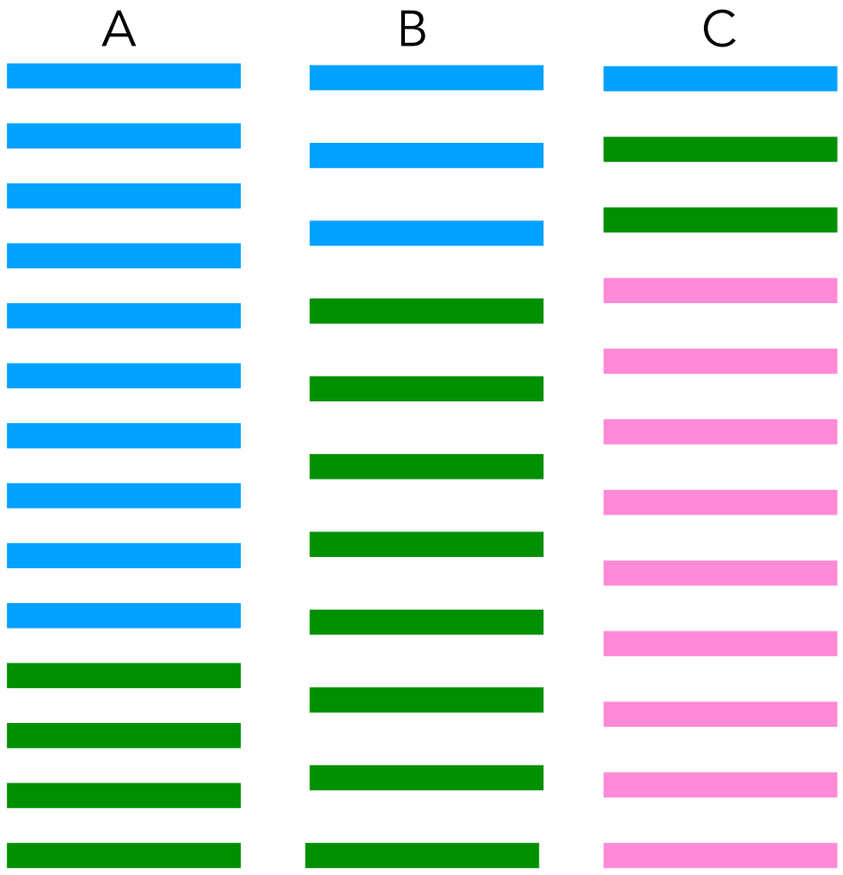
SampleC





Genetic Diversity Centre Zurich



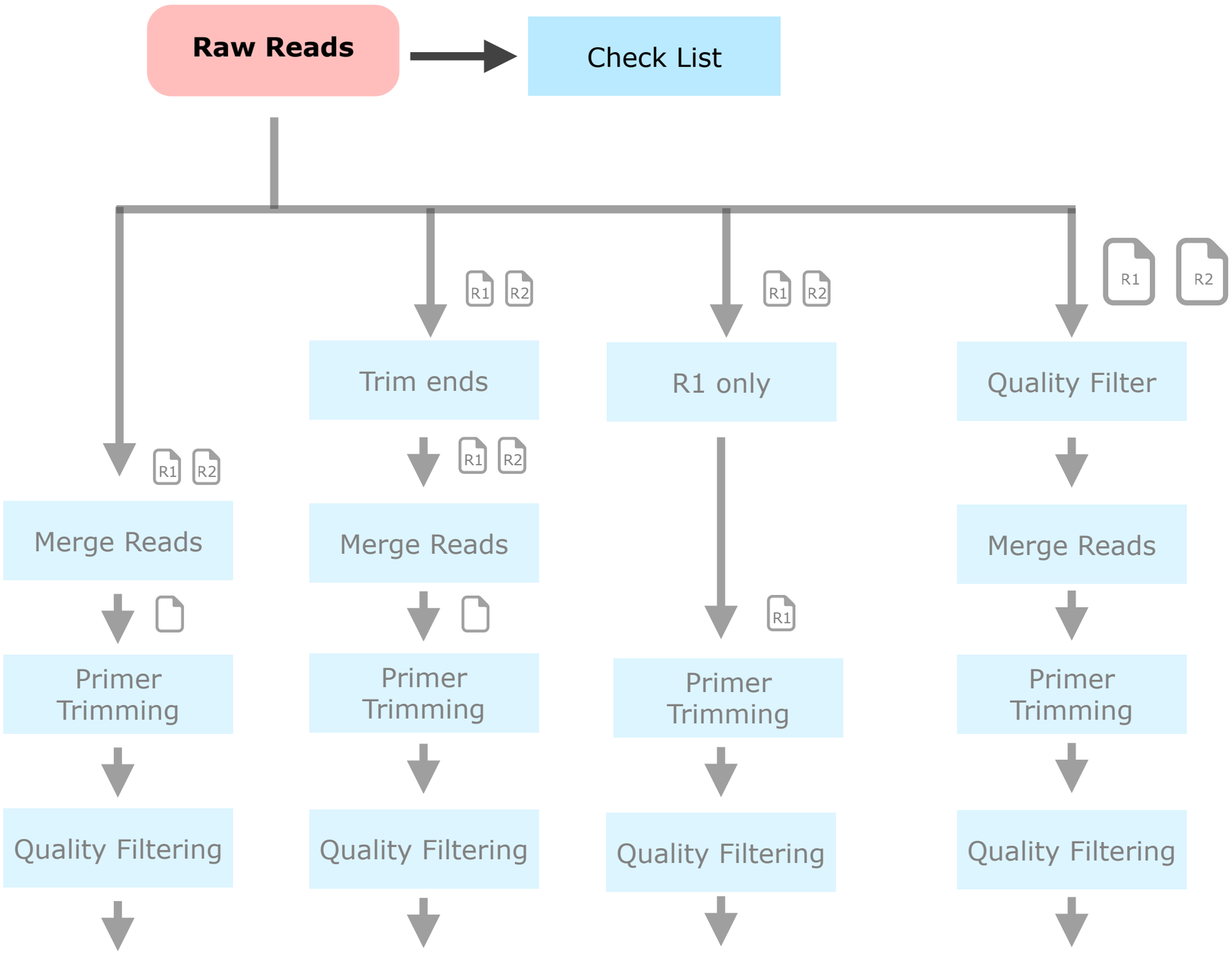


	A	B	C
GO	10	3	1
EC	4	8	2
MP	0	0	9
Total	14	11	12

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

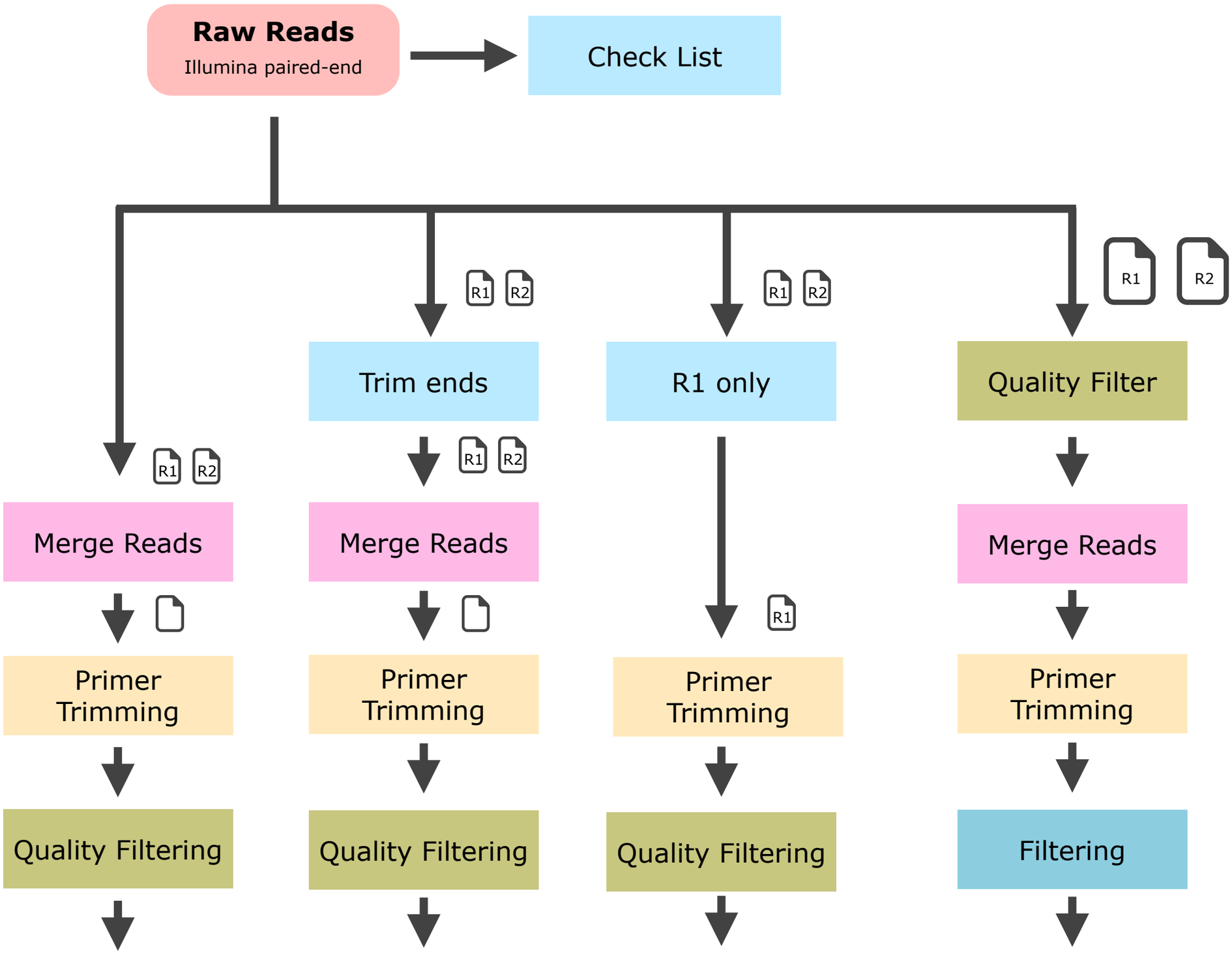
	A	B	C
	7	4	2
	0	0	9
	7	7	1
	14	11	11

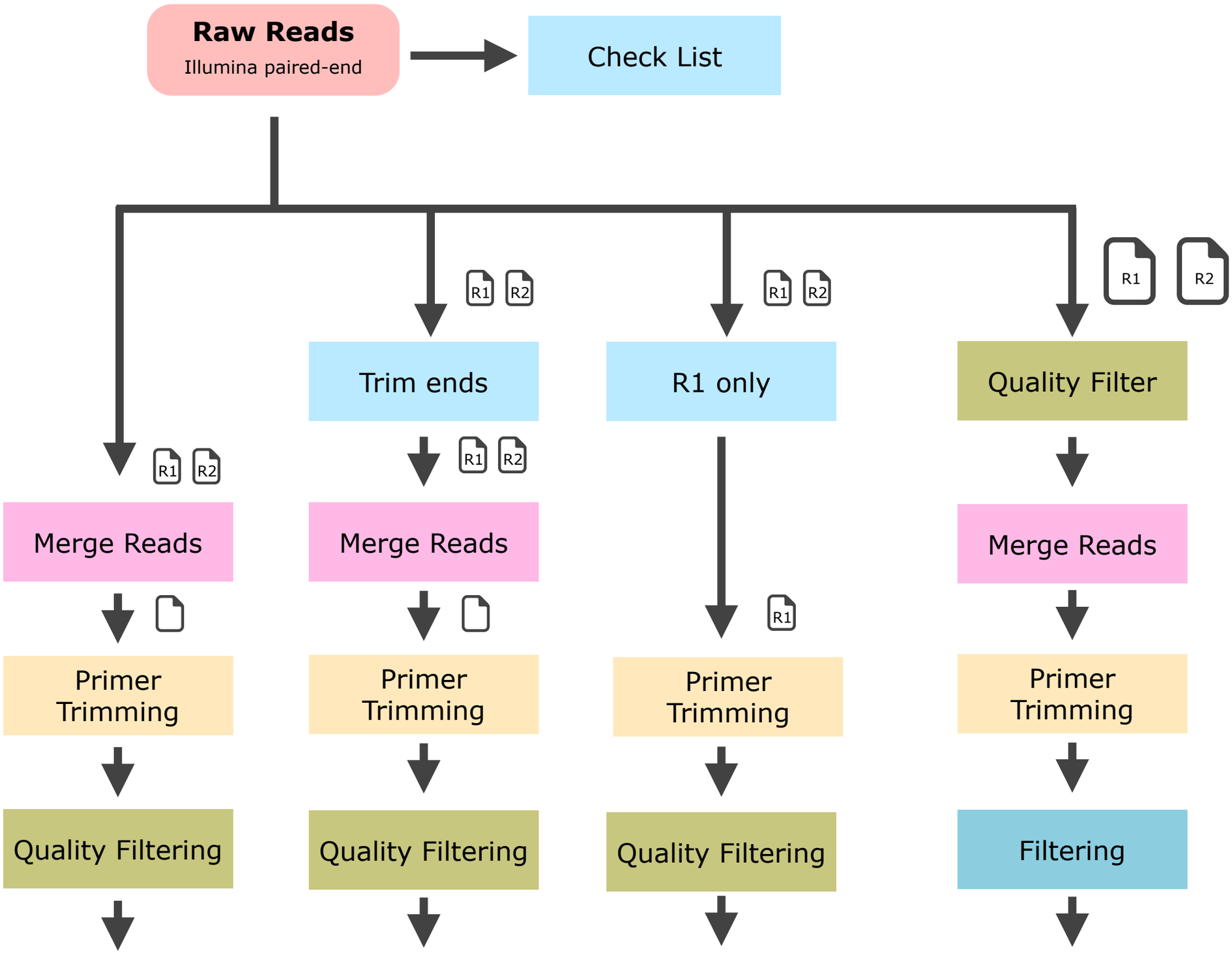
DATA PREPARATION


Raw Reads
Check List
R1
R2
Trim ends
R1
R2
R1 only
R1
R2
Quality Filter
Merge Reads
Merge Reads
Merge Reads
Primer Trimming
Primer Trimming
Primer Trimming
Primer Trimming
Quality Filtering
Quality Filtering
Quality Filtering
Quality Filtering

Check-List

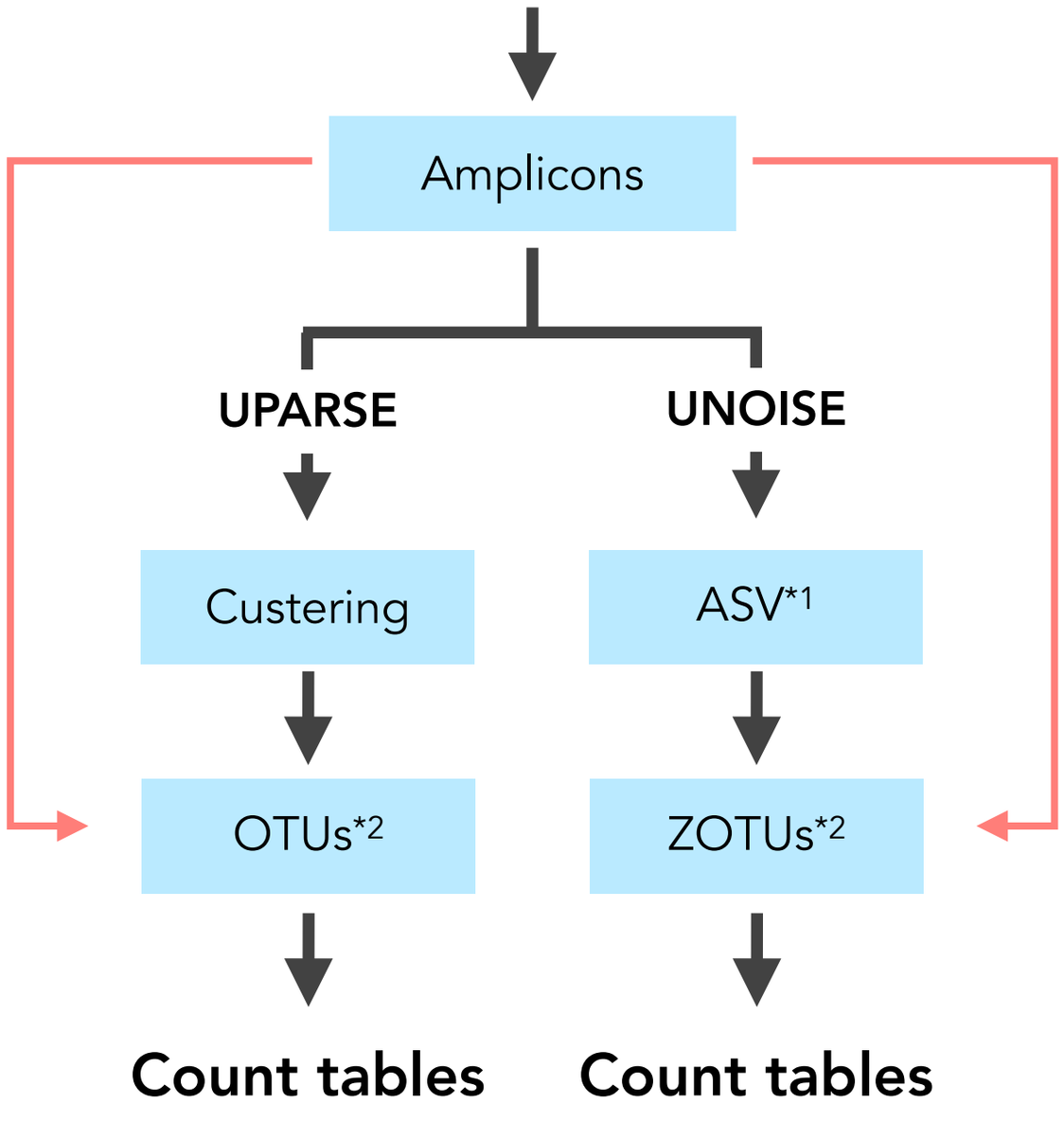
1. Download data (if possible via terminal e.g. sftp, wget)
2. Verify file integrity (md5sum)
3. Check data: $N_{samples} = N_{R1} = N_{R2}$
4. Blast a few random reads
5. Run a quality control (e.g. FastQC, FastScreen)
6. Look at the read size distribution
7. Check fastq header - how many runs?
8. Check 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
R1
R2
Trim ends
R1 only
Quality Filter
Merge Reads
Primer Trimming
Quality Filtering
Filtering


