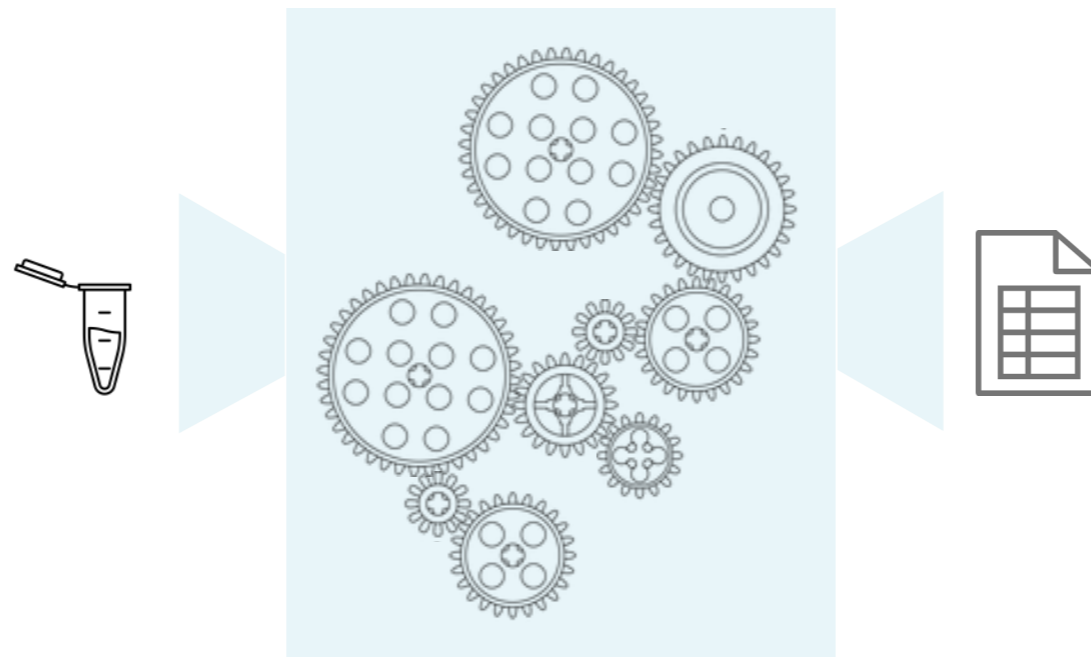
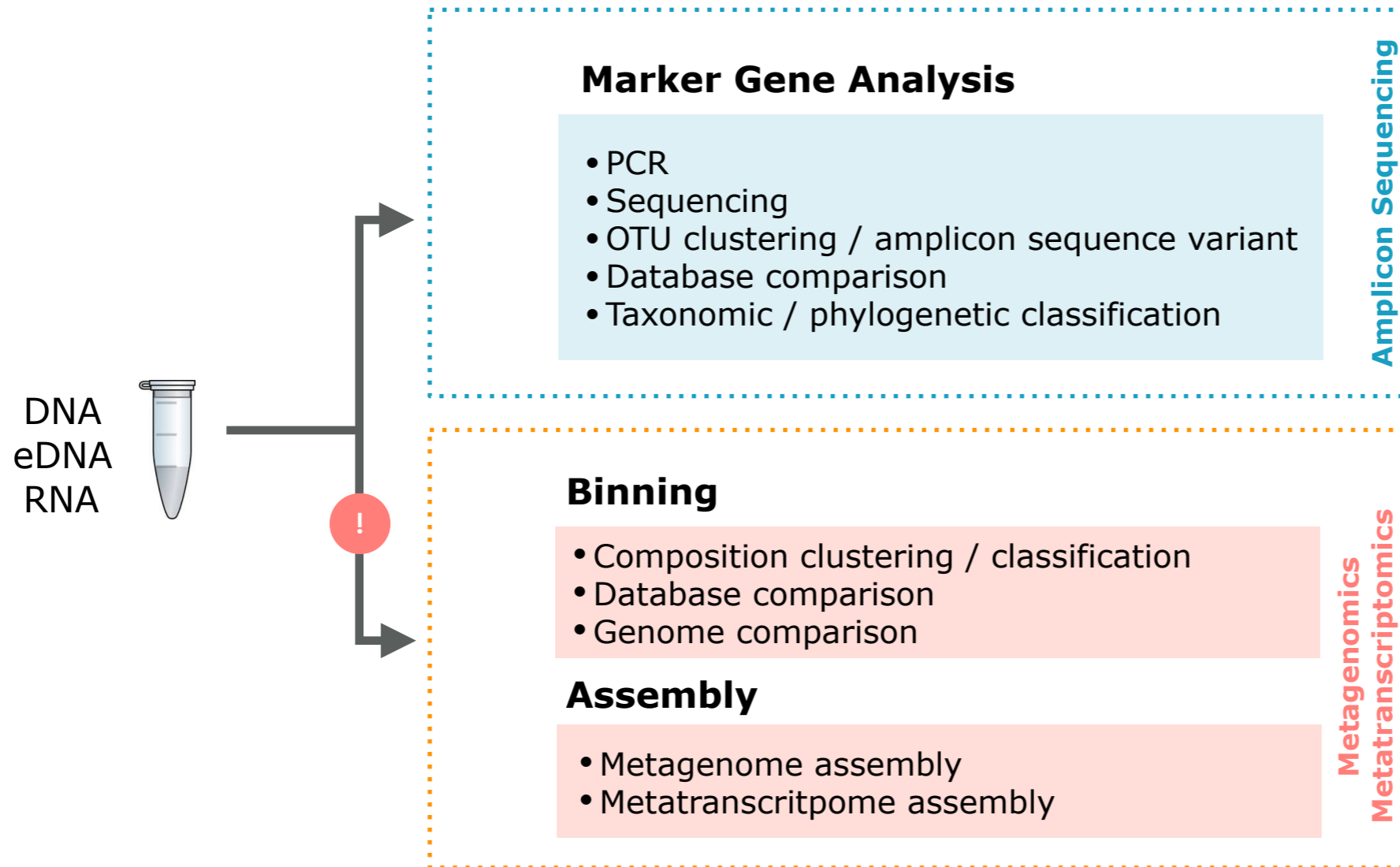



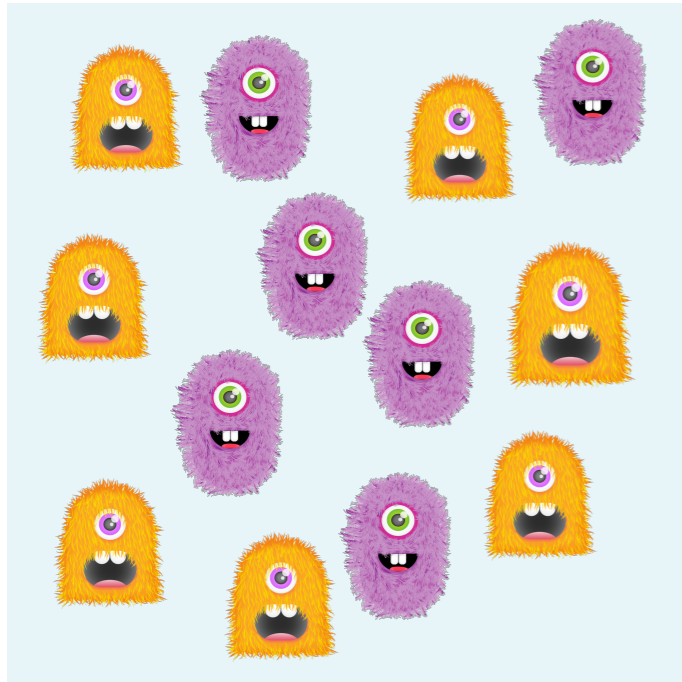
Workflow(s)



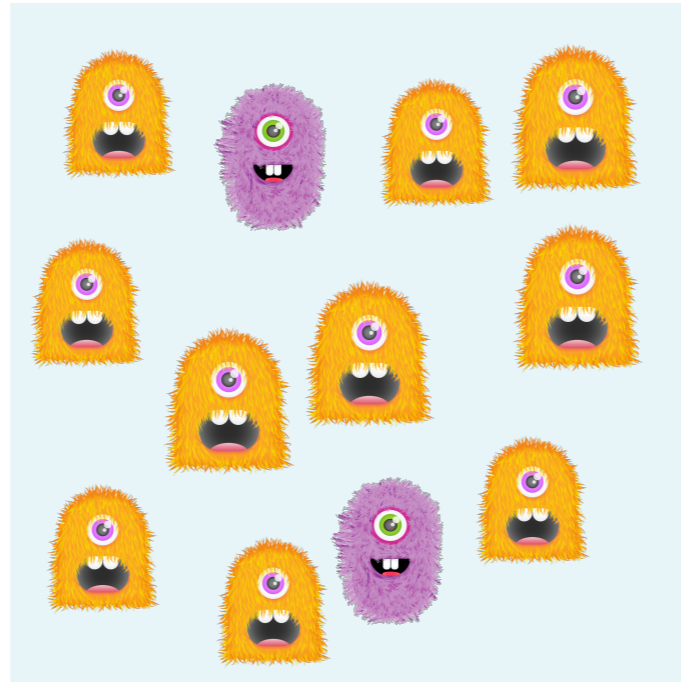


 Please think! Cleaning and/or filtering your raw data might save you some troubles.

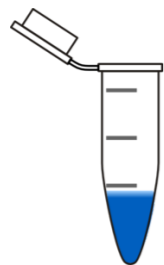
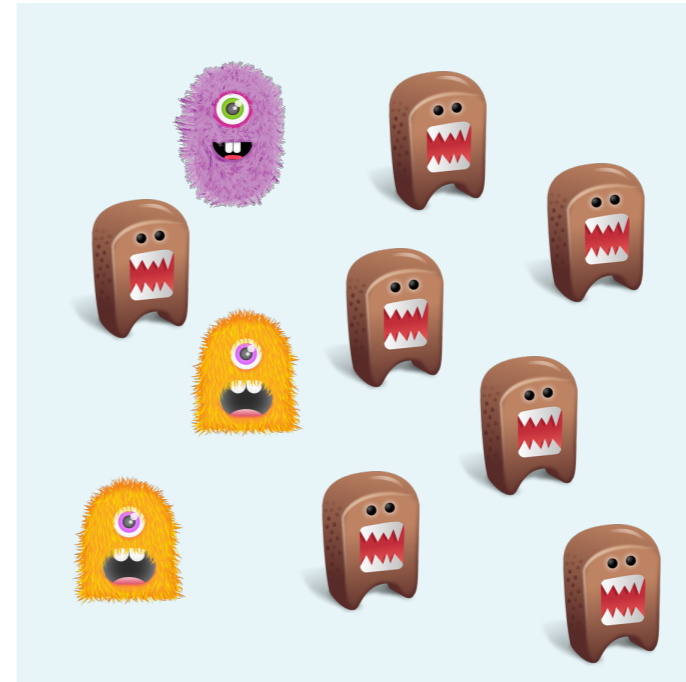
SampleA

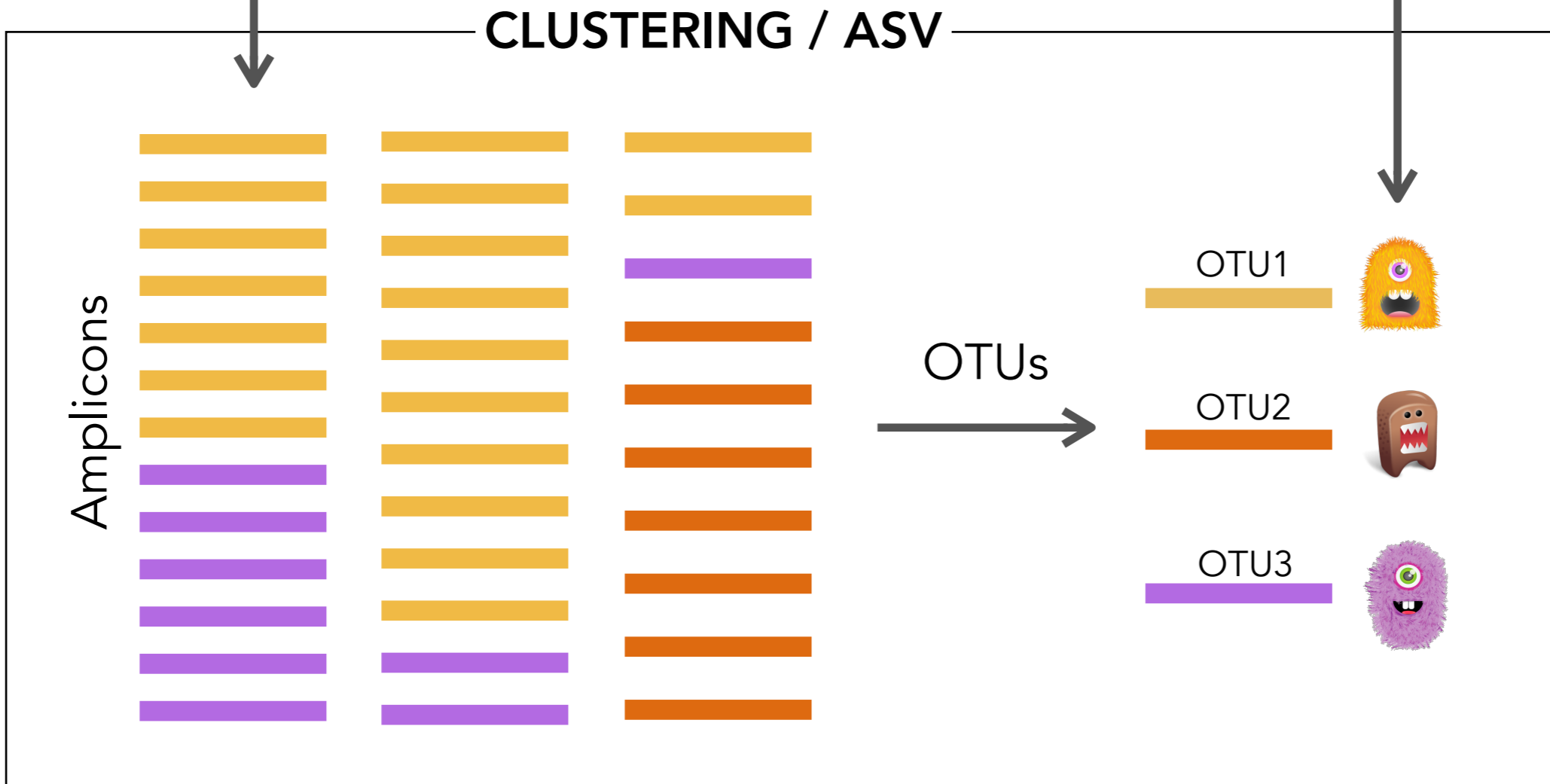
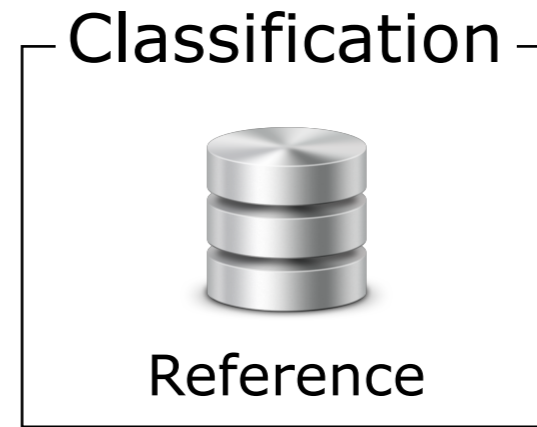
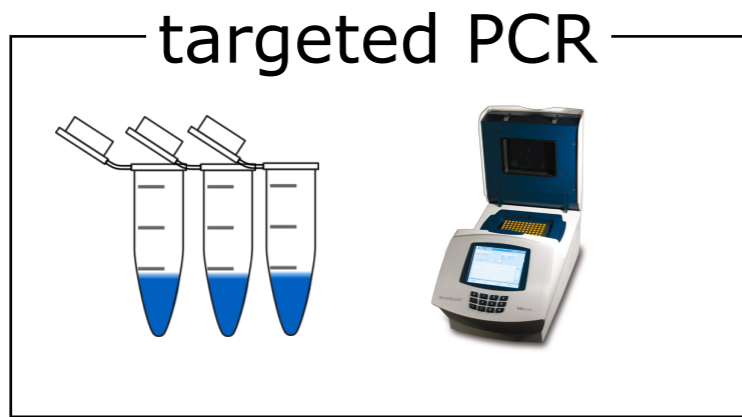


