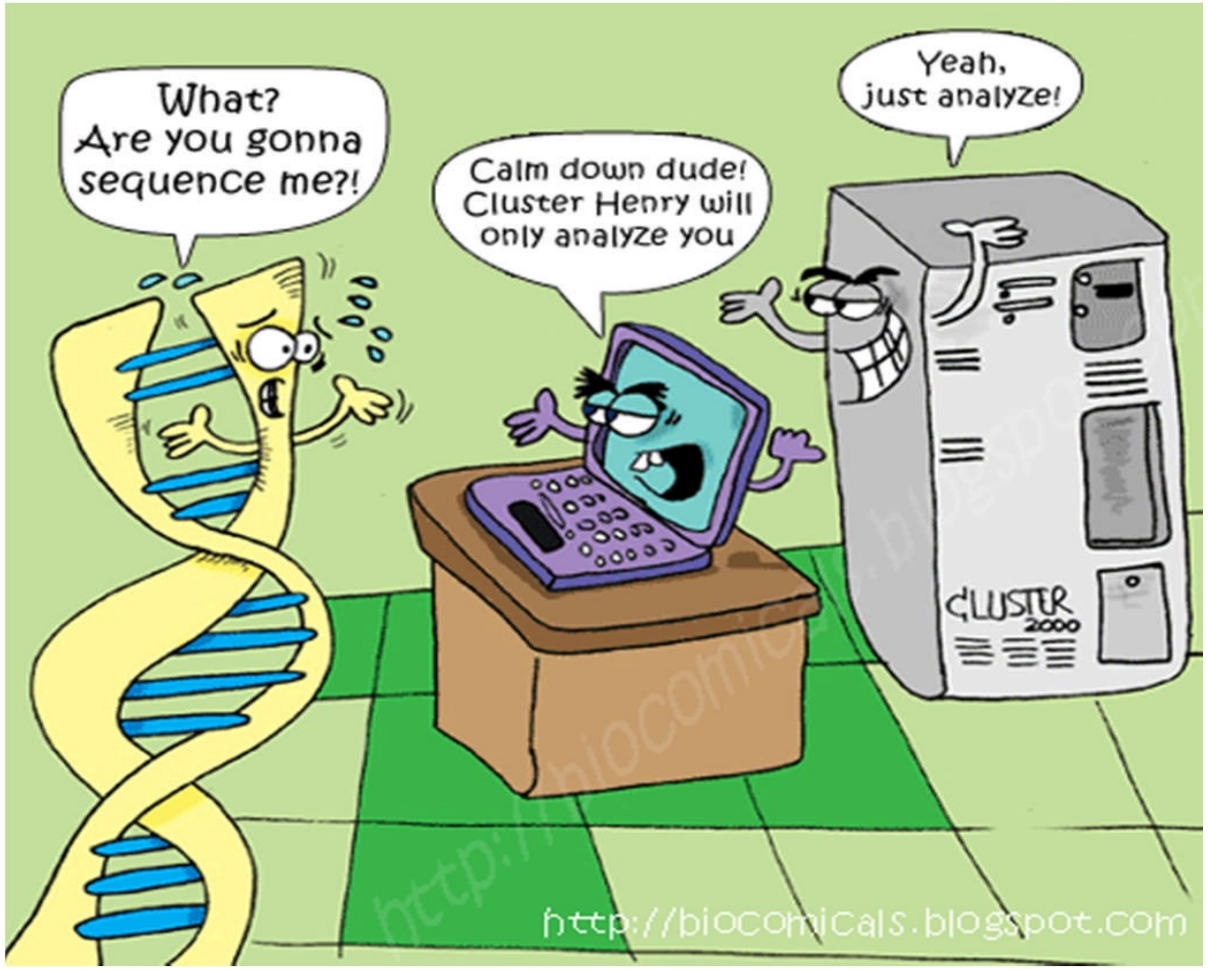
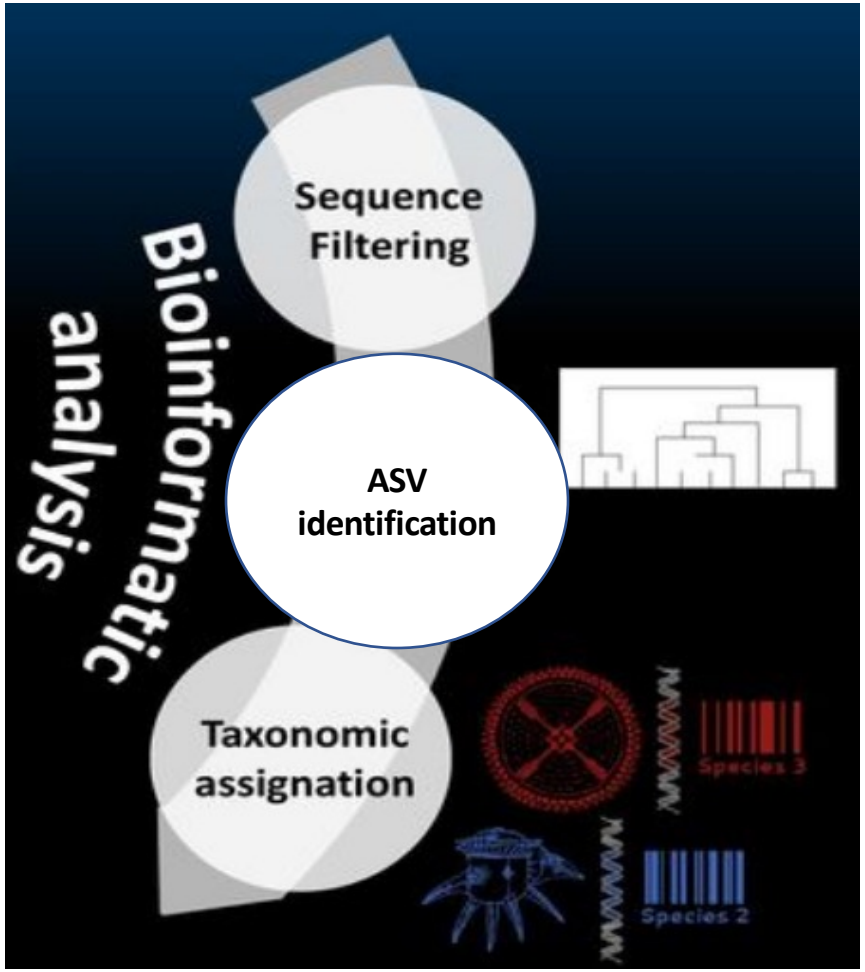
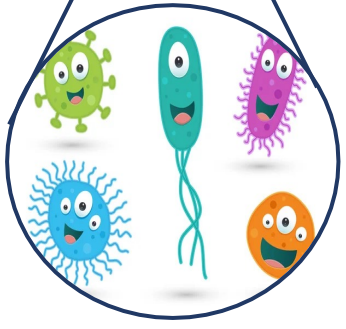
Bioinformatics processing





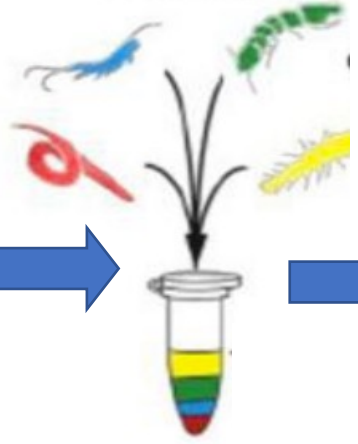
From sample to sequences ….



Microbial community

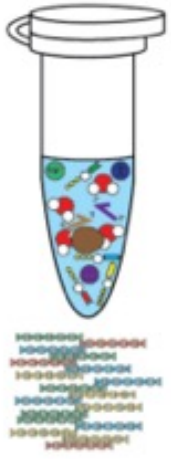
1

DNA extraction from « soup »



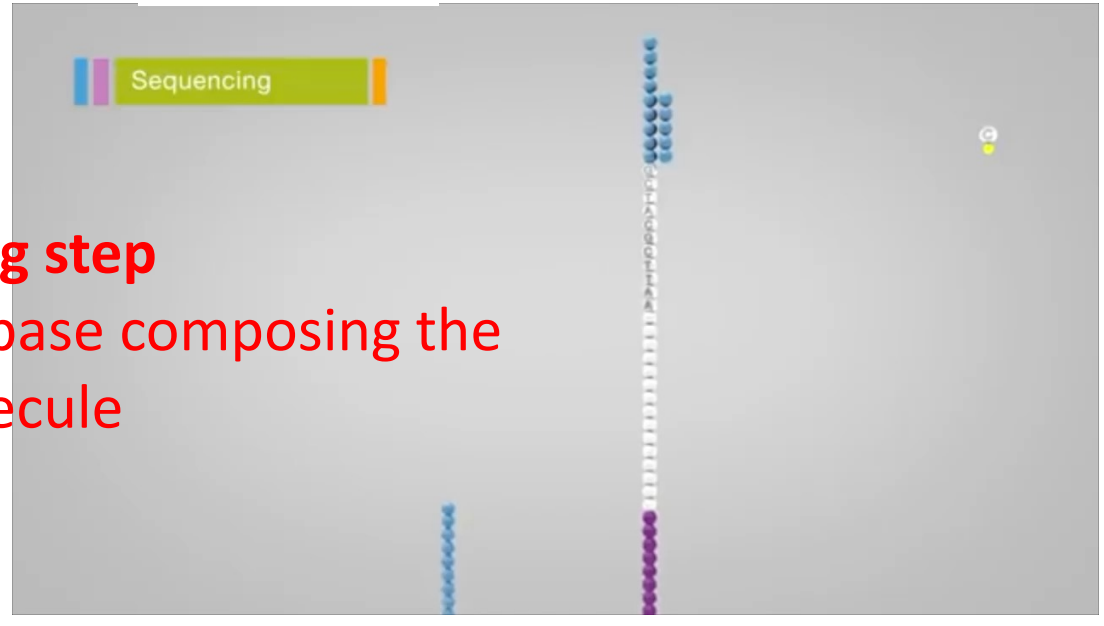
2

Amplify DNA markers
= **genes coding for 16S rRNA**



3

Sequencing of amplified genes



Base calling step
= identification of each base composing the DNA molecule

Quality score = **Q score** = Score Phred

→ The base calling step is associated to a **measure of accuracy** = A **quality score**

→ **This Q score** corresponds to the probability that a given nucleotide base is **called incorrectly** by the sequencer

It allows you to:

- Identify **bases** with a **high probability of error**
- Identify **regions of reads** or reads of **poor quality** in order to eliminate them …

Quality score = **Q score** = Score Phred

→ The base calling step is associated to a **measure of accuracy** = A **quality score**

→ **This Q score** corresponds to the probability that a given nucleotide base is **called incorrectly** by the sequencer

It allows you to:

- Identify **bases** with a **high probability of error**
- Identify **regions of reads** or reads of **poor quality** in order to eliminate them …

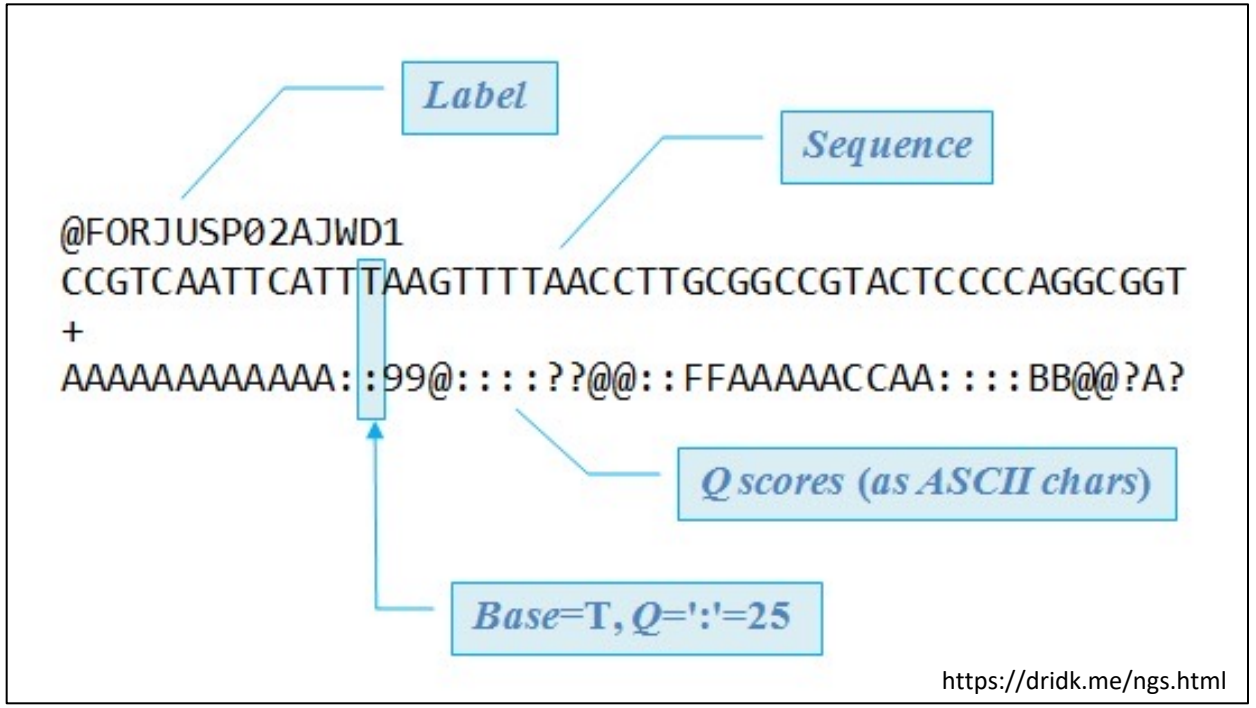
Logarithmically related to the base call error probabilities (P)