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

Preparing NGS reads for OTU and denoising analysis

(7 minutes)



*1 ASV: Amplicon Sequence Variant

*2 ZOTU: (Zero-Radius) Operational Taxonomic Units

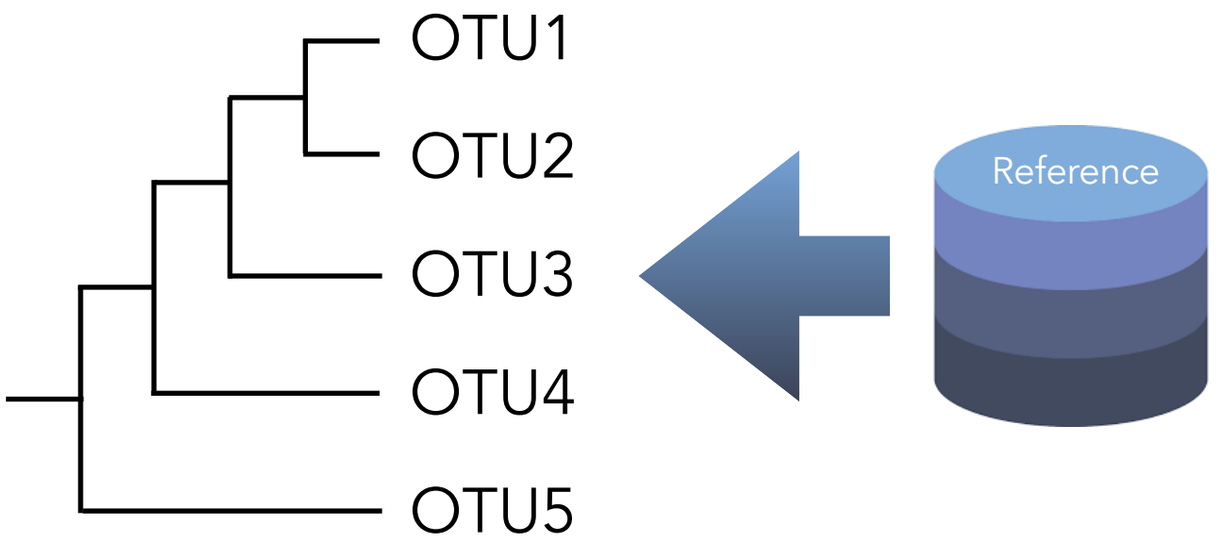
Measuring diversity by 16S sequencing

(14 minutes)

Taxonomy reference databases for 16S

(17 minutes)

OTU - Annoation (-Prediction)

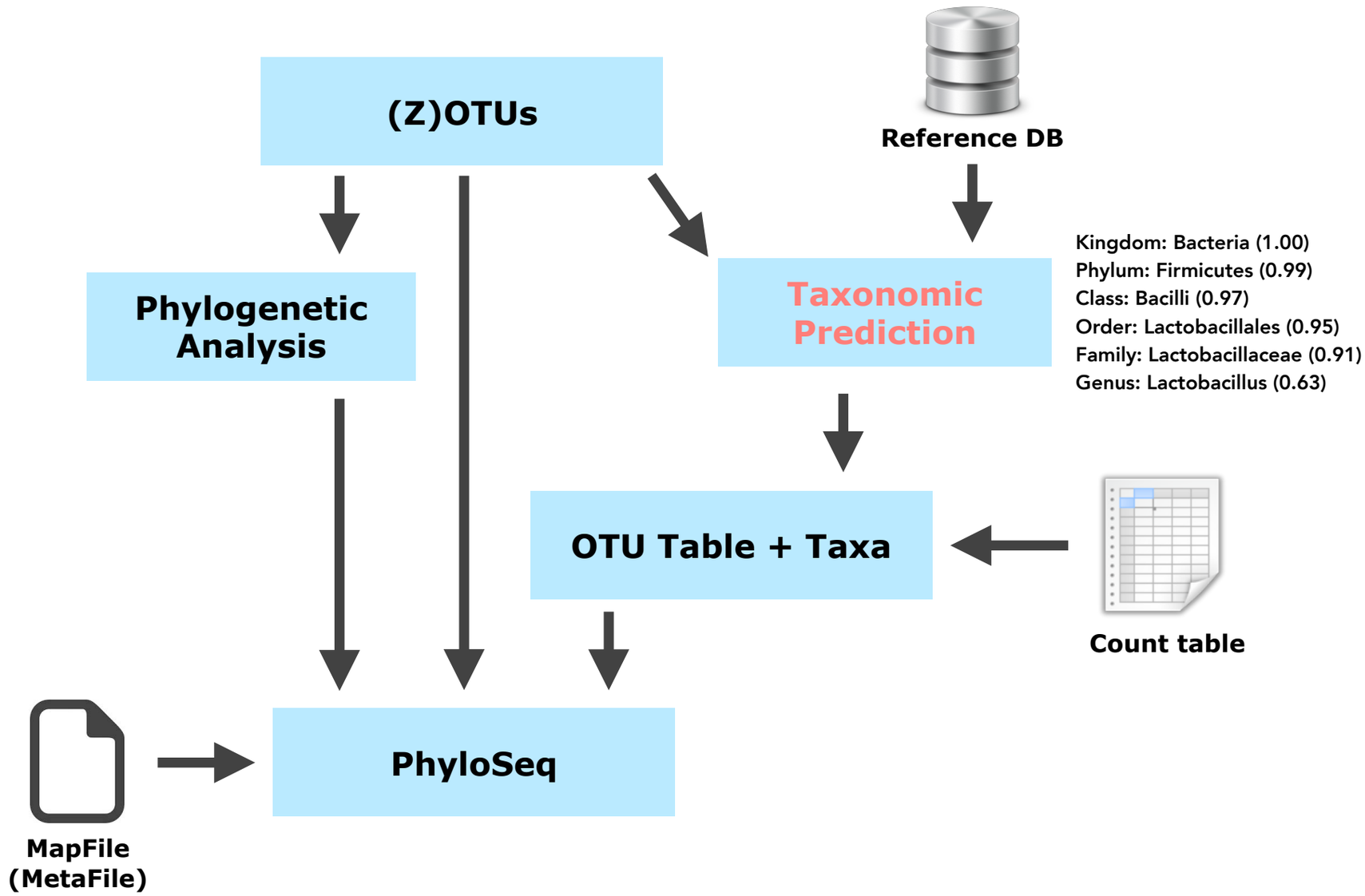


- NCBI 16S
- SILVA SSU & LSU
- Ribosomal Database Project RDP
- GreenGenes
- EzBioCloud
- …

16S taxonomy and sequence identity

(8 minutes)

(Z)OTU - Annoation (-Prediction)



Taxonomy prediction methods for 16S sequences

(10 minutes)

WORKFLOWS

GDC Genetic Diversity Centre Zurich

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)

Genetic Diversity Centre Zurich

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)

Genetic Diversity Centre Zurich

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

Amplicon Sequencing

DNA

▽

Primer Design

▽

PCR(s)

▽

Library Prep

▽

Sequencing

▽

Data Processing

▽

Data Analysis

Metagenomics

DNA

▽

DNA Fragmentation

▽

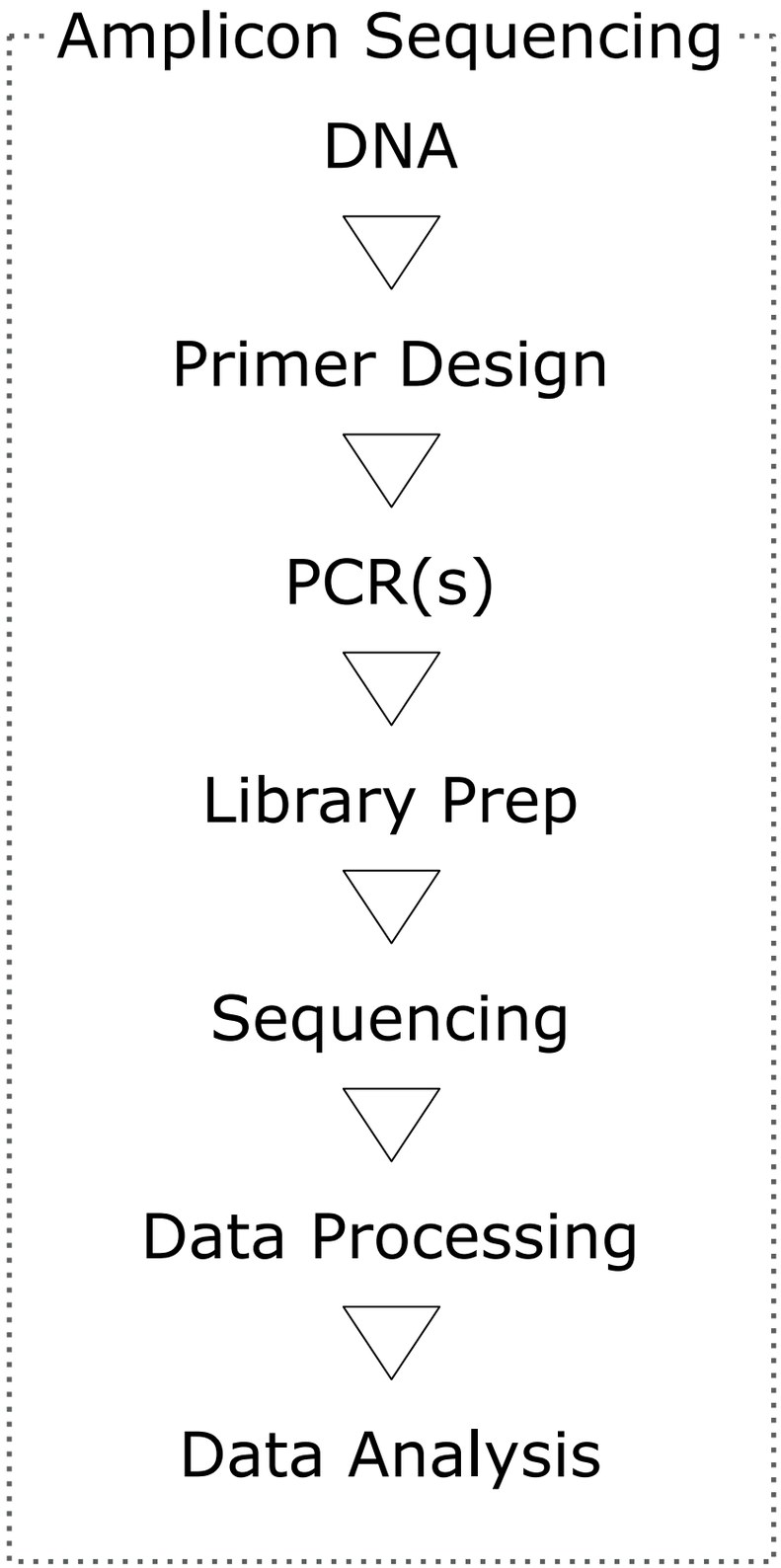
Ligation

▽

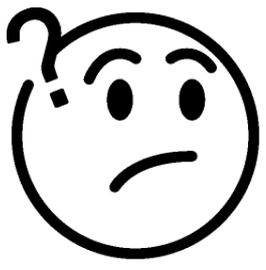
Sequencing

▽

Data Analysis



- sampling / sampling design
- extraction method
- contamination



What do I collect and when do I collect my **samples**?

How do I **store** my samples?

What is the **expected diversity** of my sample(s)?

What is the required **depth of sequencing** per sample?

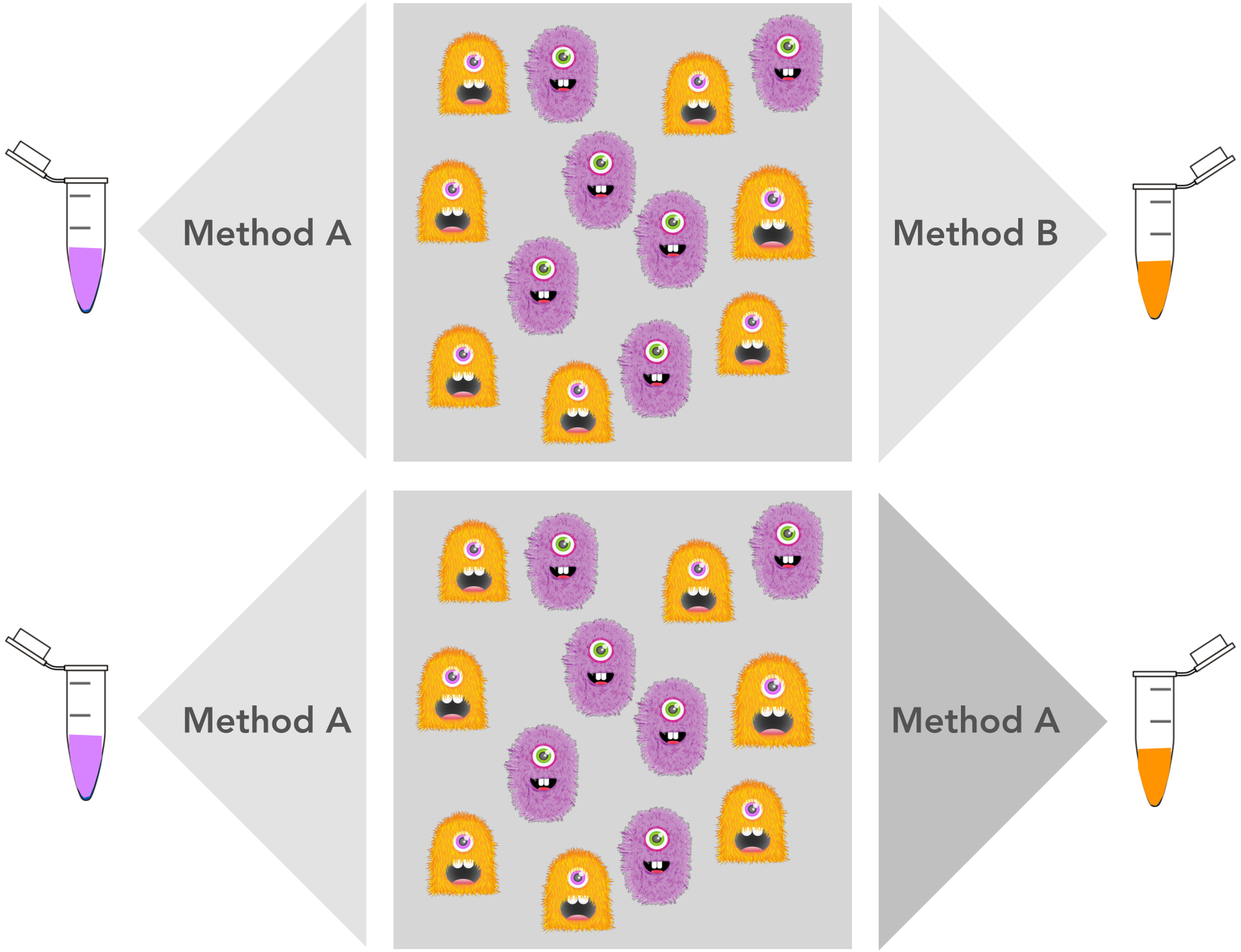
How many **samples / replicates** will be needed?