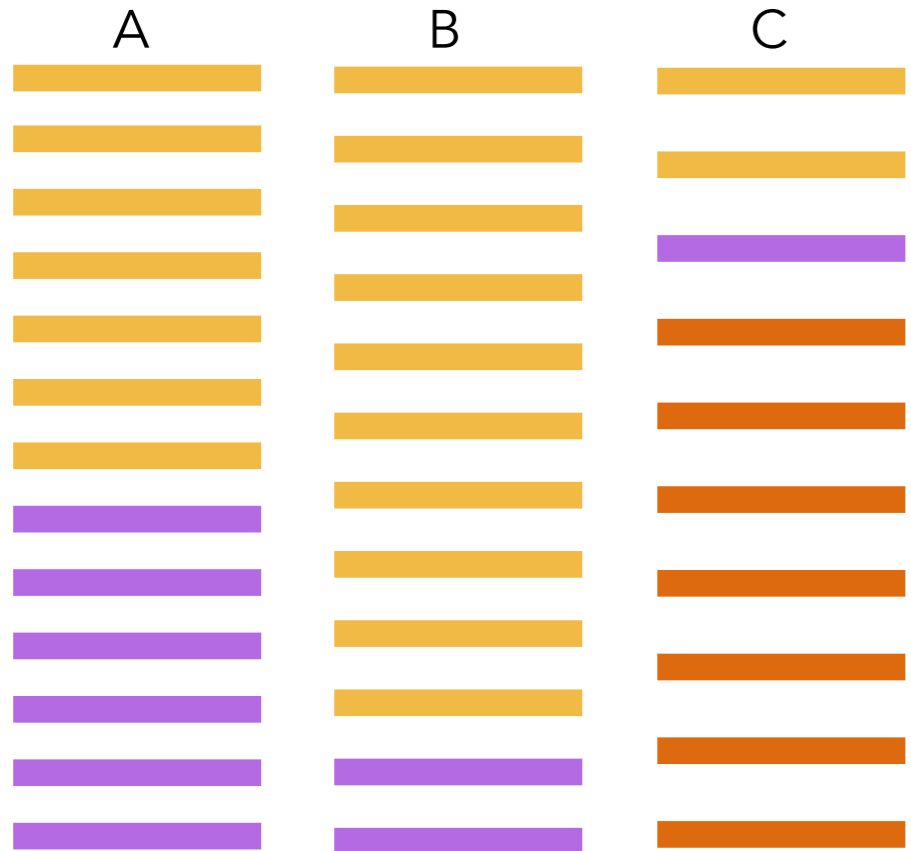
SampleB






SampleC





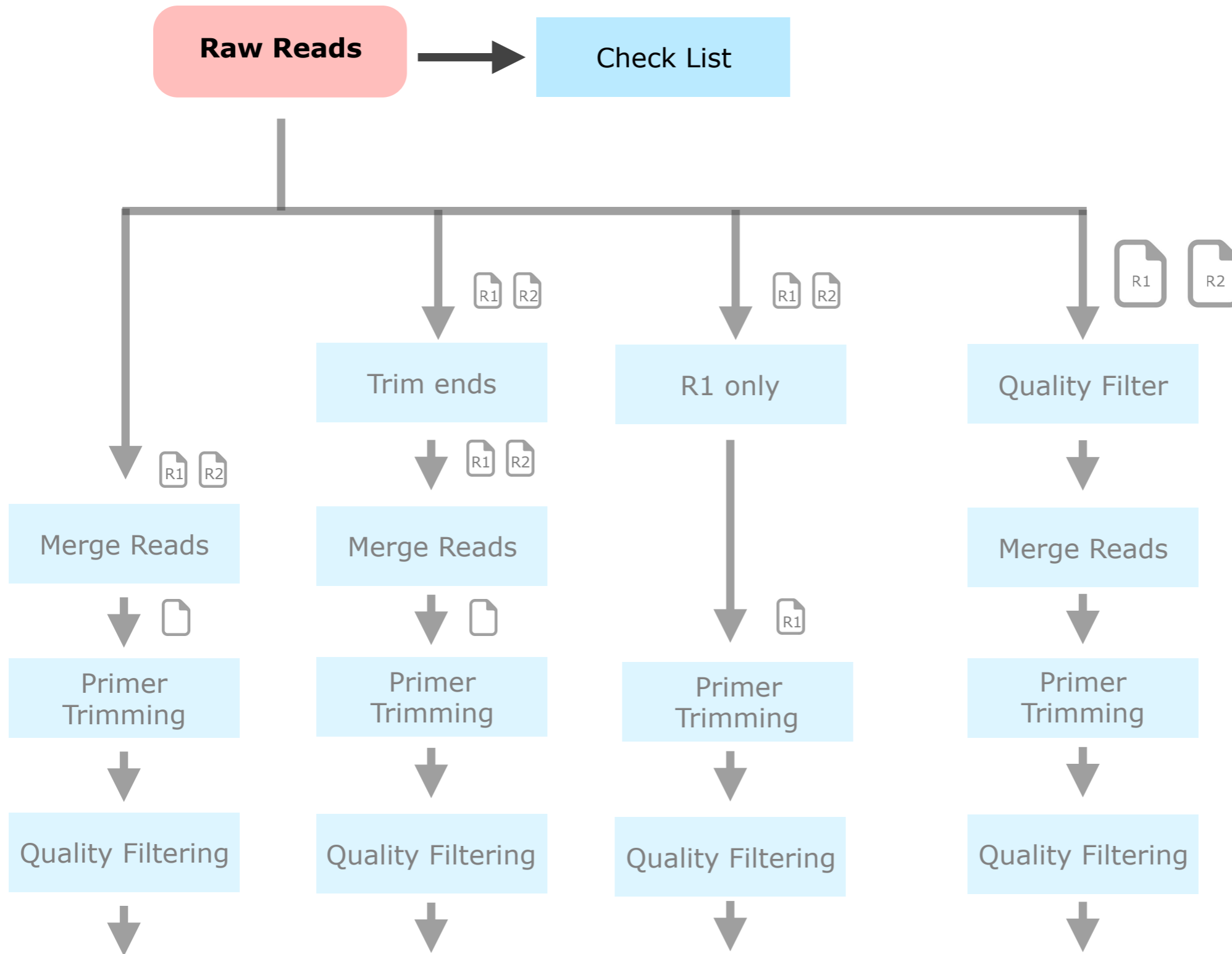


Count Table

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

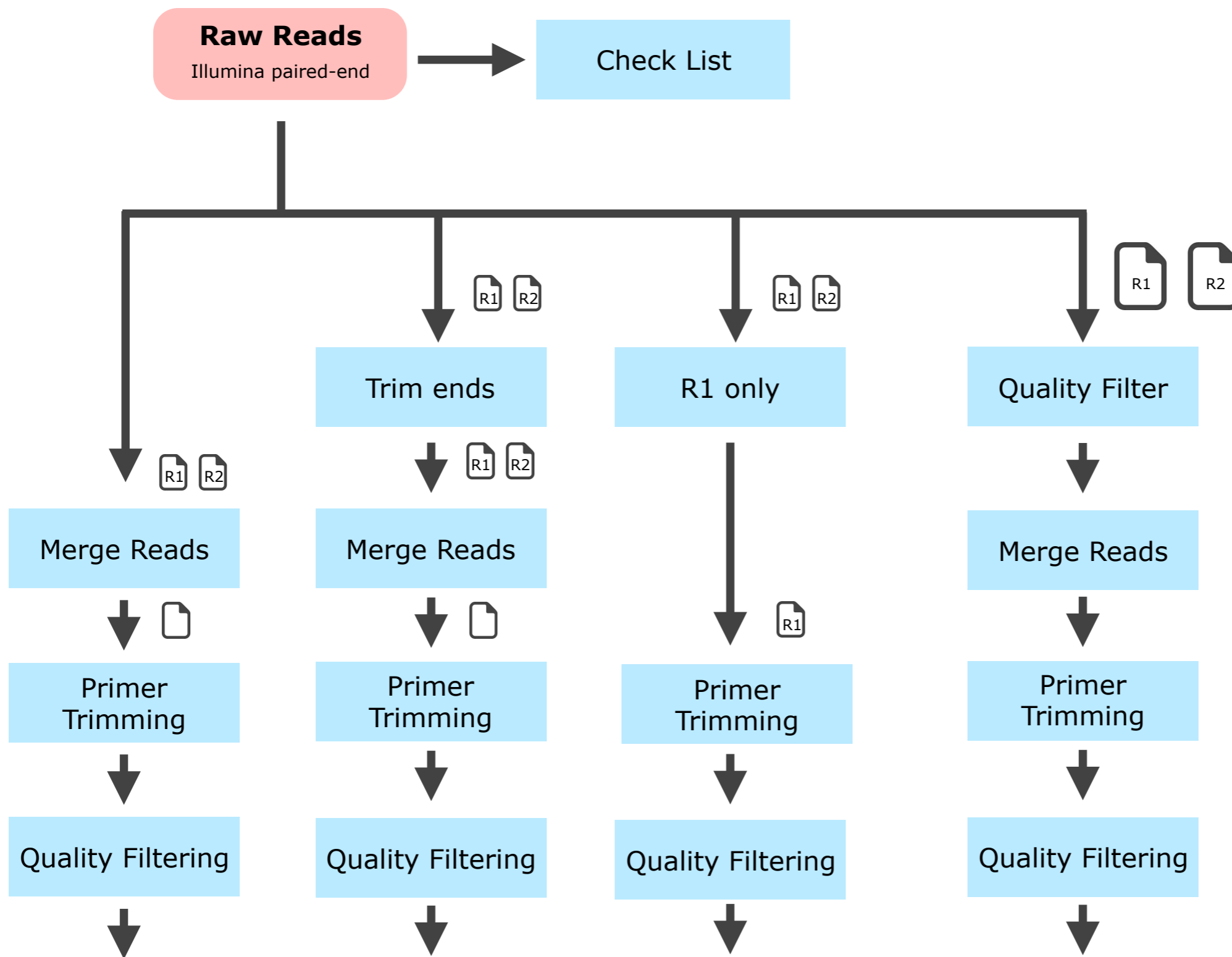


- 🎬 Measuring diversity by 16S sequencing (14 min)
- 🎬 Preparing NGS reads for OTU and denoising analysis (7 min)
- 🎬 Taxonomy reference databases for 16S (17 min)
- 🎬 16S taxonomy and sequence identity (8 min)
- 🎬 Taxonomy prediction methods for 16S sequences (10 min)

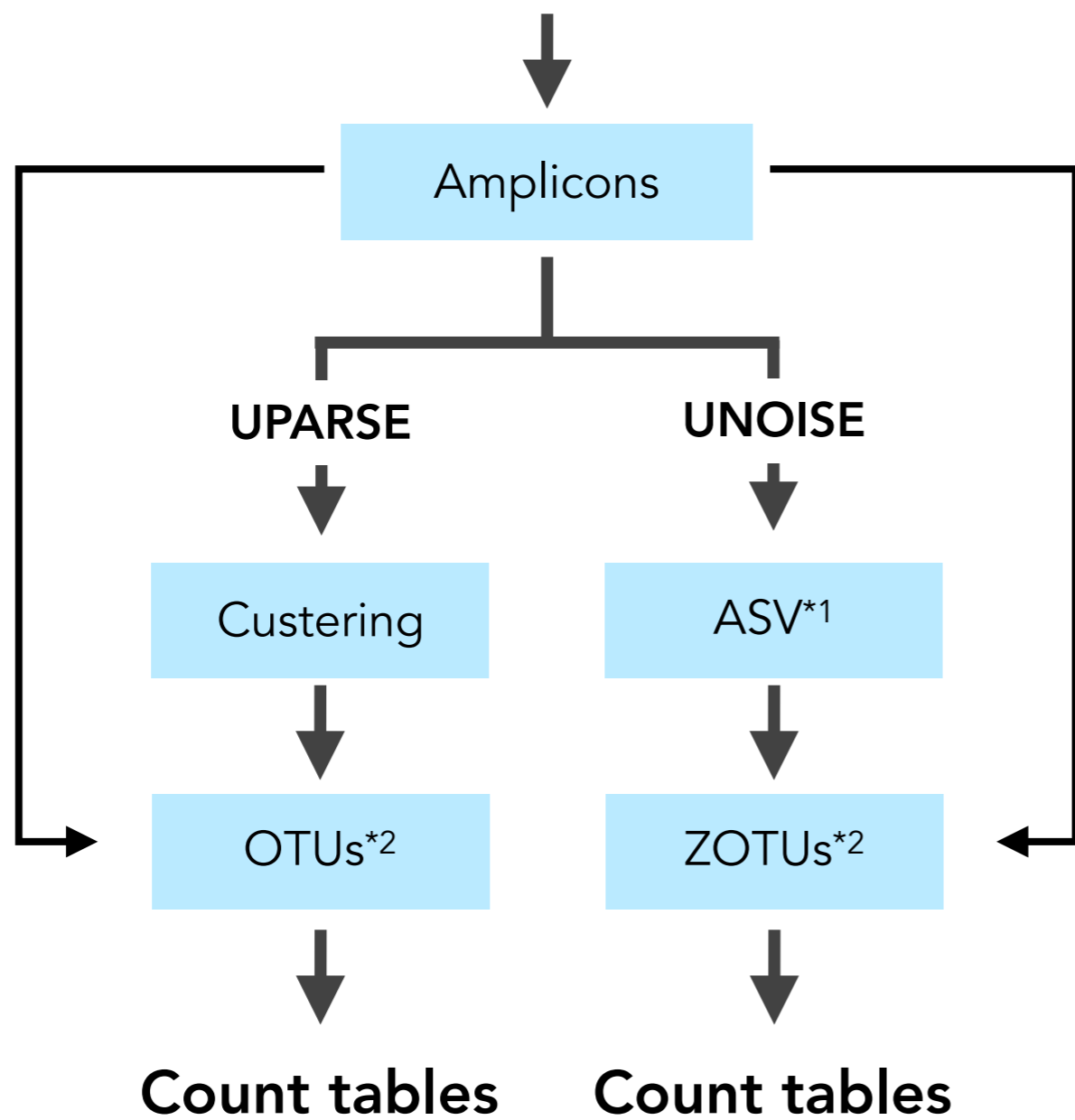


Check-List

1. Download data (if possible via terminal e.g. ftp)
2. Verify file integrity (md5sum)
3. Check data: $N_{\text{samples}} = N_{R1} + N_{R2}$
4. Blast a few random reads
5. Run a quality control (e.g. USEARCH, FastQC)
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)



Preparing NGS reads for OTU and denoising analysis

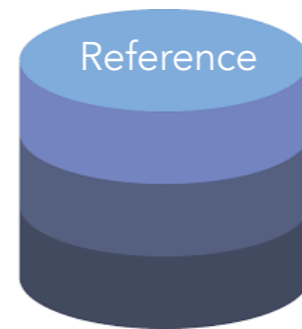
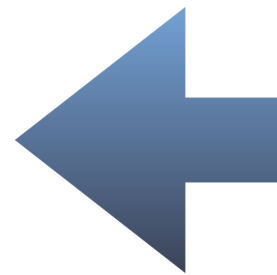
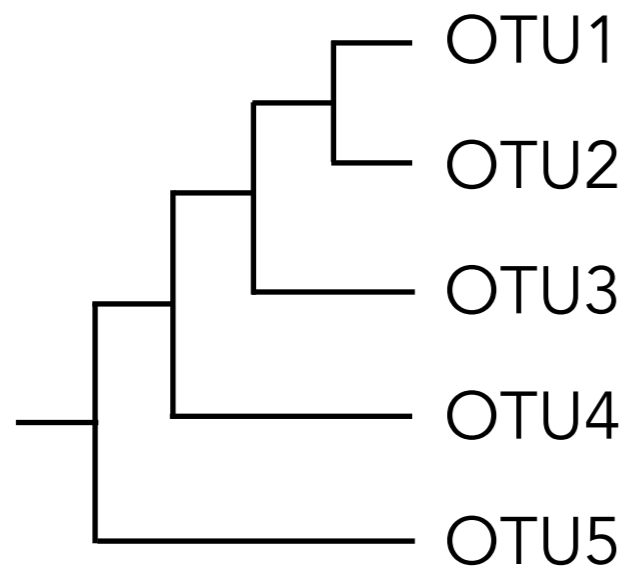


*1 ASV: Amplicon Sequence Variant

*2 ZOTU: (Zero) Operational Taxonomic Units

Measuring microbial diversity by 16S sequencing

OTU - Annoation (-Prediction)

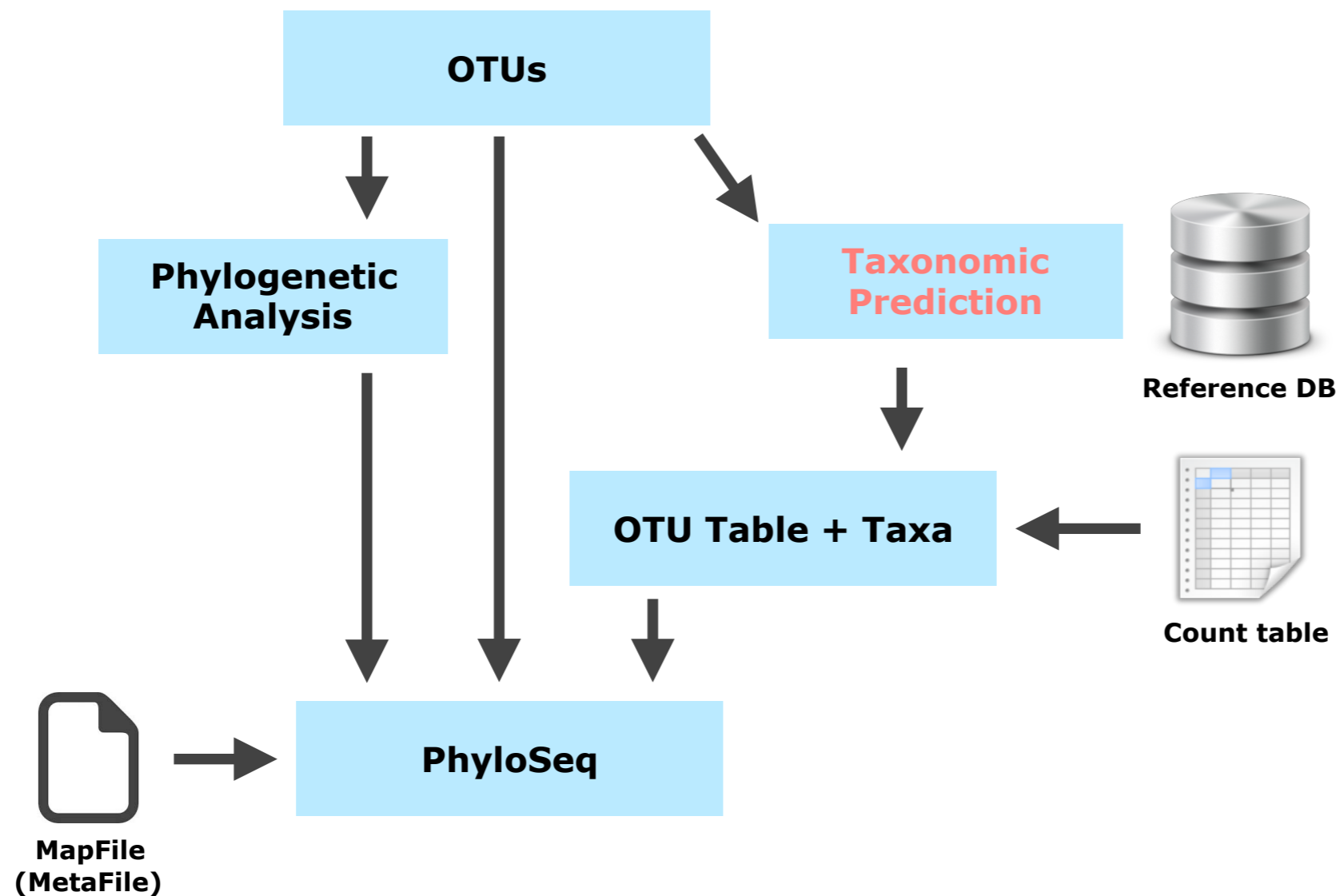


- NCBI 16S
- SILVA SSU & LSU
- Ribosomal Database Project RDP
- GreenGenes
- EzBioCloud
- ...