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

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

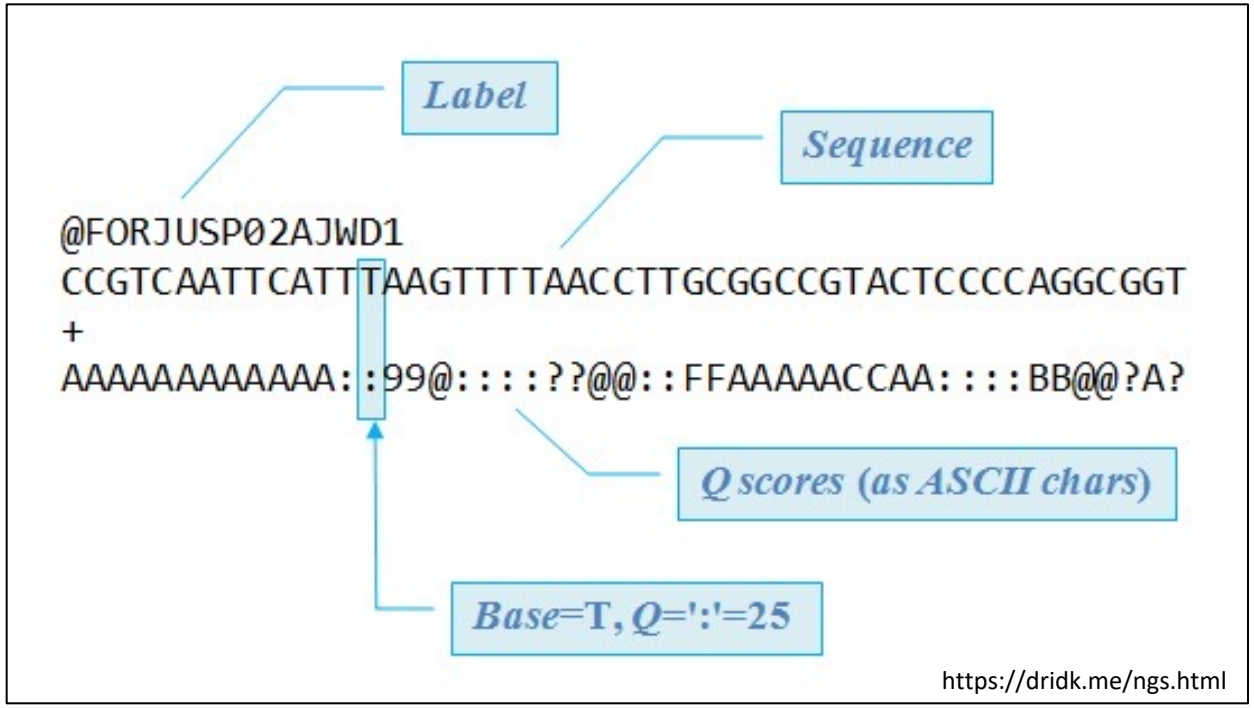
THEN ?

Delivery of sequences in a file in **fastq format**:
Text file containing the **nucleotide sequences** and the **quality scores** associated with each base (scores coded with ASCII characters)



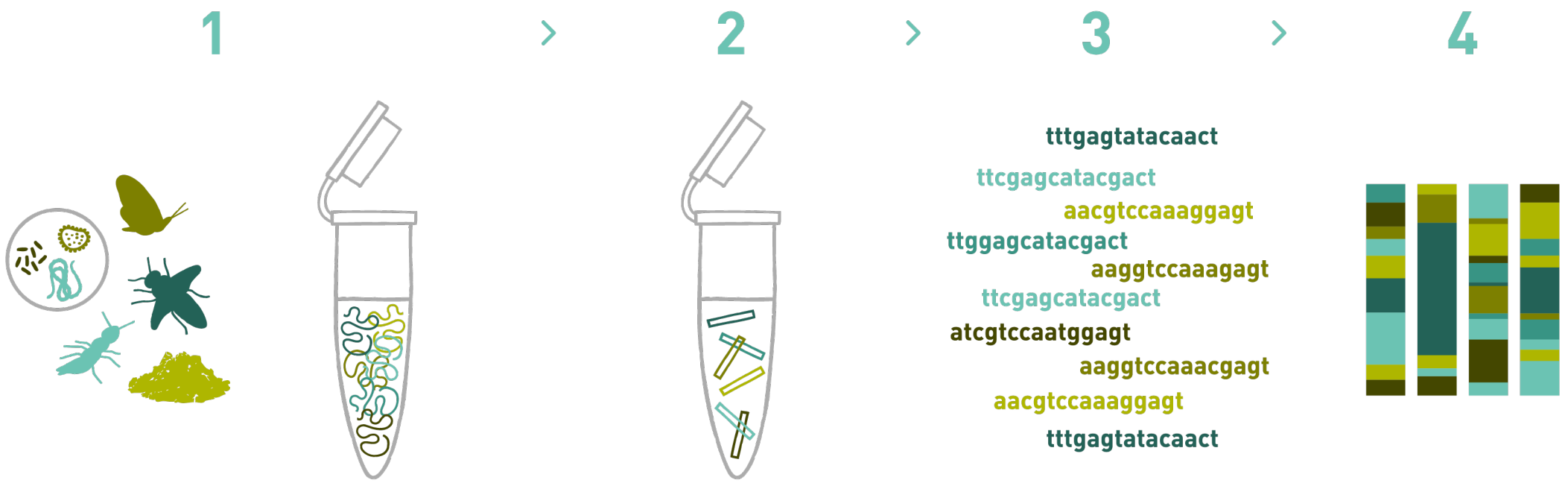
THEN ?

Delivery of sequences in a file in **fastq format**:
Text file containing the **nucleotide sequences** and the **quality scores** associated with each base (scores coded with ASCII characters)



Molecular biology VS *Bioinformatic*

Nucleotide sequences **« reads »**



Metabarcoding : an infaillible method ?

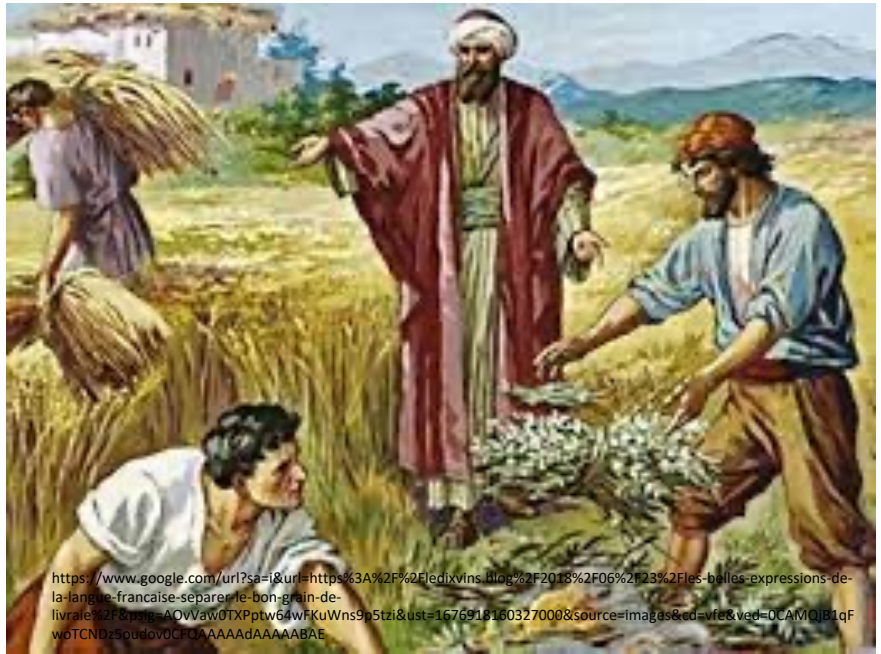
Séparer le bon grain de l'ivraie …

Ecological Applications, 30(2), 2020, e02036

From metabarcoding to metaphylogeography: separating the wheat from the chaff

XAVIER TURON,[1,4] ADRIÀ ANTICH,[1] CREU PALACÍN,[2] KIM PRÆBEL,[3] AND OWEN SIMON WANGENSTEEN[3]

[1]*Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB, CSIC), Blanes, Catalonia, Spain*
[2]*Department of Evolutionary Biology, Ecology and Environmental Sciences, and Institute of Biodiversity Research (IRBio), University of Barcelona, Barcelona, Catalonia, Spain*
[3]*Norwegian College of Fishery Science, UiT the Arctic University of Norway, Tromsø, Norway*



https://www.google.com/url?sa=i&url=https%3A%2F%2Fledixvins.blog%2F2018%2F06%2F23%2Fles-belles-expressions-de-la-langue-francaise-separer-le-bon-grain-de-livraie%2F&psig=AOvVaw0TXPptw64wFKuWns9p5tzi&ust=1676918160327000&source=images&cd=vfe&ved=0CAMQjB1qFwoTCND25oudov0CFQAAAAAdAAAAABAE

Séparer le bon grain de l'ivraie …

Ecological Applications, 30(2), 2020, e02036

True sequences

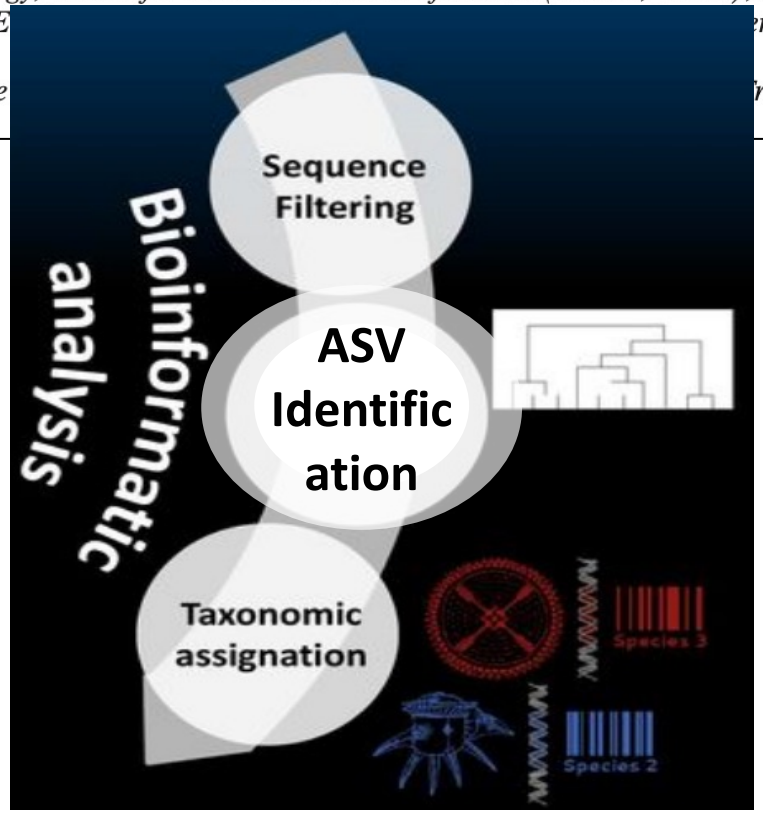
From metabarcoding to metaphylogeography: separating the wheat from the chaff

Artifacts

XAVIER TURON,[1,4] ADRIÀ ANTICH,[1] CREU PALACÍN,[2] KIM PRÆBEL,[3] AND OWEN SIMON WANGENSTEEN[3]

[1]*Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB, CSIC), Blanes, Catalonia, Spain*
[2]*Department of Evolutionary Biology, E… ersity Research (IRBio), University*