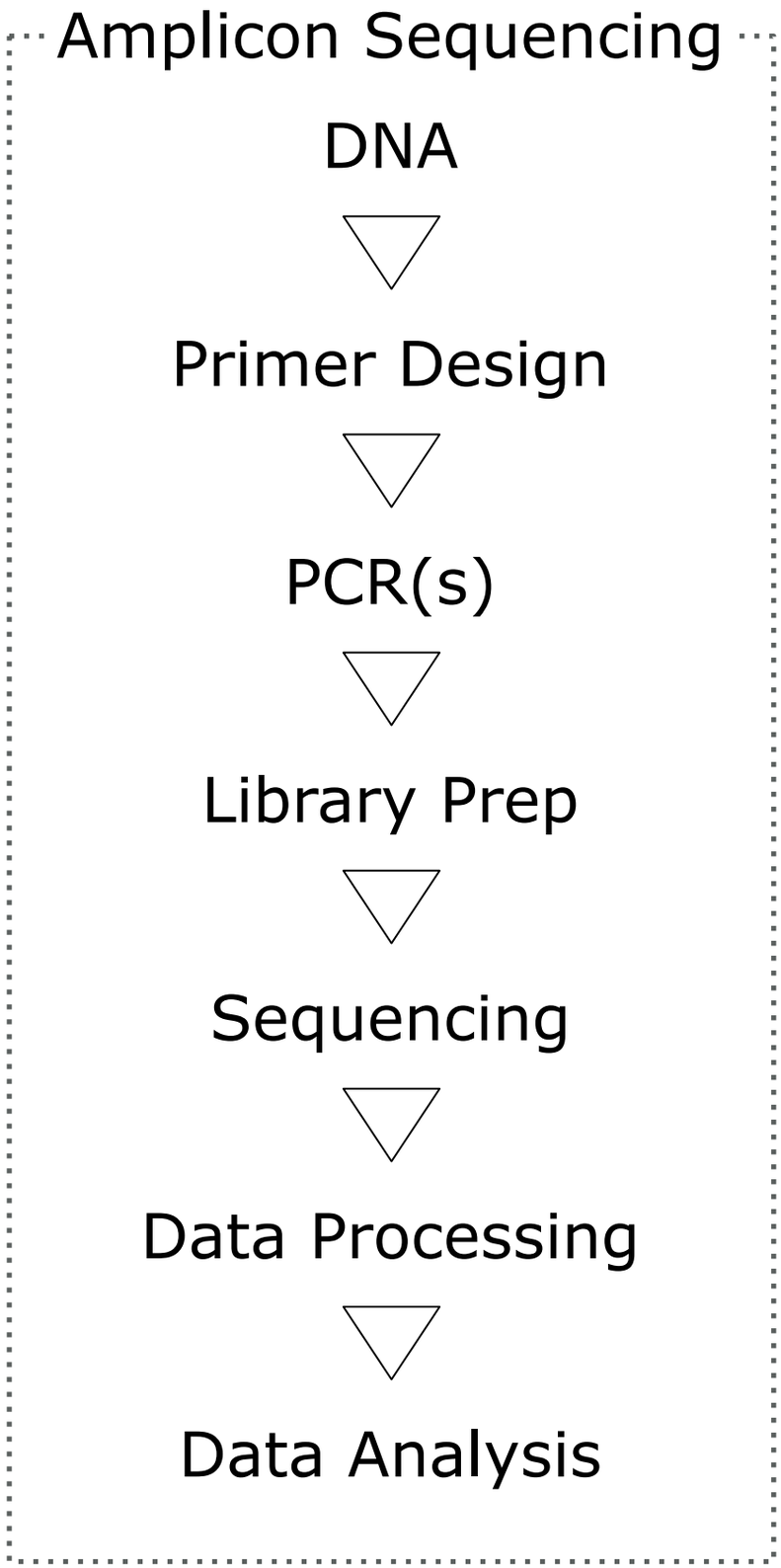
What is the **trade off** between coverage and replicates?

How much **money** do we have?

▶ Pilots are very useful

DNA Extraction Method Bias

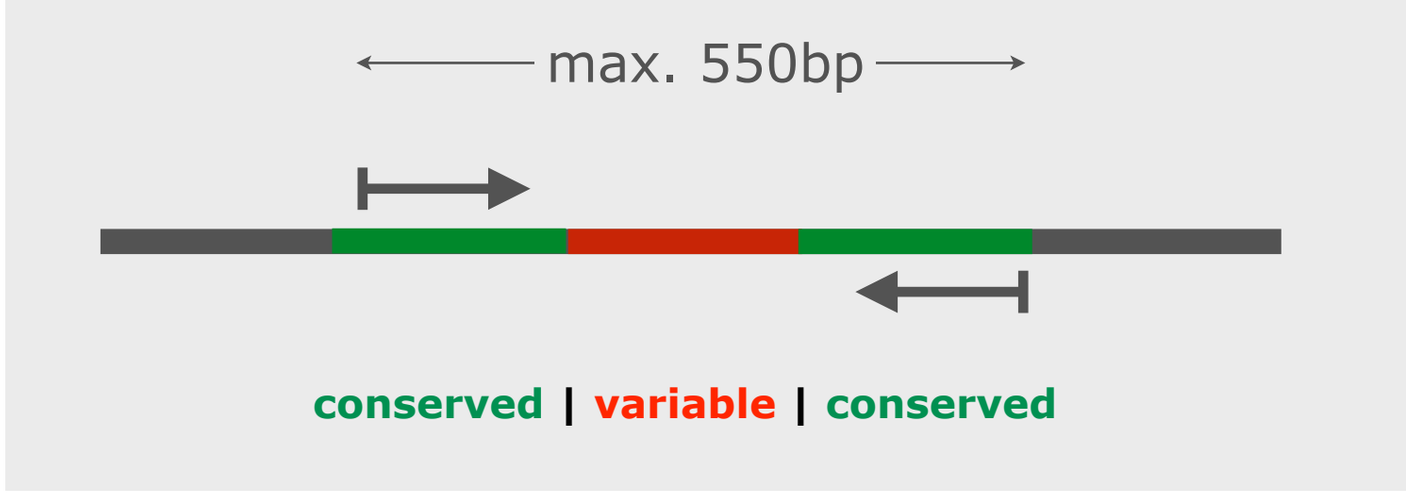




- sampling / sampling design
- extraction method
- contamination

- specificity
- quality

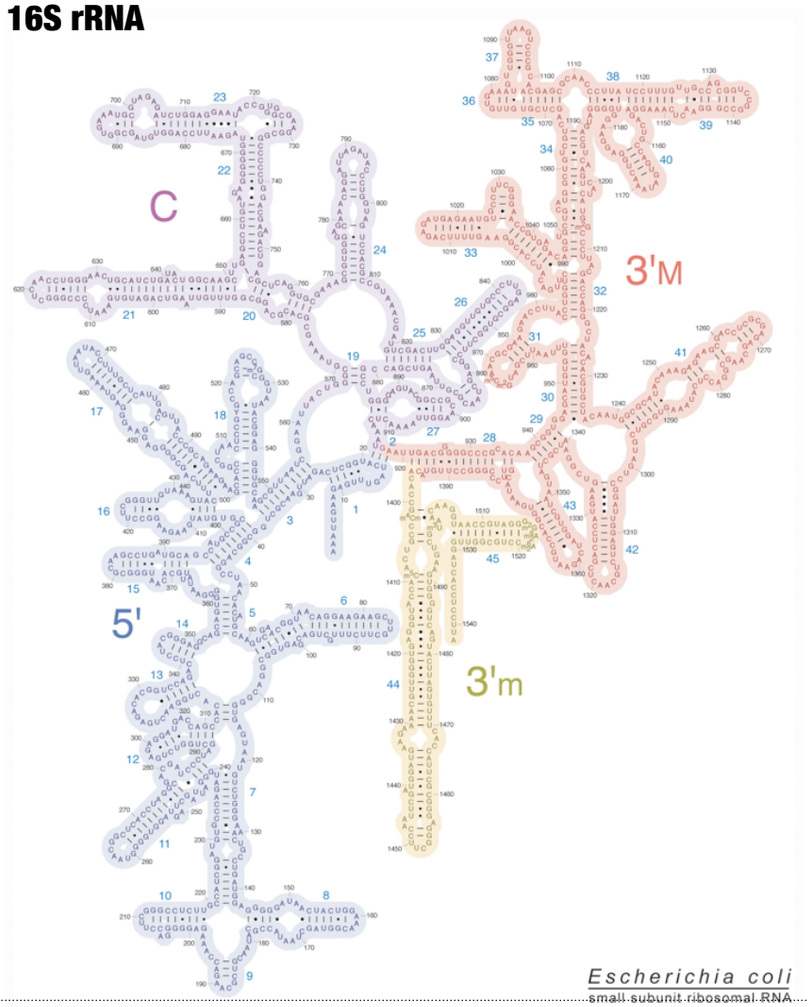
Primer Design



- traget specific - avoid false priming
- universal for all trageted species
- amplicon size
- avoid / limit amplicon size variation
- optimize PCR condition

GDC Genetic Diversity Centre Zurich

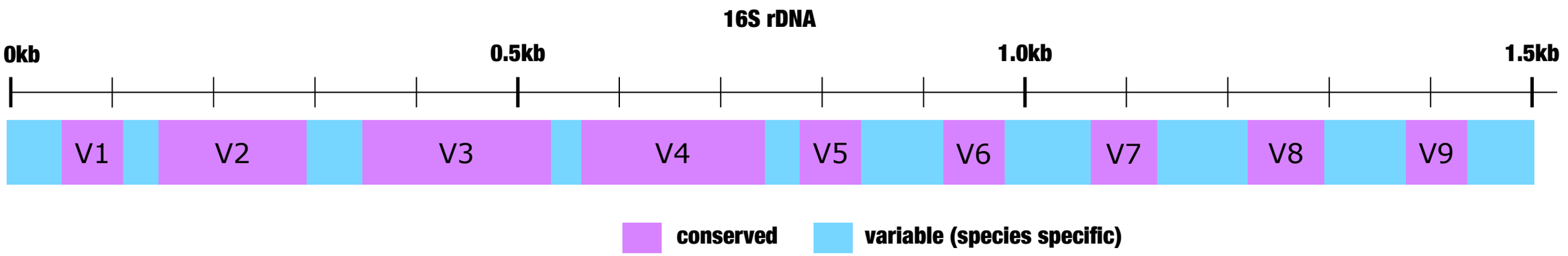
16S rRNA



Escherichia coli
small subunit ribosomal RNA

Ribosomal Database Project

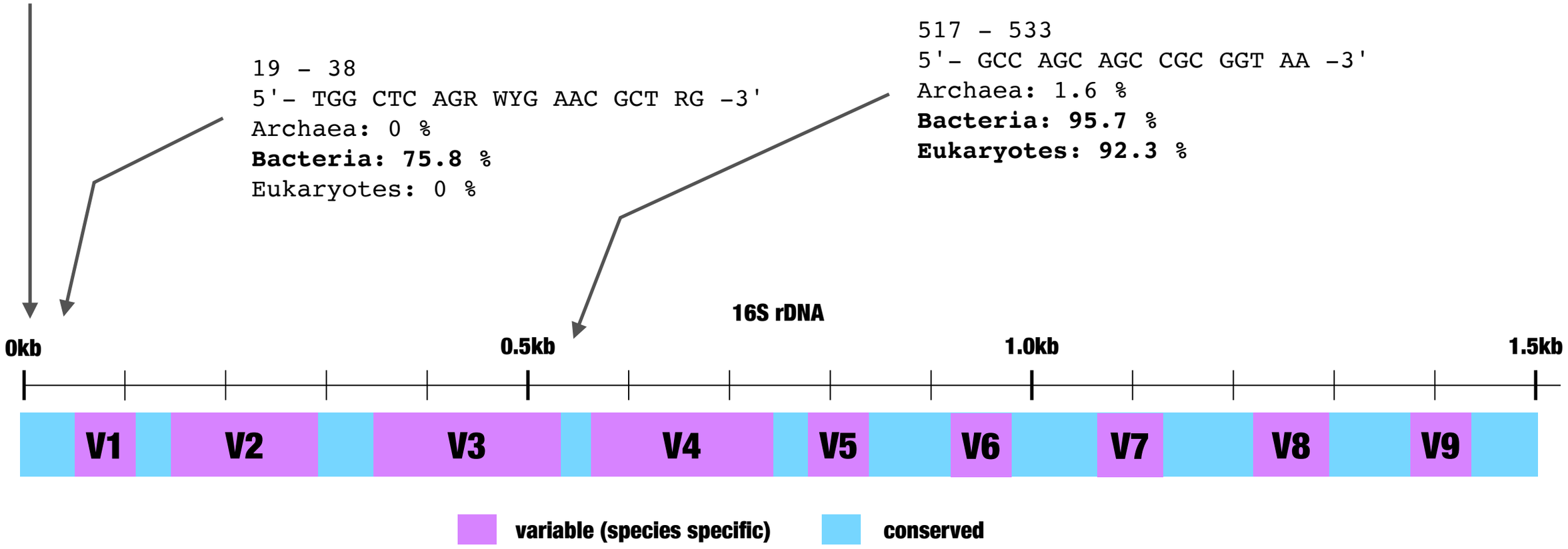
http://rdp.cme.msu.edu/



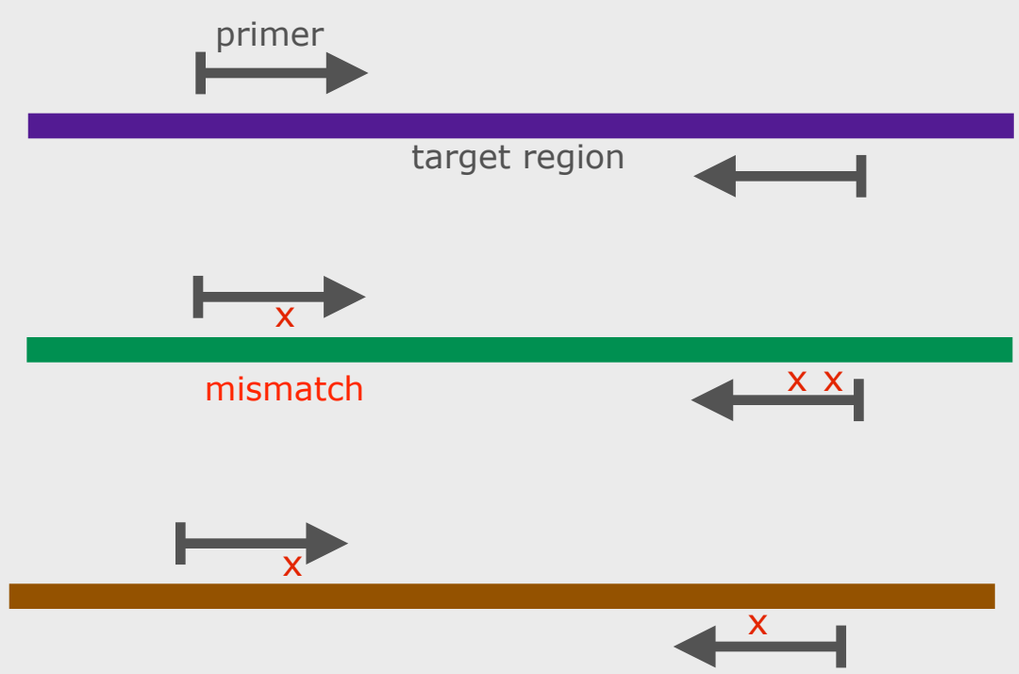
8 - 23
5'- TCY GGT TGA TCC TGC C -3'
Archaea: 65.1 %
Bacteria: 0 %
Eukaryotes: 11.1 %

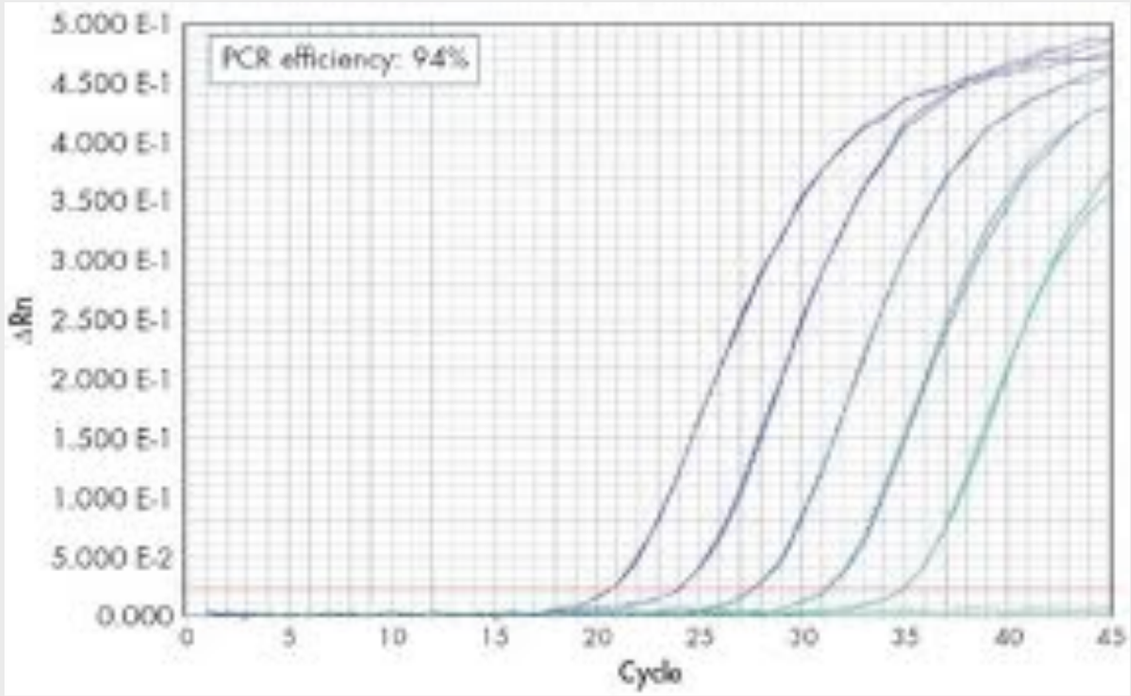
19 - 38
5'- TGG CTC AGR WYG AAC GCT RG -3'
Archaea: 0 %
Bacteria: 75.8 %
Eukaryotes: 0 %

517 - 533
5'- GCC AGC AGC CGC GGT AA -3'
Archaea: 1.6 %
Bacteria: 95.7 %
Eukaryotes: 92.3 %

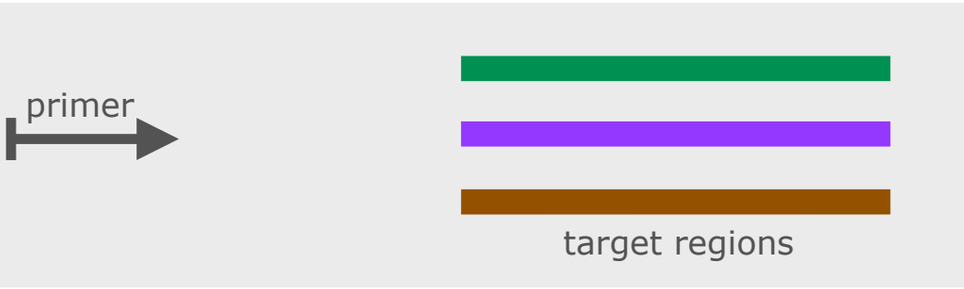


PCR efficency and primer mismatches





The primer(s) should equally amplify the target regions of the different species/individuals!