16S taxonomy

16S sequence identity and taxonomy

OTU - Annoation (-Prediction)



Predicting taxonomy for 16S OTUs



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

PacBio - CCS Data

P1 De-multiplexing

P2 In-silicon PCR with Size Selection

P3 Quality Filtering

P4 Clustering / Amplicon Sequence Variants

P5 Count Table

Metagenomic

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

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

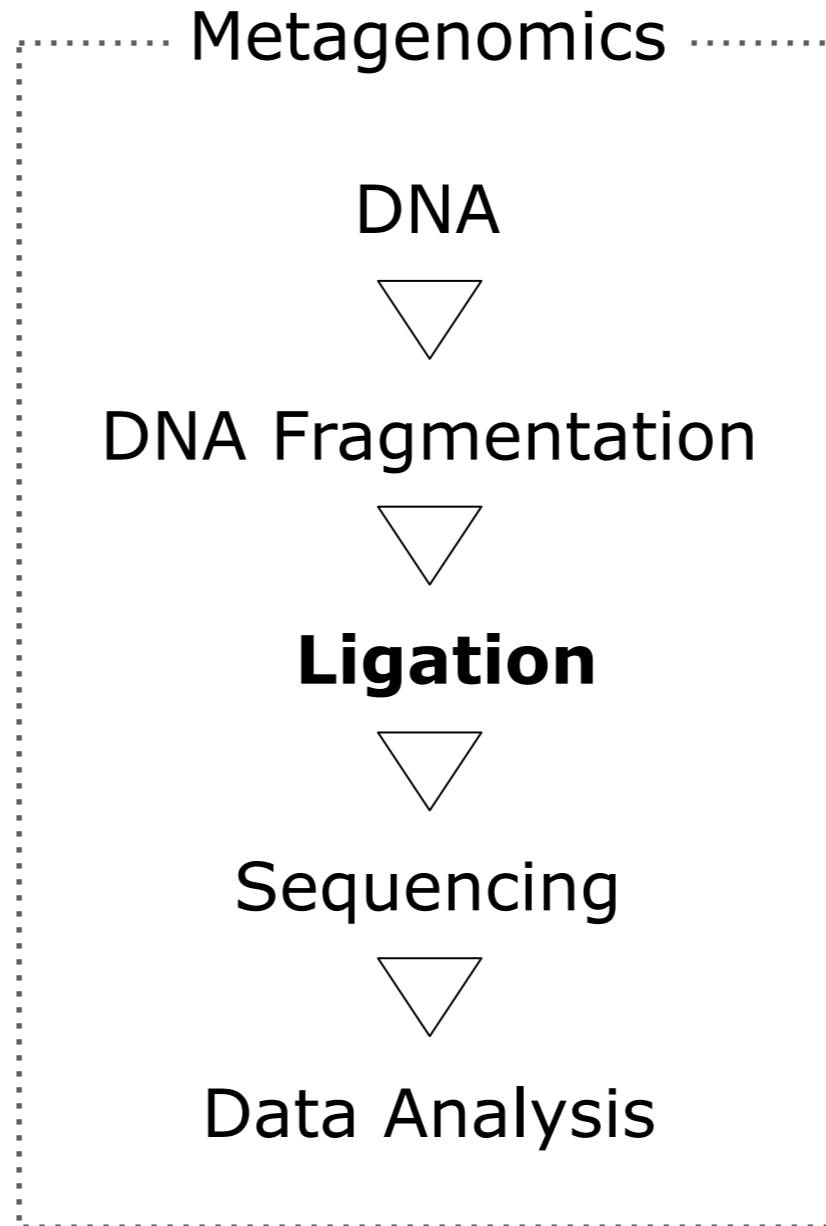
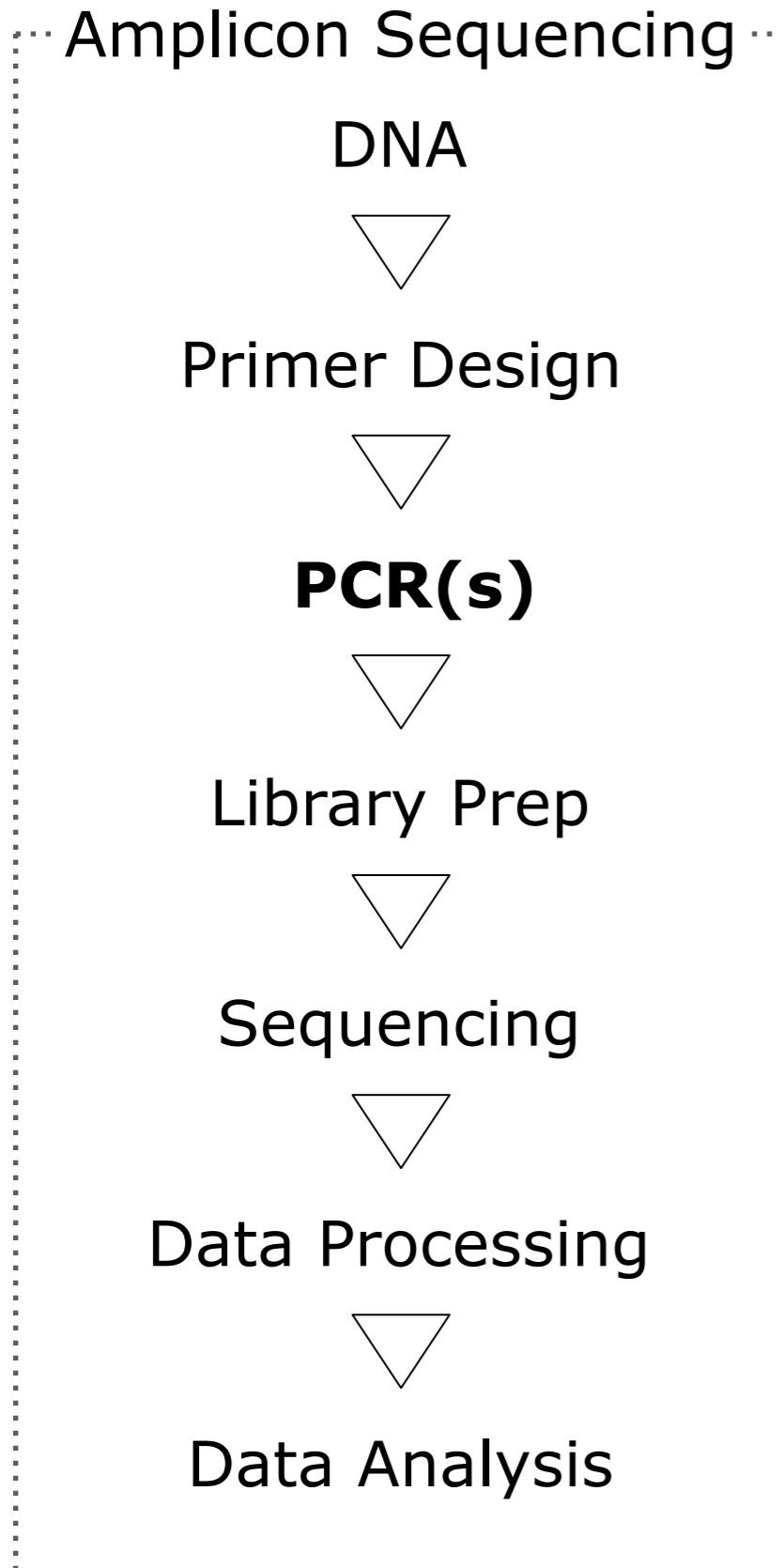
M3 Meta - Genome/Transcriptome Assembly

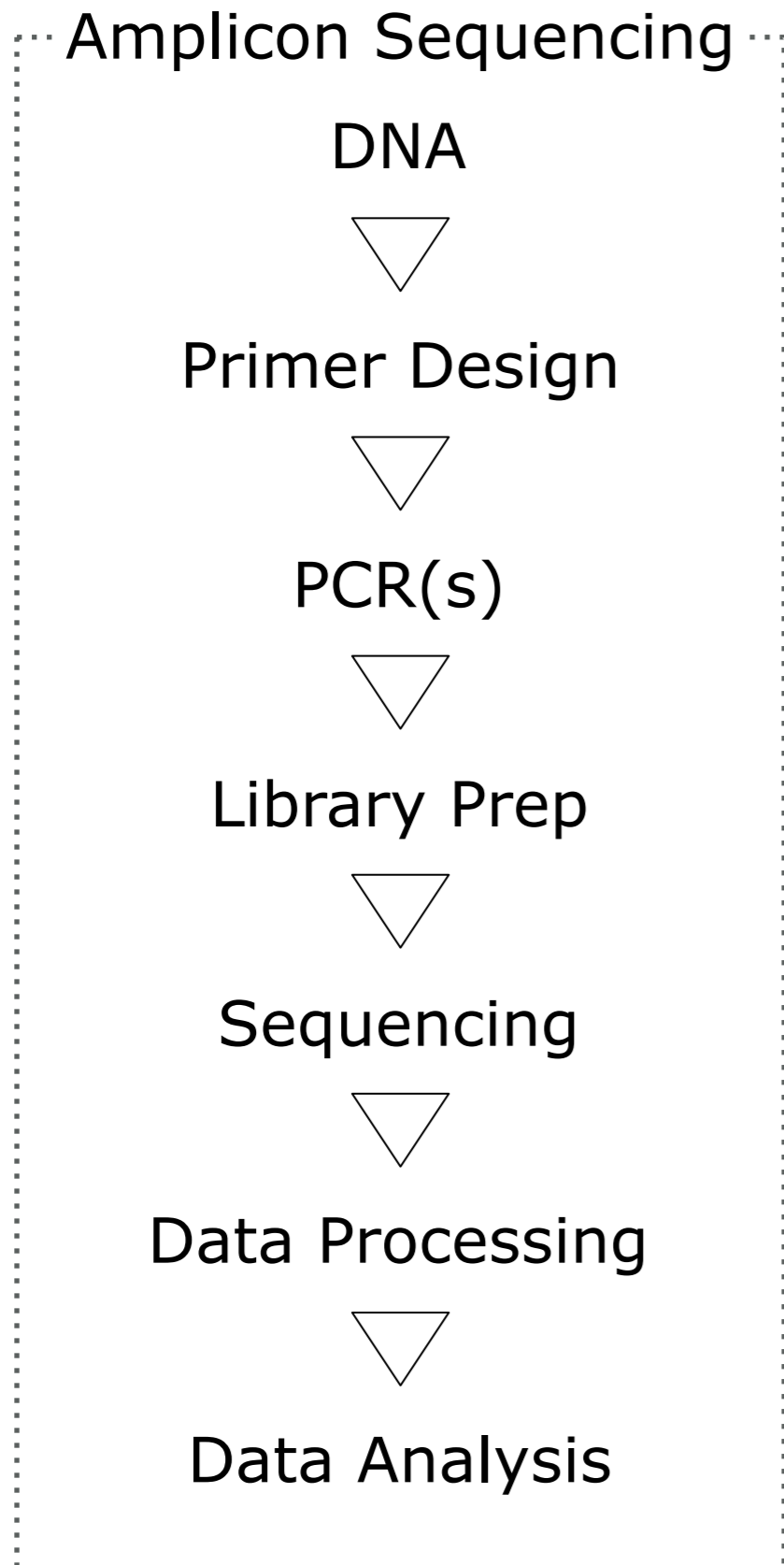
M4 Taxonomic Annotation

M5 Functional Annotation

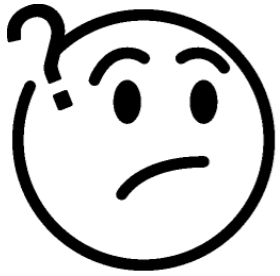
- ▶ Woese CR, Fox GE (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms". PNAS. 74 (11): 5088–90.
- ▶ Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991). "16S ribosomal DNA amplification for phylogenetic study". Journal of Bacteriology. 173 (2): 697–703.
- ▶ Schmidt TM, 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.

**MORE
THINGS
CONSIDERED**





- 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?

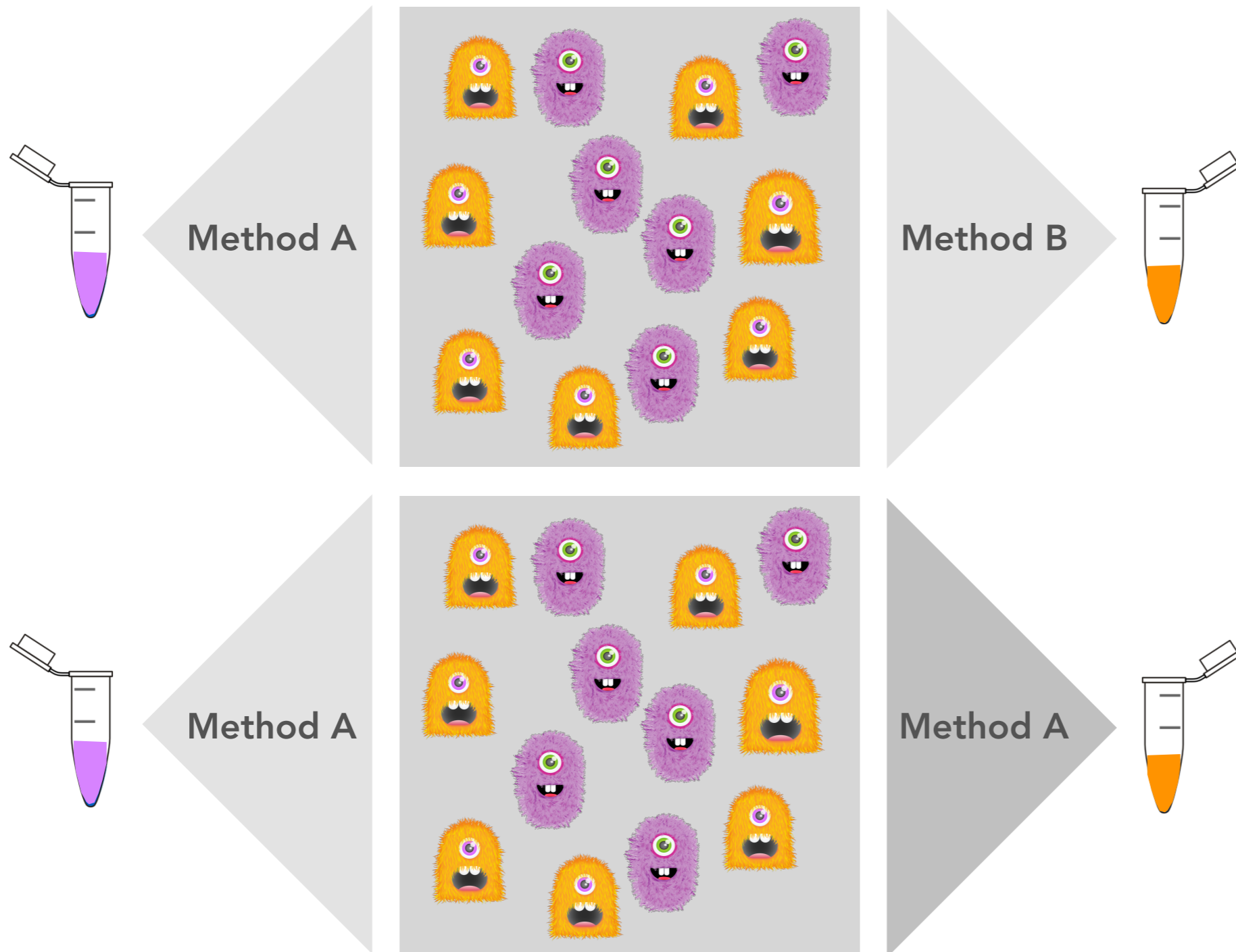
How many **samples / replicates** will be needed?

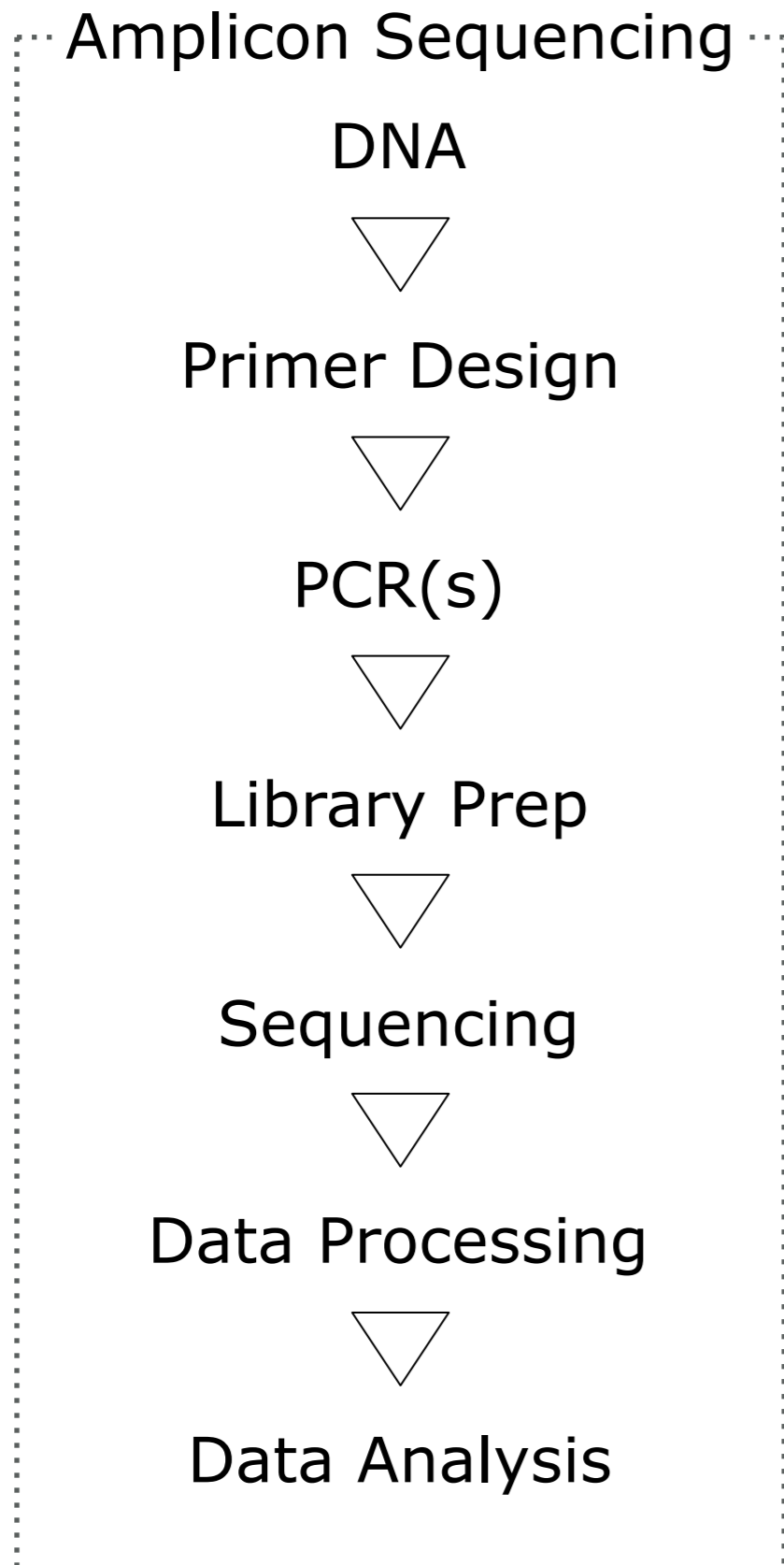
What is the **trade off** between coverage and replicates?

How much **money** do we have?