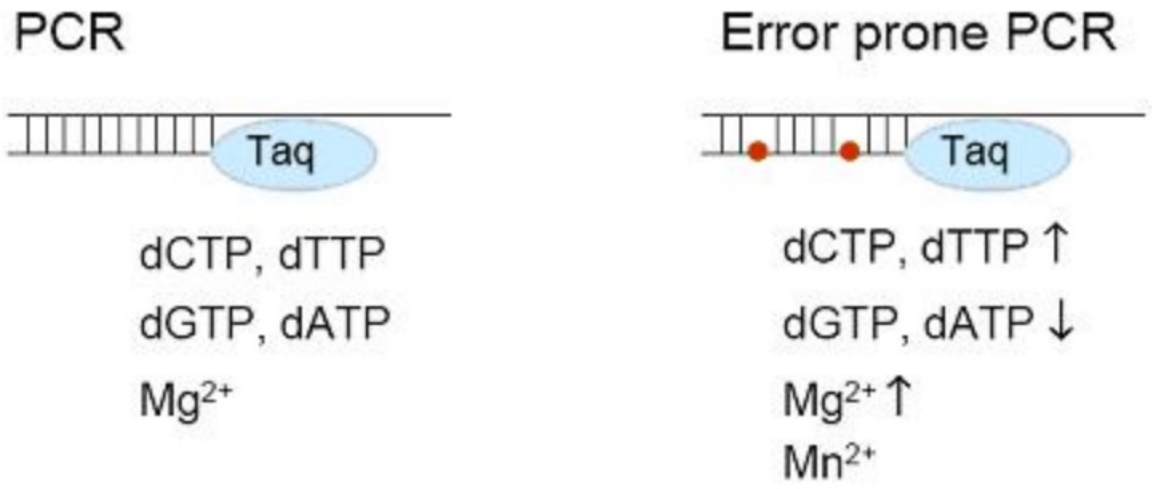
[3]*Norwegian College… Tromsø, Norway*



Error Sources in Metabarcoding

PCR errors

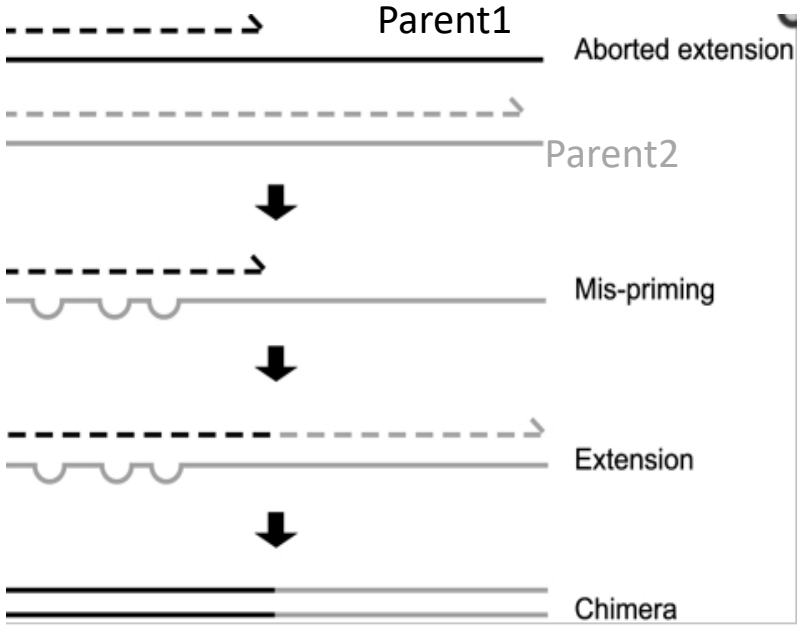
→ mostly substitution error caused by Polymerase

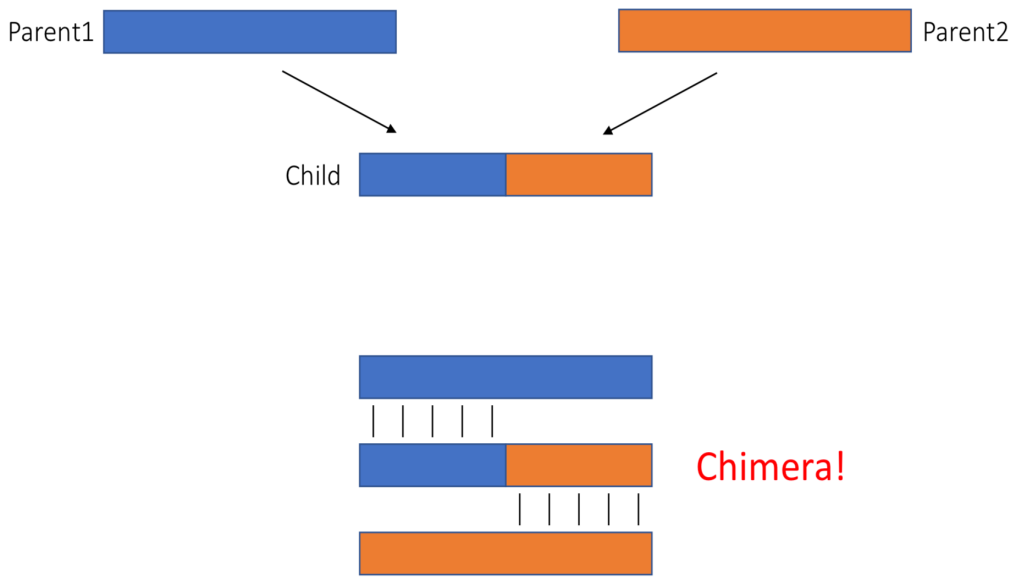


Chimera

→**Polymerase template-switching** on closely-related templates

→ Merge (a least) 2 sequences that belong to 2 different species

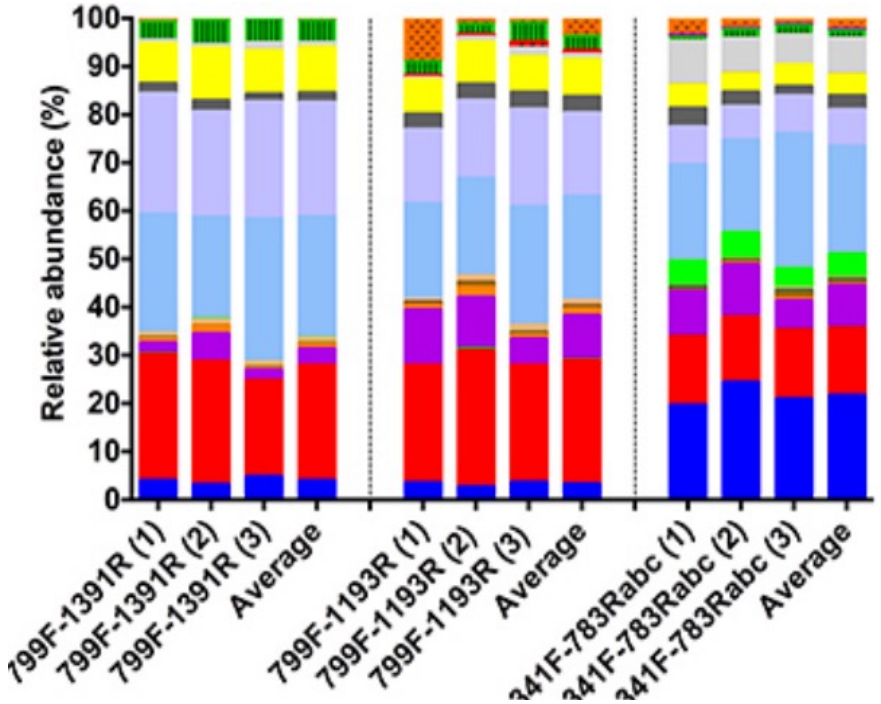




= Are not real biological entities but PCR artifacts !!

Primer bias

→ Different primer sets provide difference in abundance at taxonomic level
→ Variability of primer sensitivity according phyla, genera etc



Beckers et al. 2016

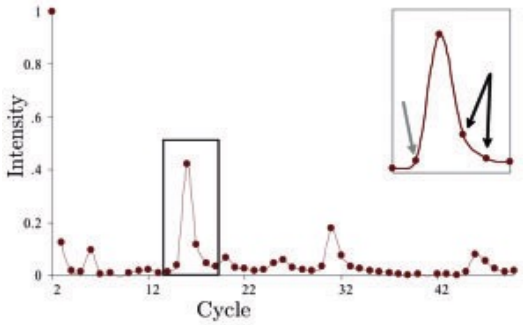
Depth bias

→ No equally amplification of the different templates (preferential targets)
→ The initial more abundant template are more amplified …

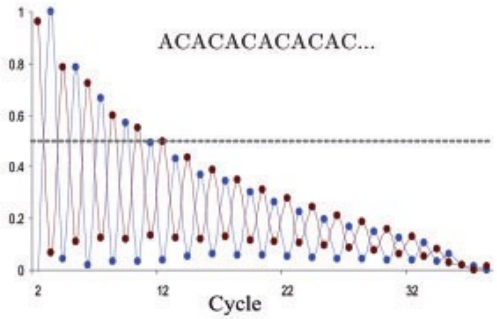
Sequencing errors

Commonly modeled biases in base-callers for the Illumina platform

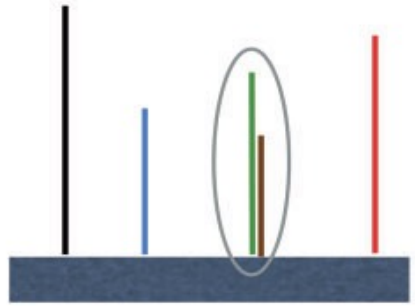
Phasing noise ϕ



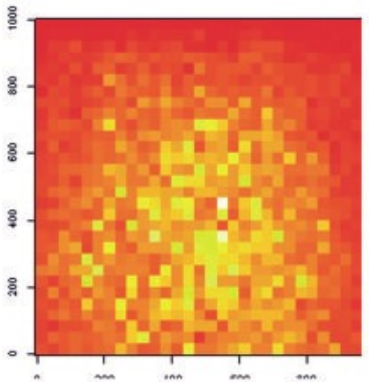
Signal Decay δ



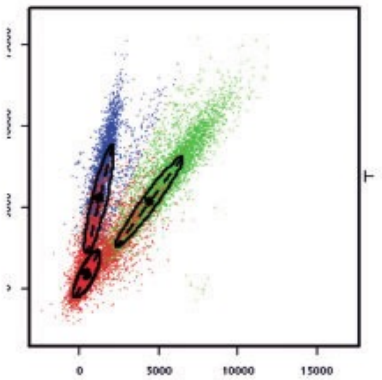
Mixed Cluster μ



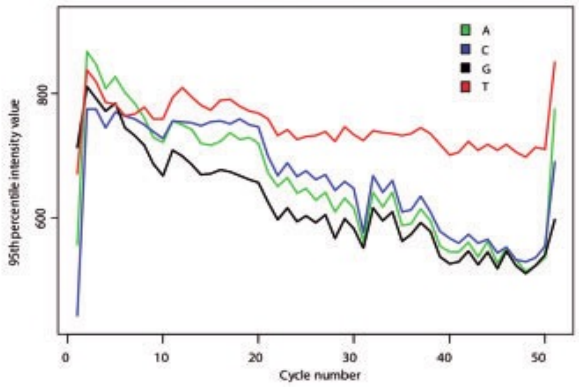
Boundary effects ω



Cross-talk Σ



T fluophore accumulation τ



Ledergerber et al. 2010

16S Sequencing errors

Make it difficult to distinguish biologically real nucleotide differences from sequencing artefacts

Consequence?

Impact the taxonomy assignment resolution (i.e species level, misidentification)

Solution?

The Denoising : correct sequencing errors!

Denoiser Tools

- Deblur *Amir et al. 2017*
- Unoise3 *Edgar et al. 2016*
- Dada2 *Callahan et al. 2016. Nat. Meth.*

New bioinformatic sequence "denoising" approaches have been developed to correct sequencing errors thus improving taxonomic resolution

Processing marker-gene data with…

This workflow assumes that your sequencing data meets certain criteria:

- Samples have been demultiplexed, i.e. split into individual per-sample fastq files.
- Non-biological nucleotides have been removed, e.g. primers, adapters, linkers, etc.
- If paired-end sequencing data, the forward and reverse fastq files contain reads in matched order.

DADA2
Divisive Amplicon Denoising Algorithm

- The goal is **NOT** to find OTU clusters

BUT

- Determine if a sequence read came from **True Variation** or **Sequencing Error**

→ Introduced a model-based approach for correcting amplicon errors

DADA2

Generates a **parametric error model** that is trained on the entire sequencing run and then **applies that model to correct and collapse the sequence errors** into what the authors call **amplicon sequence variants** (ASVs)

ASV vs. OTU

Sequence 1 : ATCGT-------
Sequence 2: GCATC-------
Sequence 3: GCAT**G**-------

Seq.1 = ASV1 (15 read)
Seq.2 = ASV2 (40 reads)
Seq.3 = ASV3 (25 reads)

ASV1 = *Listeria monocytogenes*
ASV2 = *Streptoccocus pneumoniae*
ASV3 = *Streptoccocus pneumoniae*

Same species

Same strain
But different copies

2 different strains

ASV = Variant d'amplicon (i.e. : Une séquence **UNIQUE**)

ASV vs. OTU

Sequence 1 : ATCGT-------
Sequence 2: GCATC-------
Sequence 3: GCAT**G**-------

Seq.1 = ASV1 (15 read)
Seq.2 = ASV2 (40 reads)
Seq.3 = ASV3 (25 reads)

ASV1 = *Listeria monocytogenes*
ASV2 = *Streptoccocus pneumoniae*
ASV3 = *Streptoccocus pneumoniae*

Same species

Same strain
But different copies

2 different strains

ASV = Variant d'amplicon (i.e. : Une séquence **UNIQUE**)

Choice of an **identity threshold** (=similarity) between sequences (>97% for rank of species)

Sequence clustering

Streptococcus pneumoniae

OTU = Unité taxonomique opérationnelle
= regroupement de séquence sur la base d'un seuil de similitude fixé

The model relies on

Read abundances : True reads are likely to be more abundant

Distances : Less abundant reads with only a few base-differences away from a more abundant sequence are likely error-derived (Hamming distance)

Q-scores : Calculate a substitution model, estimating a probability for each possible base substitution (e.g. A replacing G, G replacing T, etc).

→ DADA2 uses a **probability threshold** to decide whether to assign counts from a less abundant, “error-derived” read to a more abundant, “true” sequence

True Variation or Sequencing Error?

Sequence Read 1: acttcatgataccacatgatacg

Sequence Read 2: acttcatgctaccacatgatacg

True Variation or Sequencing Error?

Sequence Read 1: acttcatgataccacatgatacg

Sequence Read 2: acttcatgctaccacatgatacg

	Abundance	Quality Score	Base Transitions
Sequence 1	50,000	42	C -> A
Sequence 2	400	14	A -> C

True Variation or Sequencing Error?

Sequence Read 1: acttcatgataccacatgatacg

Sequence Read 2: acttcatgctaccacatgatacg

	Abundance	Quality Score	Base Transitions
Sequence 1	50,000	42	C -> A
Sequence 2	40,000	35	A -> C

True Variation or Sequencing Error?