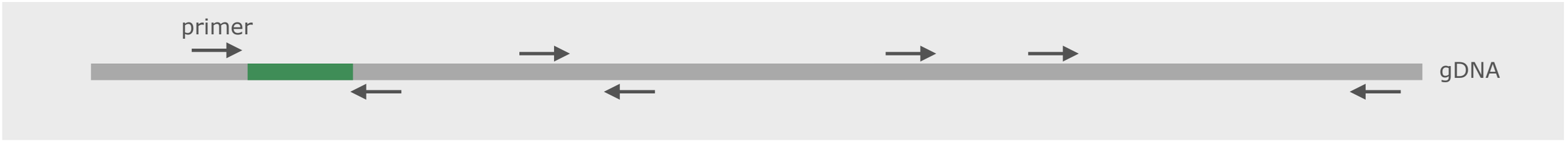


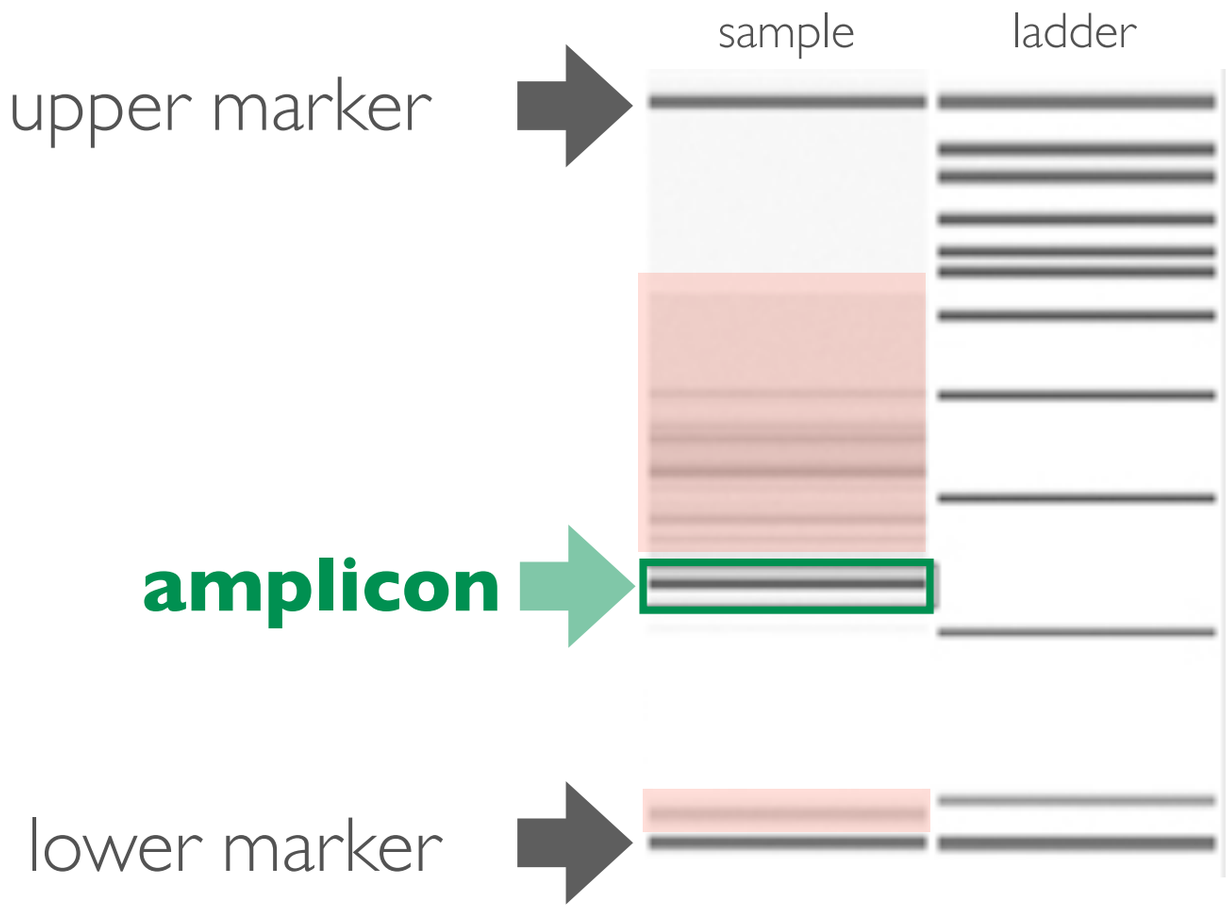
Degenerated primers

A group of degenerate oligonucleotides contain **related sequences** with differences at specific locations. These are used simultaneously in the hope that one of the sequences of the oligonucleotides will be perfectly complementary to a target DNA sequence.

http://www.bioinformatics.org/sms2/iupac.html

Primer Quality and PCR Contitions





Desalted Oligonucleotides

Residual of low molecular by-products arising and accumulating from the frequent chemical reactions during synthesis are removed. Such purification is sufficient for oligonucleotides shorter than 30 and/or oligonucleotides used for non-critical applications such as PCR, sequencing, probing, mobility shift or hybridization. However, desalted oligos are not recommended for use in molecular cloning projects.

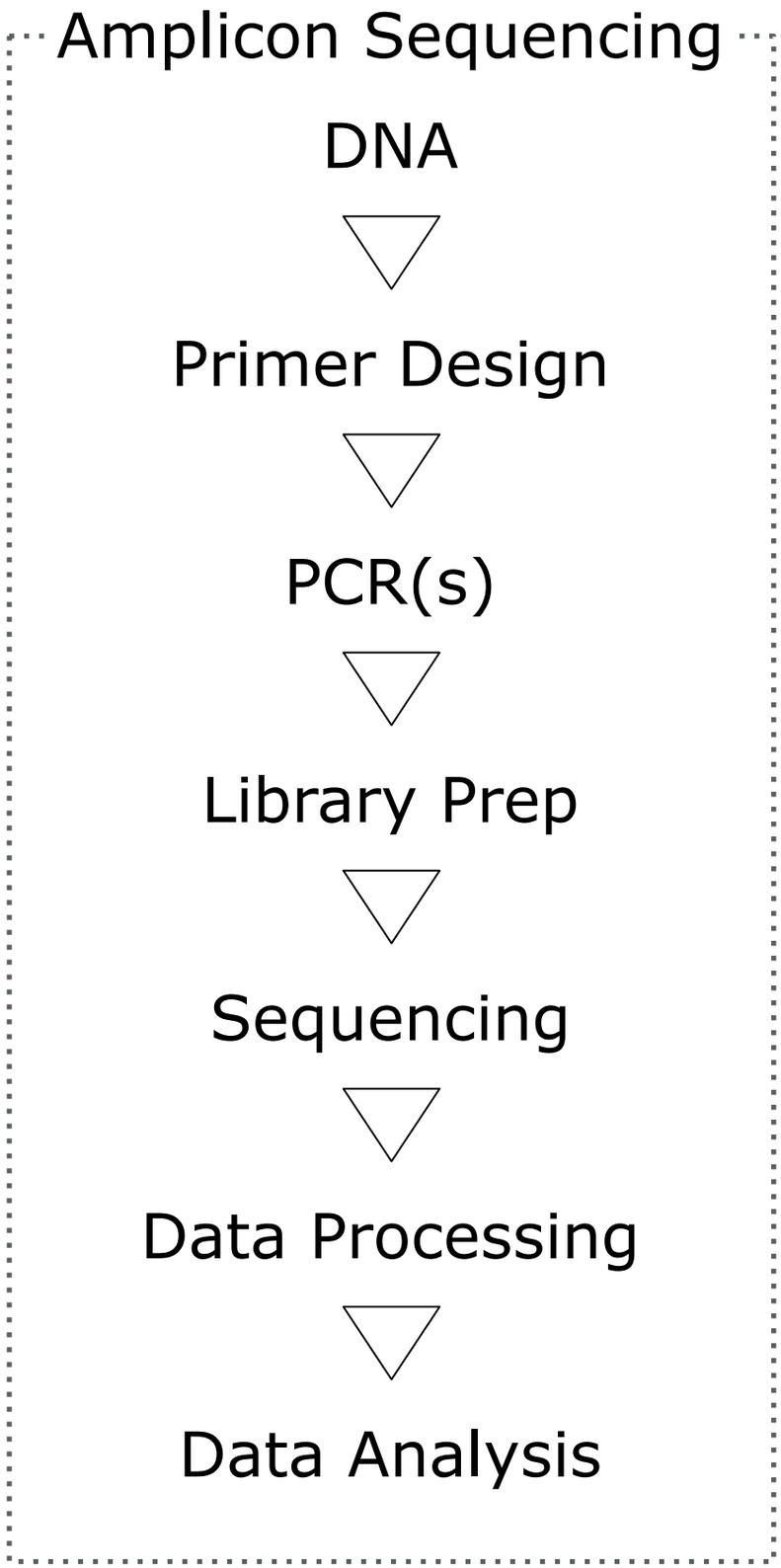
HPLC-Purified Oligonucleotides

Oligos <50 bases in length can be well purified via Reverse Phase HPLC. Through this purification approach, preferably residual, n-x truncated oligos (lacking the hydrophobic DMT protection group at the 5’ end) are removed. This results in a **90-95% purity** of the targeted oligonucleotide. RP-HPLC is useful for a higher level of purity required for more demanding applications such as cloning, DNA fingerprinting, real-time PCR, FISH, etc.

PAGE-Purified Oligonucleotides

Polyacrylamide gel electrophoresis (PAGE) purification is generally necessary for long oligos (>50 bases) and for all those primers with critical 5' sequences (restriction endonuclease sites, RNA promoters). It is the best method to differentiate full-length oligos from aborted sequences (n-1 oligos), based on size, conformation and charge. PAGE purification has an excellent resolution and yields a product that is, on average, **95-99% pure**. This type of purification is highly recommended for sensitive experiments such as cloning, mutagenesis, DNA fingerprinting, in situ hybridization, gene synthesis, etc.

Source: http://www.microsynth.ch



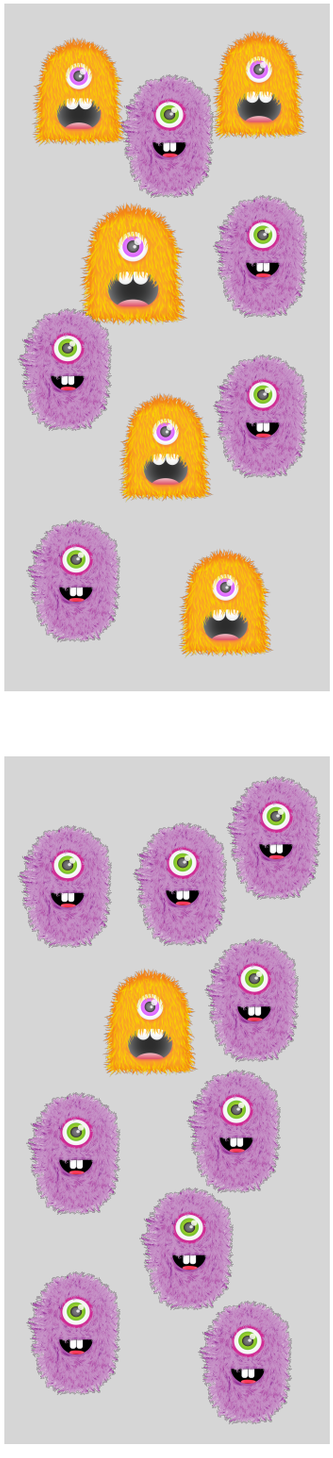
- sampling / sampling design
- extraction method
- contamination

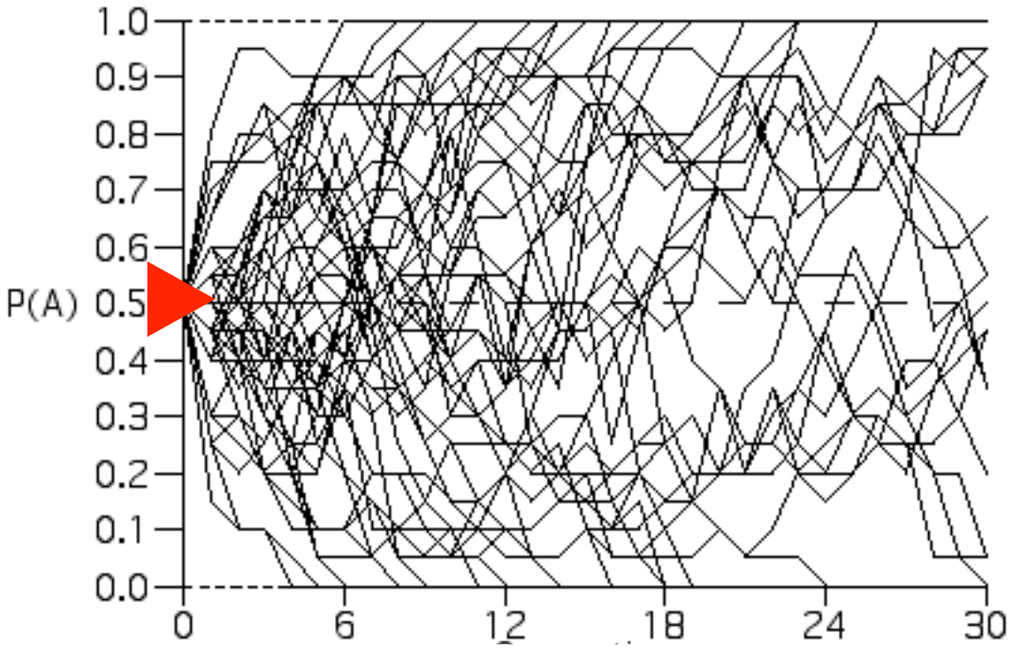
- specificity
- quality

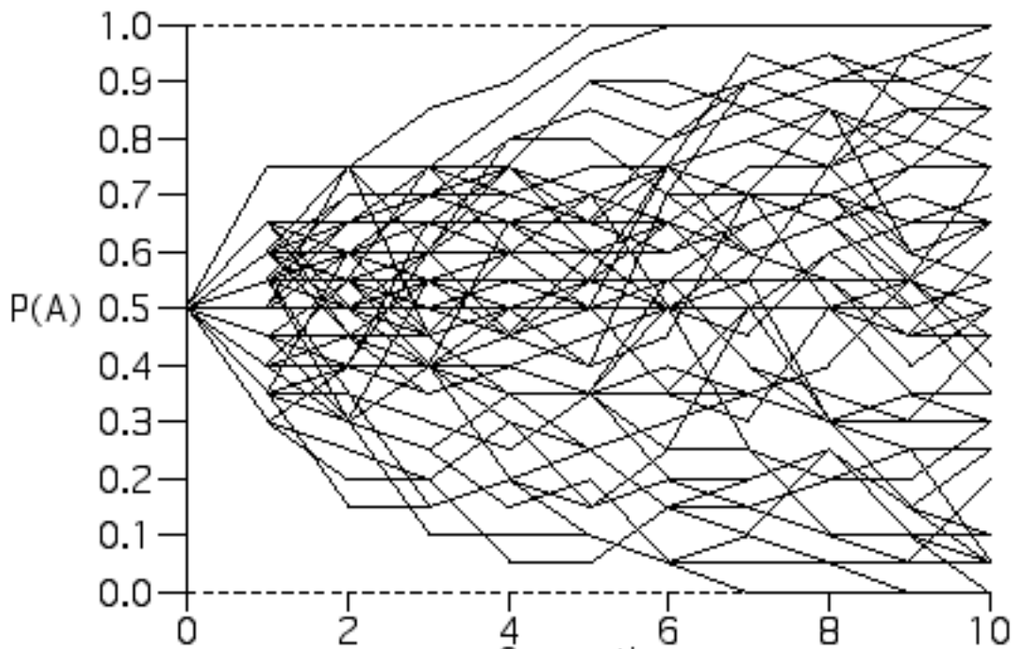
- PCR conditions and setup
- number of cycles and PCRs