▶ **Pilots are very useful**

DNA Extraction Method Bias

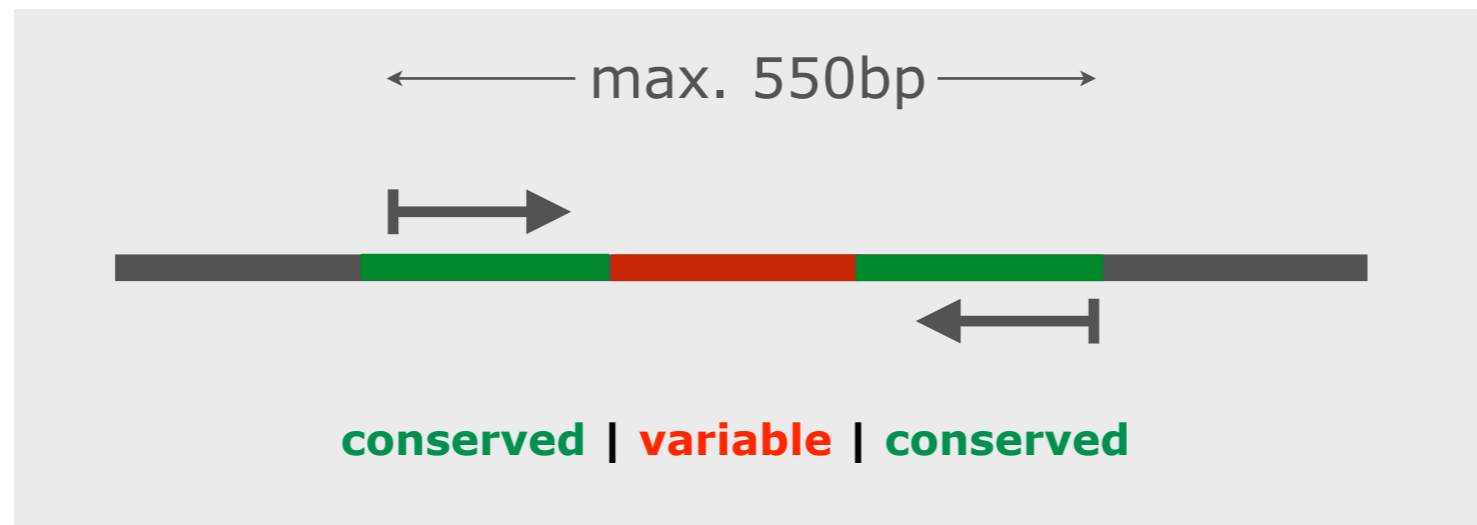




- sampling / sampling design
- extraction method
- contamination

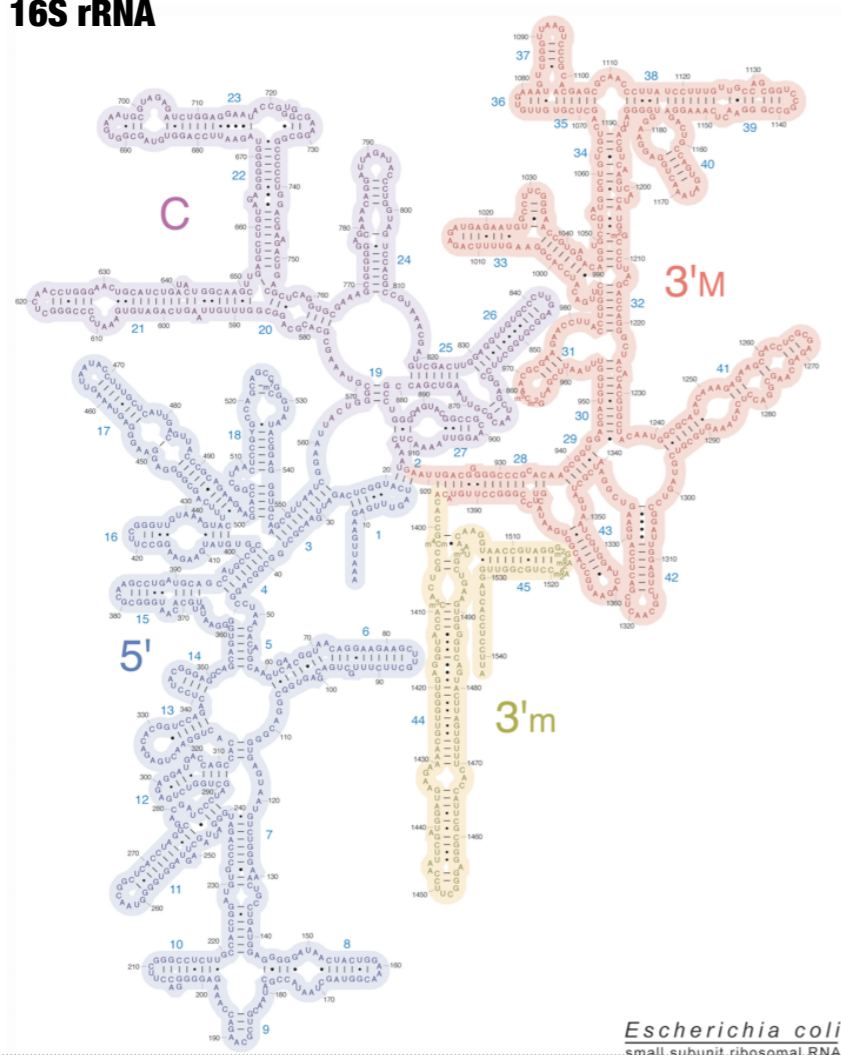
- specificity
- quality

Primer Design



- ▶ target specific - avoid false priming
- ▶ universal for all targeted species
- ▶ amplicon size
- ▶ avoid / limit amplicon size variation
- ▶ optimize PCR condition

16S rRNA

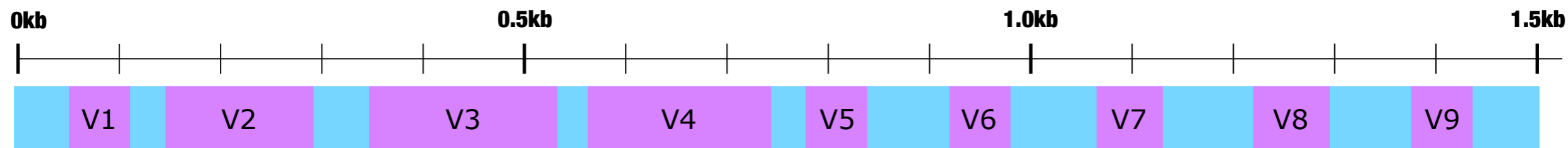


Ribosomal Database Project

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



16S rDNA

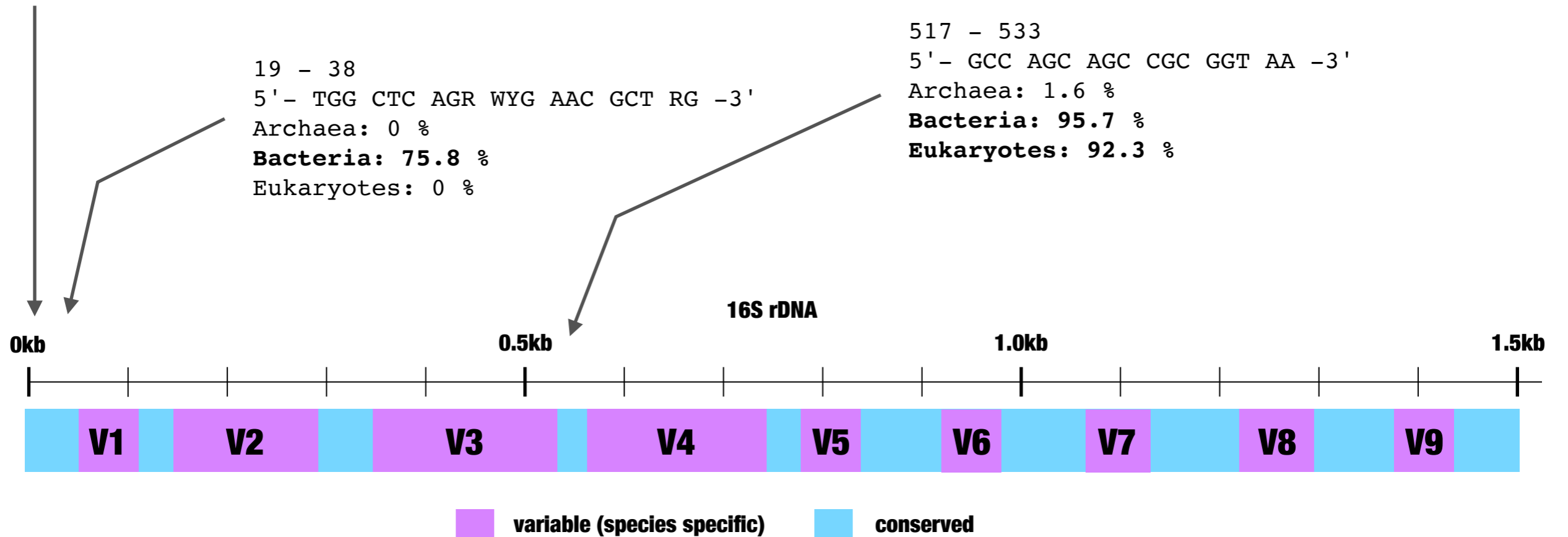


 conserved  variable (species specific)

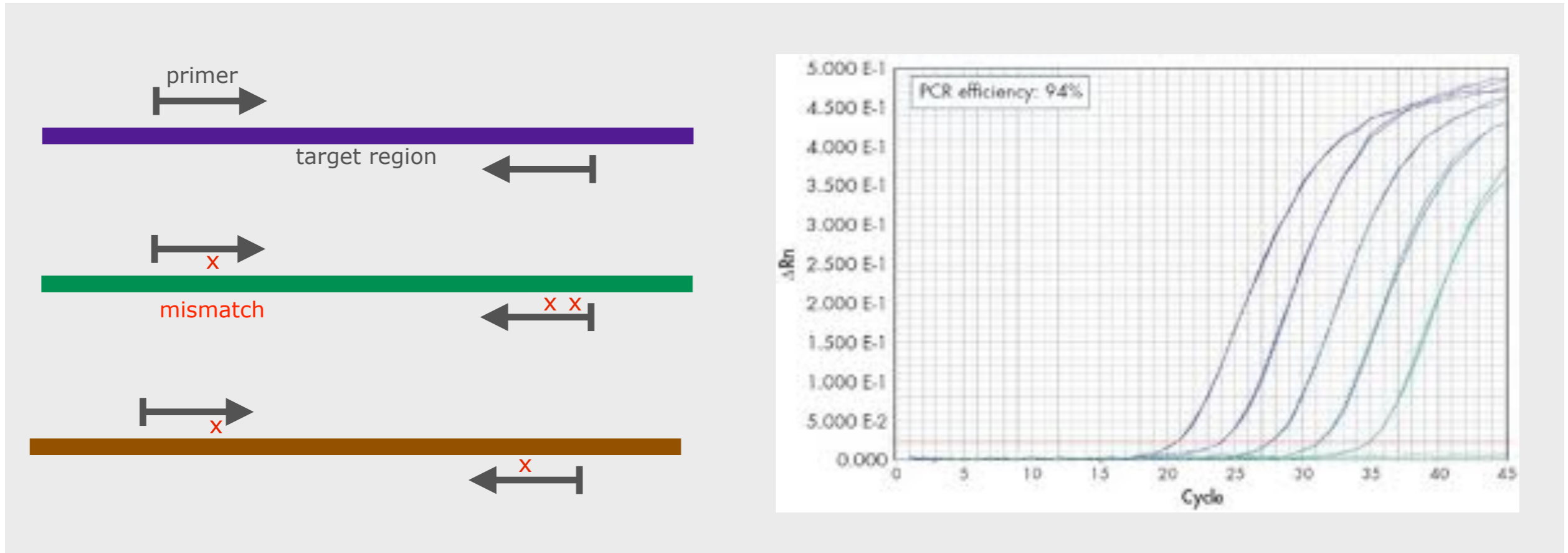
8 - 23
5'- TCY GGT TGA TCC TGC C -3'
Archaea: 65.1 %
Bacteria: 0 %
Eukaryotes: 11.1 %

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

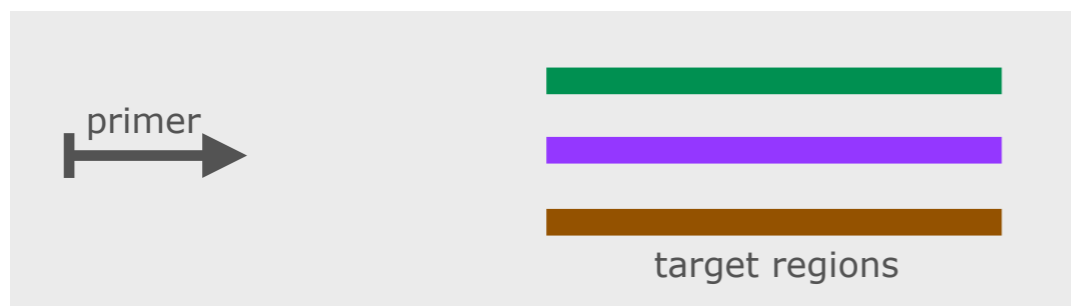
517 - 533
5'- GCC AGC AGC CGC GGT AA -3'
Archaea: 1.6 %
Bacteria: 95.7 %
Eukaryotes: 92.3 %



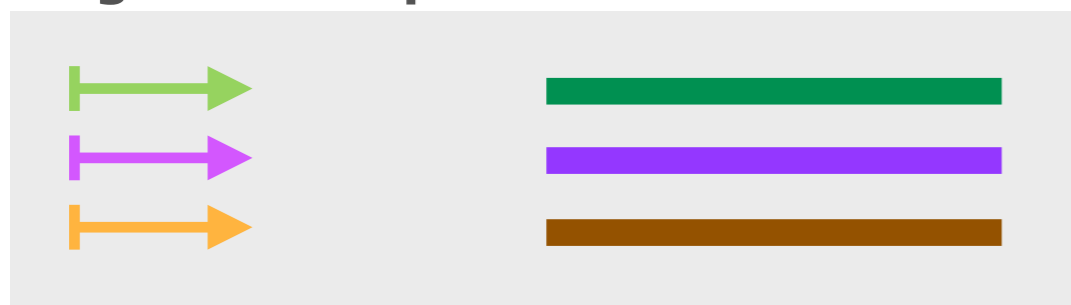
PCR efficiency and primer mismatches



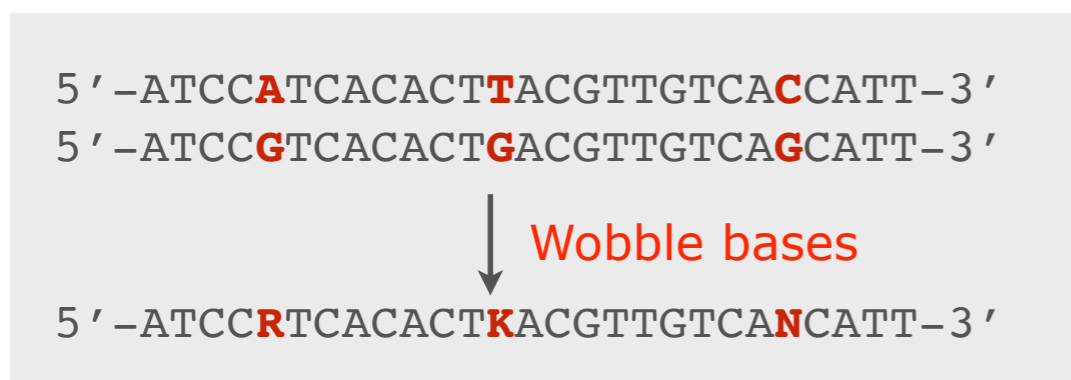
The primer(s) should equally amplify the target regions of the different species/individuals!



Degenerated primers

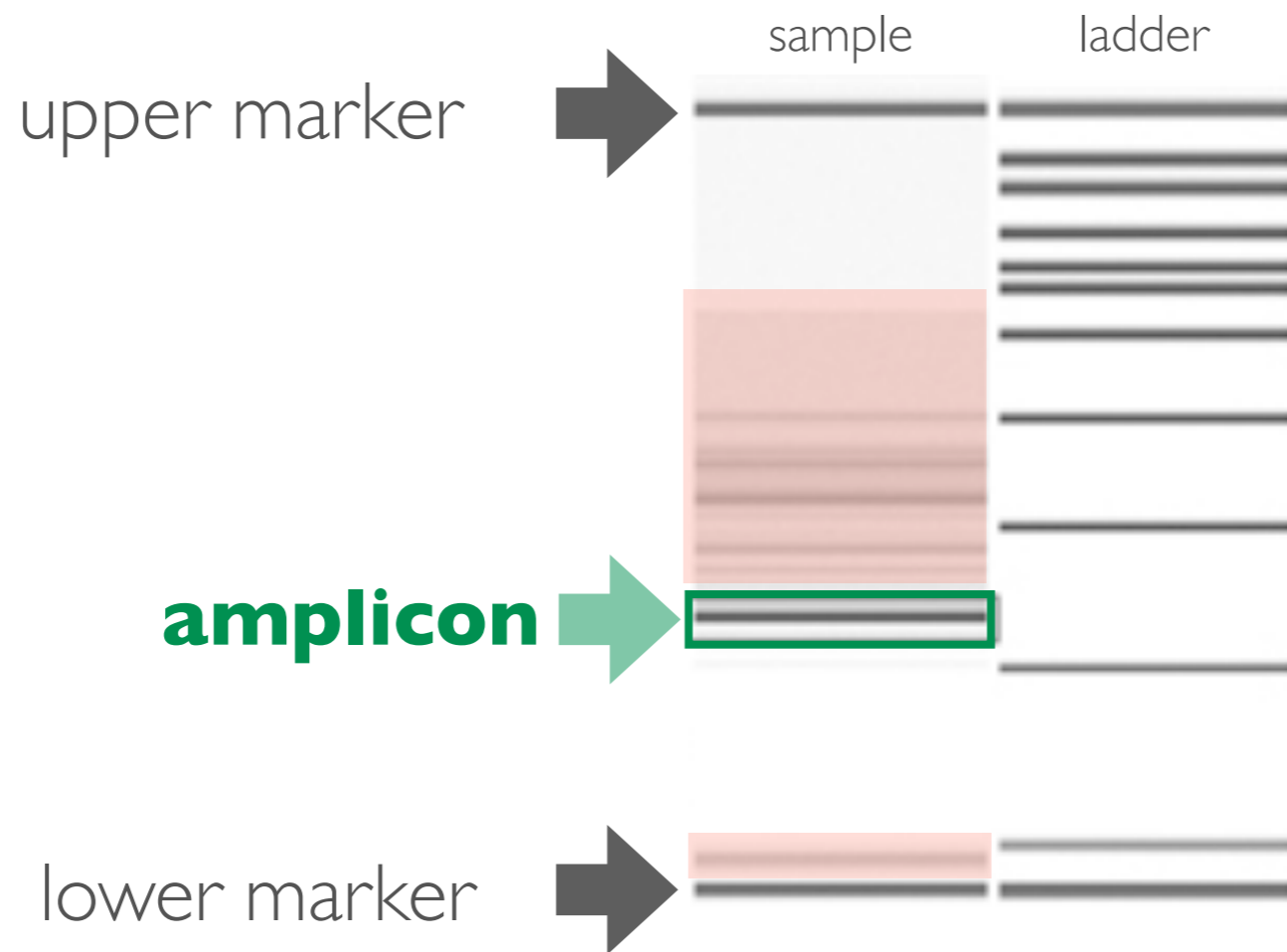


A group of degenerate oligonucleotides contain **related sequences** with differences at specific locations. These are used simultaneously in the hope that one of the sequences of the oligonucleotides will be perfectly complementary to a target DNA sequence.



<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.

