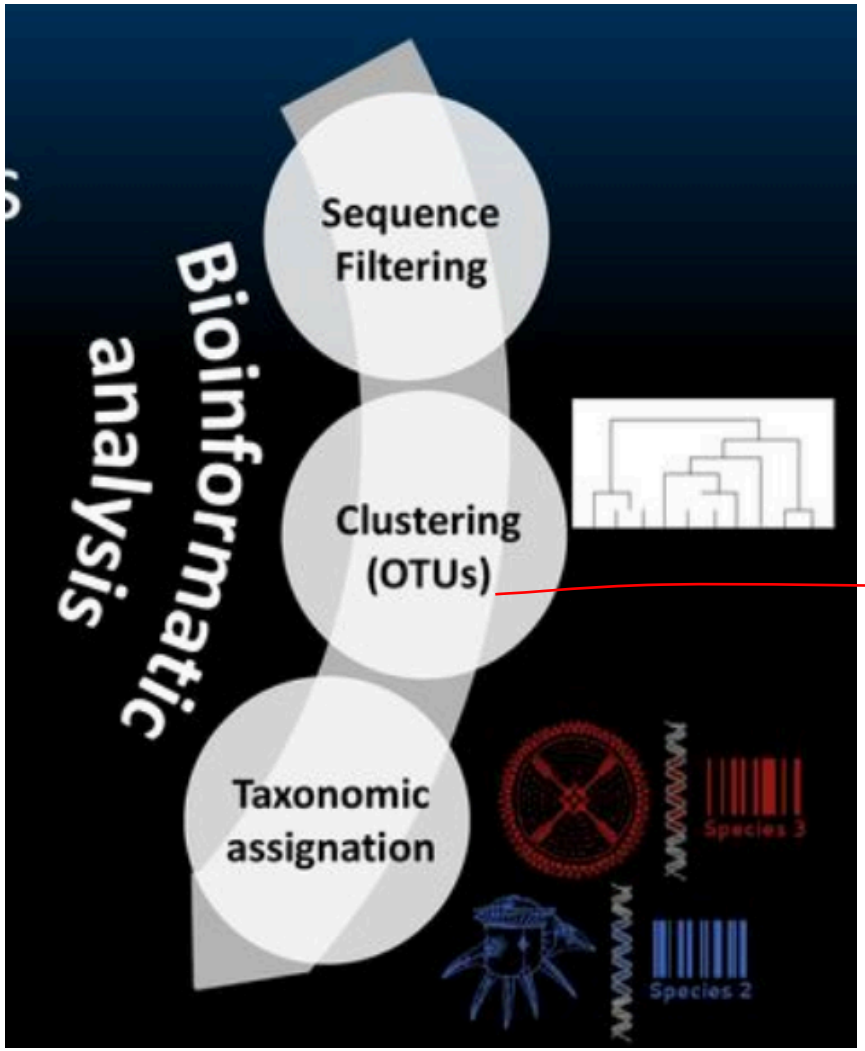

Bioinformatic analysis
Sequence Filtering
Clustering (OTUs)
Taxonomic assignation
Species 3
Species 2

vs ASV

Concept : OTU clustering

- Group X sequences in N Clusters using a sequence identity threshold between sequences (>97%)

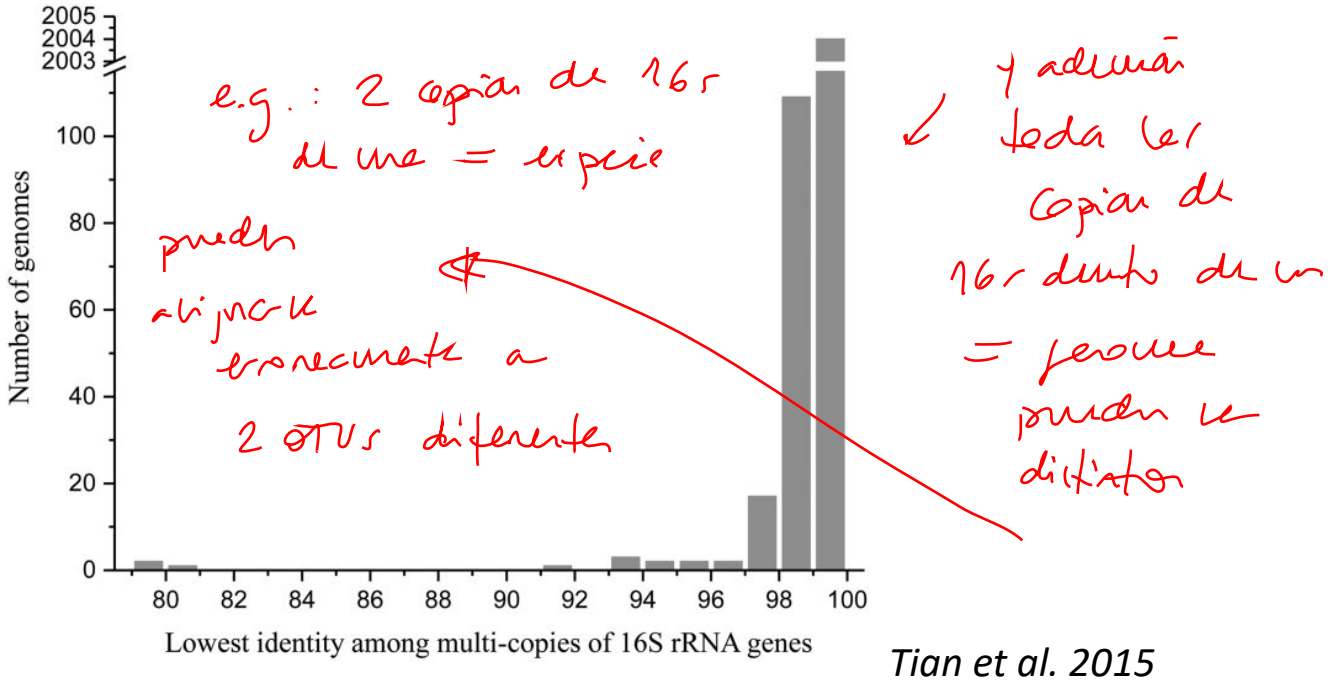
1 cluster = 1 Operational Taxonomic Unit (OTU)