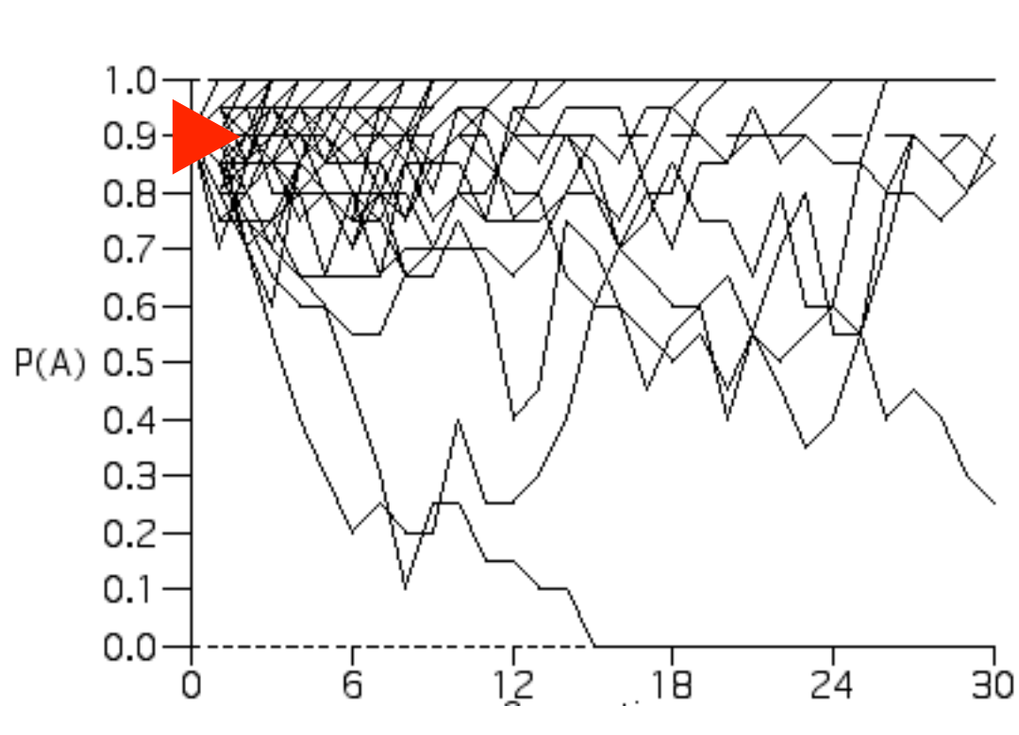
GDC Genetic Diversity Centre Zurich

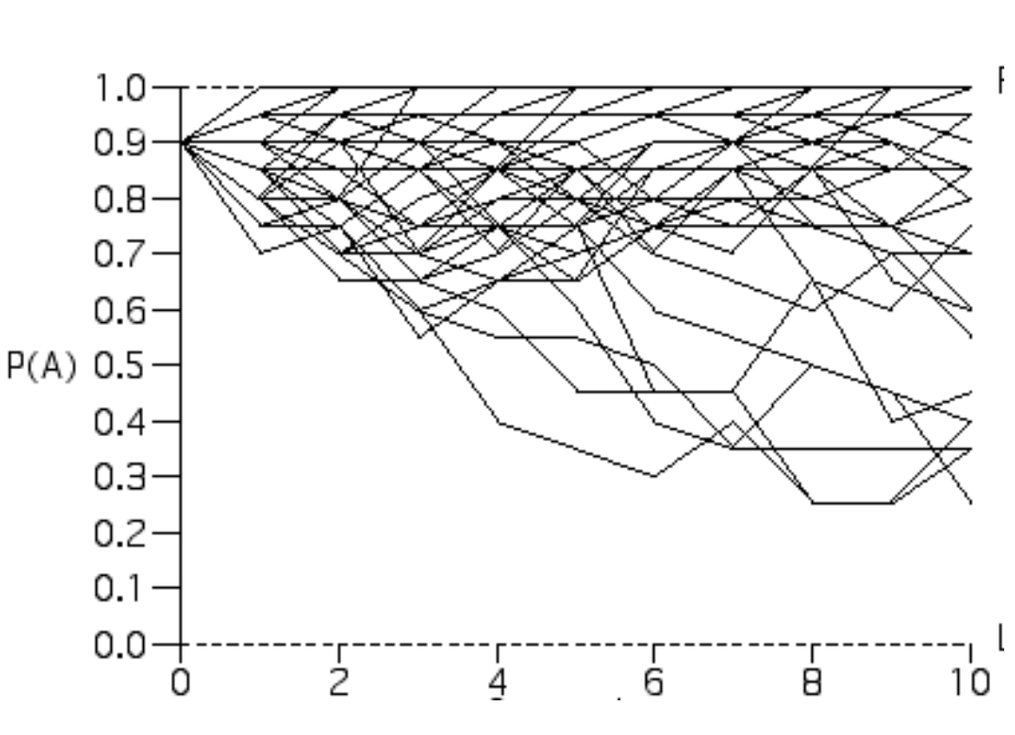
Number of PCR Cycles

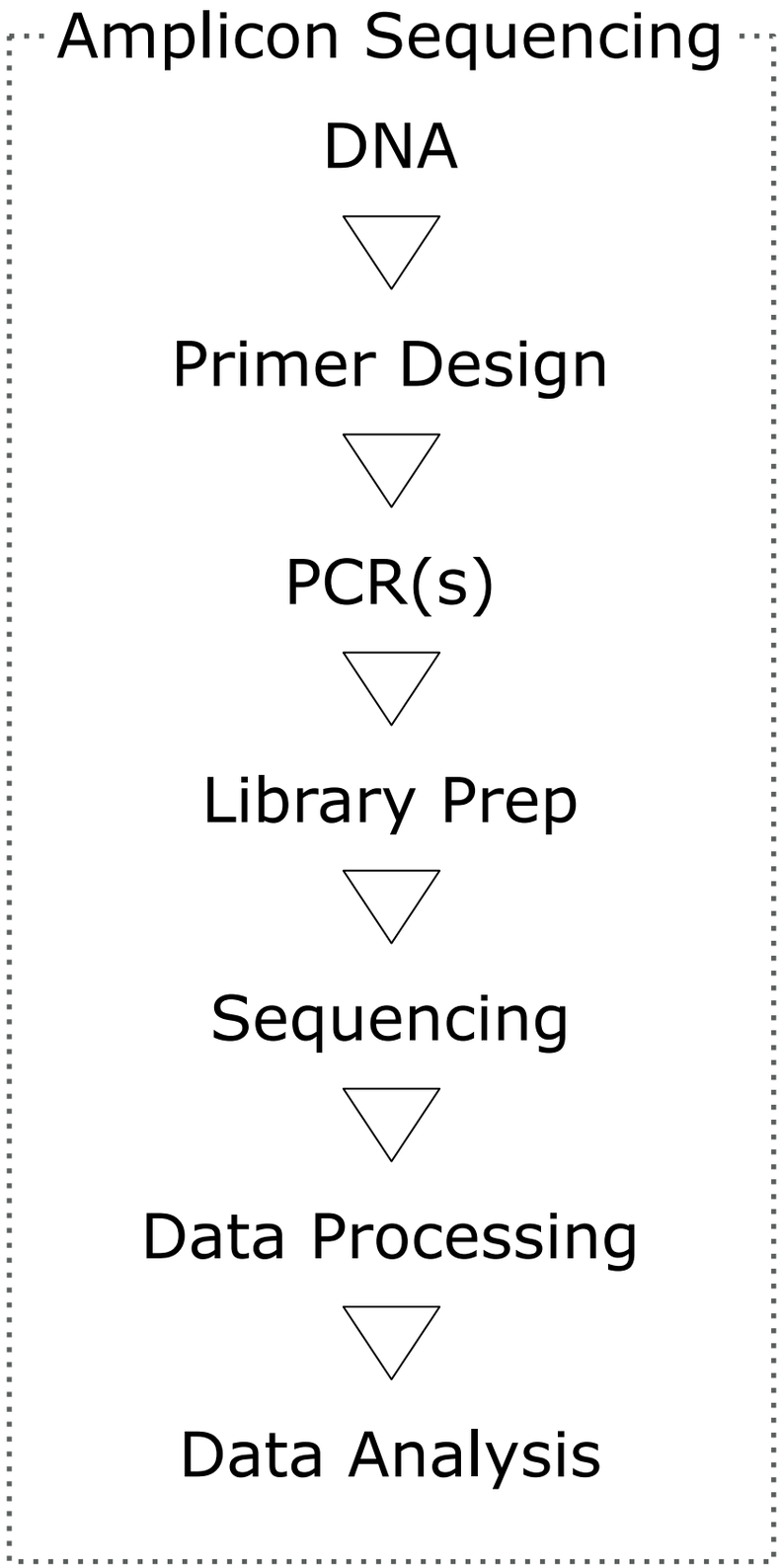












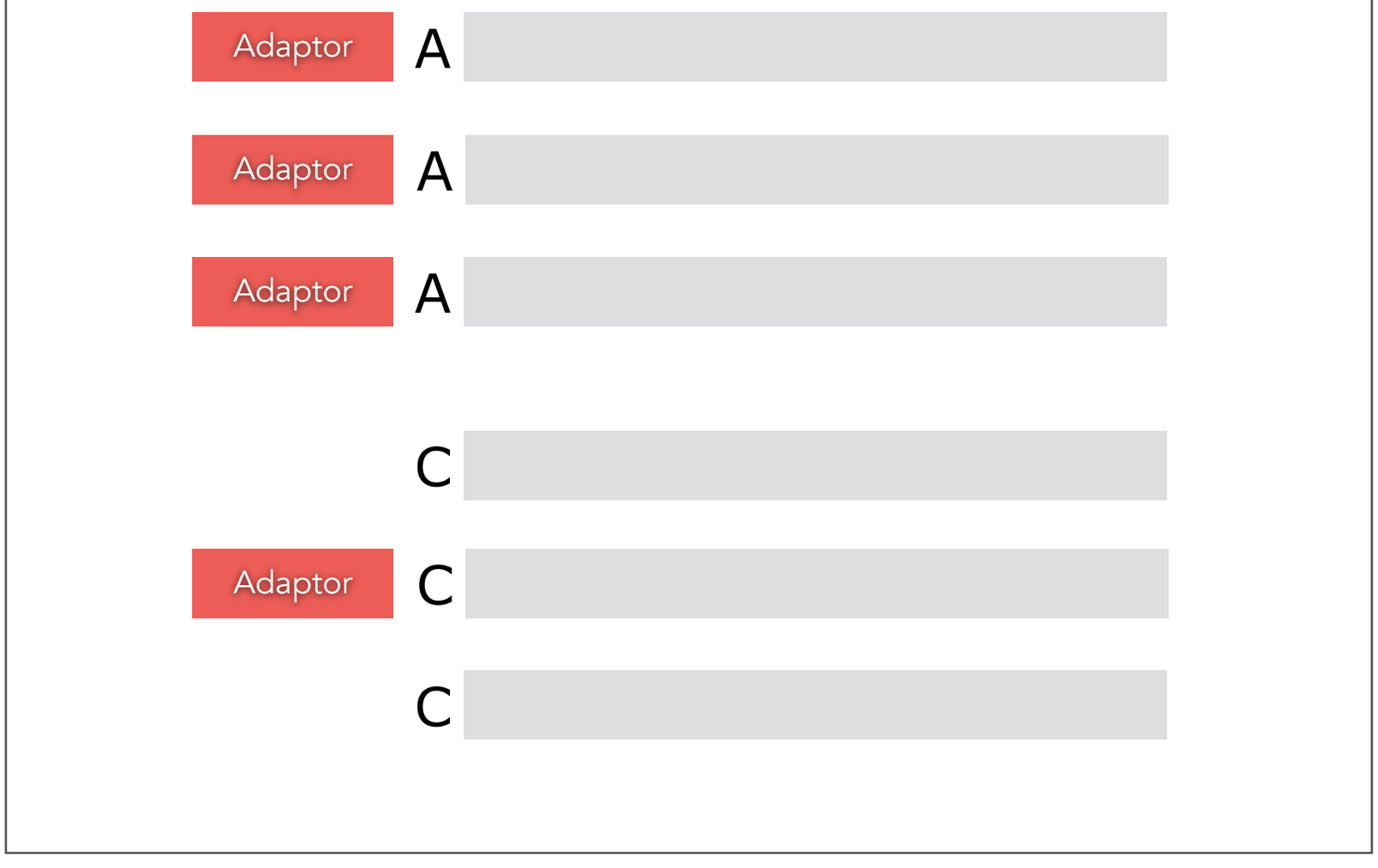
- sampling / sampling design
- extraction method
- contamination