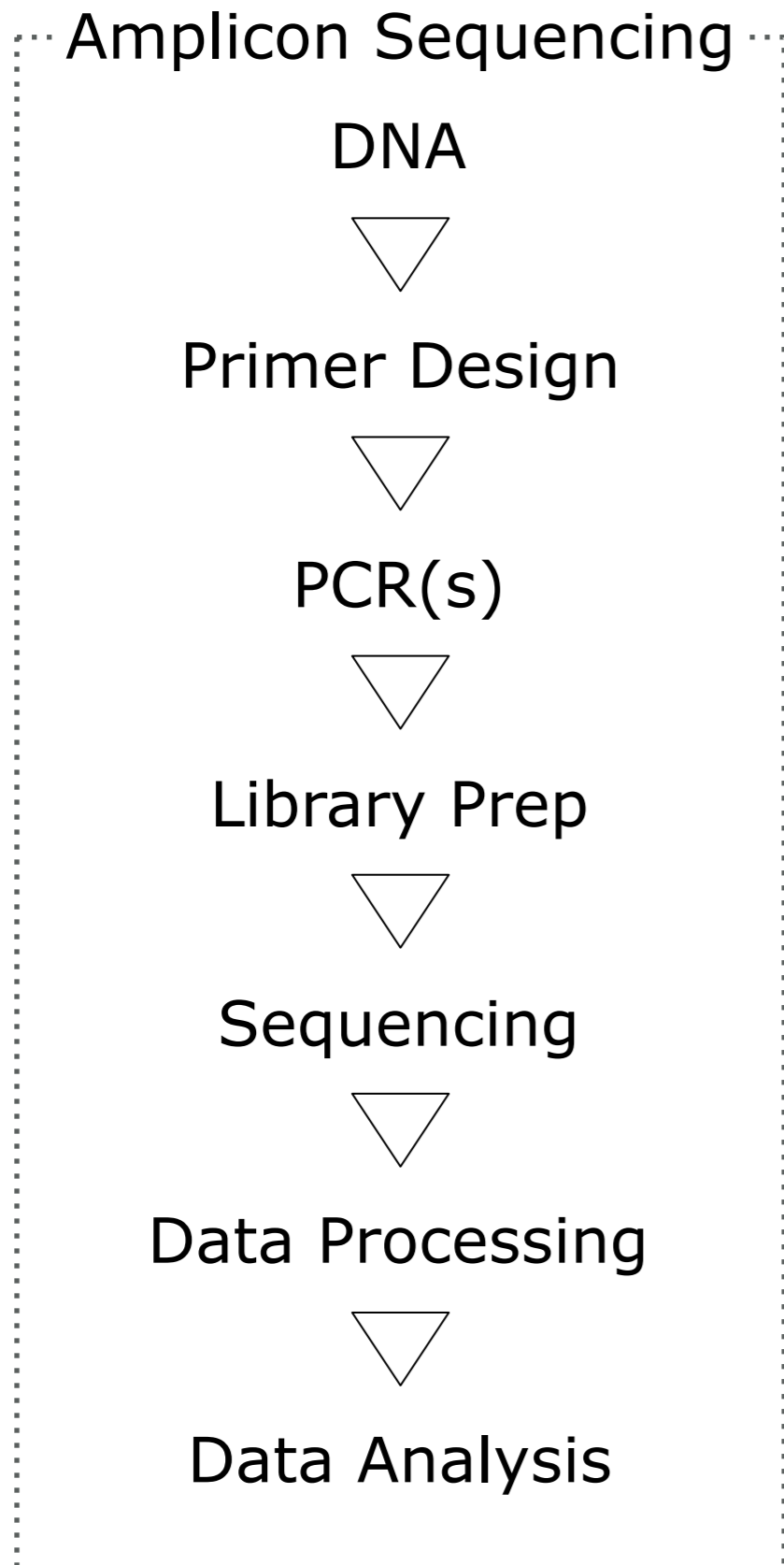
HPLC-Purified Oligonucleotides

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>

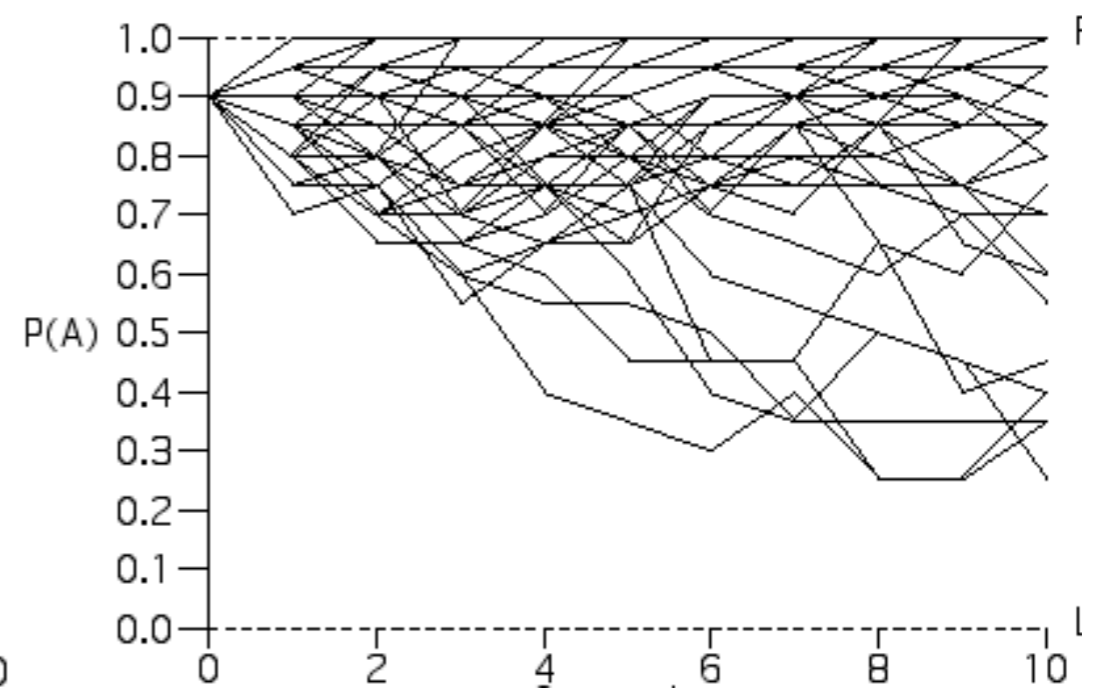
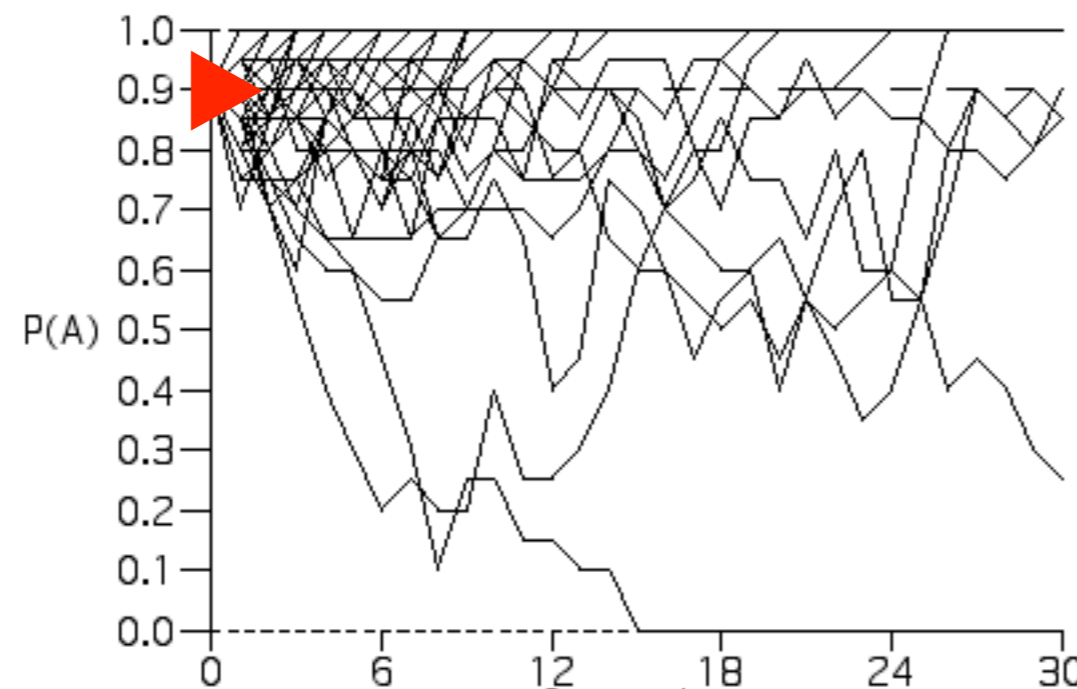
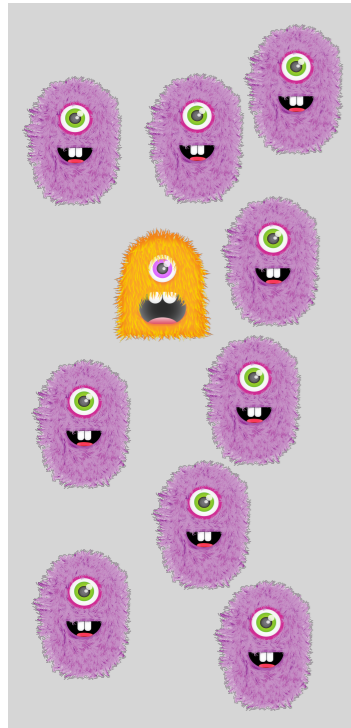
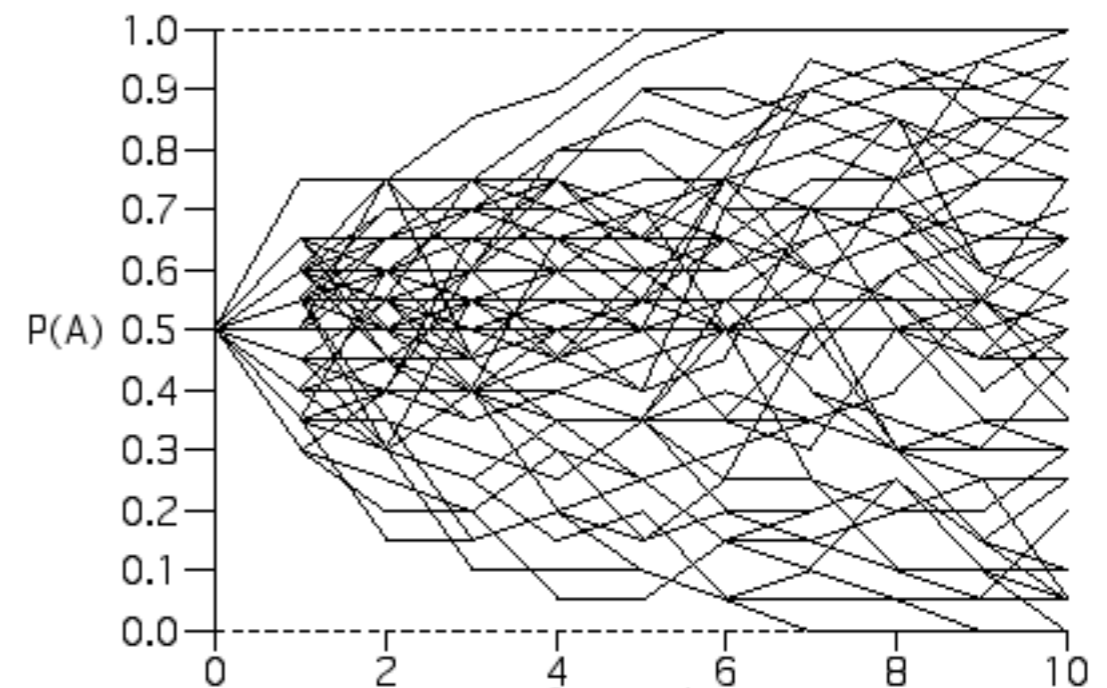
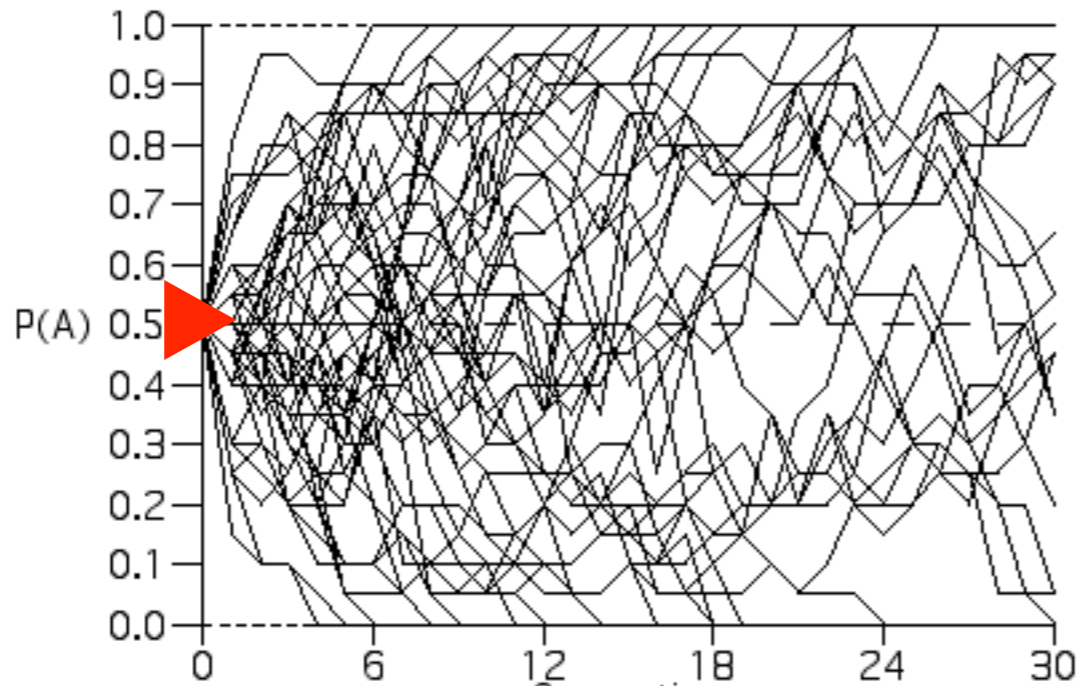
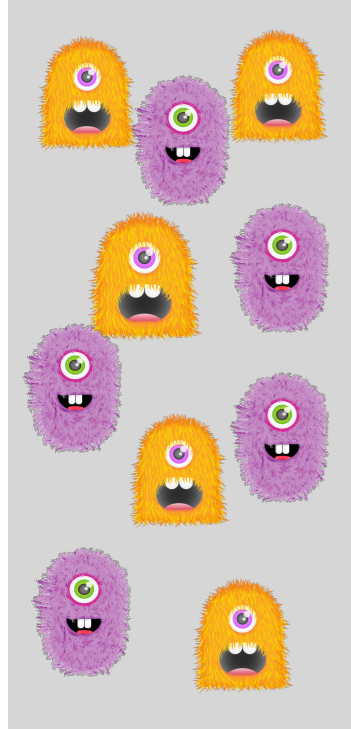


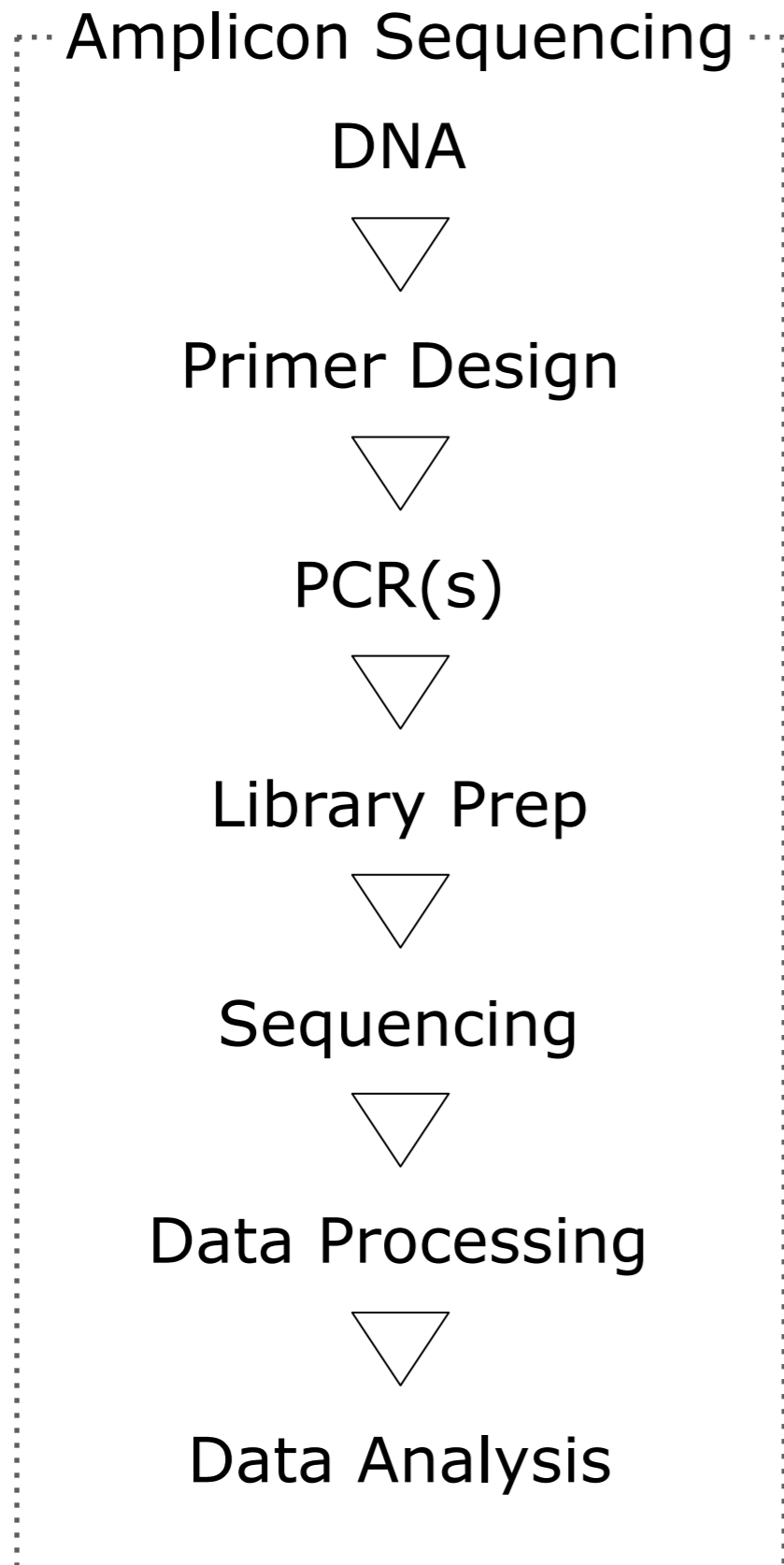
- sampling / sampling design
- extraction method
- contamination