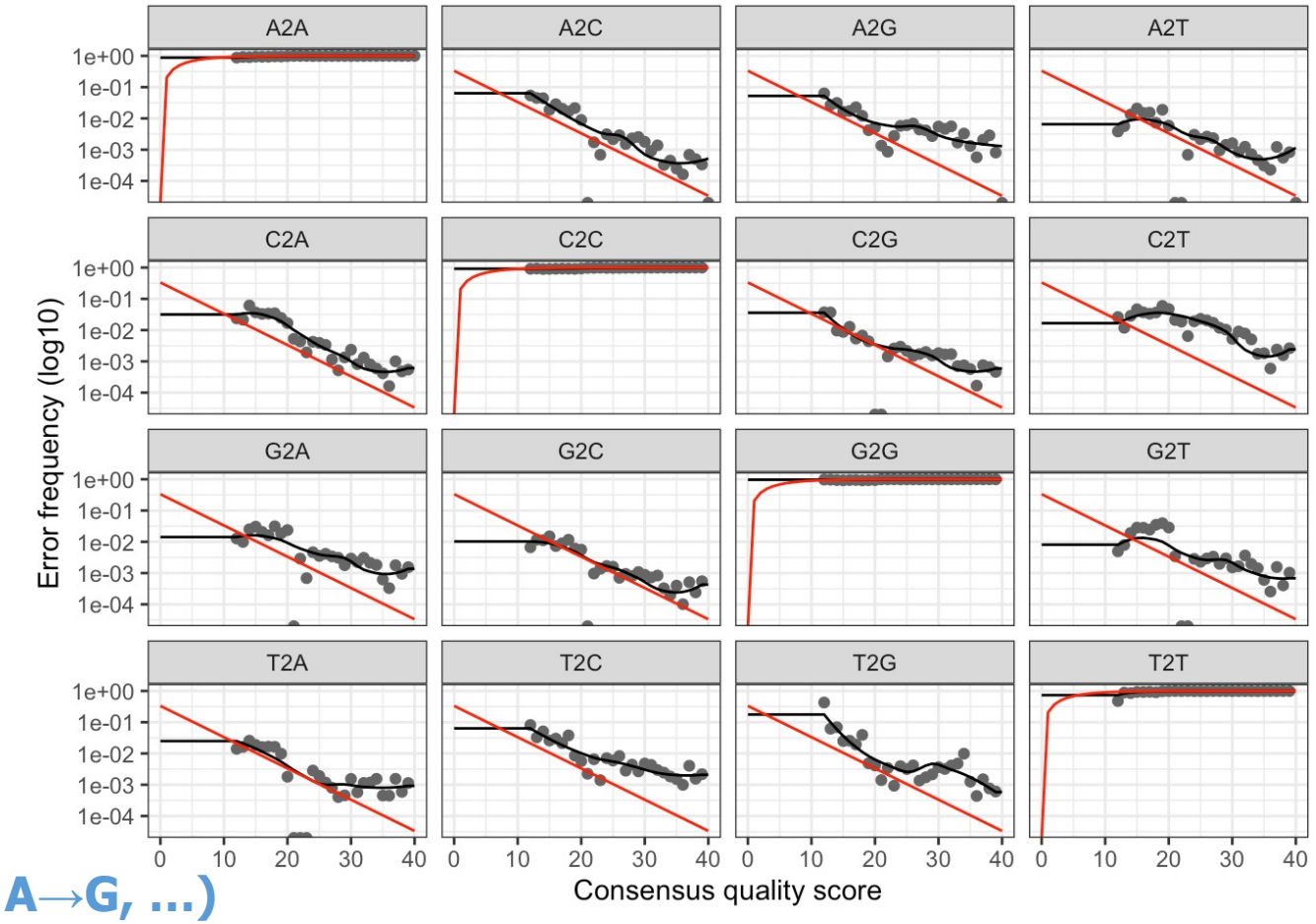
Sequence Read 1: acttcatgataccacatgatacg

Sequence Read 2: acttcatgctaccacatgatacg

	Abundance	Quality Score	Base Transitions
Sequence 1	50,000	42	C -> A
Sequence 2	40,000	35	A -> C

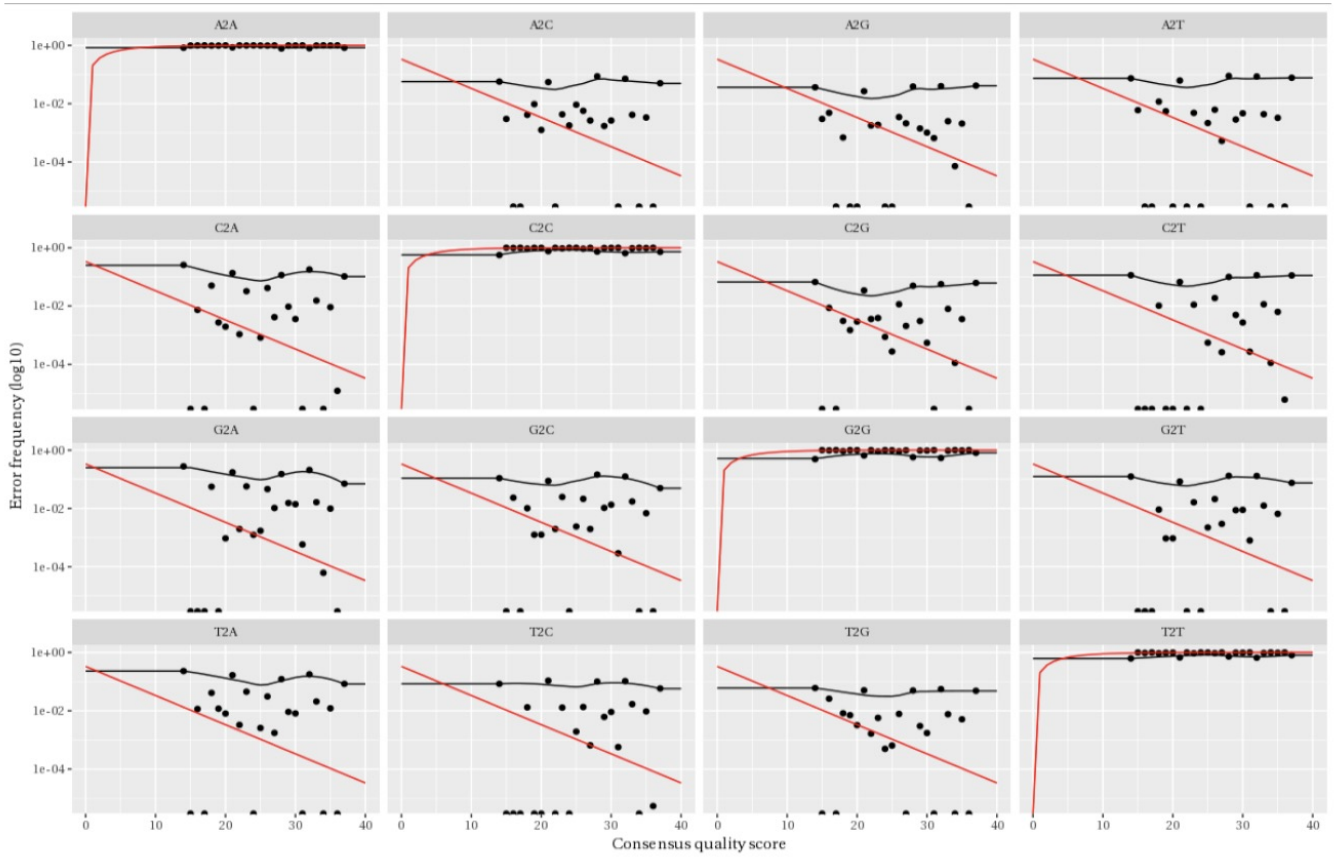
Base Correction is function of read abundance, Quality score and transition probability

Error model Estimation



- Error for each possible transition (A→C, A→G, …)
- **Points** are the observed error rates for each consensus quality score
- **The black line** shows the estimated error rates after convergence of the machine-learning algorithm
- **Important** → error rates drop with increase quality as expected

How to be confident with the model error estimation??



The estimated error rates (black line) are NOT a good fit to the observed rates (points)!! ➔ Bad data!

ASV and heterogeneity of 16S within same genome!

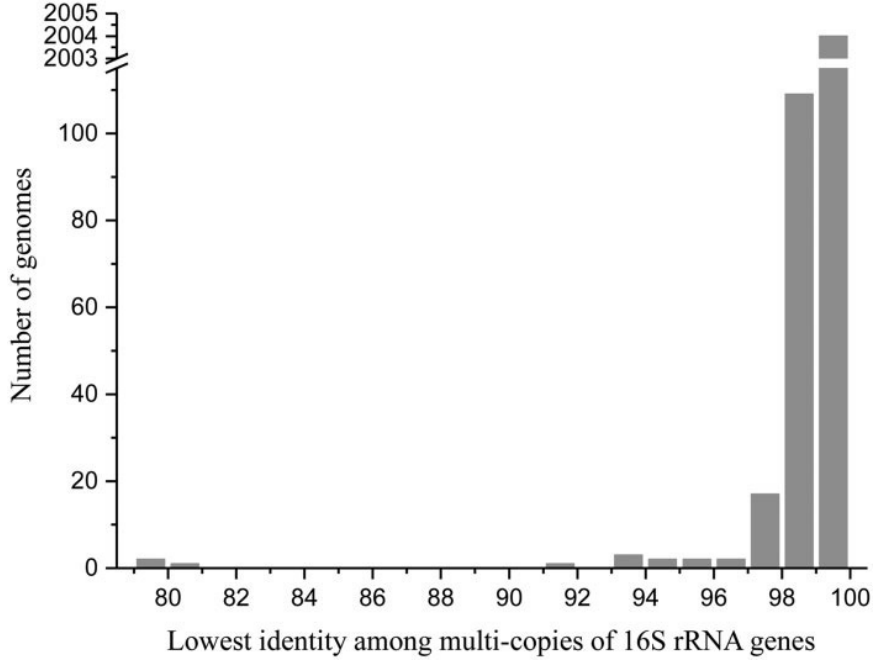
- Heterogeneity: Multicopy of 16S whitin species/strain genome

→ ranging from 1 to 15 copies depending on the species (Klappenbach *et al.* 2000)

- Variability of the 16S from different strains of species

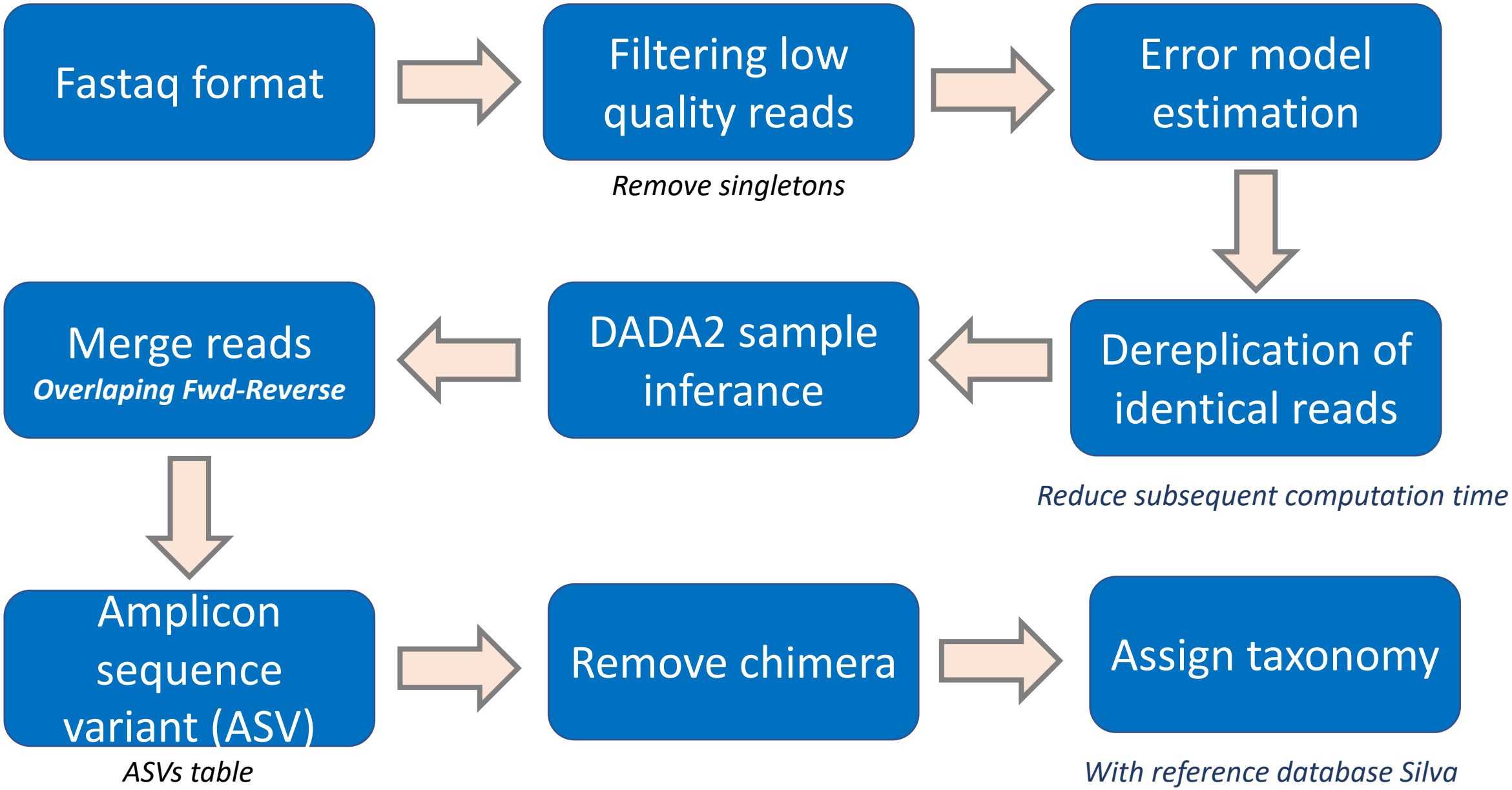
→Some genomes have a ribosomal sequence variation of up to 11% !