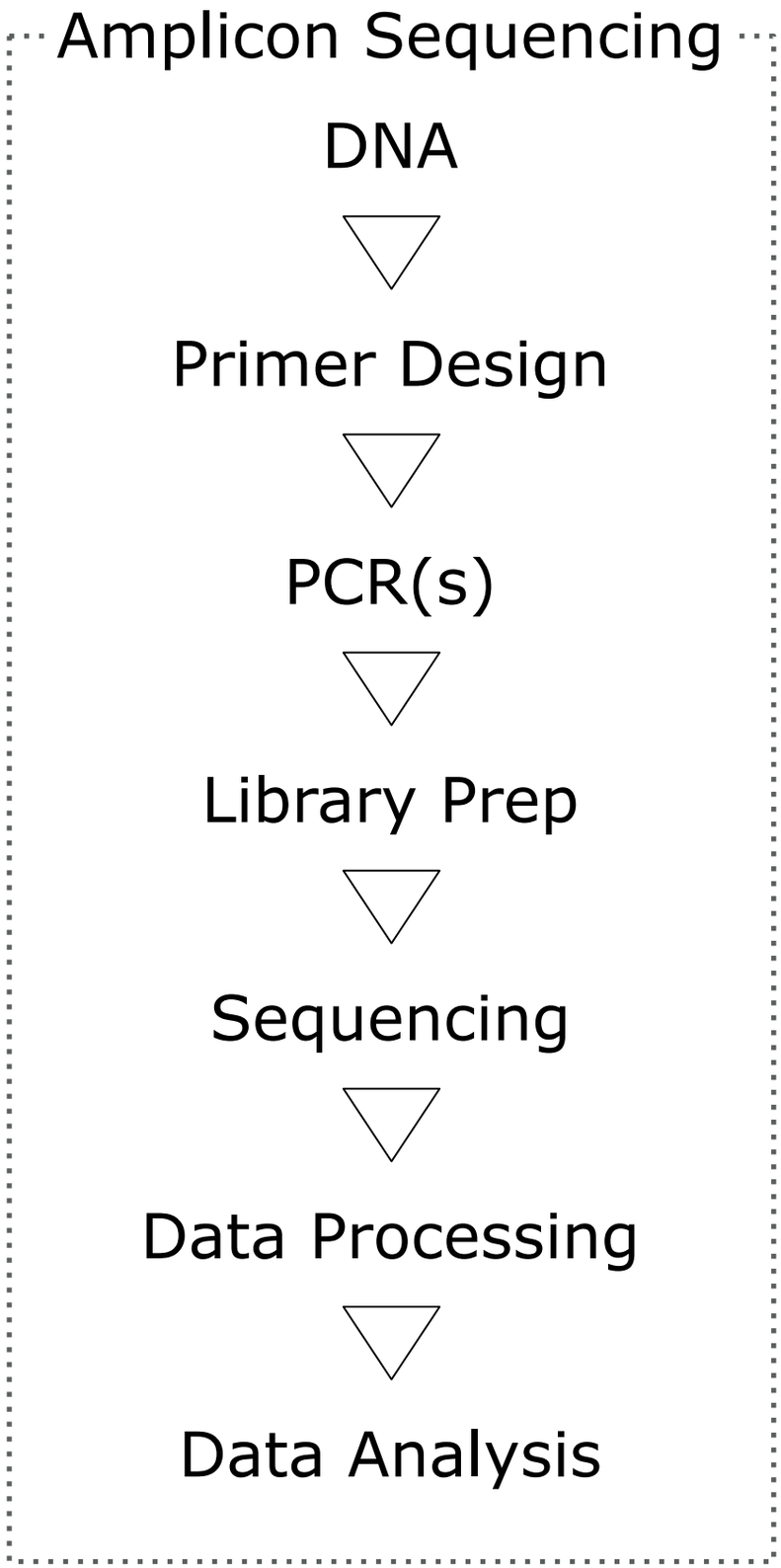
- specificity
- quality

- PCR conditions and setup
- number of cycles and PCRs

- Ligation bias
- Clean-up

Ligation Bias





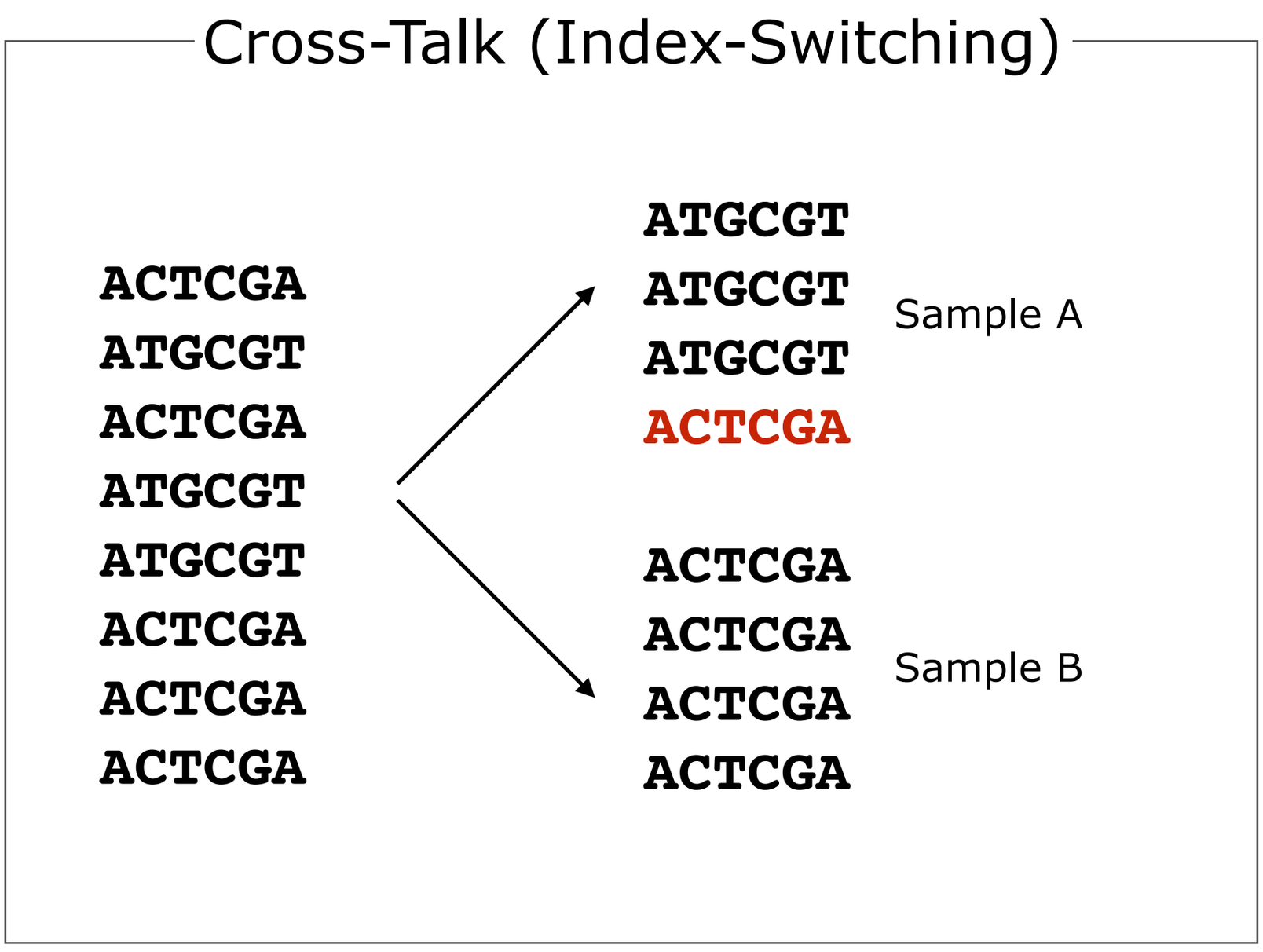
- sampling / sampling design
- extraction method
- contamination

- specificity
- quality

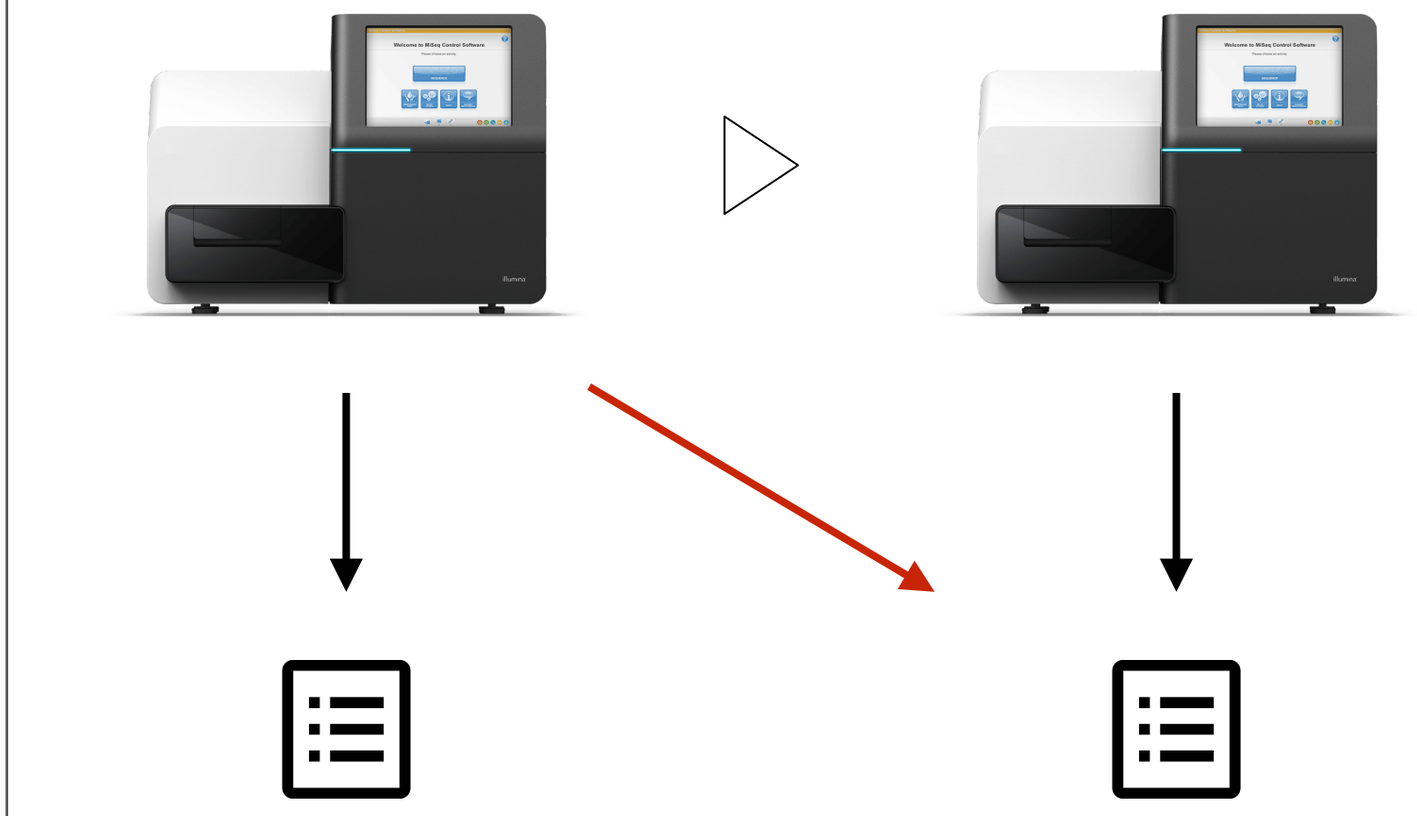
- PCR conditions and setup
- number of cycles and PCRs

- Ligation bias
- Clean-up

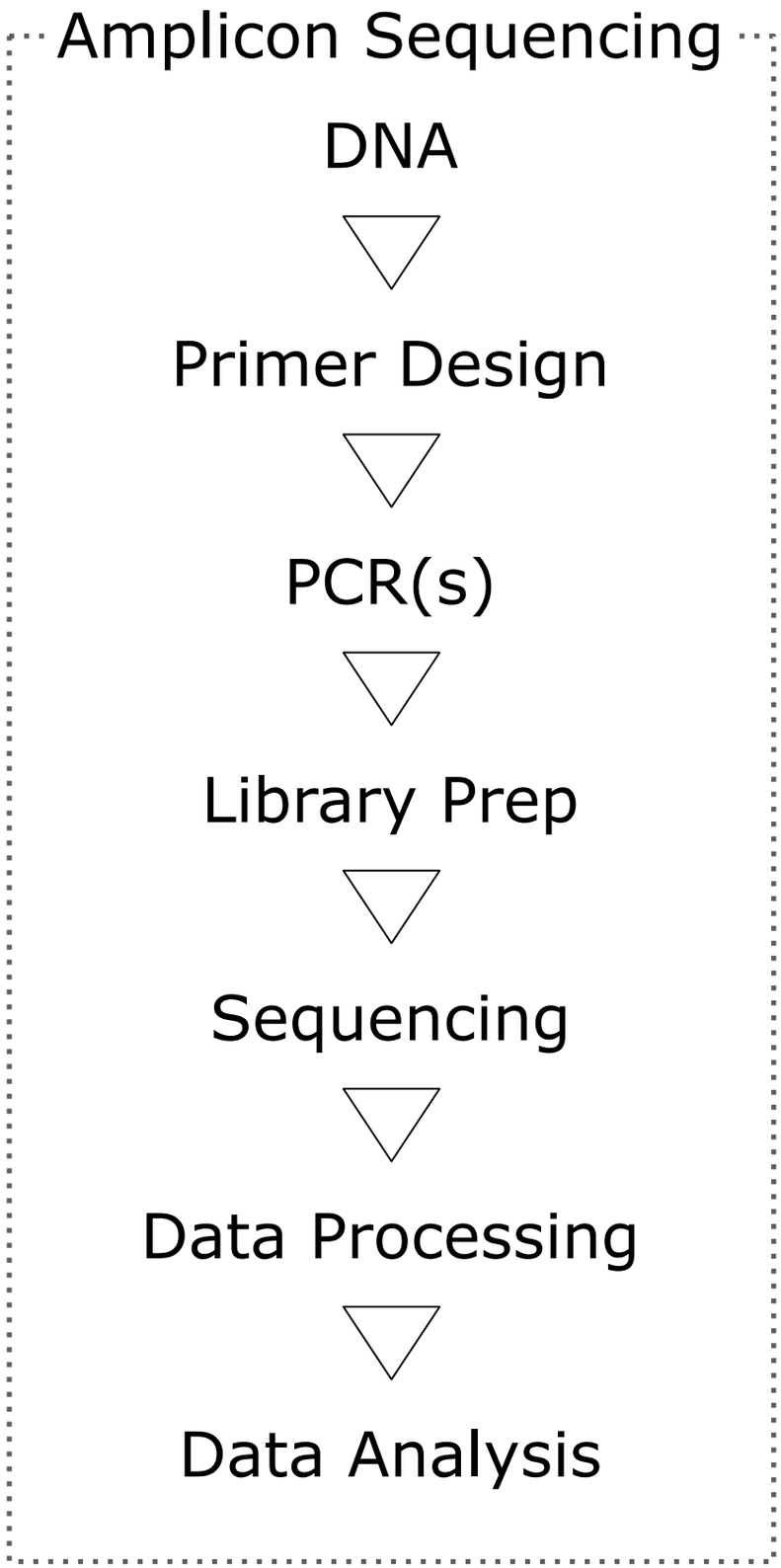
- quality / complexity / depth
- read-length / amplicon-length
- cross-talk, carry-over


Cross-Talk (Index-Switching)
ACTCGA
ATGCGT
ACTCGA
ATGCGT
ATGCGT
ACTCGA
ACTCGA
ACTCGA
ATGCGT
ATGCGT
ATGCGT
ACTCGA
Sample A
ACTCGA
ACTCGA
ACTCGA
ACTCGA
Sample B

Carry-over (Contamination)



GDC Genetic Diversity Centre Zurich



- sampling / sampling design
- extraction method
- contamination

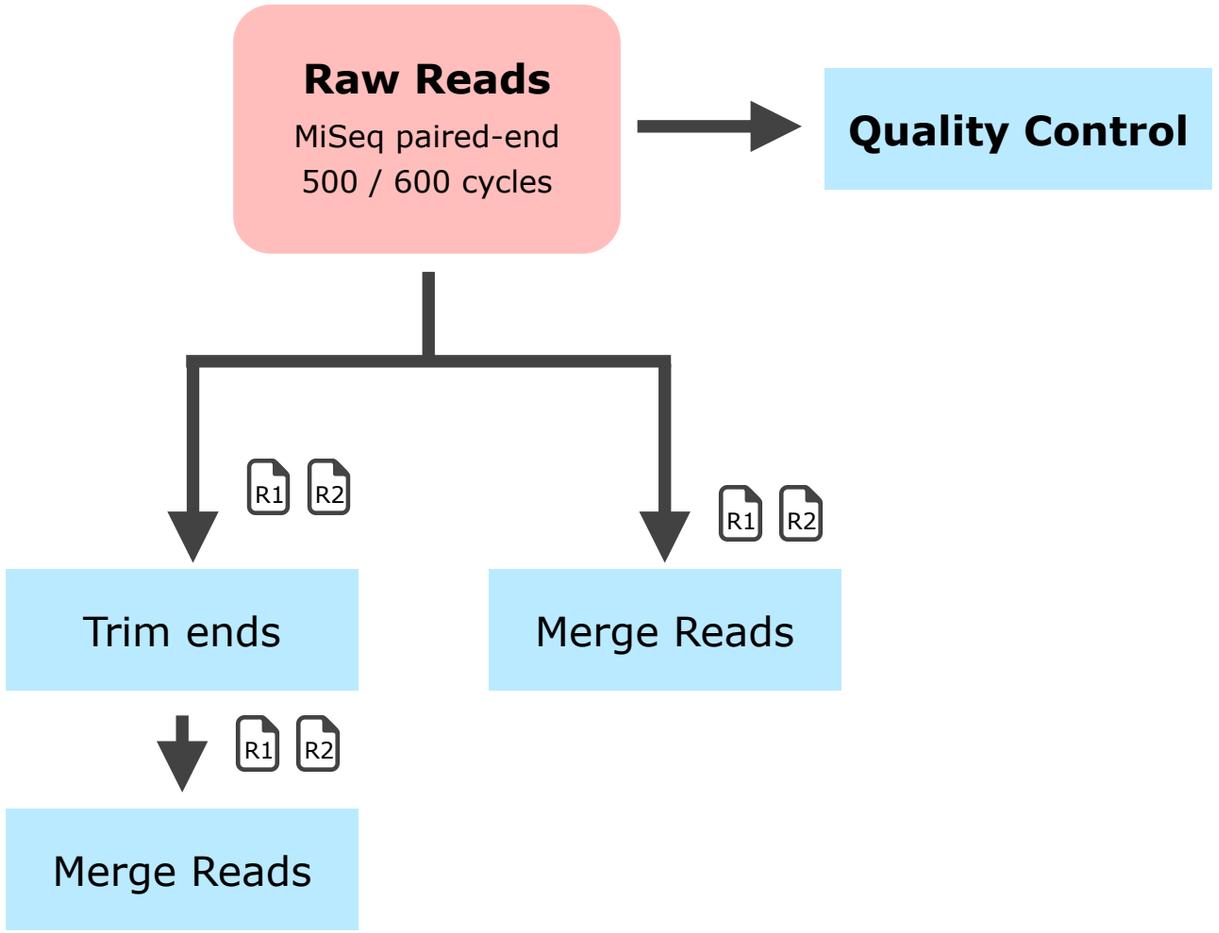
- specificity
- quality

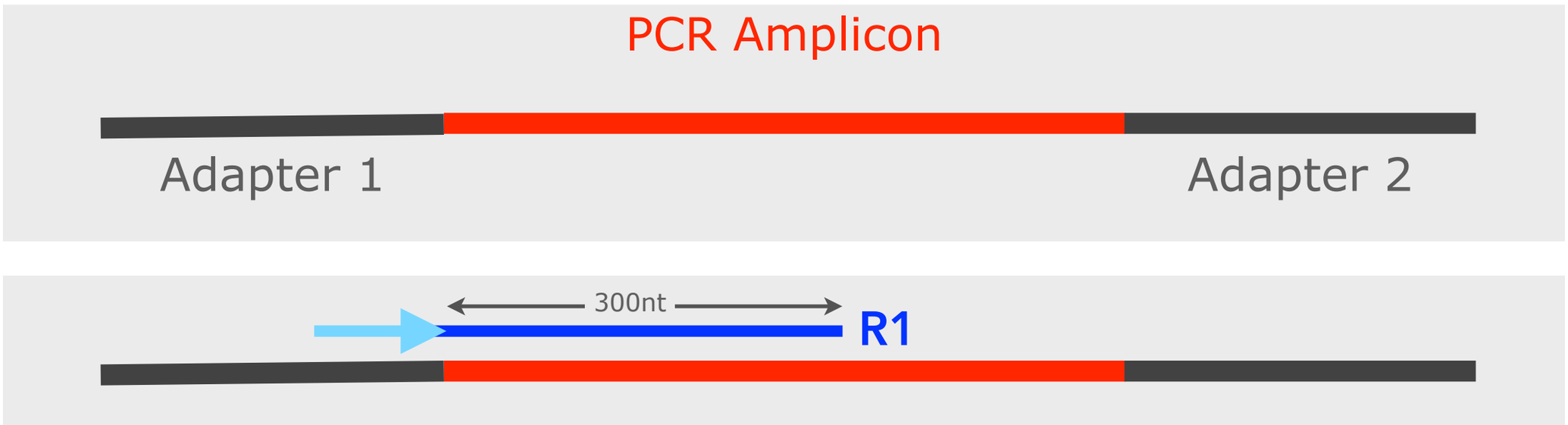
- PCR conditions and setup
- number of cycles and PCRs