↳ into regn de >97% identidad

→ Represent all the organisms sharing phenotypic similarities and so define a taxa/species (Sneath & Sokal, 1973)

97% OTU clustering absorbs

- Heterogeneity of Multicopy 16S of the species genome
- Variability of the 16S from different strains of species



Tian et al. 2015

OTU Clustering: The magic number 97

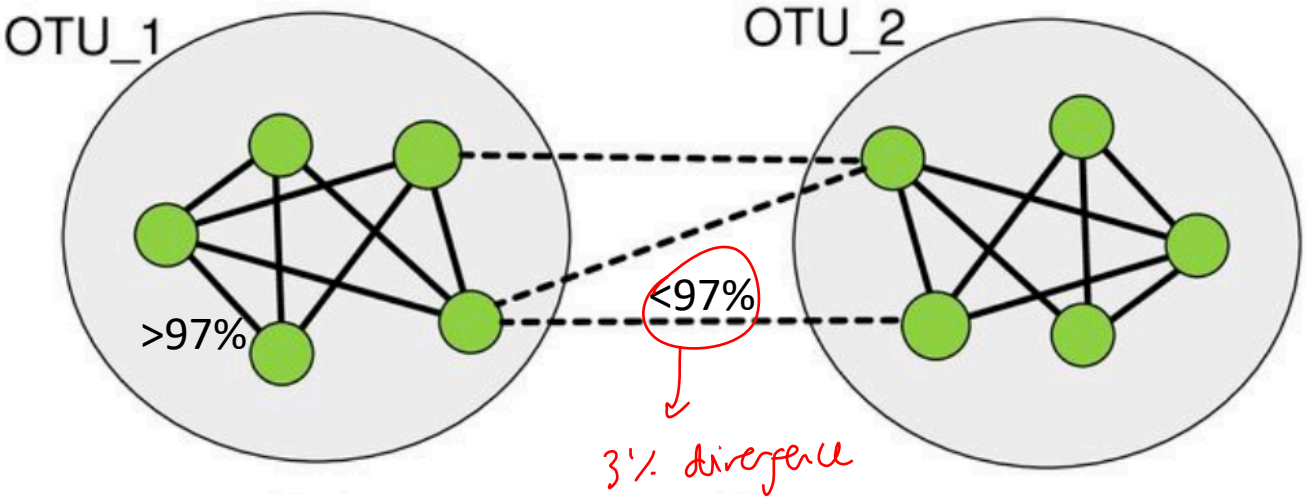
→ Correspondence between 97% sequence identity cut-off & the 70% DNA-DNA Hybridization (DDH) one (Stackebrandt & goebel 1994)

<97 % -> New species (16S), so >97% -> same species

<70% -> New species (DDH)

- Clustering method : Uclust, Usearch, Swarm → more evolué

- Ideal 97% clustering
  - All pairs in **same** cluster >97%
  - All pairs in **different** clusters <97%