- specificity
- quality

- PCR conditions and setup
- number of cycles and PCRs

Number of PCR Cycles





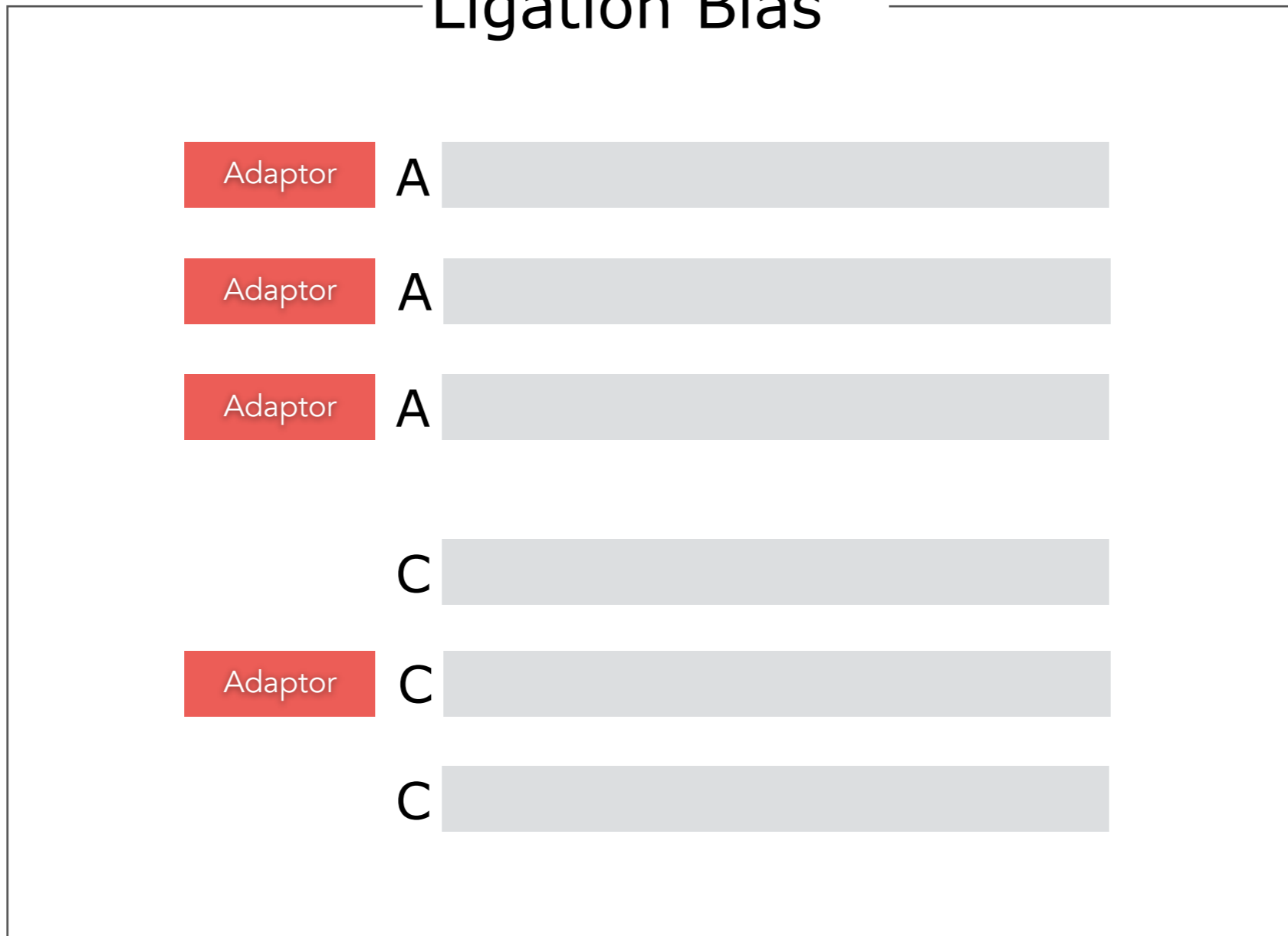
- sampling / sampling design
- extraction method
- contamination

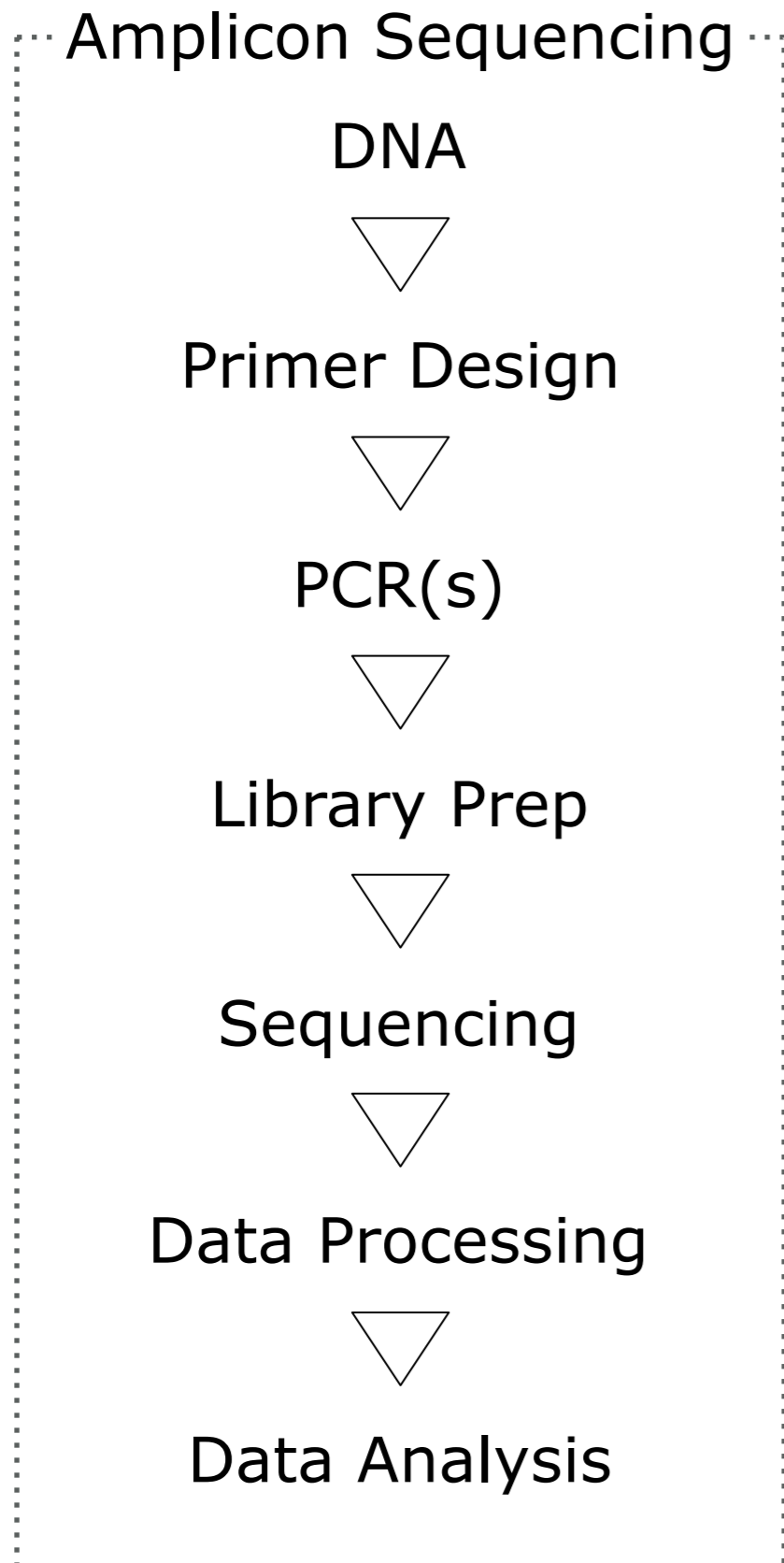
- specificity
- quality

- PCR conditions and setup
- number of cycles and PCRs

- Ligation bias
- Clean-up

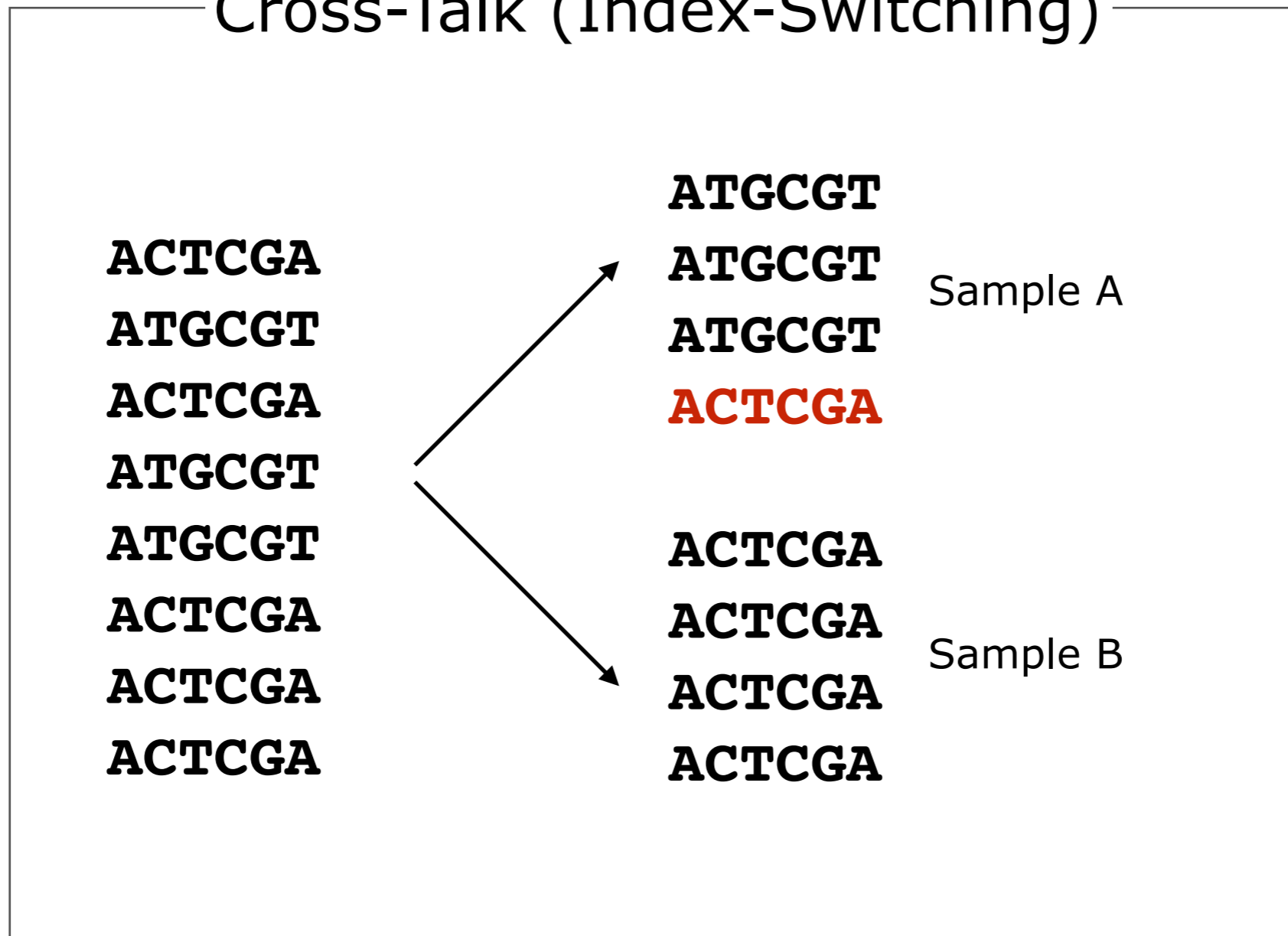
Ligation Bias



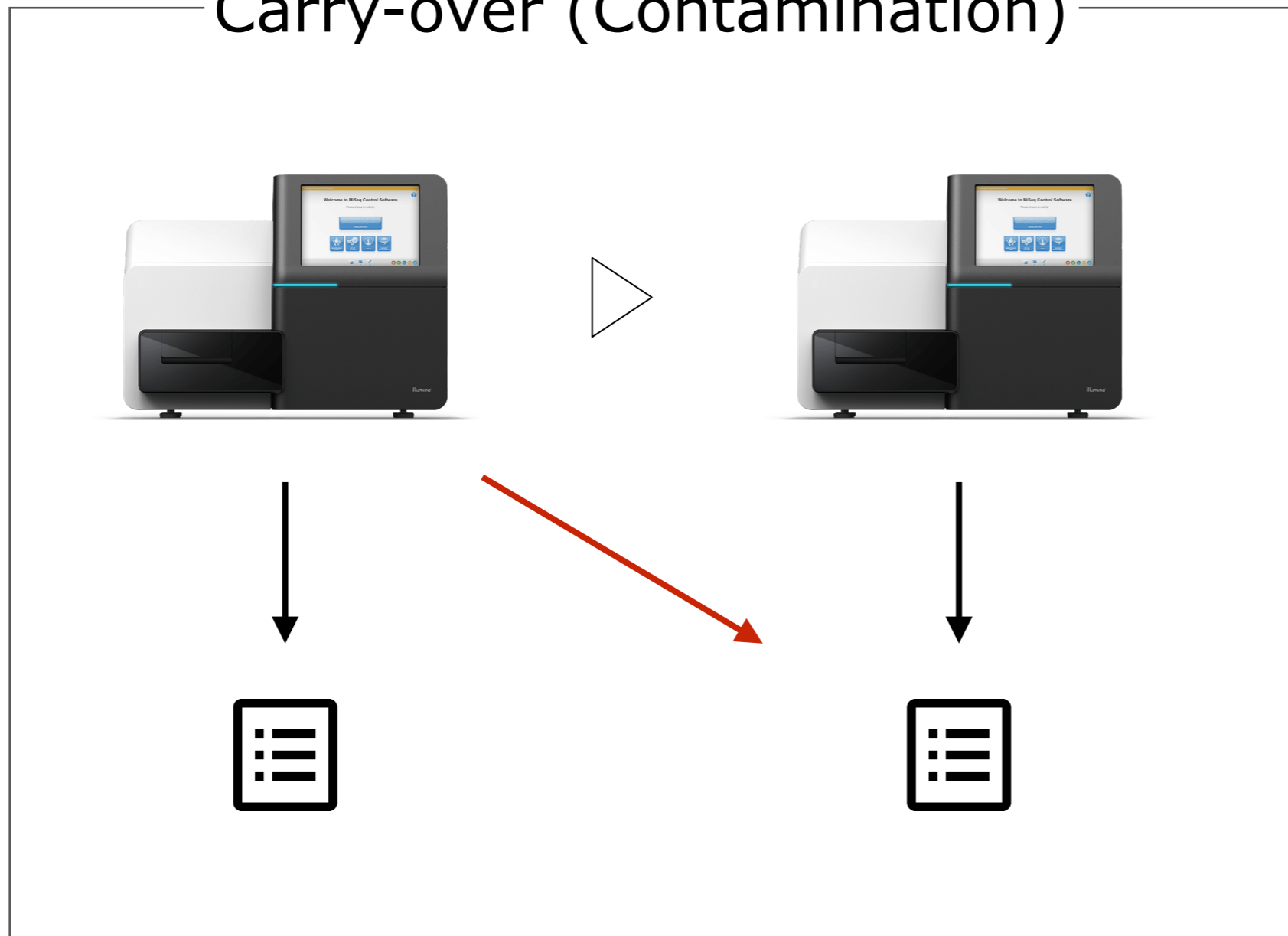


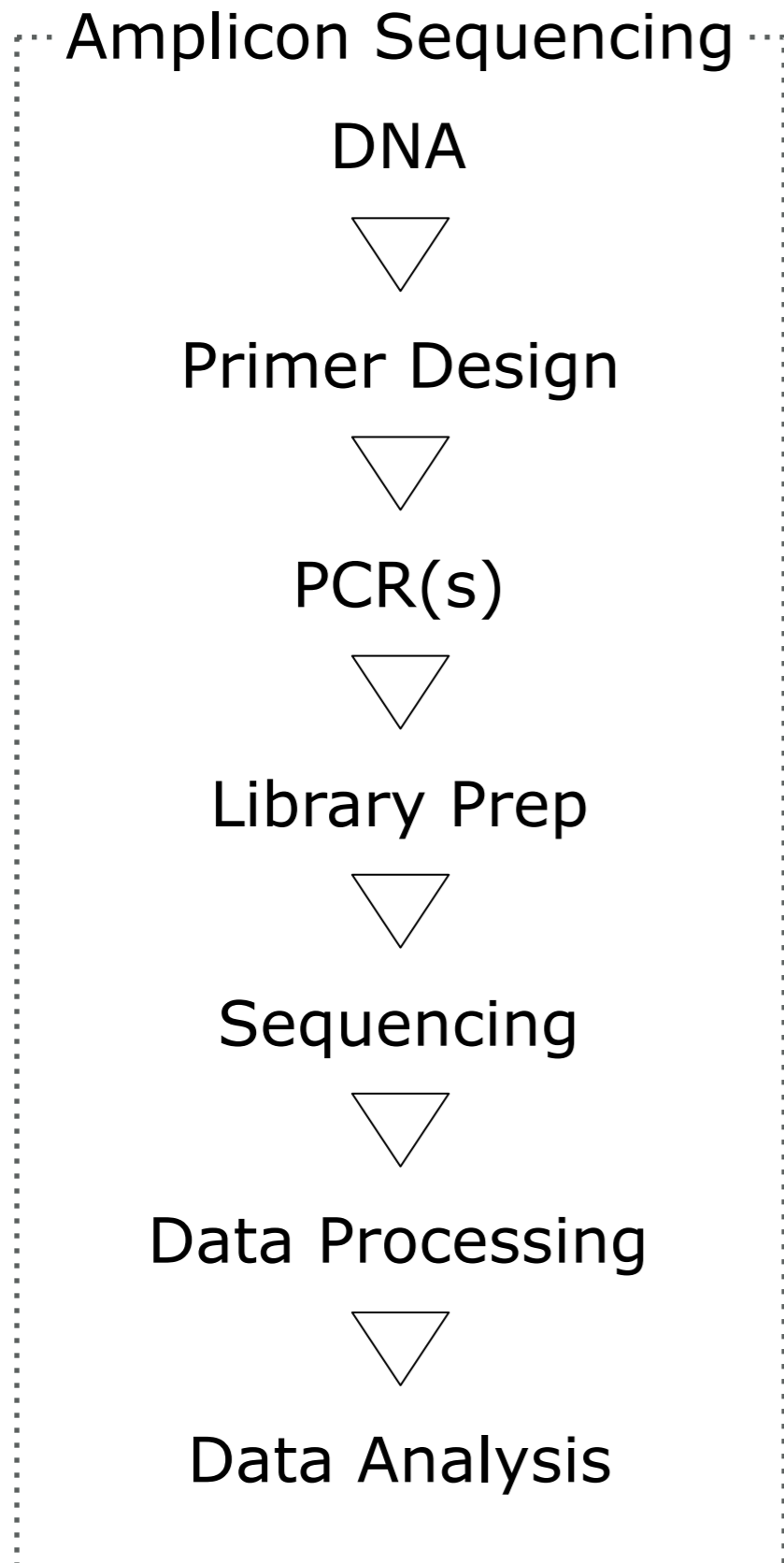
- sampling / sampling design
- extraction method
- contamination
- specificity
- quality
- PCR conditions and setup
- number of cycles and PCRs
- Ligation bias
- Clean-up
- quality / complexity / depth
- read-length / amplicon-length
- cross-talk, carry-over