Multicopy of 16S introduces a bias in estimating the relative abundance of different organisms in the sample

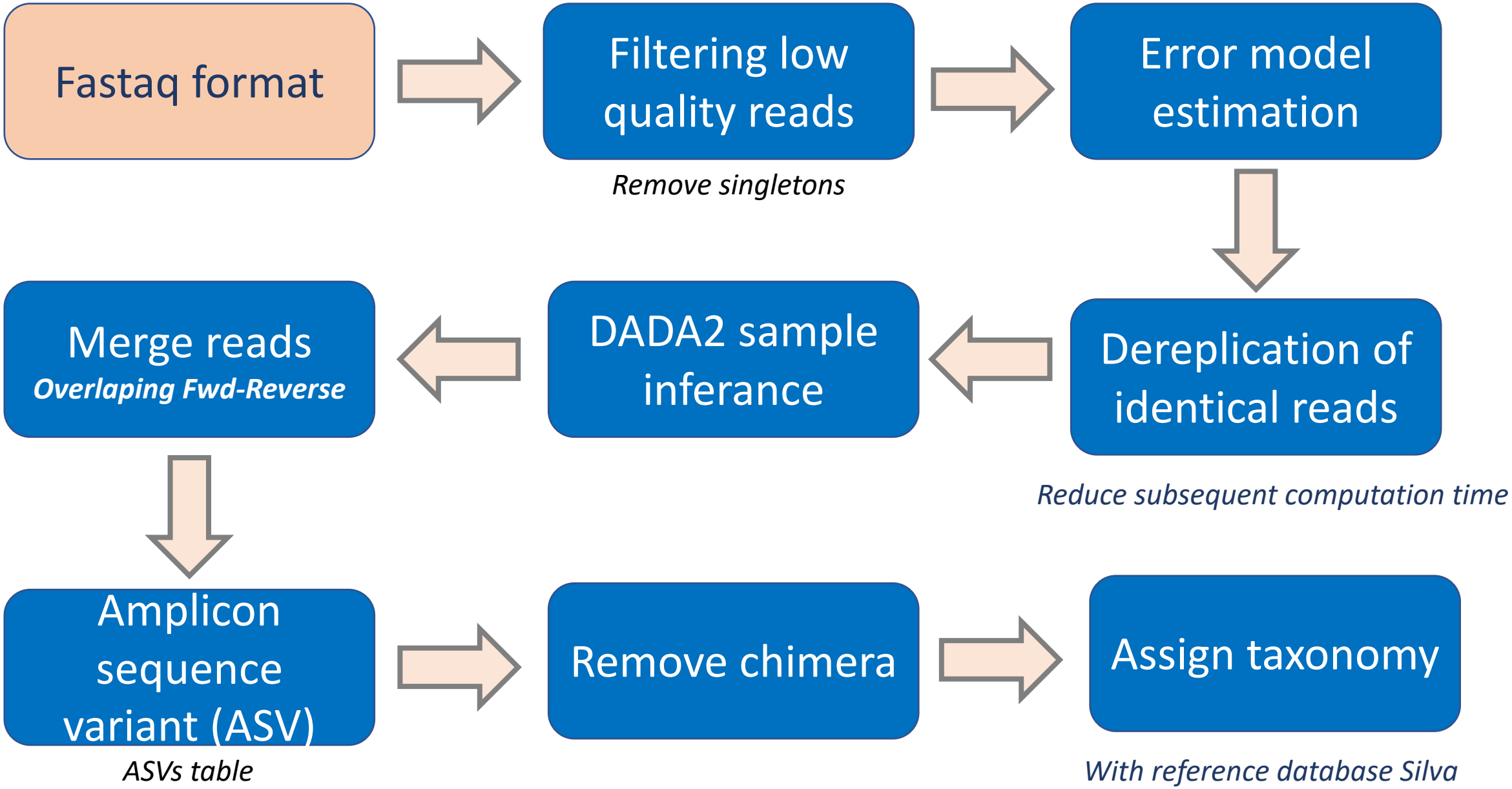


Tian et al. 2015

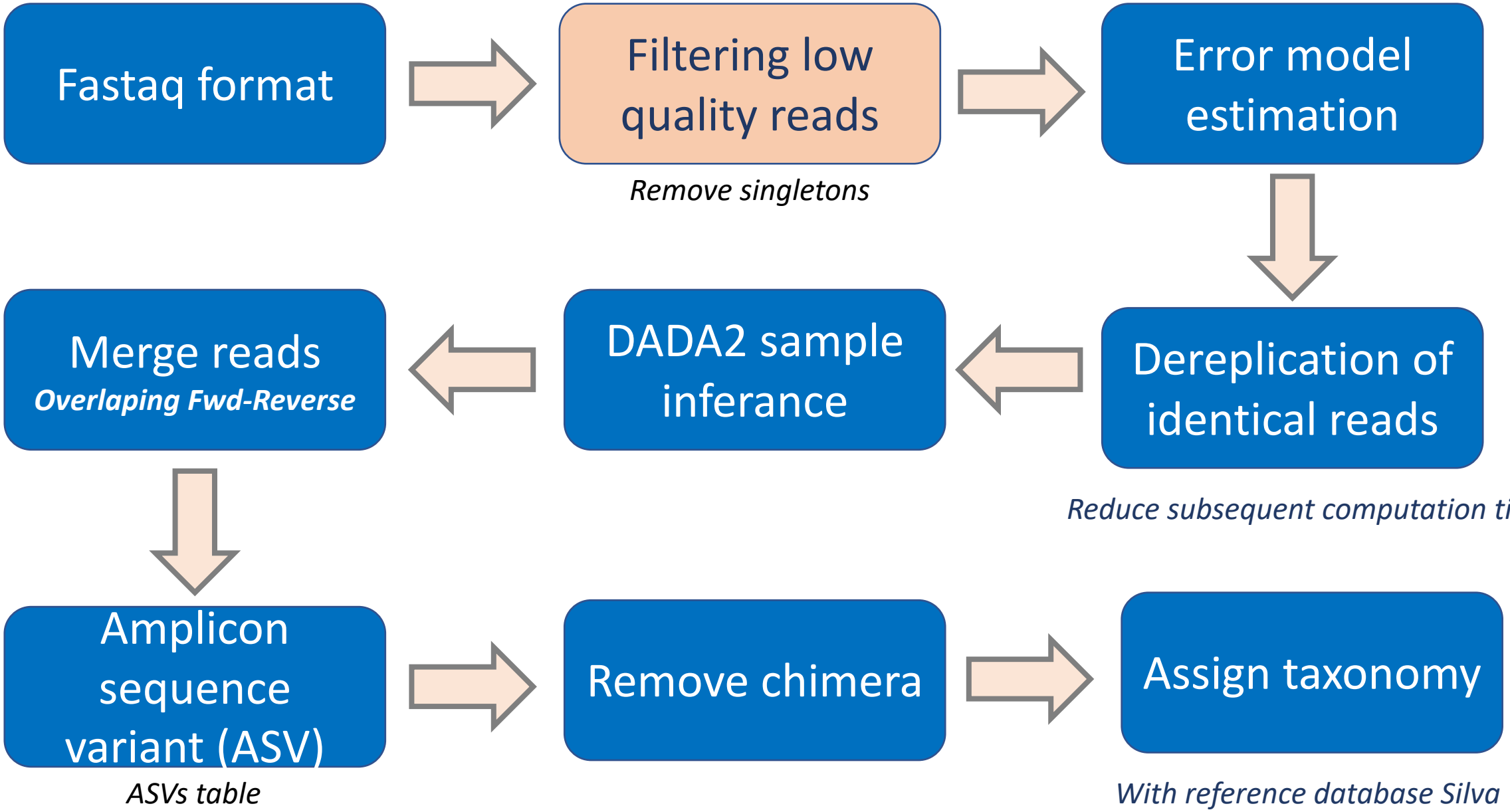
DADA2 Workflow



DADA2 Workflow



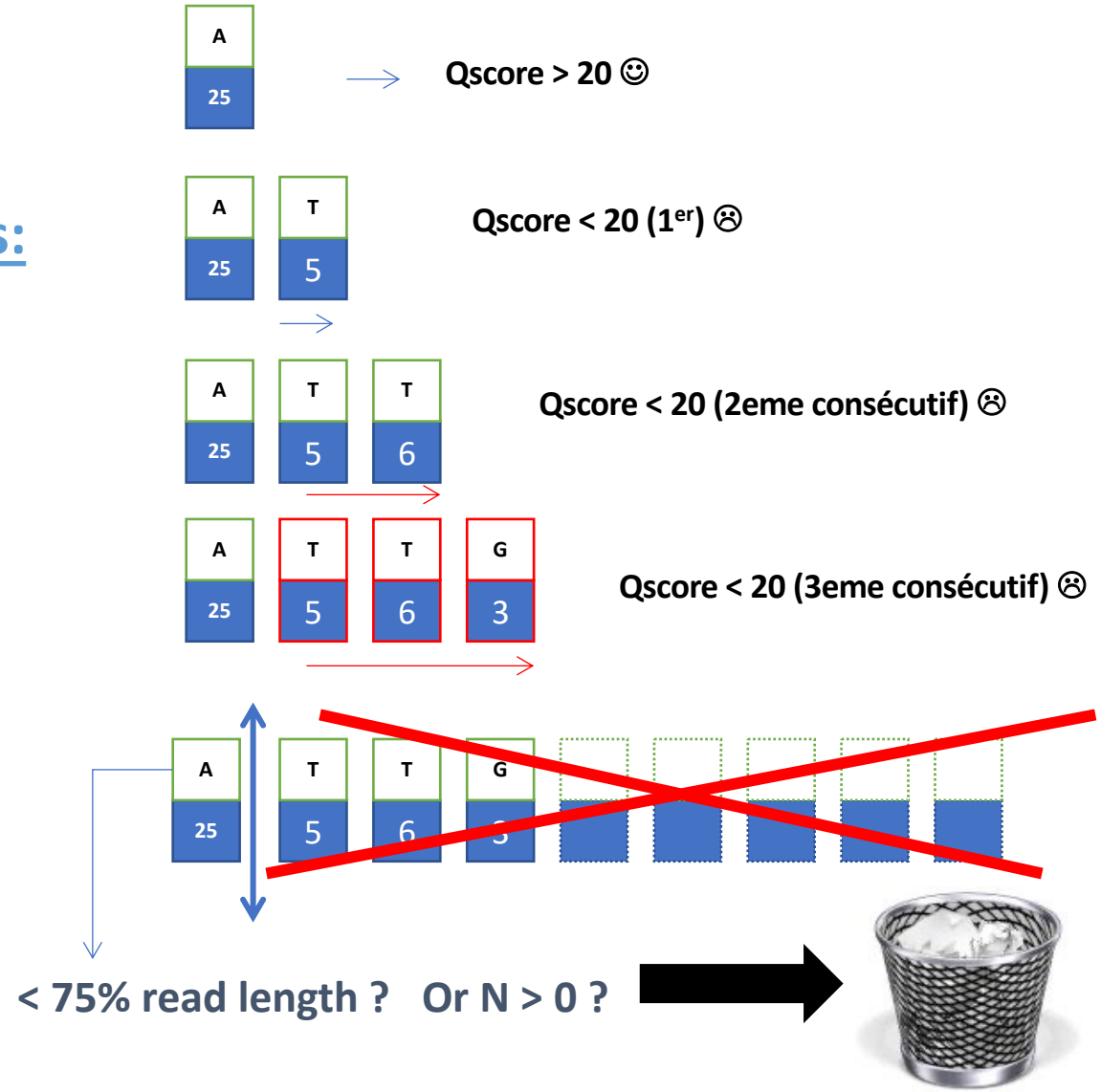
DADA2 Workflow



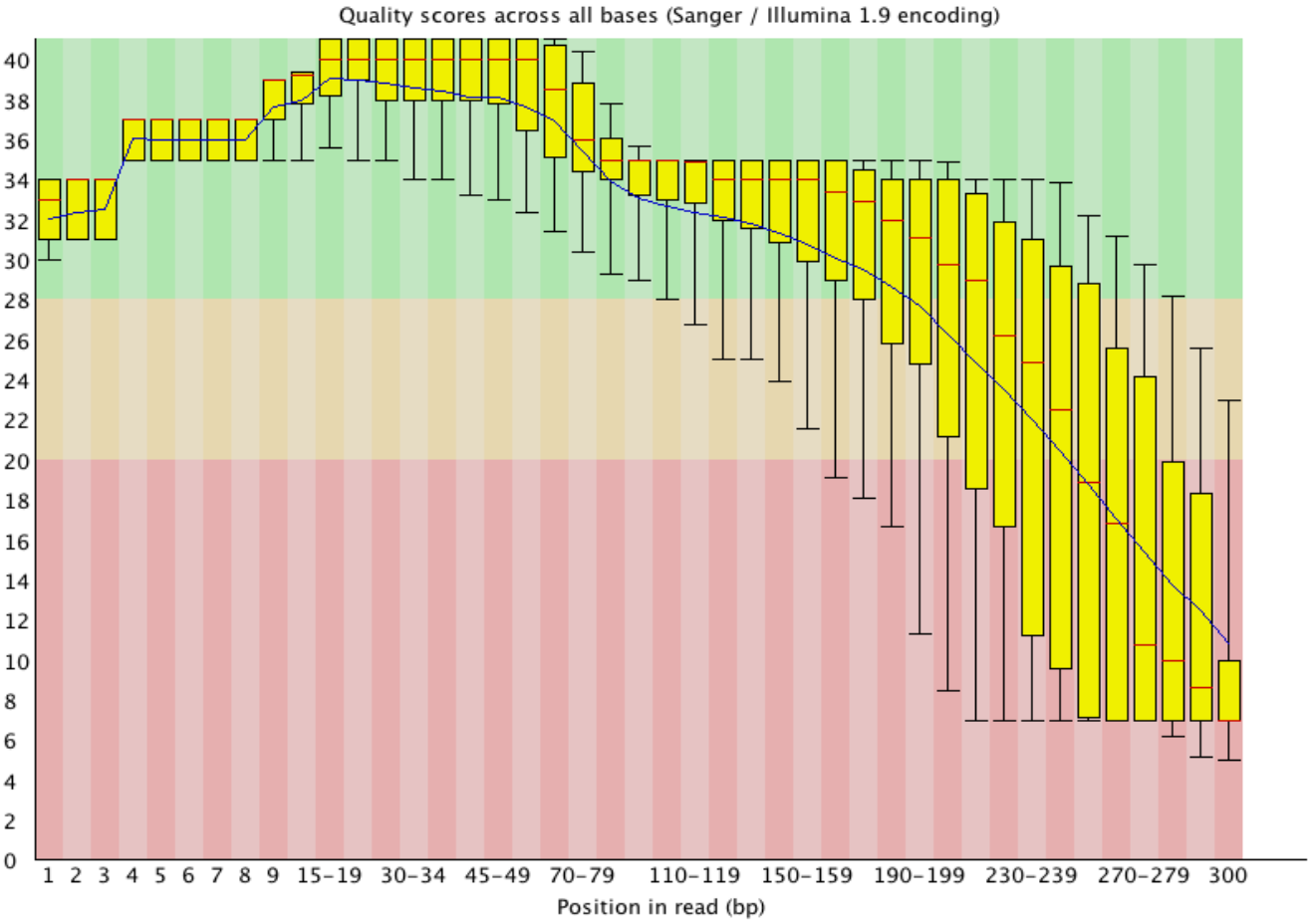