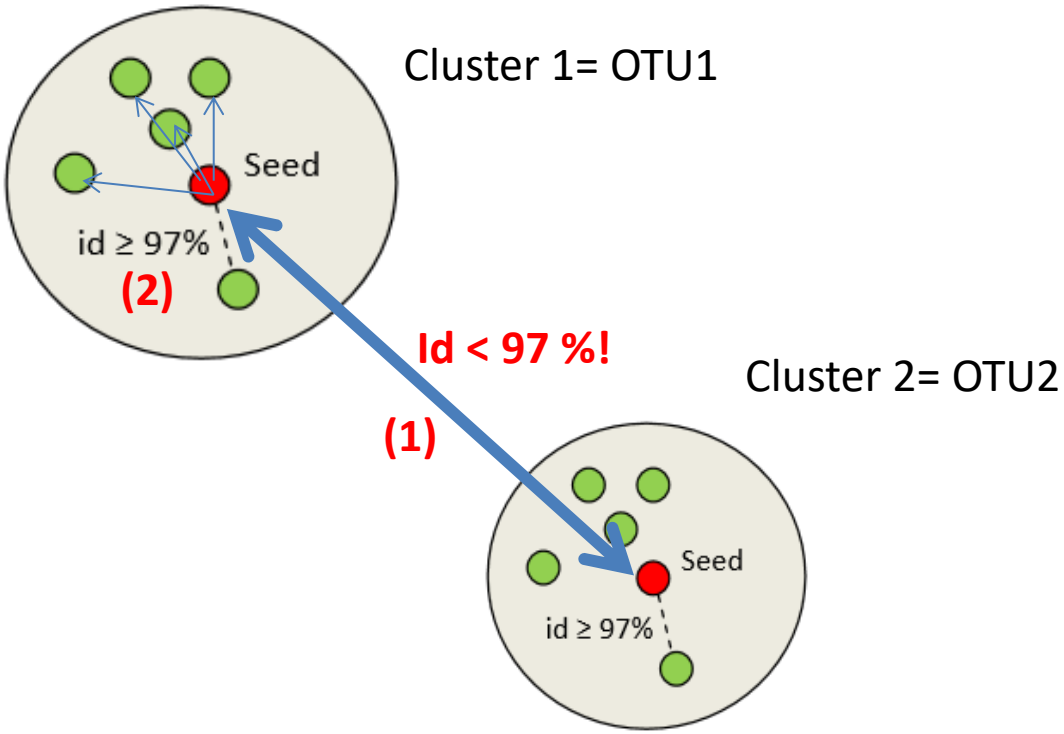


Centroid sequence= Pick representative OTU sequence from Cluster

Find set of clusters such as:

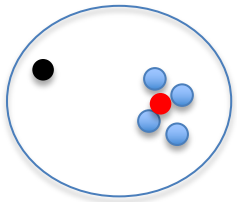
(1) All centroids have similarity < 97% to each other, and

(2) All member sequences have similarity >= 97% to a centroid

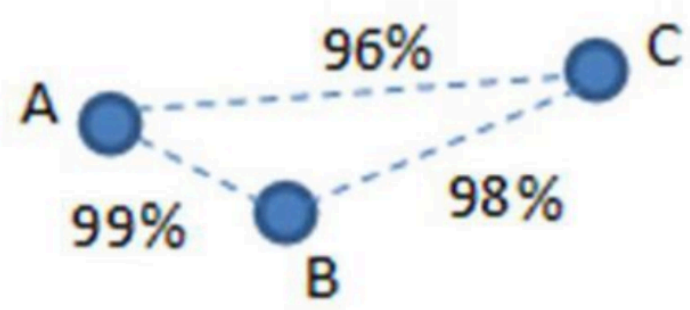


Centroid sequence is the sequence that minimizes the sum of distances to the others sequences in the cluster