- Ligation bias
- Clean-up

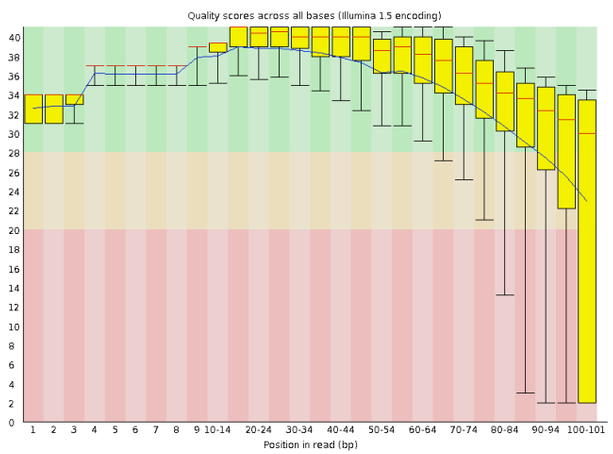
- quality / complexity / depth
- read-length / amplicon-length
- cross-talk, carry-over

- quality filtering
- error removal (e.g. chimeras, sequencing errors)
- diversity

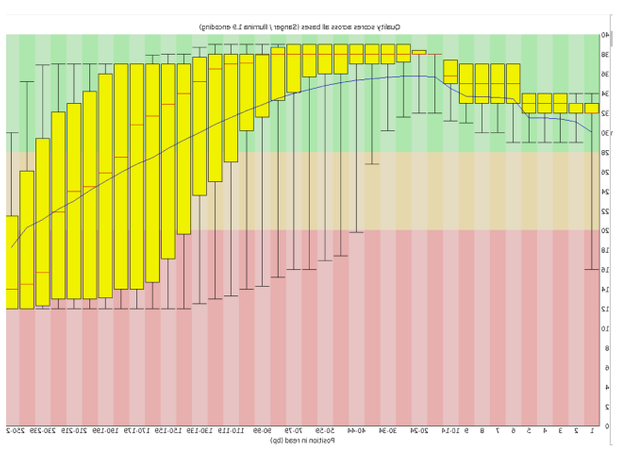

Raw Reads
MiSeq paired-end
500 / 600 cycles
Quality Control
R1
R2
R1
R2
Trim ends
Merge Reads
R1
R2
Merge Reads


PCR Amplicon
Adapter 1
Adapter 2
300nt
R1
300nt
R2

R1

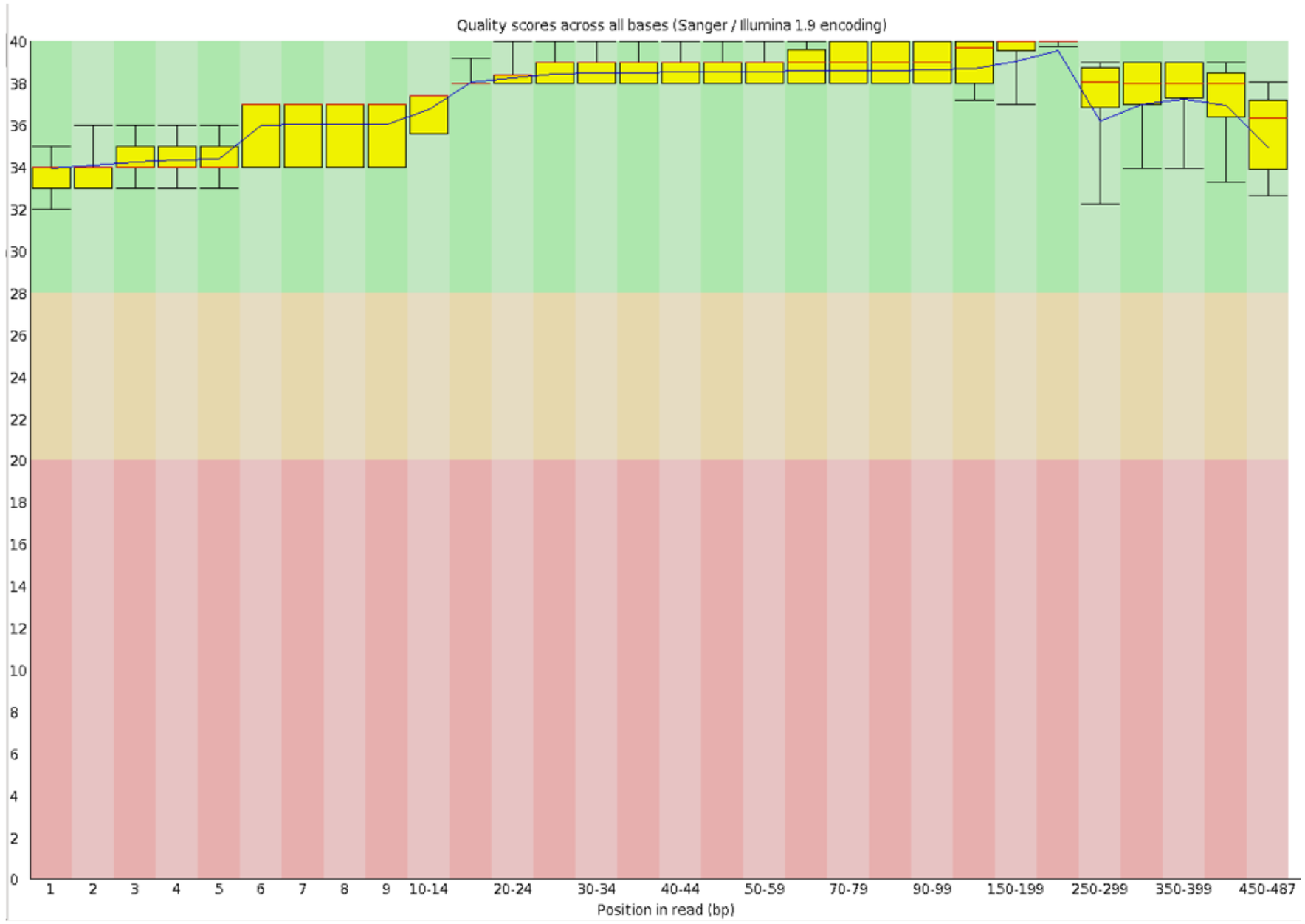

Quality scores across all bases (Illumina 1.5 encoding)
Position in read (bp)

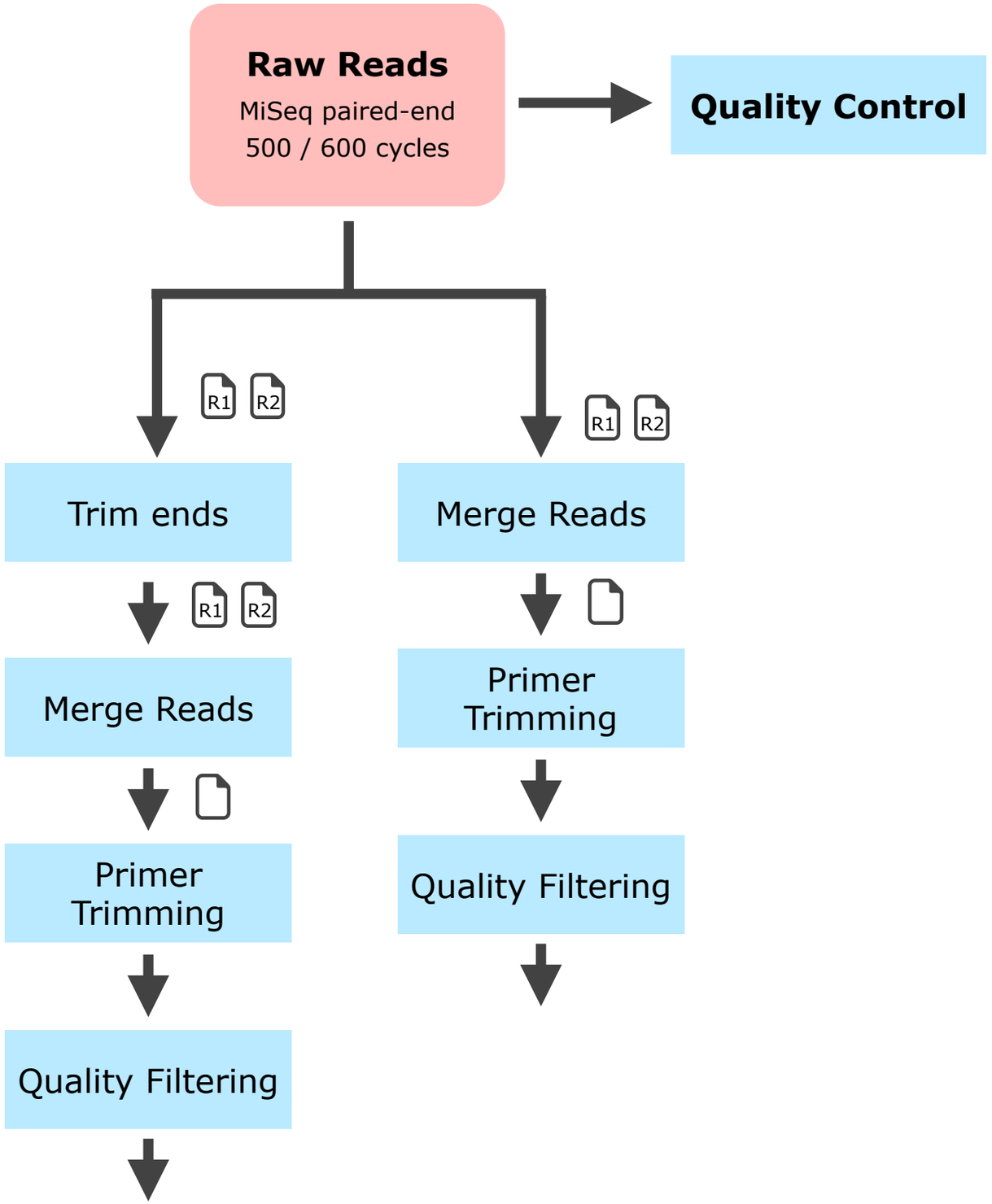
Q(R1) > Q(R2)

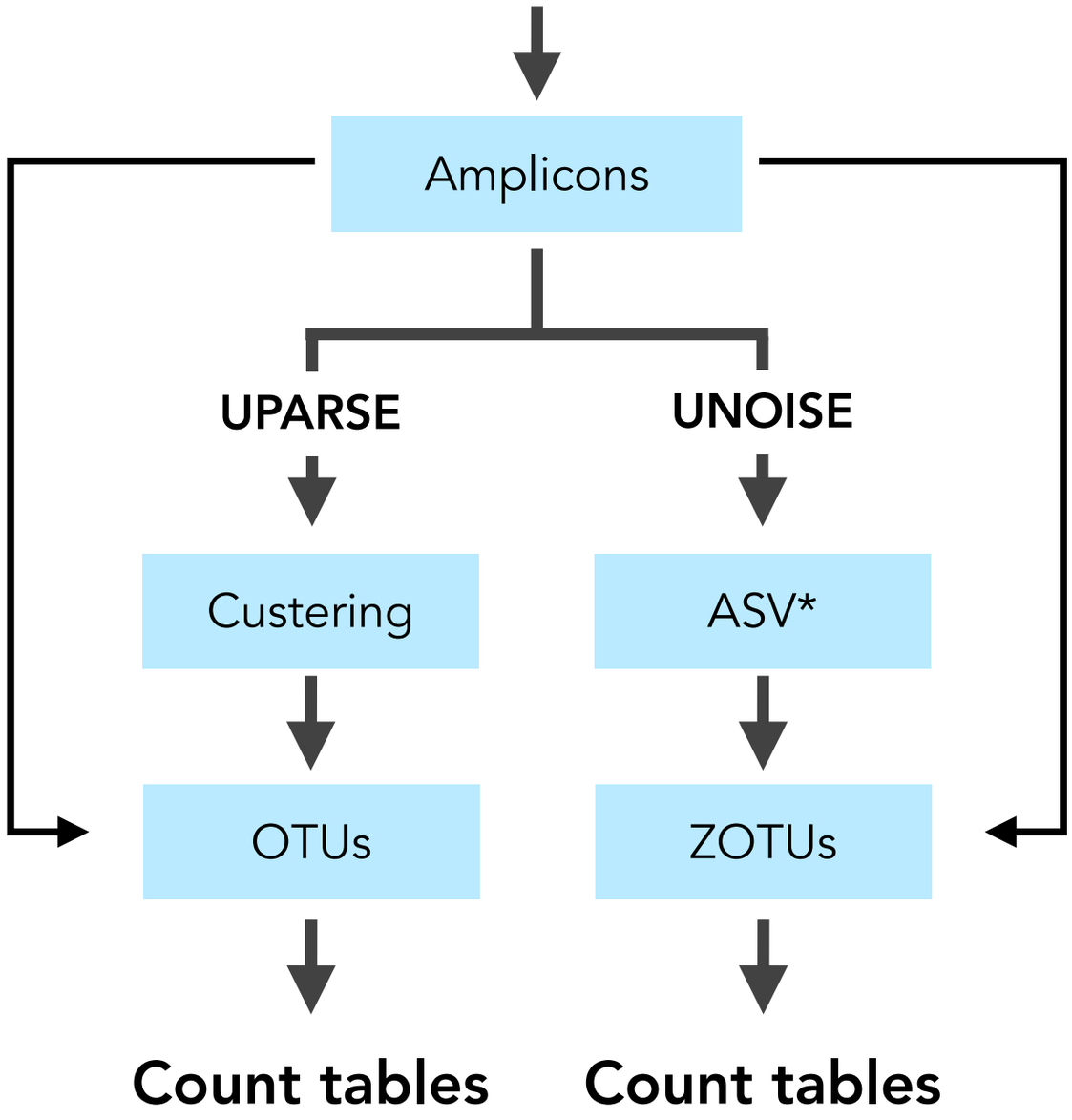


R2

Quality scores for merged paired-end reads




Raw Reads
MiSeq paired-end
500 / 600 cycles
Quality Control
R1
R2
R1
R2
Trim ends
Merge Reads
Merge Reads
Primer Trimming
Primer Trimming
Quality Filtering
Quality Filtering



* ASV: Amplicon Sequence Variant

Chimera Formation

```
SpeciesA: ACCGGGCATACGTAGCAACAACGGTGTATA
          |||   ||||||||||||||||   ||  |  |  |
SpeciesB: ACCATGCATACGTAGCAACTTCGCTCTTTC
```

Chimera Formation

```
SpeciesA: ACCGGGCATACGTAGCAACAACGGTGTATA
          |||   ||||||||||||||||   ||  |  |  |
SpeciesB: ACCATGCATACGTAGCAACTTCGCTCTTTC
```