Quality Filtering: Global vs. local

Remove bad quality sequencing reads:

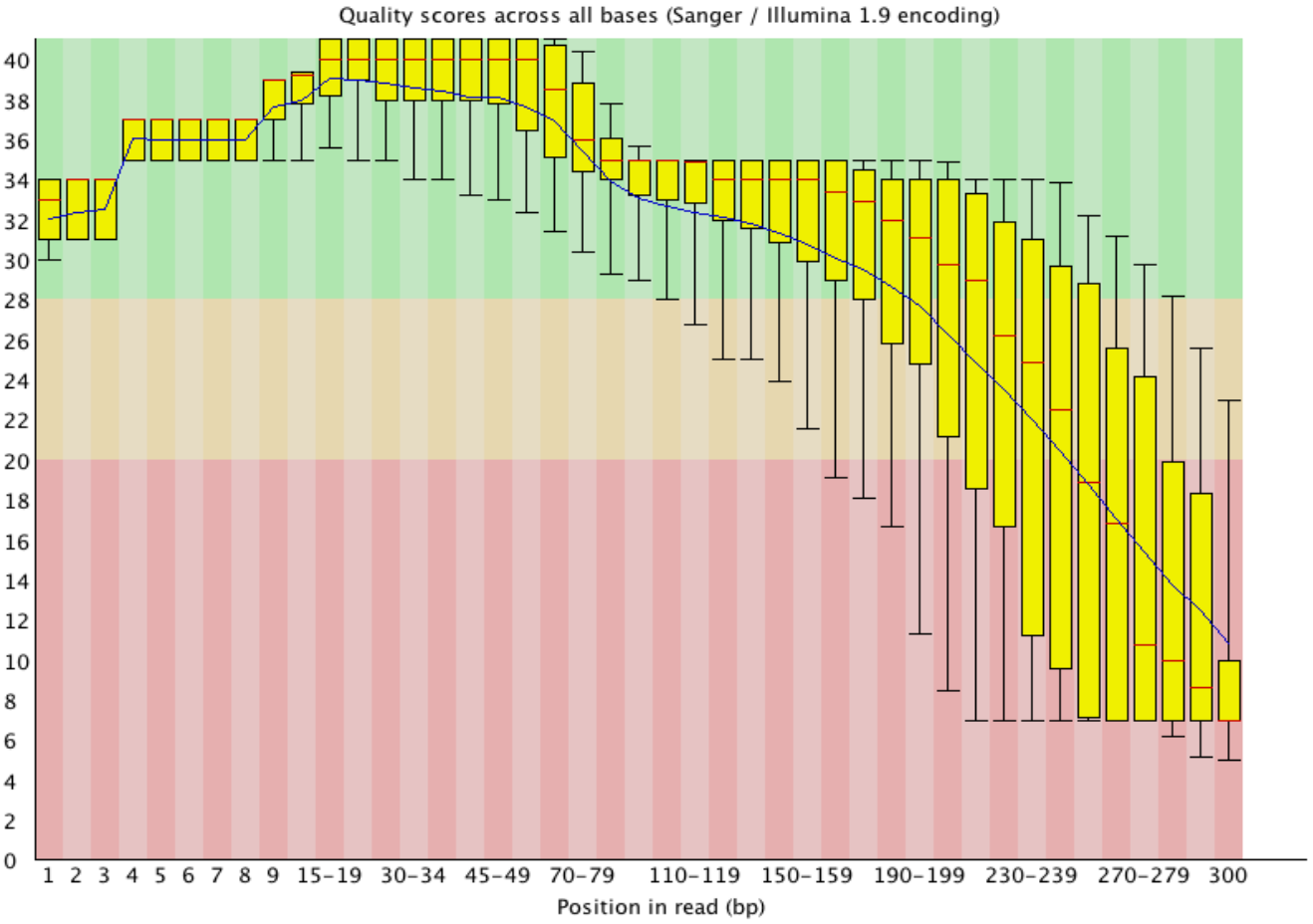
- Avoid ambiguous bases « N »
- Define Min/max length of reads
- Define cut-off for base Quality (Qscore max=40; min=0)



Filtering bad reads....

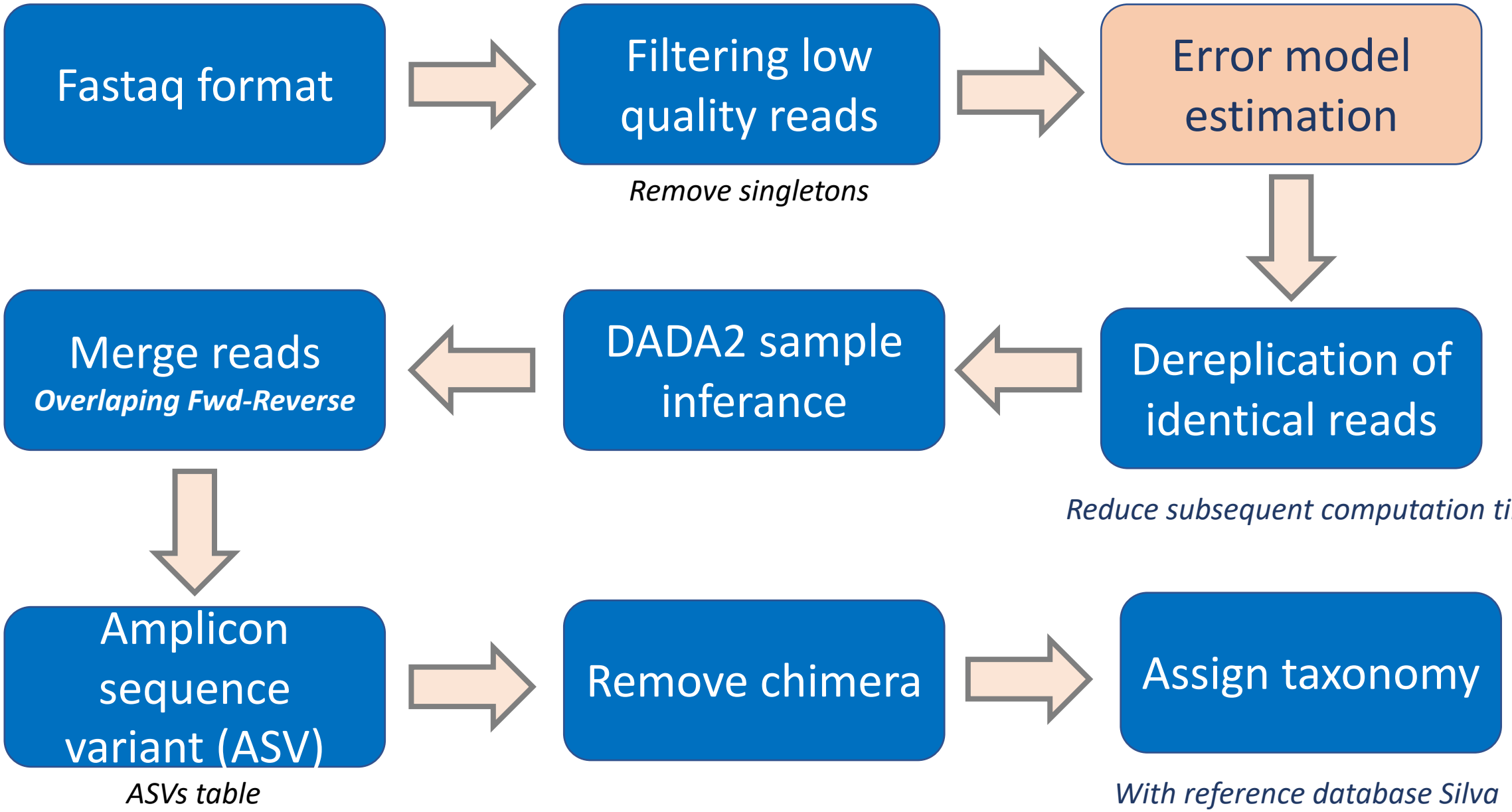


Filtering bad reads....