→ Commonly is the most abundant one

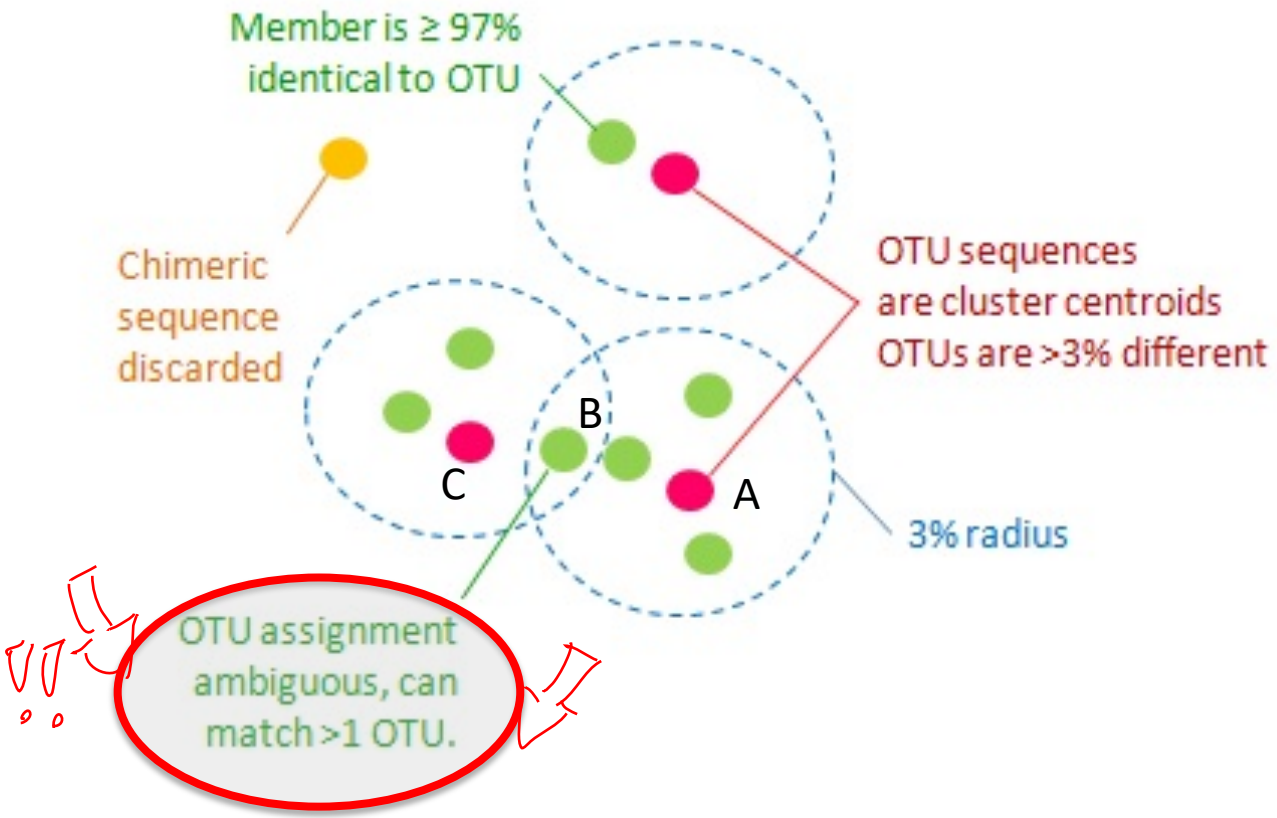


97% rule doesn't work



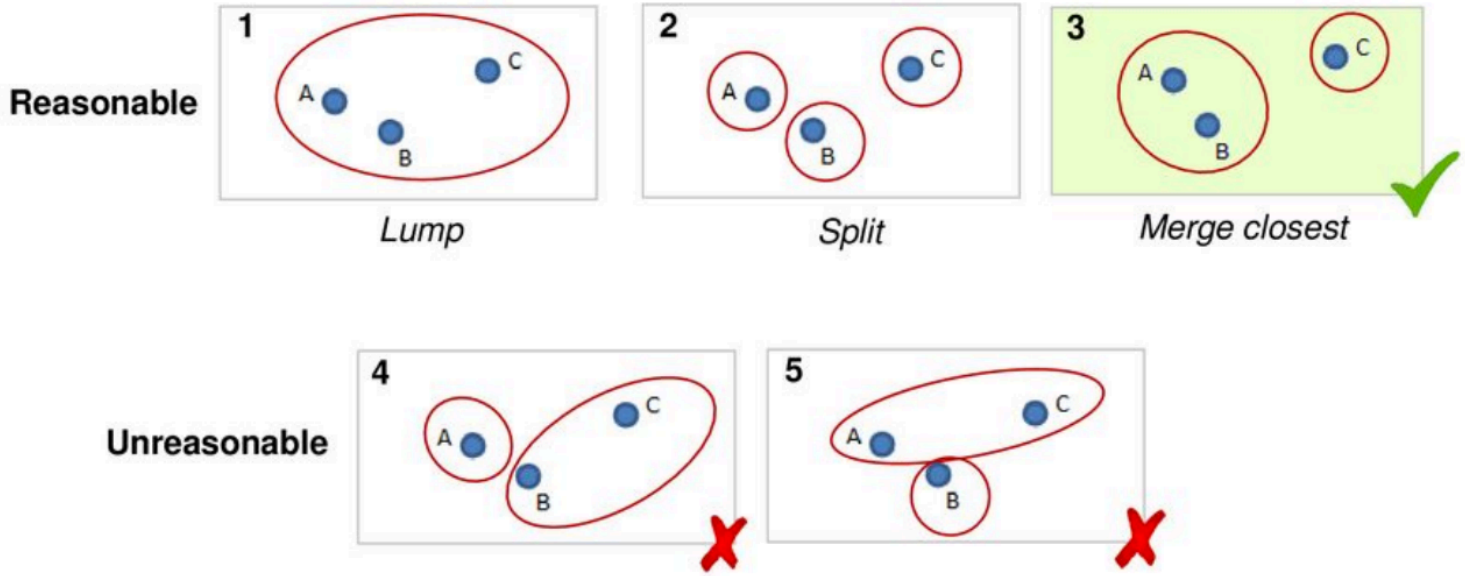
By the 97% rule…
A+B **same** OTU
B+C **same** OTU
A+C **different** OTU
… oops!

This means ….


Member is ≥ 97%
identical to OTU
Chimeric
sequence
discarded
OTU sequences
are cluster centroids
OTUs are >3% different
B
C
A
3% radius
OTU assignment
ambiguous, can
match >1 OTU.