Cross-Talk (Index-Switching)



Carry-over (Contamination)





- sampling / sampling design
- extraction method
- contamination

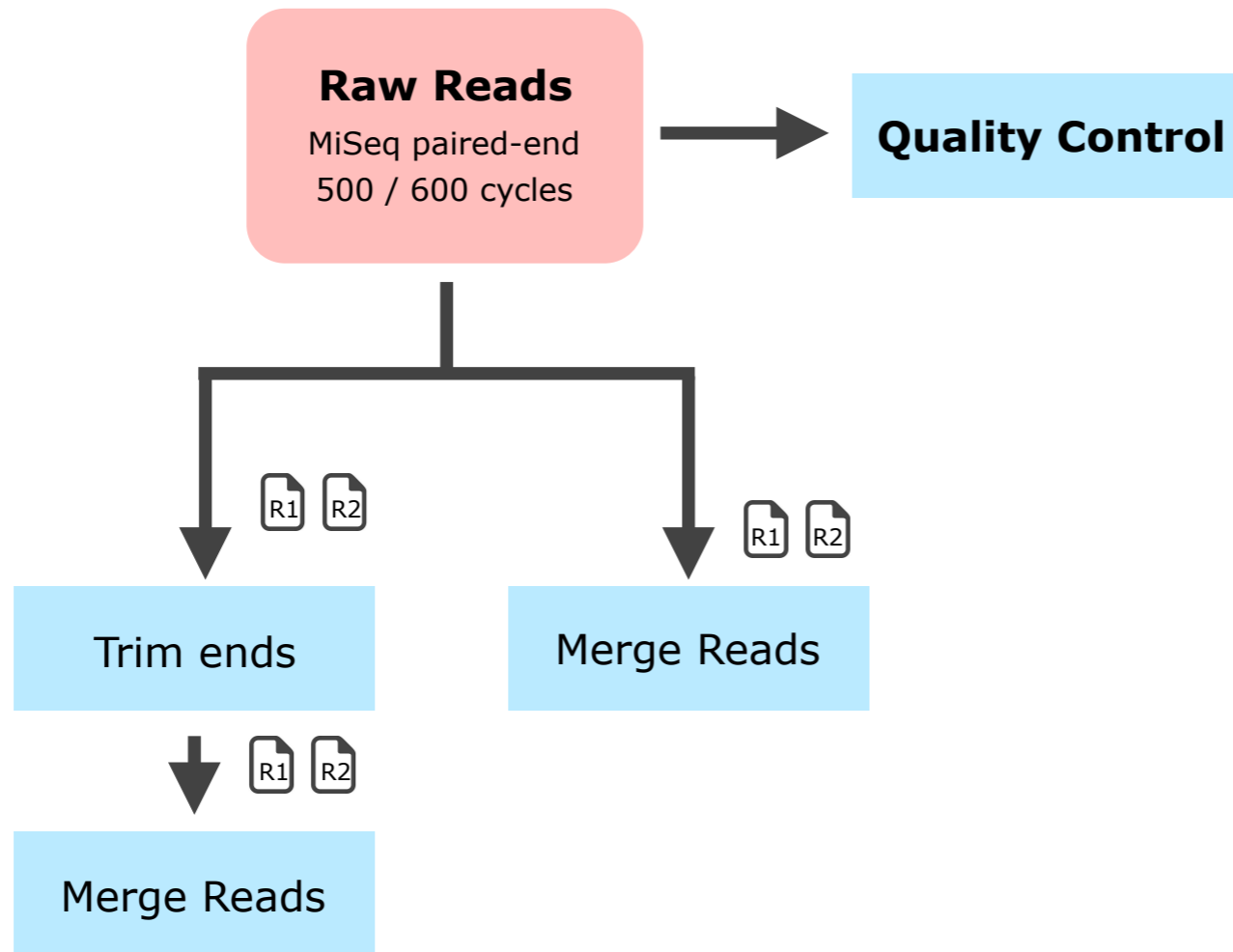
- specificity
- quality

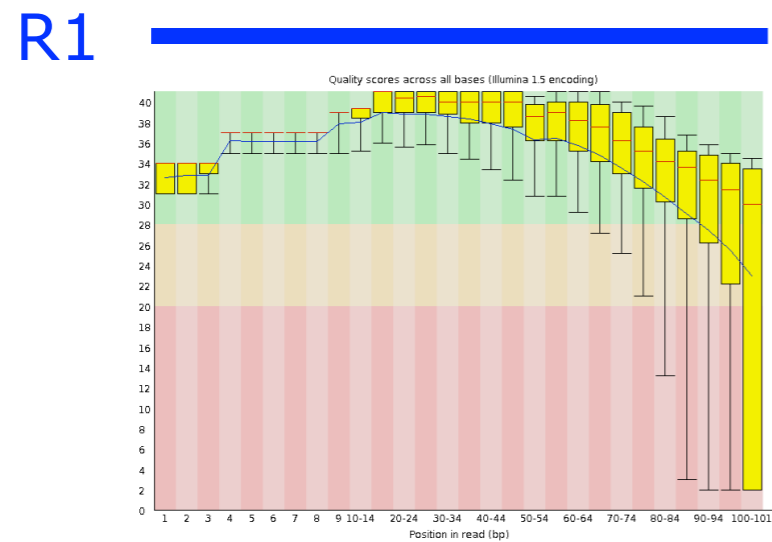
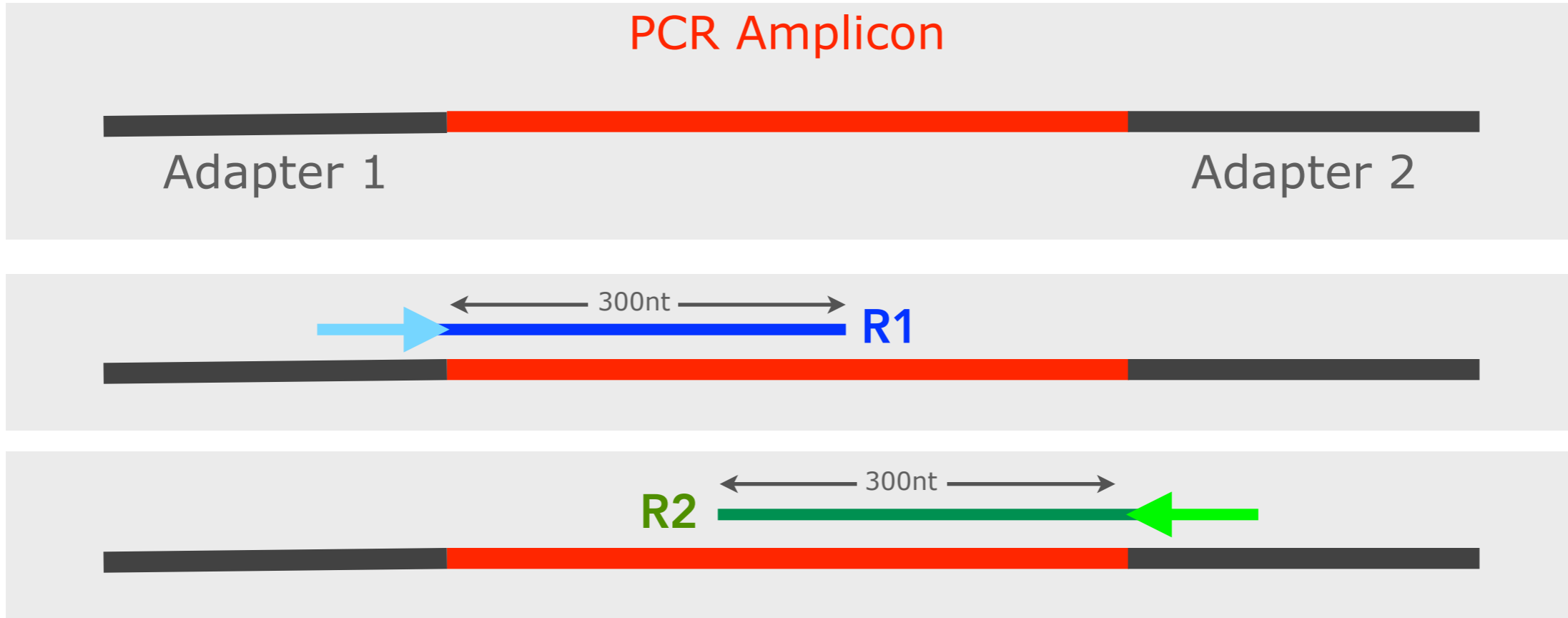
- PCR conditions and setup
- number of cycles and PCRs

- Ligation bias
- Clean-up

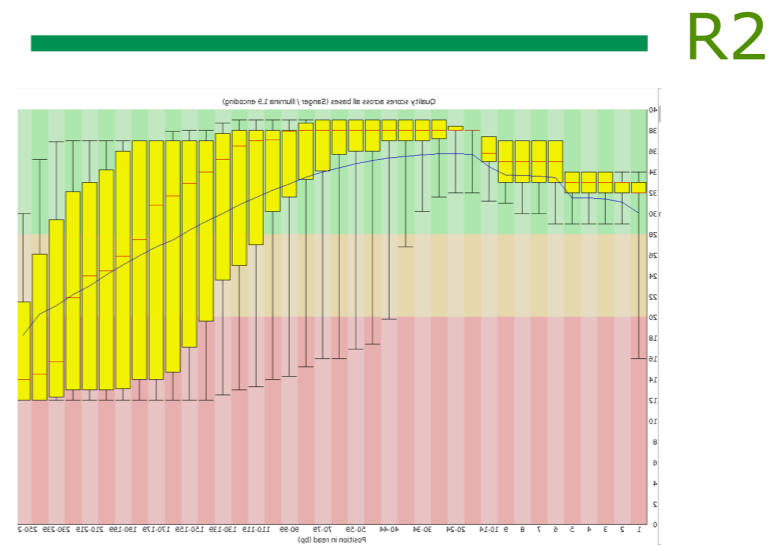
- quality / complexity / depth
- read-length / amplicon-length
- cross-talk, carry-over