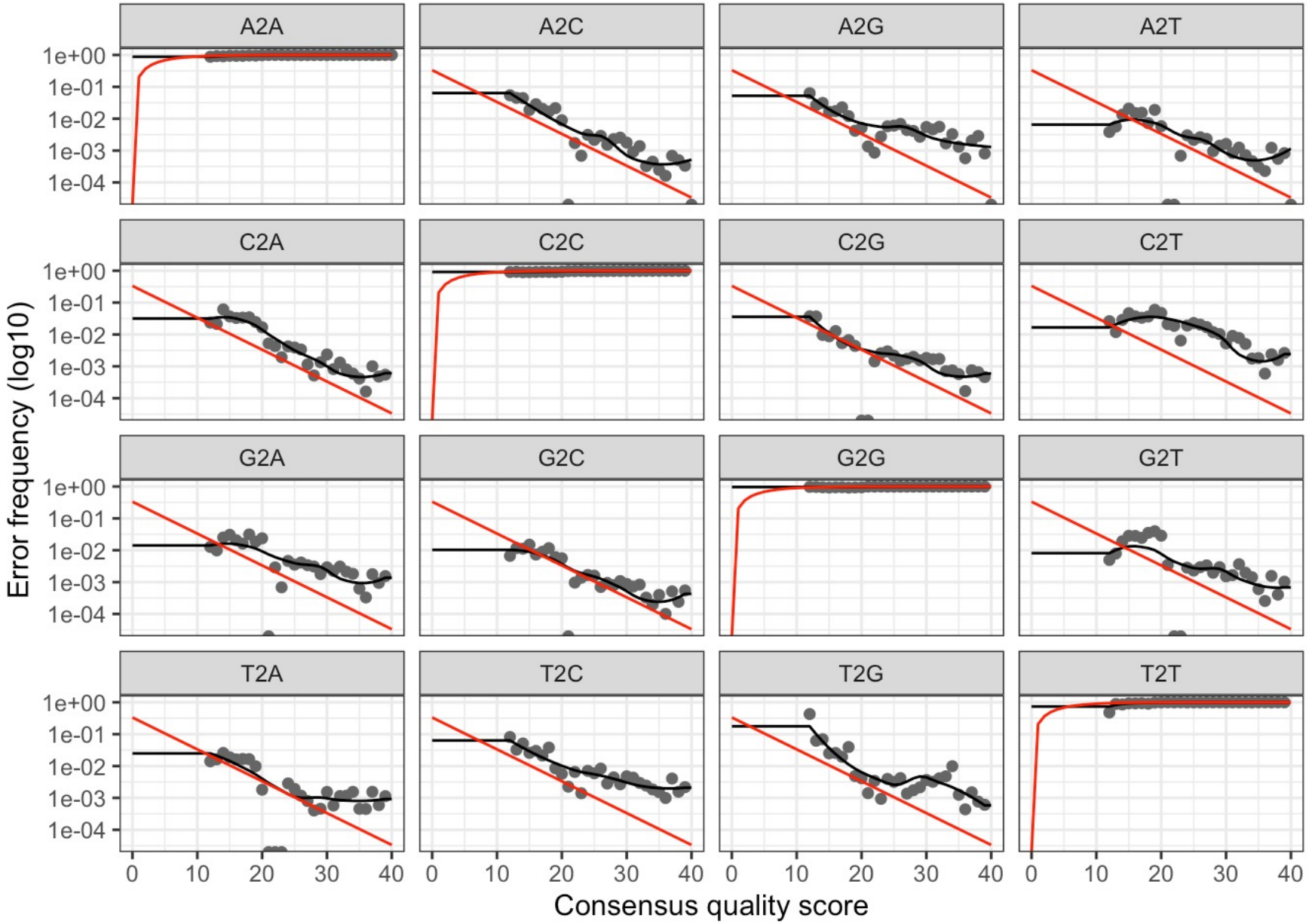


... Remove Singleton

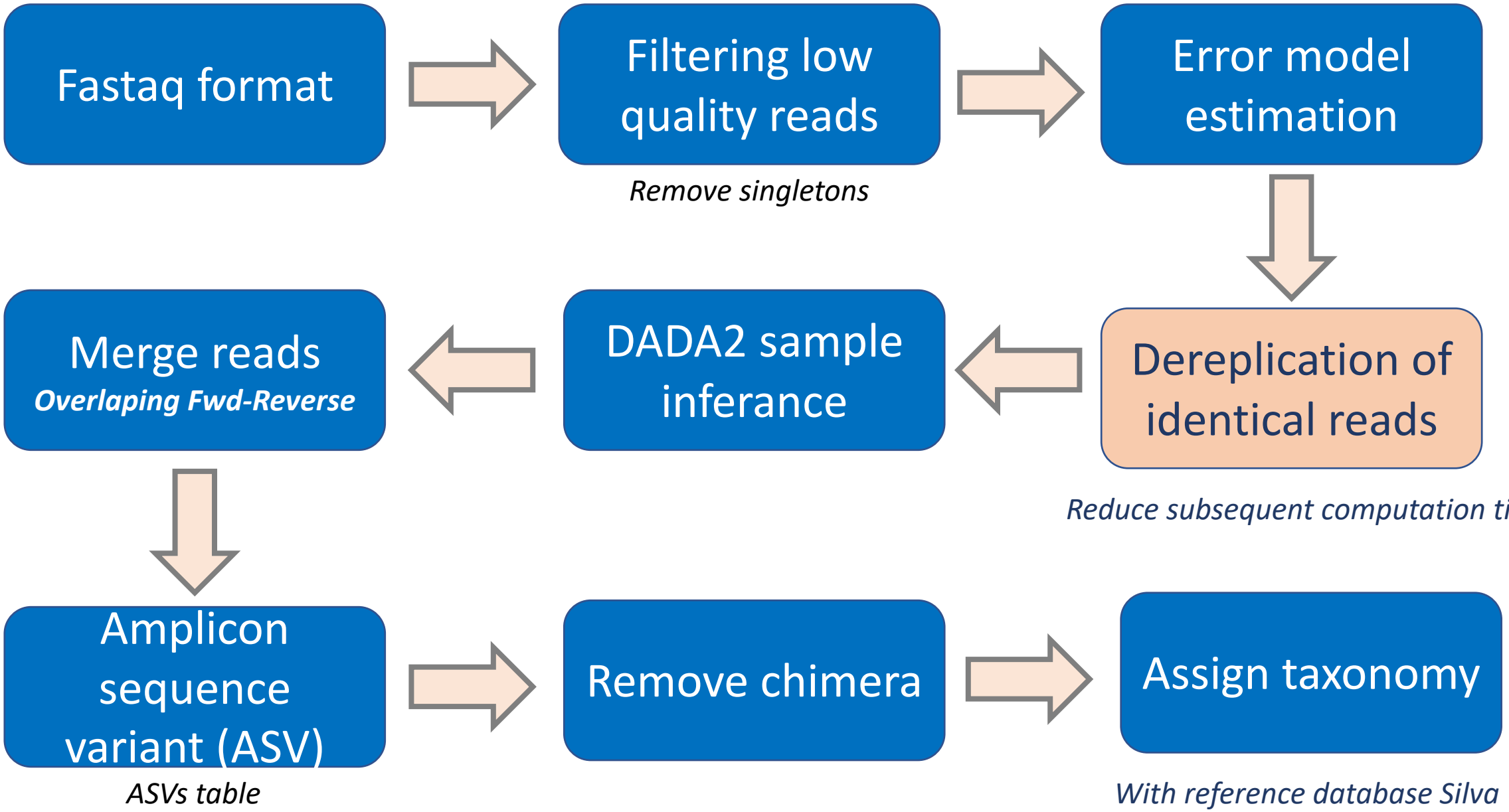
DADA2 Workflow



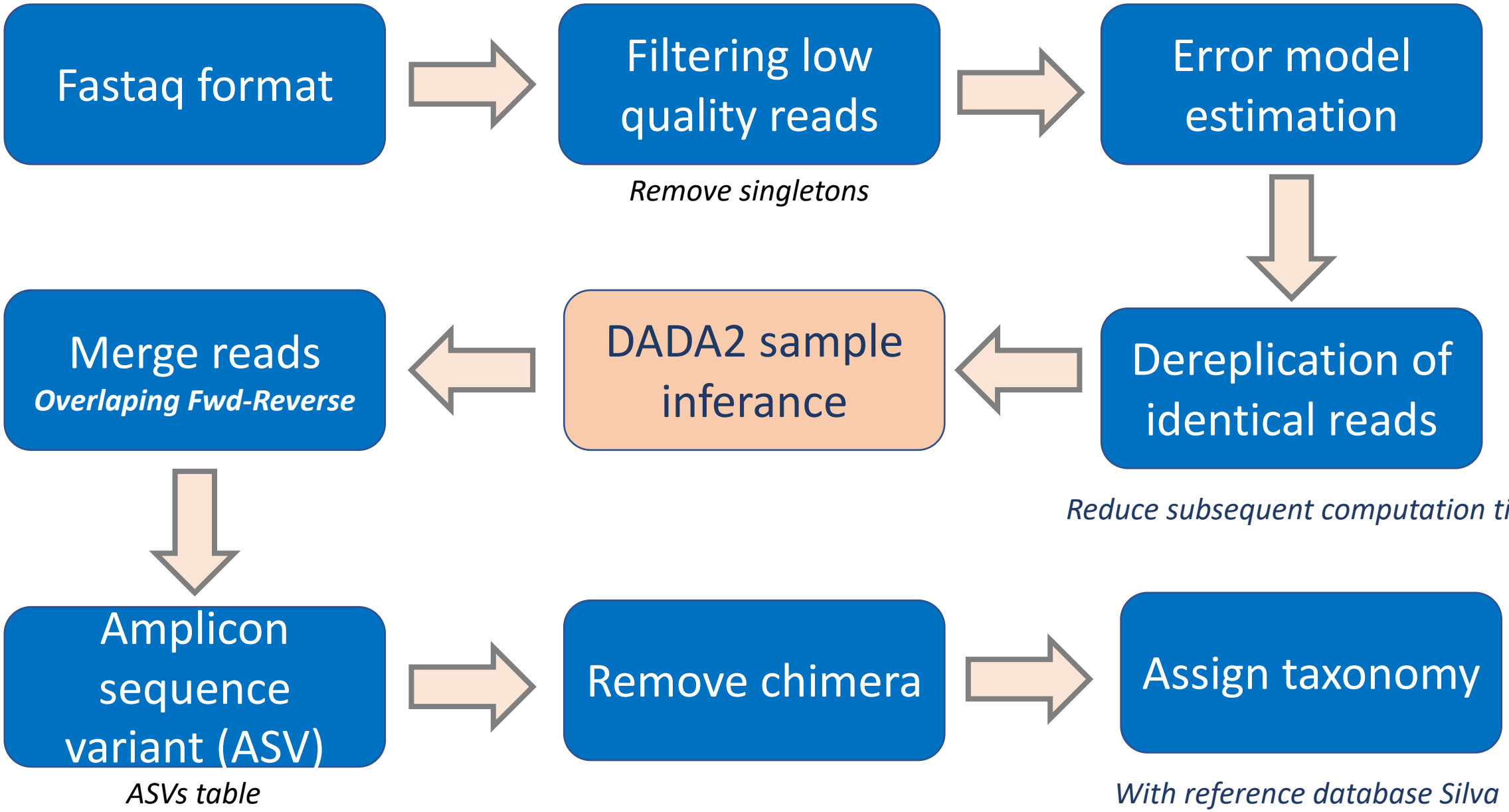
DADA2 Workflow



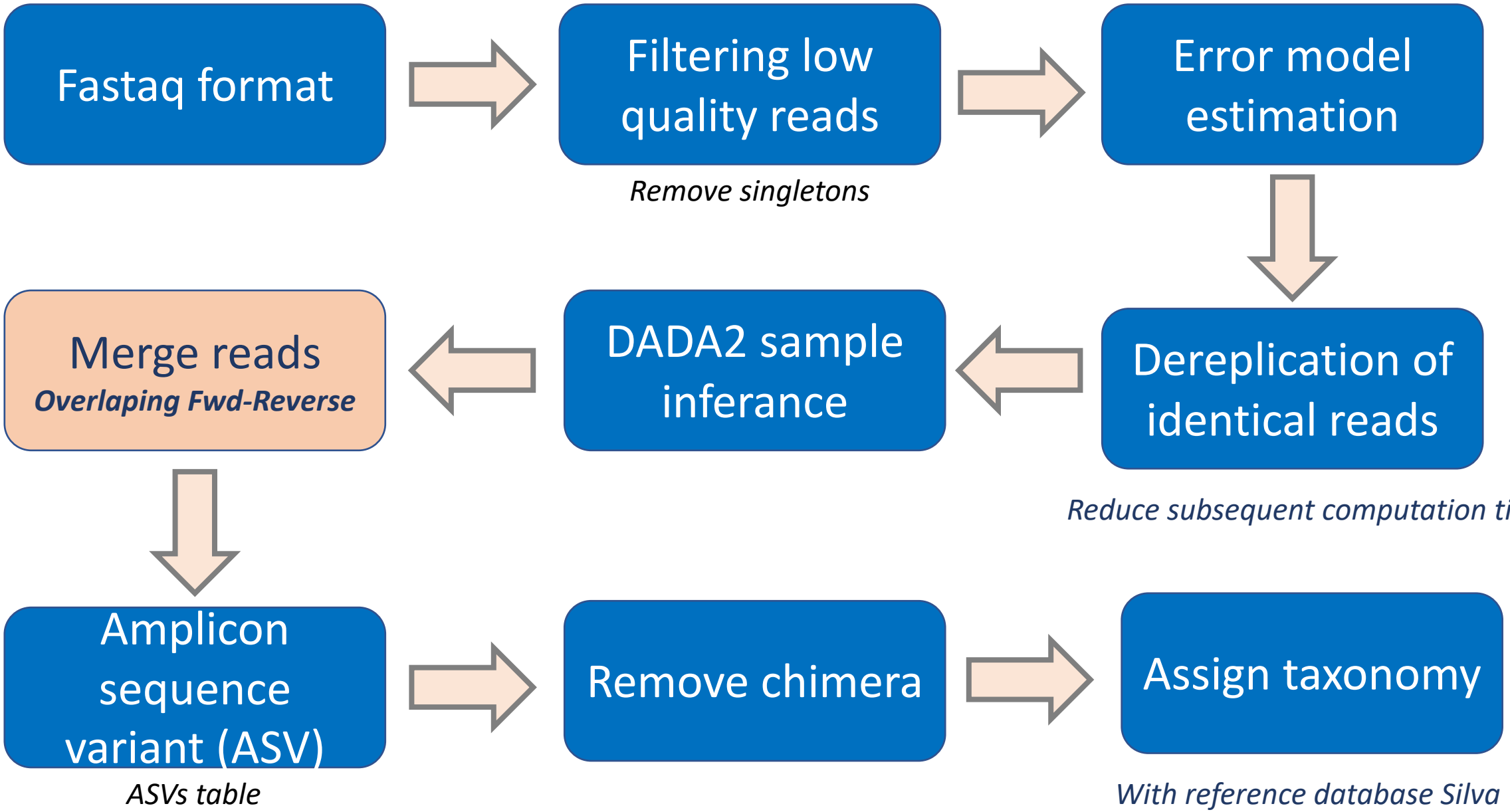
DADA2 Workflow



DADA2 Workflow

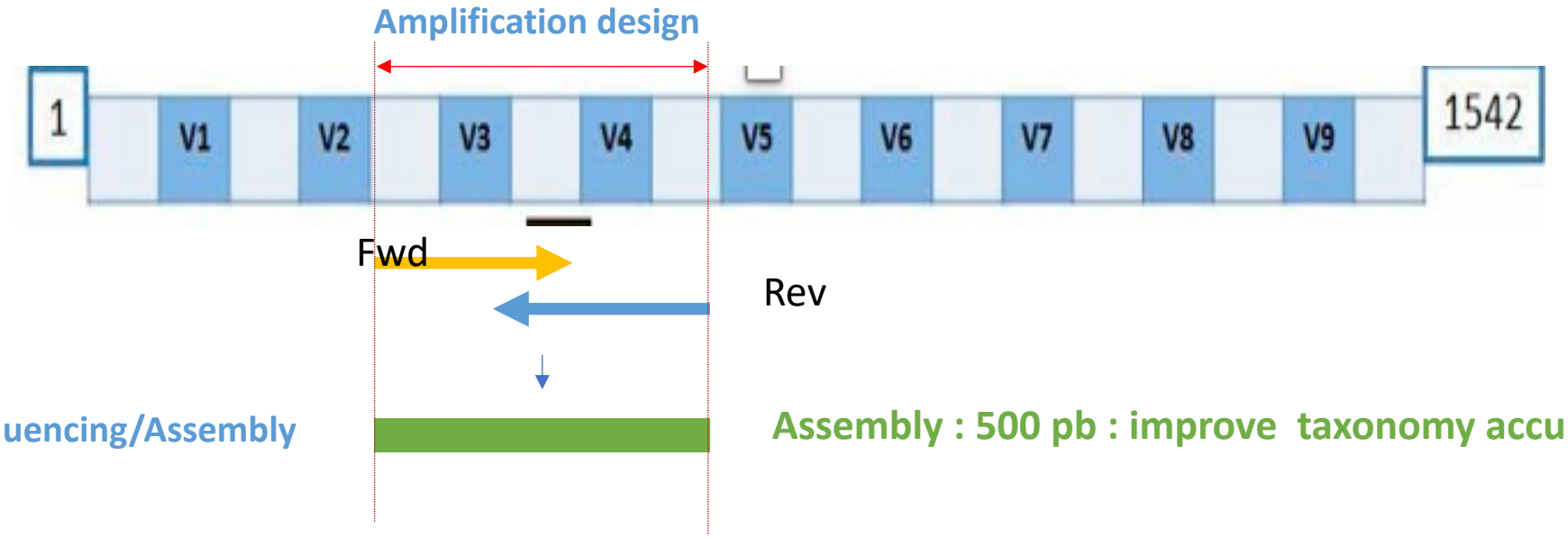


DADA2 Workflow

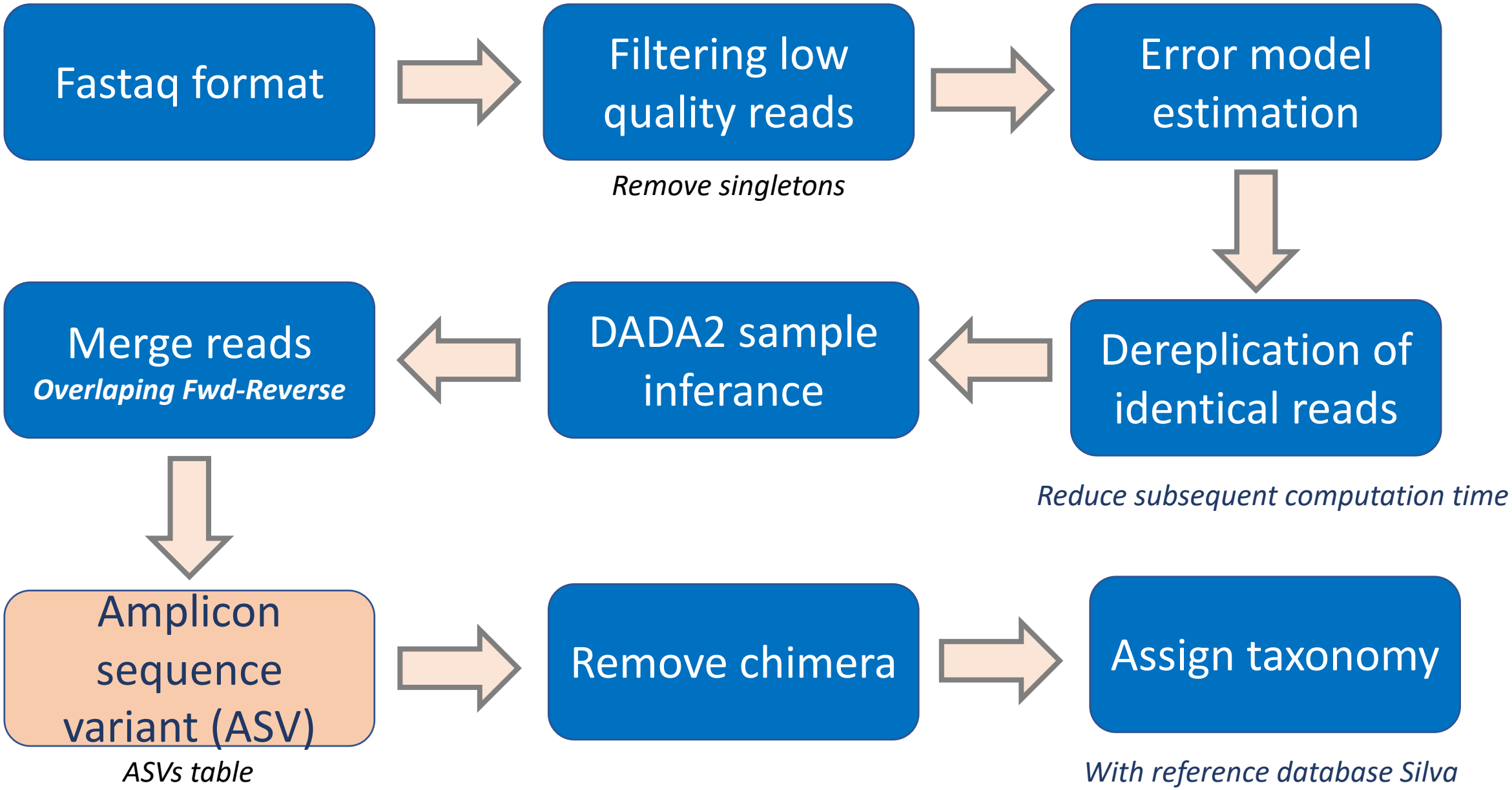


Merge : Assembly of Fwd & Reverse

Overlapping paired-end reads : Assembly is possible = increase amplicon size



DADA2 Workflow



DADA2 Workflow

Fastaq format

Filtering low quality reads

Remove singletons

Error model estimation

Dereplication of identical reads

Reduce subsequent computation time

DADA2 sample inferance

Merge reads

Overlaping Fwd-Reverse

Amplicon sequence variant (ASV)

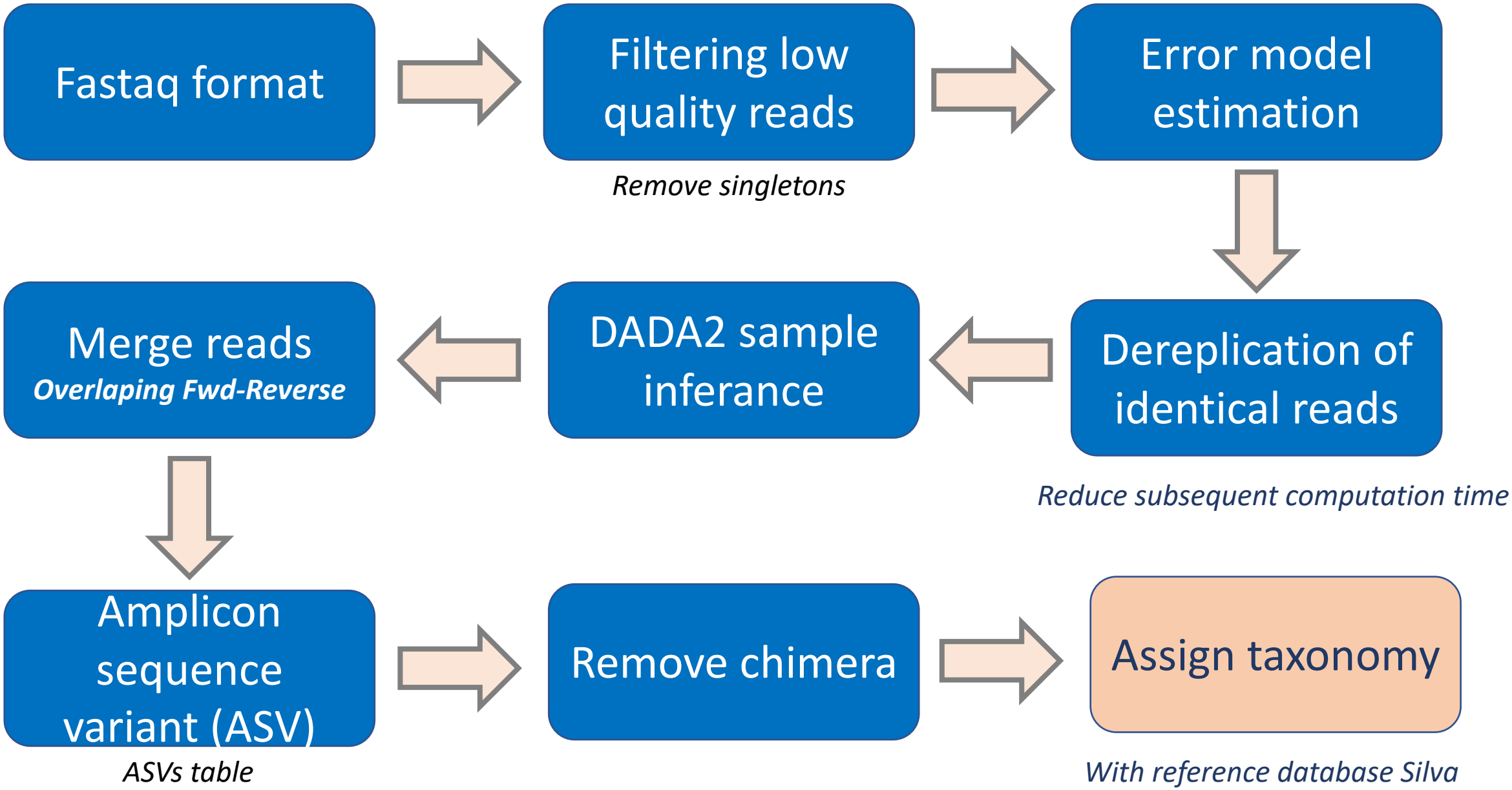
ASVs table

Remove chimera

Assign taxonomy

With reference database Silva

DADA2 Workflow



Assign Taxonomy for Amplicon Sequence Variant

- Formatted training fasta files for **RDP**, **Greengenes** and **Silva** reference databases are maintained

- Silva is probably the more complete database

Databases for Dada2

Maintained:

- Silva version 132, Silva version 128, Silva version 123 (Silva dual-license)
- RDP trainset 16, RDP trainset 14
- GreenGenes version 13.8
- UNITE (use the General Fasta releases)

Contributed:

- RefSeq + RDP (NCBI RefSeq 16S rRNA database supplemented by RDP)
  - Reference files formatted for `assignTaxonomy`
  - Reference files formatted for `assignSpecies`
- GTDB: Genome Taxonomy Database (More info: http://gtdb.ecogenomic.org/)
  - Reference files formatted for `assignTaxonomy`
  - Reference files formatted for `assignSpecies`
- HitDB version 1 (Human InTestinal 16S rRNA)
- RDP fungi LSU trainset 11
- Silva Eukaryotic 18S, v132 & v128
- PR2 version 4.7.2+. SEE NOTE BELOW.