Five possible sets of clusters

with conflicting OTUs



case of neither OTUs

→ Tend to overestimate the true diversity

→ Increase spurious OTUs … especially low abundant OTUs

97% OTU clustering reasonable

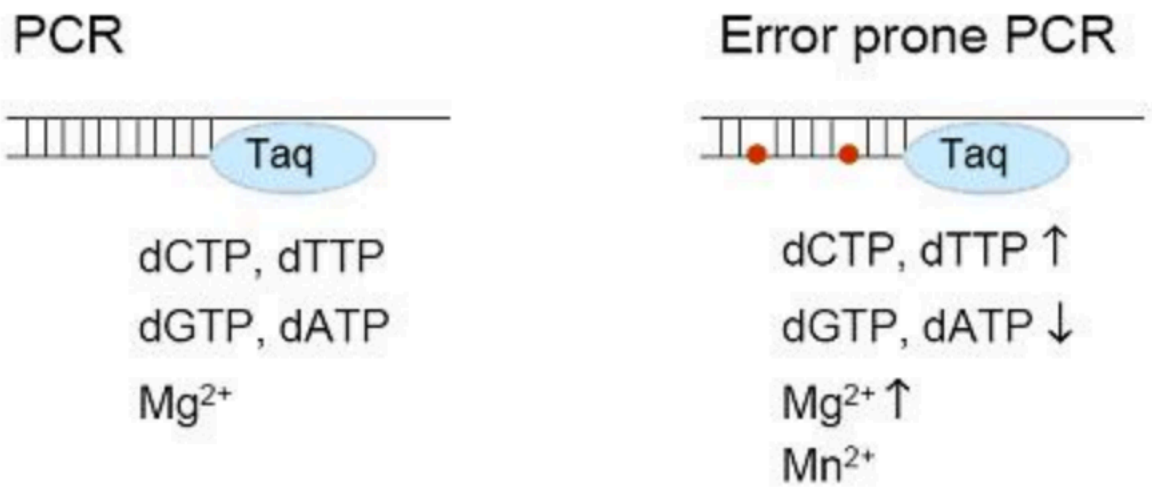
If sequencing errors are not important…

But, PCR + NGS cause Huge diversity of errors!!!

Error Sources in Metabarcoding

- **PCR errors**

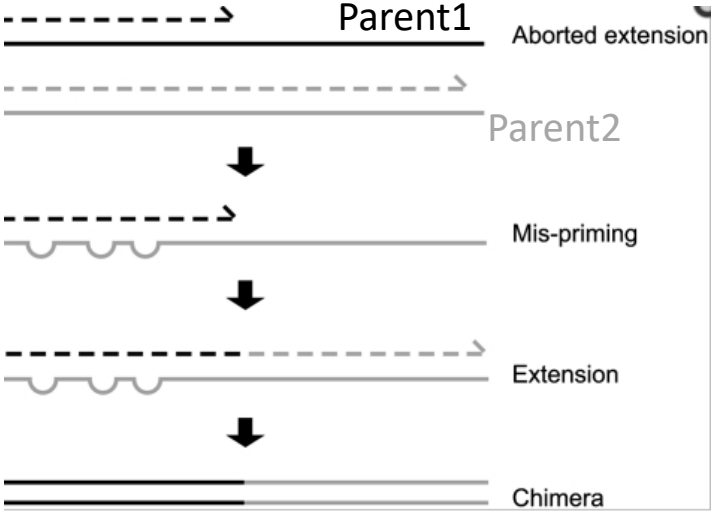
→ mostly substitution error caused by Polymerase

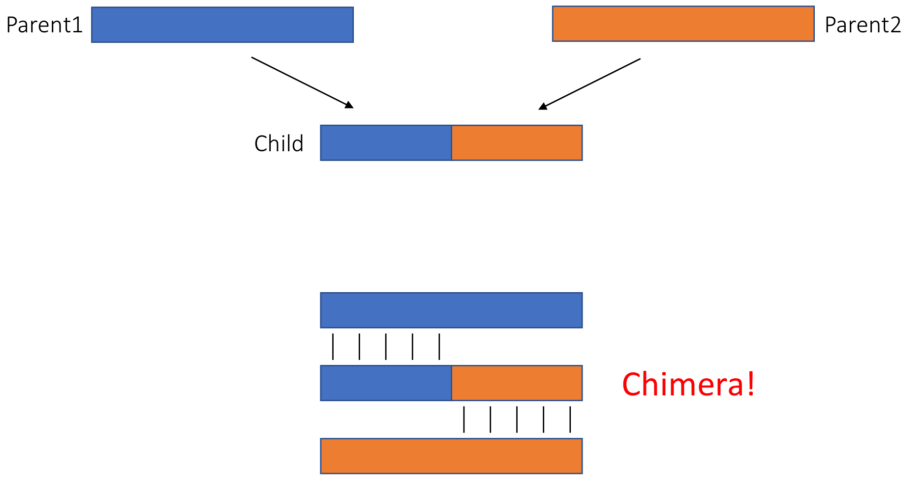


- **Chimera**

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

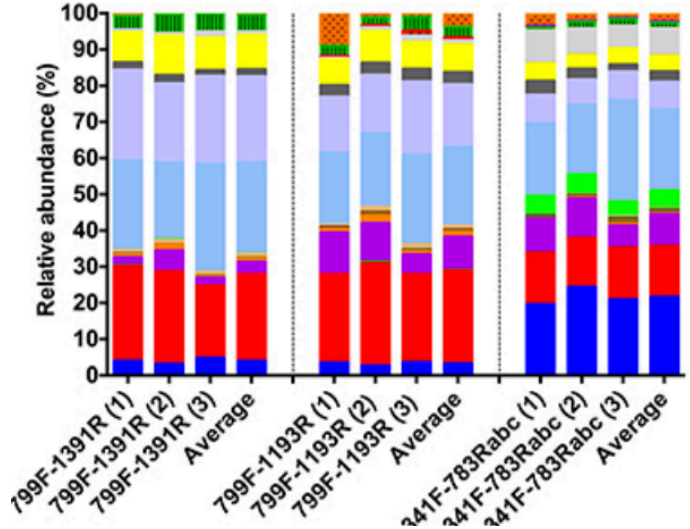
→ Merge (a least) 2 sequences that belong to 2 different species