PCR

Primer

TGGCCC→

ACCGGGCATACGTAGCAACAACGGTGTATA

TGG CCC

ACCATGCATACGTAGCAACTTCGCTCTTTC

GDC Genetic Diversity Centre Zurich

TGGCCCGTATGCAT➜
ACCGGGCATACGTAGCAACAACGGTGTATA

Extention

TGGCCCGTATGCAT

ACCGGGCATACGTAGCAACAACGGTGTATA

Denaturation

TGGCCCGTATGCAT➜
ACCATGCATACGTAGCAACTTCGCTCTTTC

ACCGGGCATACGTAGCAACTTCGCTCTTTC

PCR / Sequencing Errors

ACCATGCATACGTAGCAACA

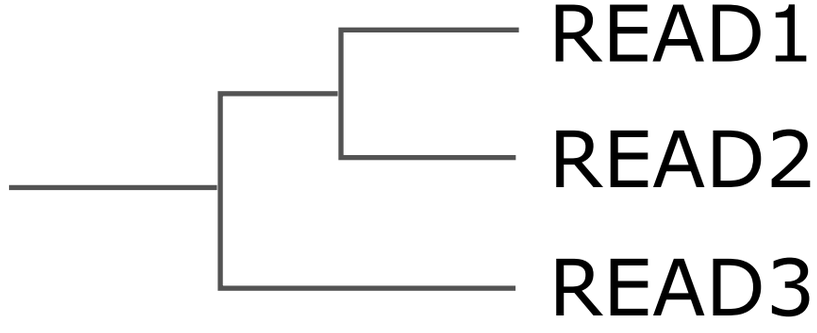
▽ PCR

ACCATGCATACG**A**AGCAACA

▽ Sequencing

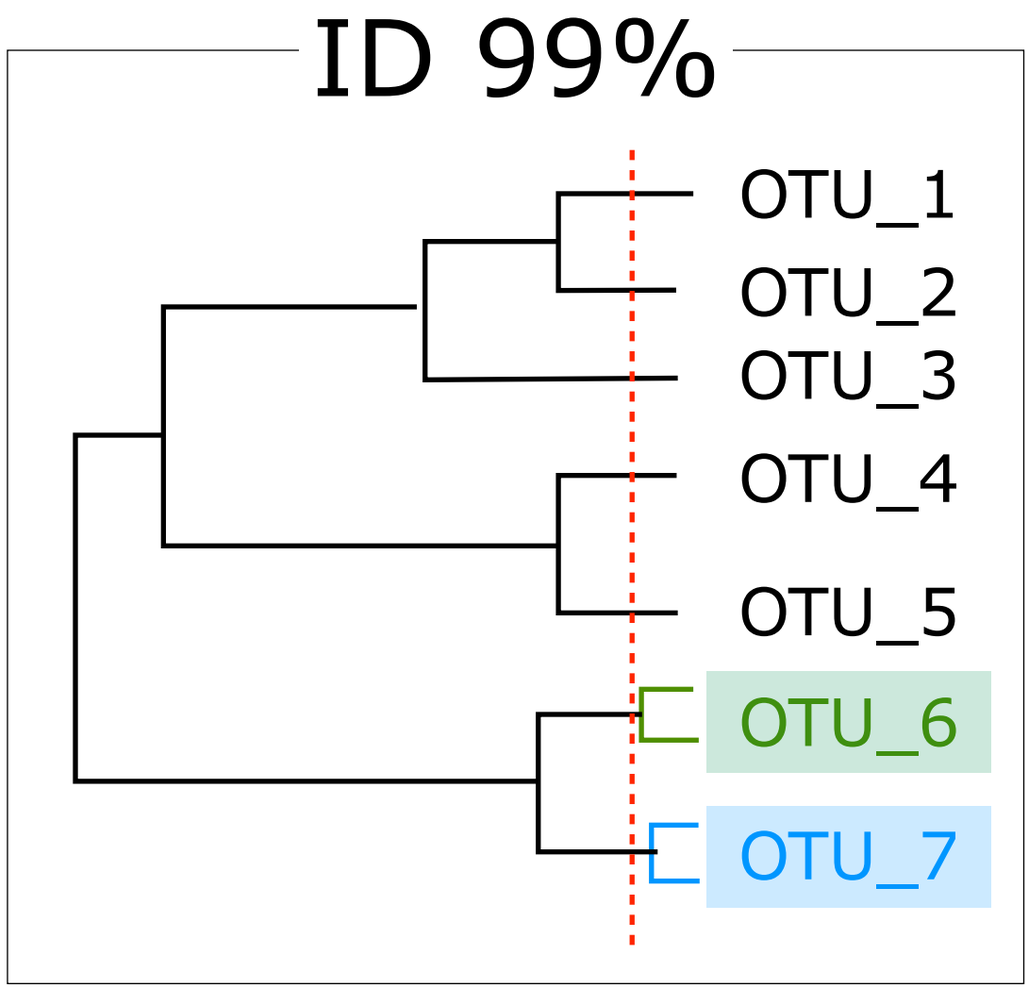
ACCAT**C**CATACG**A**AGCAACA

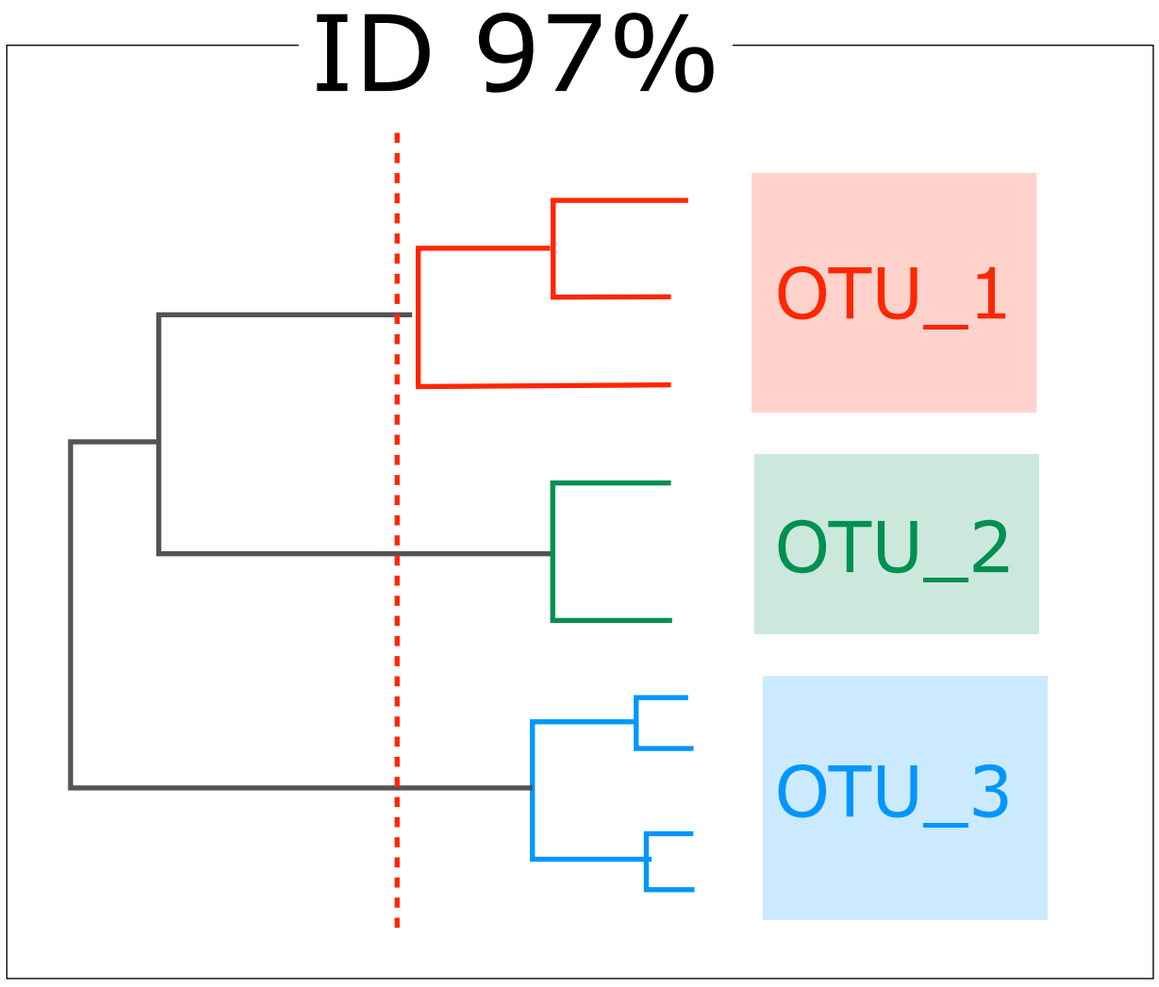
```
READ1: ACCATGCAAAACGTAGCCAACA
READ2: ACCATGCAAAACGTAGCCAACA
READ3: ACCATGCATTACGTAGGGAACA
```

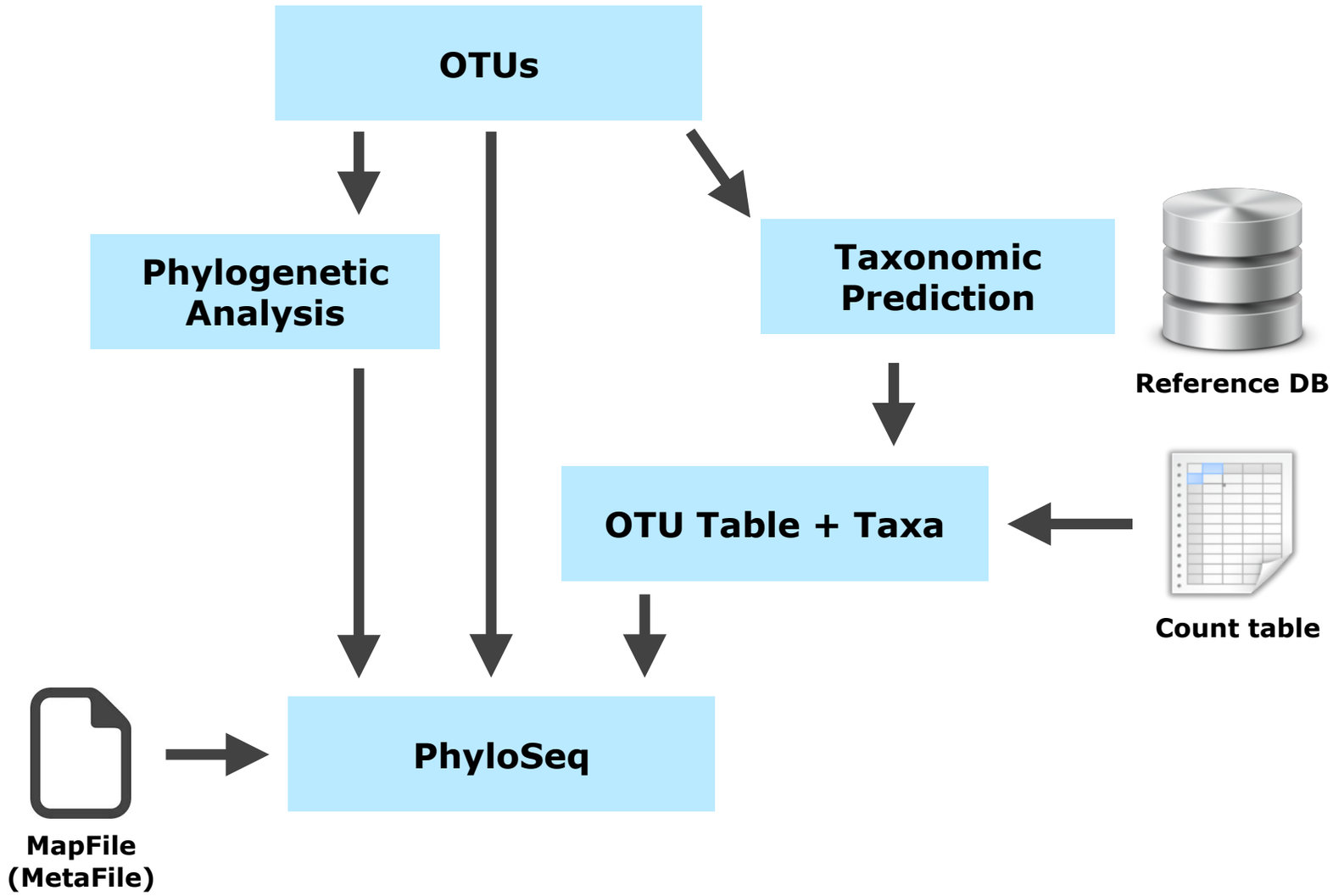


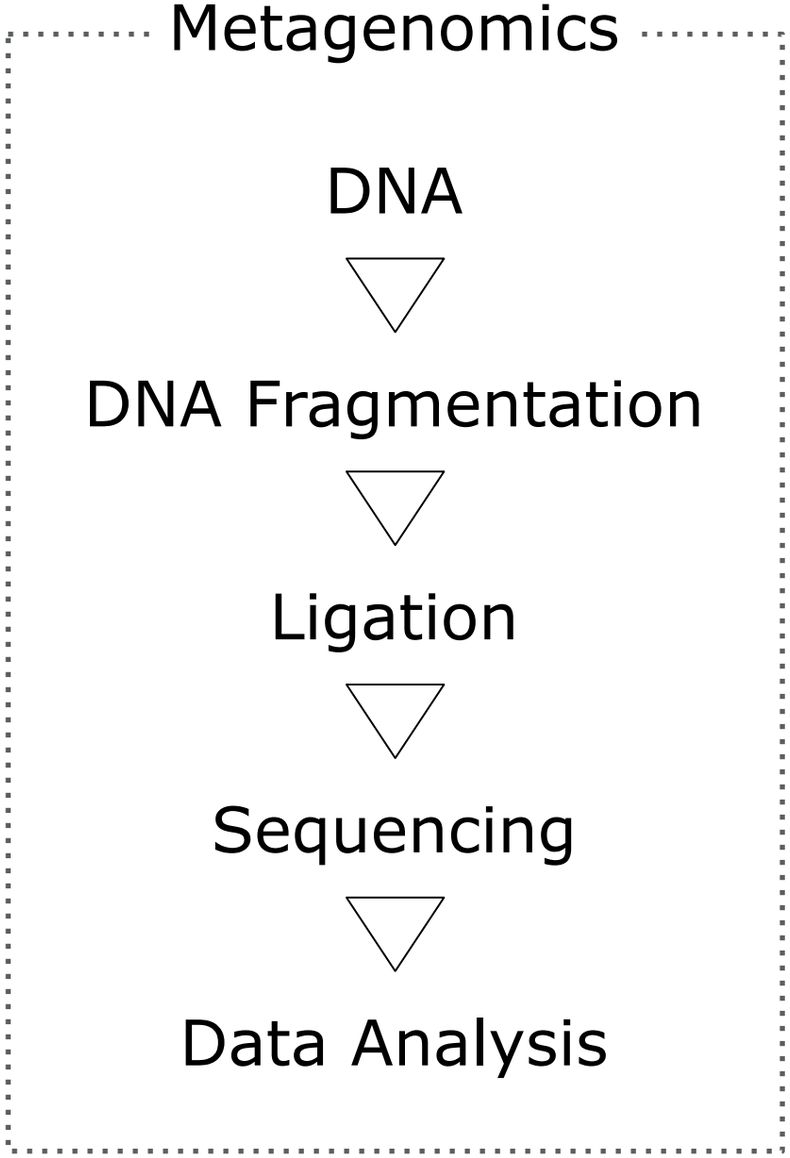
=> abundance filtering

=> cluster threshold


ID 99%
OTU_1
OTU_2
OTU_3
OTU_4
OTU_5
OTU_6
OTU_7


ID 97%
OTU_1
OTU_2
OTU_3


OTUs
Phylogenetic Analysis
Taxonomic Prediction
Reference DB
OTU Table + Taxa
Count table
PhyloSeq
MapFile (MetaFile)



- sampling / sampling design
- extraction method
- contamination

- Fragmentation bias

- Ligation bias
- Clean-up

- quality / complexity / depth
- read-length / amplicon-length
- cross-talk, carry-over

- quality filtering
- error removal (e.g. chimeras, sequencing errors)
- diversity
- assembly errors
- assignment errors

GDC Genetic Diversity Centre Zurich

- Woese CR, Fox GE (**1977**). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms". PNAS. 74 (11): 5088–90.
- Weisburg WG, Barns SM, Pelletier DA, Lane DJ (**1991**). "16S ribosomal DNA amplification for phylogenetic study". Journal of Bacteriology. 173 (2): 697–703.
- Schmidt TM, Relman DA (**1994**). Phylogenetic identification of uncultured pathogens using ribosomal RNA sequences. Methods in Enzymology. 235. pp. 205–22.
- Gray JP, Herwig RP (**1996**). "Phylogenetic analysis of the bacterial communities in marine sediments". Applied and Environmental Microbiology. 62 (11): 4049–59.
- Coenye T, Vandamme P (**2003**). "Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes". FEMS Microbiology Letters. 228 (1): 45–9.
- Clarridge JE (**2004**). "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases". Clinical Microbiology Reviews. 17 (4): 840–62.
- Jay ZJ, Inskeep WP (**2015**). "The distribution, diversity, and importance of 16S rRNA gene introns in the order Thermoproteales". Biology Direct. 10 (35): 35.
- Tsukuda M, Kitahara K, Miyazaki K (**2017**). "Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16 S rRNAs". Scientific Reports. 7 (1): 9993.