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

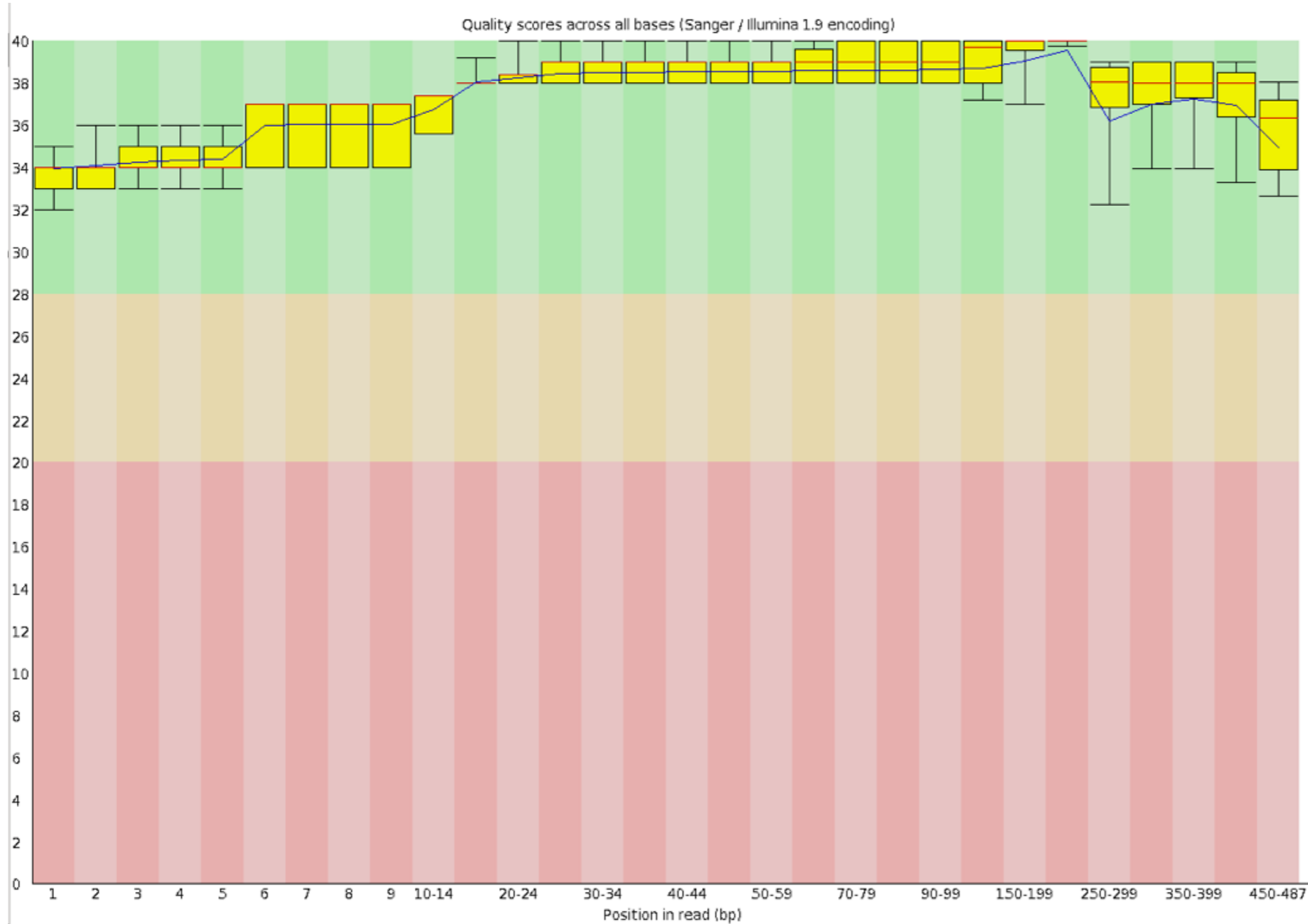


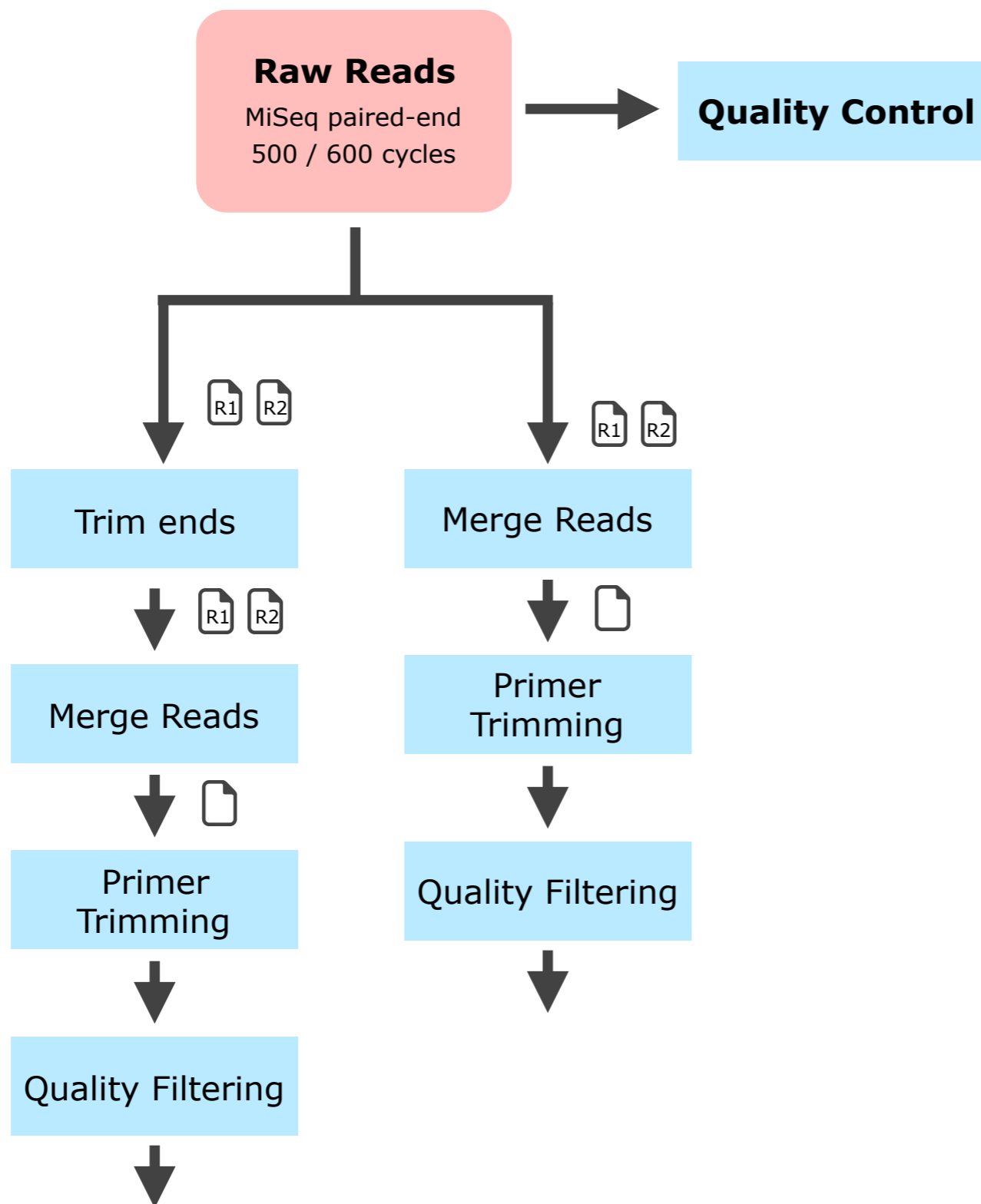


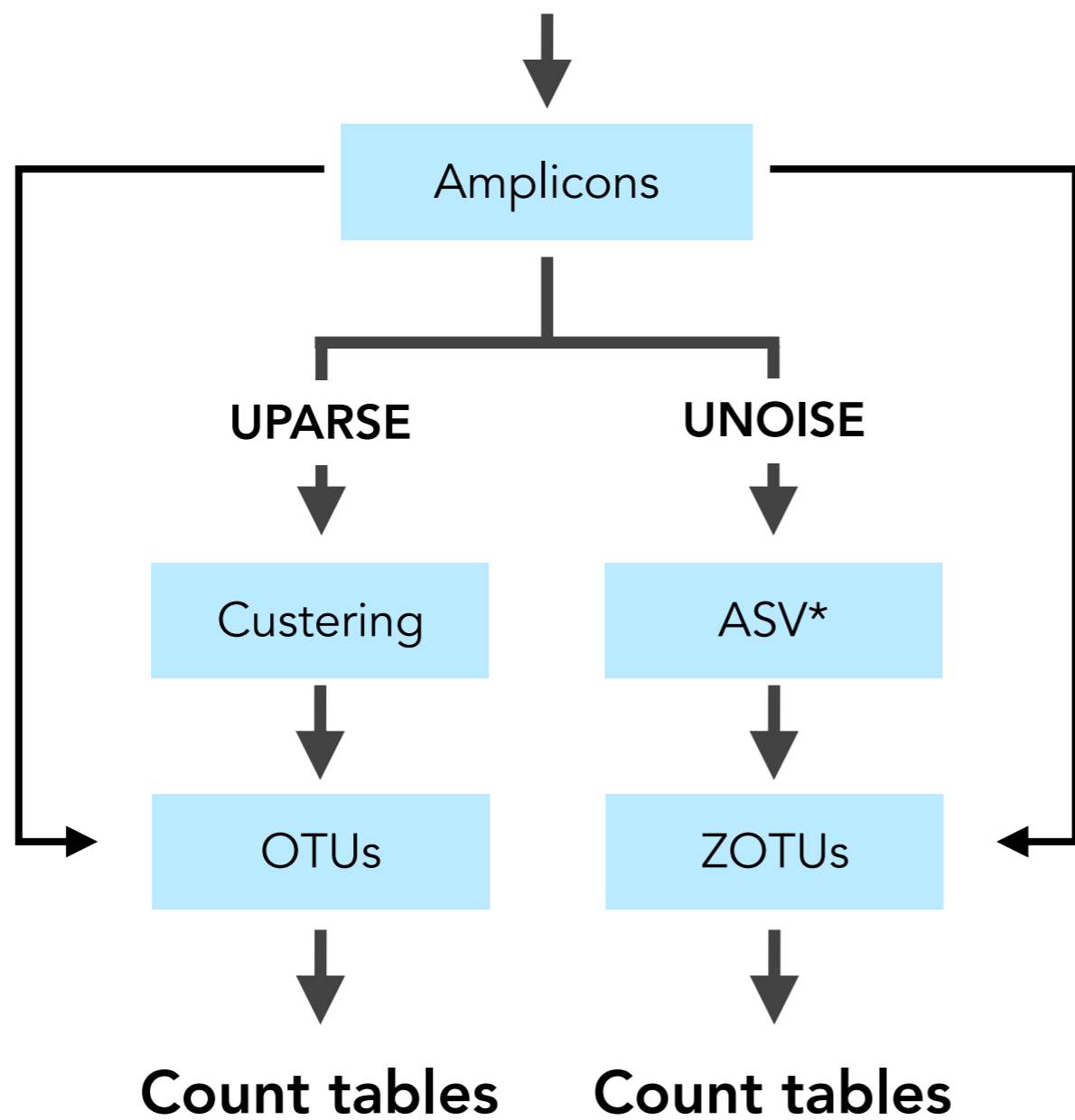
$$Q(R1) > Q(R2)$$



Quality scores for merged paired-end reads







* ASV: Amplicon Sequence Variant

Chimera Formation

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

Chimera Formation

SpeciesA: ACCGGGCATACGTAGCAACAACGGGTGTATA
 ||| ||||| ||| | |
 SpeciesB: ACCATGCATACGTAGCAACTTCGCTCTTTC

Primer PCR

TGGCCC →
 ACCGGGCATACGTAGCAACAACGGGTGTATA

TGGCCC
 ACCATGCATACGTAGCAACTTCGCTCTTTC

TGGCCCGTATGCAT→
ACCGGGCATACGTAGCAACAACGGTGTATA

Extention

TGGCCCGTATGCAT
ACCGGGCATACGTAGCAACAACGGTGTATA

Denaturation

TGGCCCGTATGCAT→
ACCATGCATACGTAGCAACTTCGCTCTTTC

ACCGGGCATACGTAGCAACTTCGCTCTTTC

PCR / Sequencing Errors

ACCATGCATACGTAGCAACA

▽ PCR

ACCATGCATACG**A**AGCAACA

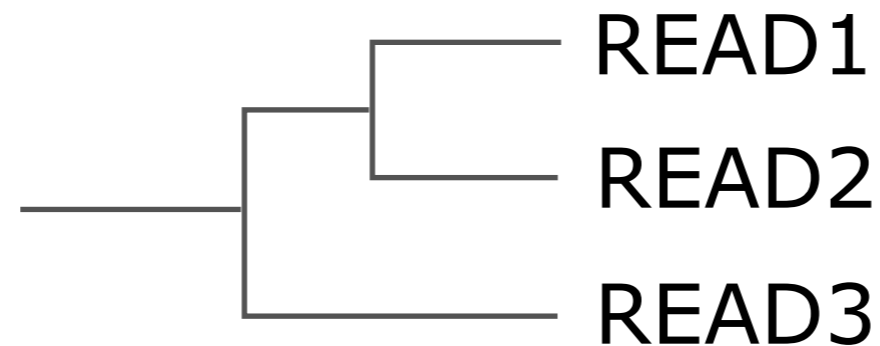
▽ Sequencing

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

READ1 : ACCATGCA**A**ACGTAG**C**AACA

READ2 : ACCATGCA**A**ACGTAG**C**AACA

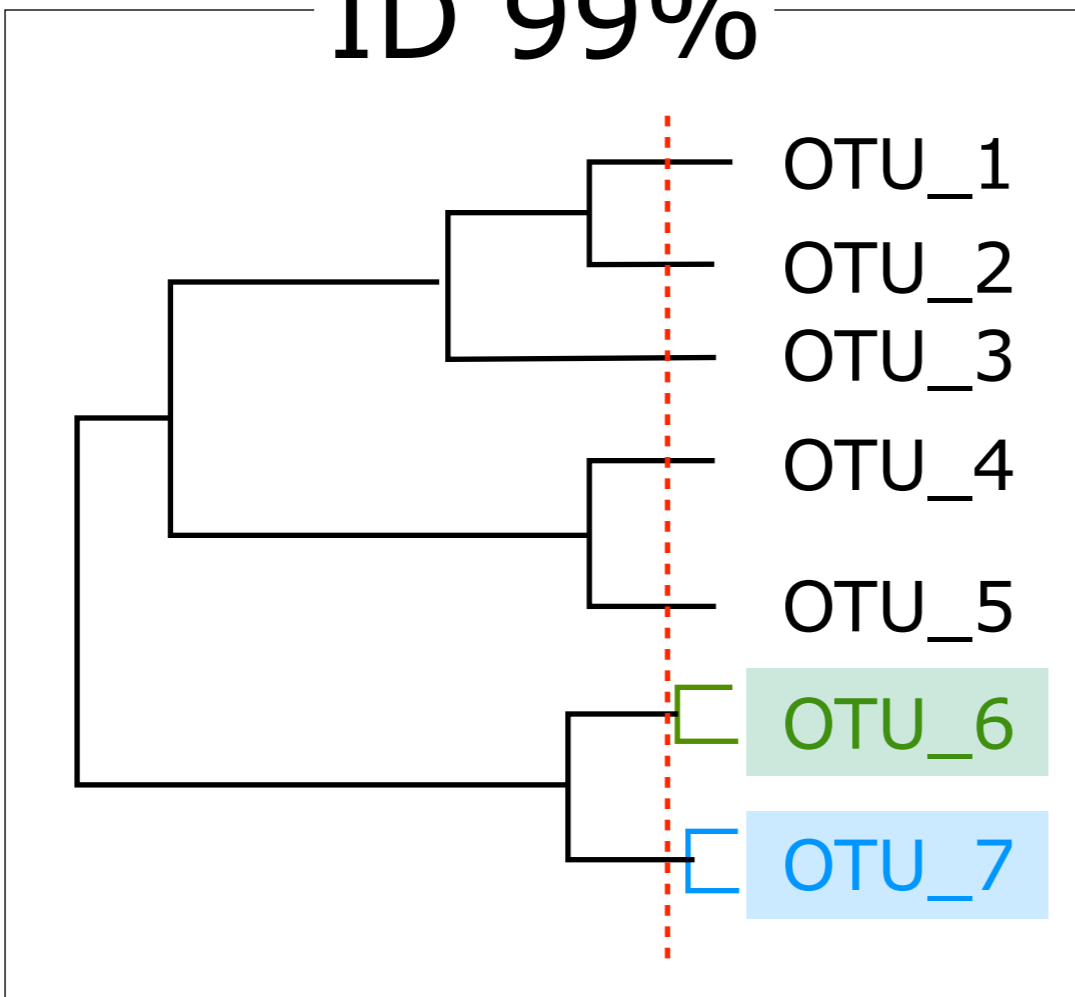
READ3 : ACCATGCAT**T**ACGTAG**G**AACA



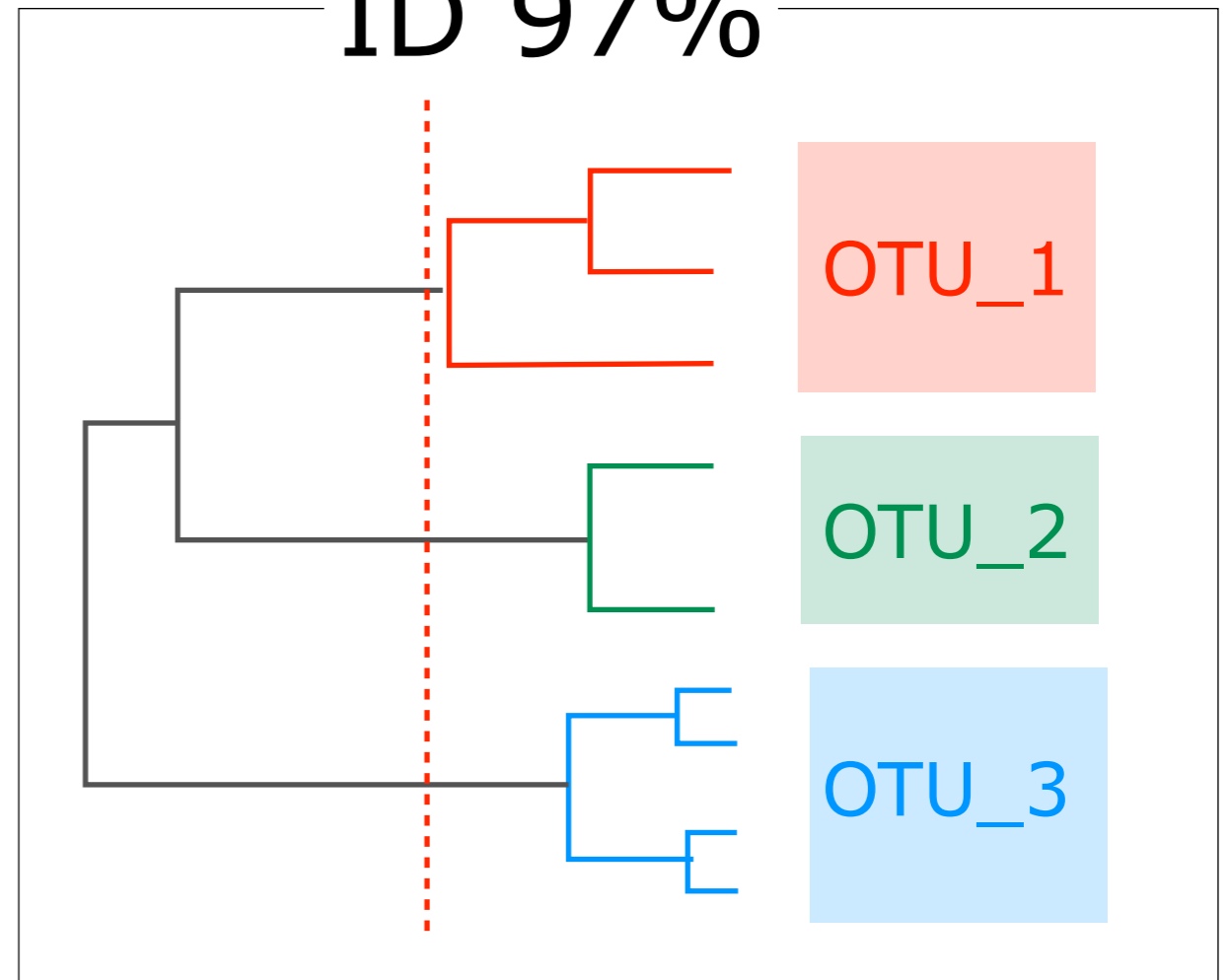
=> abundance filtering

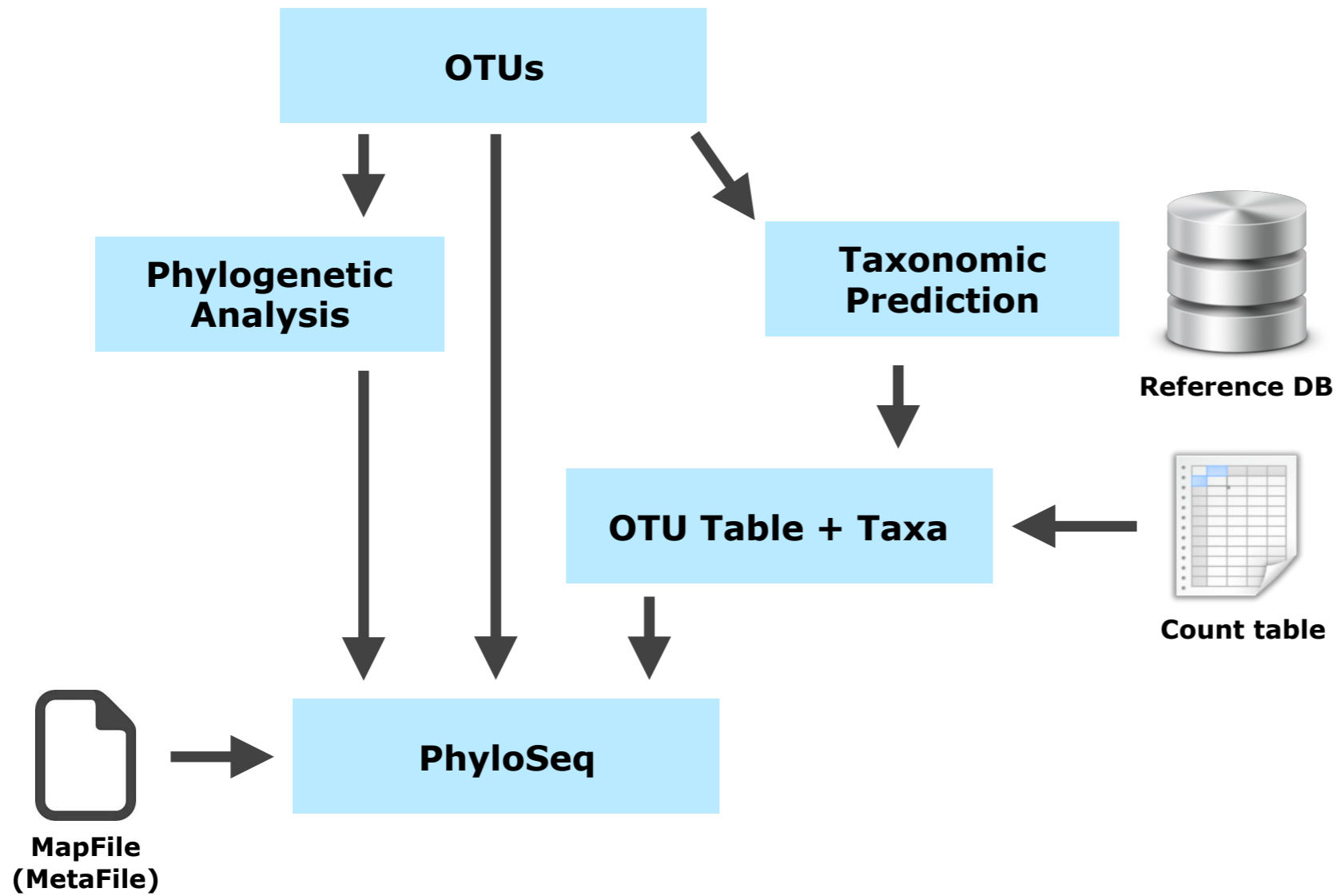
=> cluster threshold

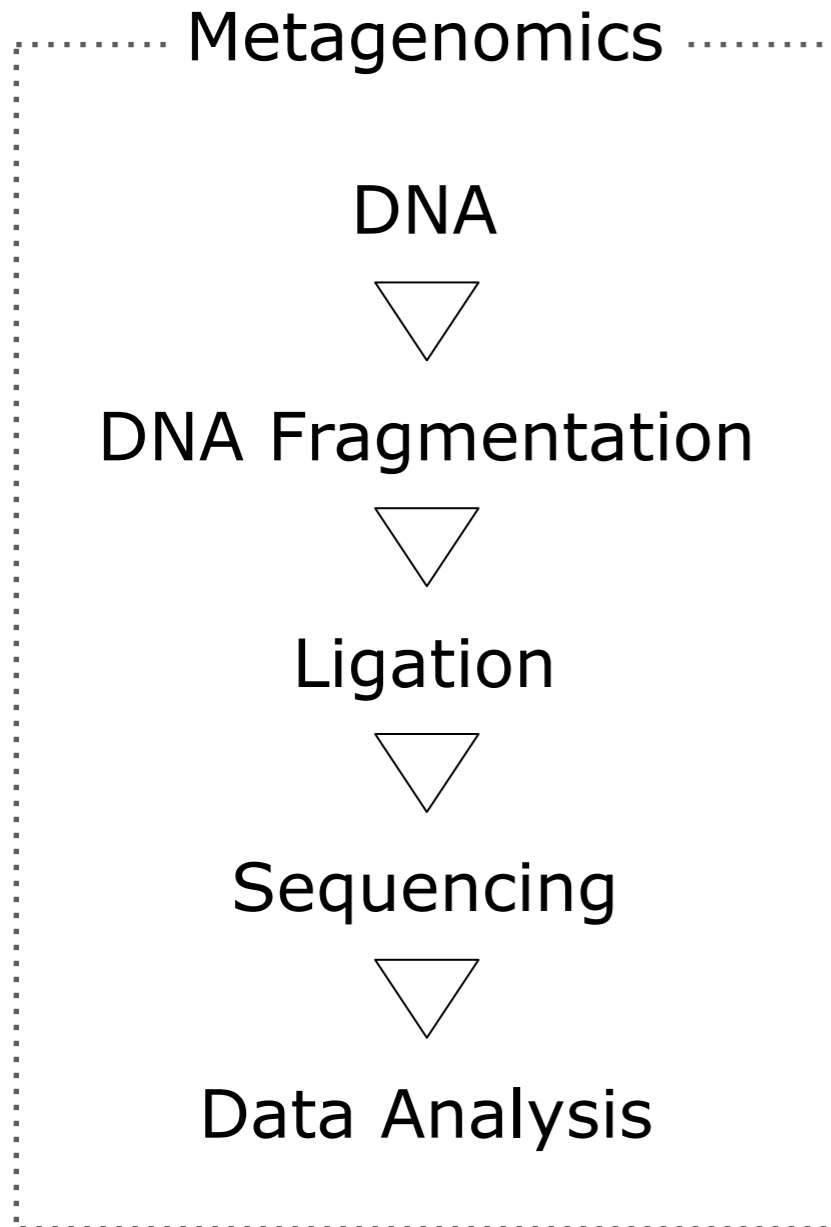
ID 99%



ID 97%







- 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

Introduction

Workflow-Worries

W1

Which data preparation step is the most difficult one for you and why?

W2

Which data preparation step is the most important one and why?