- **Primer bias**

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



Beckers et al. 2016

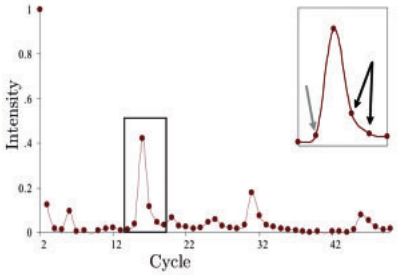
- **Depth bias**

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

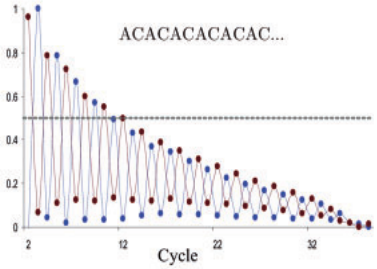
• Sequencing errors

Commonly modeled biases in base-callers for the Illumina platform

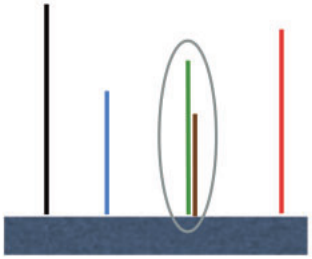
Phasing noise ϕ



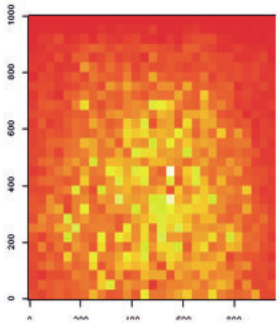
Signal Decay δ



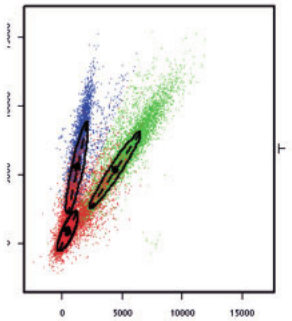
Mixed Cluster μ



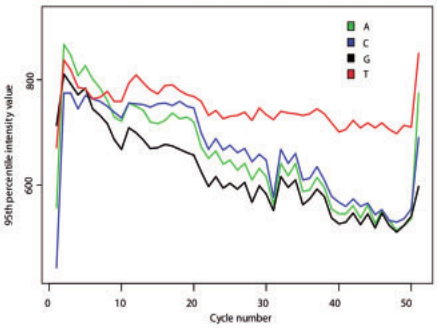
Boundary effects ω



Cross-talk Σ



T fluophore accumulation τ



Ledergerber et al. 2010

WHY KEEPING the 97% cut-off ???

- An old reference Stackebrandt 1994....

- Cut-off 98.5 % of sequence identity is used since many years in bacterial identification (before NGS)

to define new species

- New cut-off 98.70% was recently proposed by Chun et al. 2019, IJSEM

impact of Base Calling Error

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%

quality of data ↓

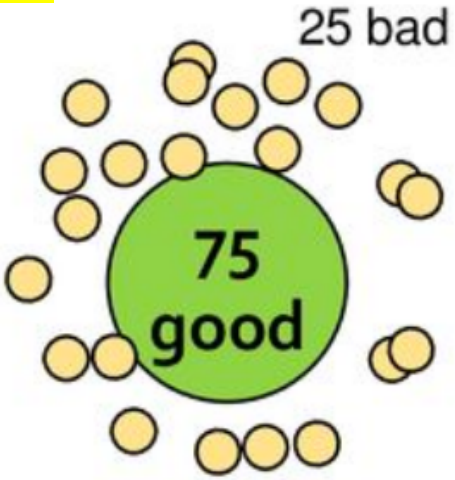
- Read length 250 nt
- Very High Quality Bases: all are Q30
- One letter in a thousand is wrong
- L=250, so, 1000 letters = four reads

→ One in four reads has a bad base

Think about …

- Make 100 reads of one Template : *T. maritima*
- At Q30, ¼ is bad means 75 correct reads, 25 bad reads