What the literature says ….

Denoising tools :
More efficiente!

Table 1 | The accuracy of DADA2, UPARSE, MED, mothur, and QIIME on three mock community data sets

			Output reads (%)	Output sequences					Reference strains
				Total	Reference	Exact	One Off	Other	
Balanced	Forward	DADA2	99.2	93	59	33	1	0	57
		UPARSE	99.1	81	48	29	2	2	53
		MED	95.5	86	59	5	22	0	57
		Mothur	96.3	249	44	25	15	165	49
		QIIME	99.2	378	51	34	3	290	54
	Merged	DADA2	96.2	87	57	29	1	0	55
		UPARSE	94.2	76	45	27	2	2	50
		MED	91.1	64	56	6	2	0	54
		Mothur	94.1	108	42	27	11	28	47
		QIIME	94.1	170	45	28	4	93	50
HMP	Forward	DADA2	95.1	151	23	112	8	8	21
		UPARSE	96.7	161	20	123	10	8	21
		MED	80.9	83	23	2	58	0	21
		Mothur	95.4	849	20	177	47	605	21
		QIIME	97.4	1,375	20	177	60	1,118	21
	Merged	DADA2	92.3	67	23	40	2	2	21
		UPARSE	67.7	94	20	59	2	13	21
		MED	64.8	32	23	3	6	0	21
		Mothur	62.1	121	20	82	9	10	21
		QIIME	67.6	290	20	71	8	191	21
Extreme	Forward	DADA2	99.5	68	26	35	3	4	23
		UPARSE	99.5	74	21	40	0	13	21
		MED	86.4	95	16	0	79	0	13
		Mothur	–	–	–	–	–	–	–
		QIIME	99.5	3,237	20	44	73	3,100	20
	Merged	DADA2	97.6	25	24	1	0	0	21
		UPARSE	69.9	23	18	4	0	1	18
		MED	67.6	32	17	0	15	0	14
		Mothur	94.3	44	23	14	0	7	23
		QIIME	69.9	36	19	8	1	8	19

False positive species :
overestimation !

Callahan *et al.* 2016

Conclusion: ASV (dada) vs. OTU (Qiime)

- Most of the differences between ASVs & OTU methods are shown with the alpha diversity analysis :

- Difference in species number & diversity
- Spurious OTU (species not expected) with Qiime (overestimation)

- **No impact of the methods for the beta diversity analysis**

- Dada2 : Good performance in the **detection of rare** without the cost of non expected (spurious)from sample **highly diversified**

For low diversity sample -> less good, increase spurious ASVs

Nearing *et al.* 2018

Now it’s your turn !