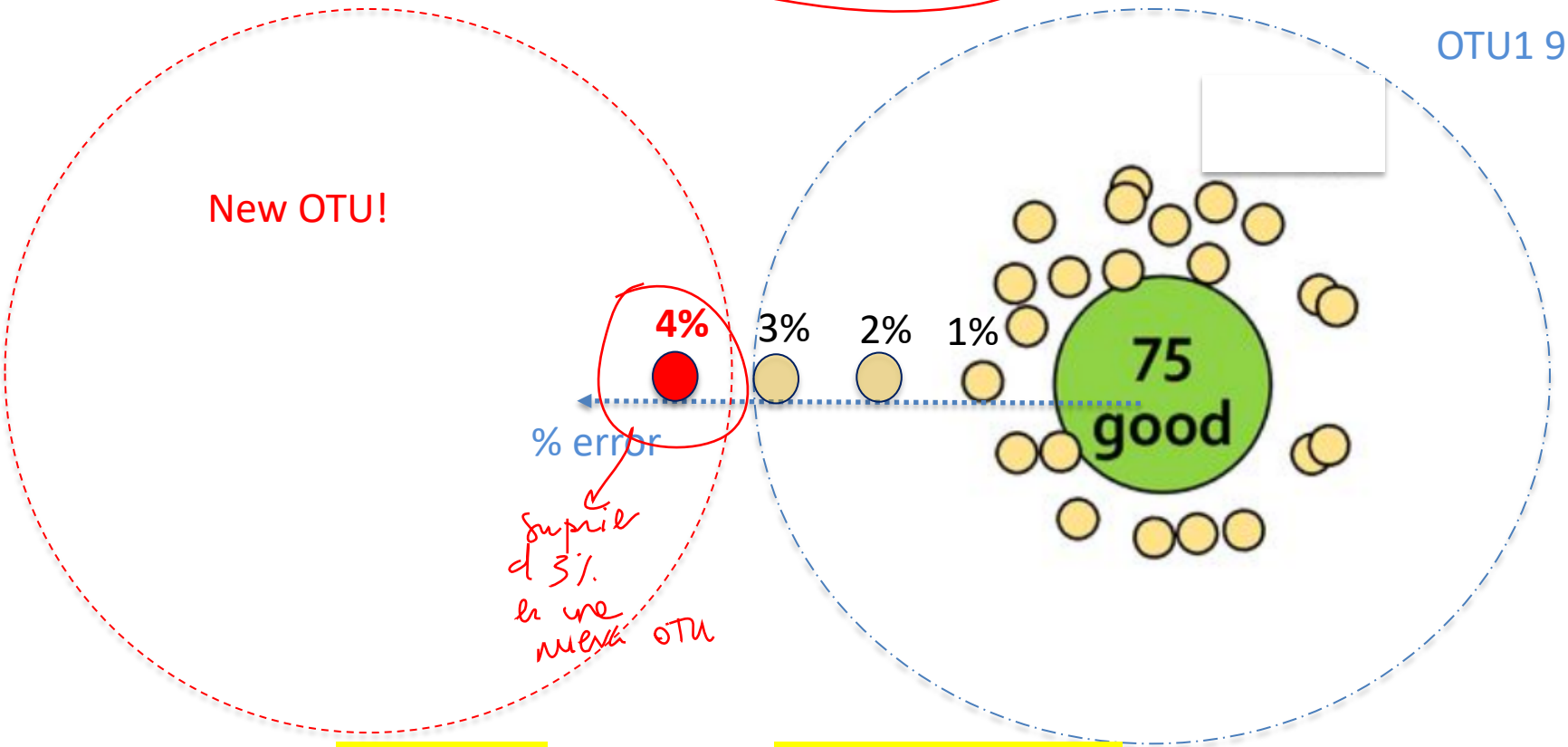
- Bad reads are differents
- So you have 26 uniques, **only one** good
  → 25/26 = 96% of bad sequences!!!



97% clustering & base calling errors…

- Generate reads for only one species : *T. maritima* (one 16S operon)
- Reads 100 bp length
- So 97% clustering can accept **3 base errors per read**

if your read is = 100bp



In theory : ONE OTU expected, 100% identical reads (*T. maritima*)
In practice: More than one OTU, different reads

97% OTU Clustering

Do not differentiate variants/strains

Threshold allow to absorb very low base calling errors

Remains sensitive to sequencing errors : generate spurious OTUs

rare stuff

OTU clustering-based methods (Usearch, Uclust, Swarm)

do not correct sequencing errors

16S Sequencing errors

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

OTU clustering cannot resolve this!!

Consequence?

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

Solution? → ASV

The Denoising : correct sequencing errors!

Denoiser Tools

→ they were used for 454 and they have been adapted to Nike much later...

- 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

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)

↳ it does not do CLUSTERS

you don't need to cluster sequences cuz you are sure who they are (error eliminated)

True Variation or Sequencing Error?

Sequence Read 1: acttcatgataccacatgatacg

Sequence Read 2: acttcatgctaccacatgatacg

Sequence Read 1: acttcatgataccacatgatacg

Sequence Read 2: acttcatgctaccacatgatacg

Some eg. transitions are more likely than others → DADA2 can use to find and down represent unlikely transitions

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

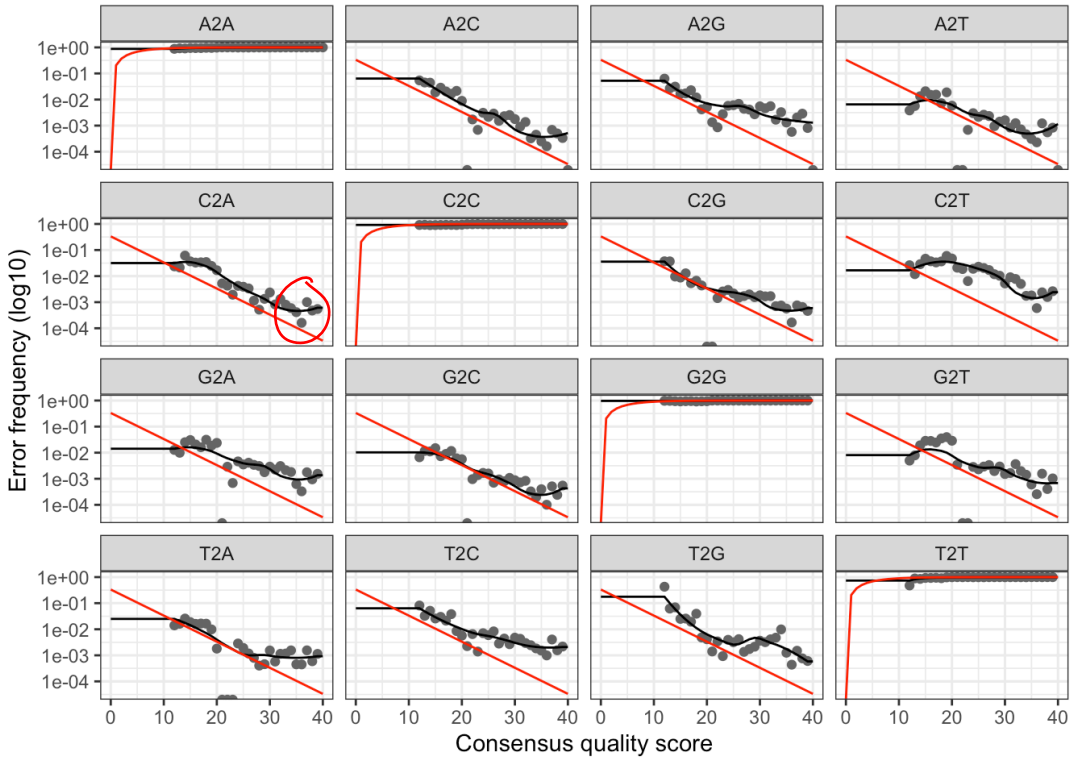
Sequence Read 1: acttcatgataccacatgatacg

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

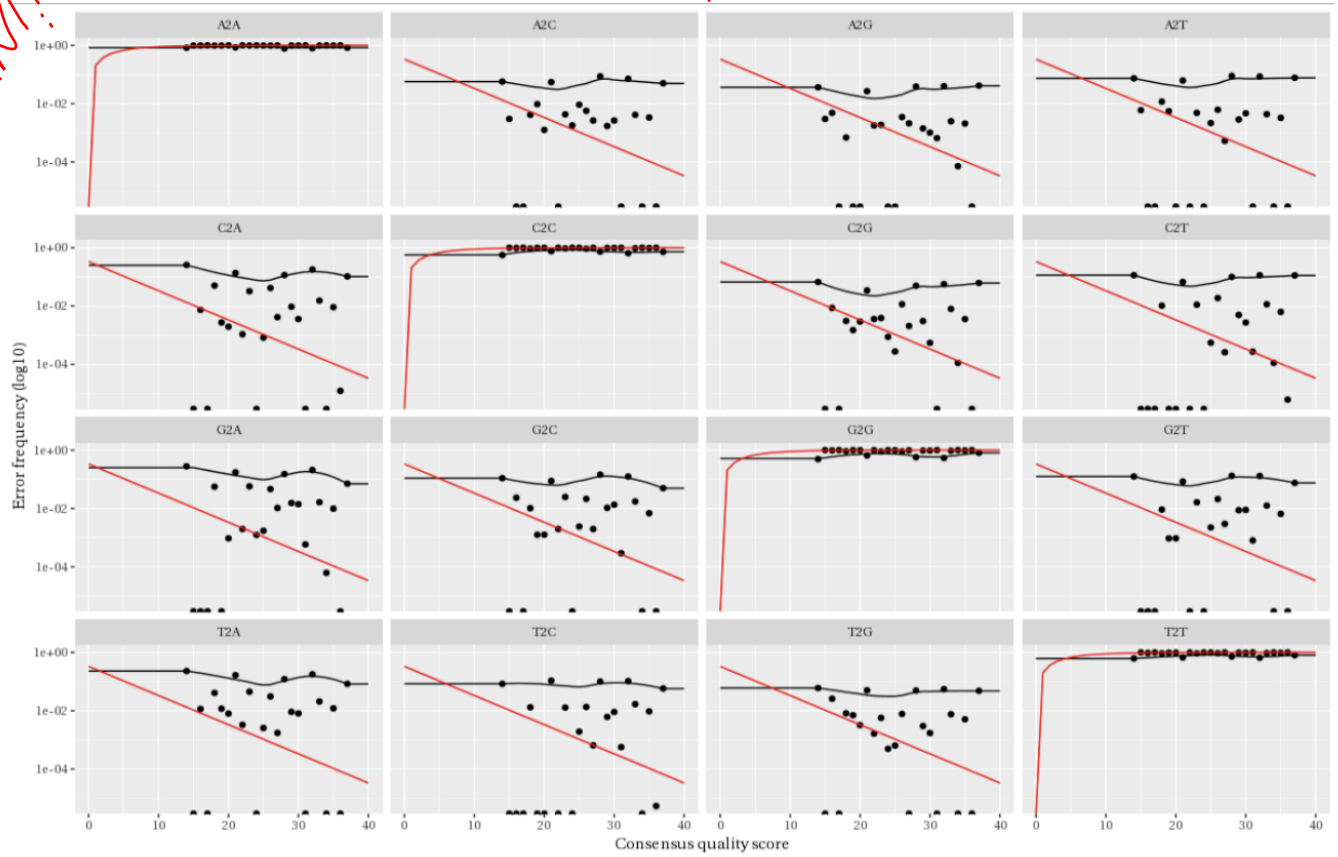
Error model Estimation



- **Error rates 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 increased quality as expected**

How to be confident with the model error estimation??

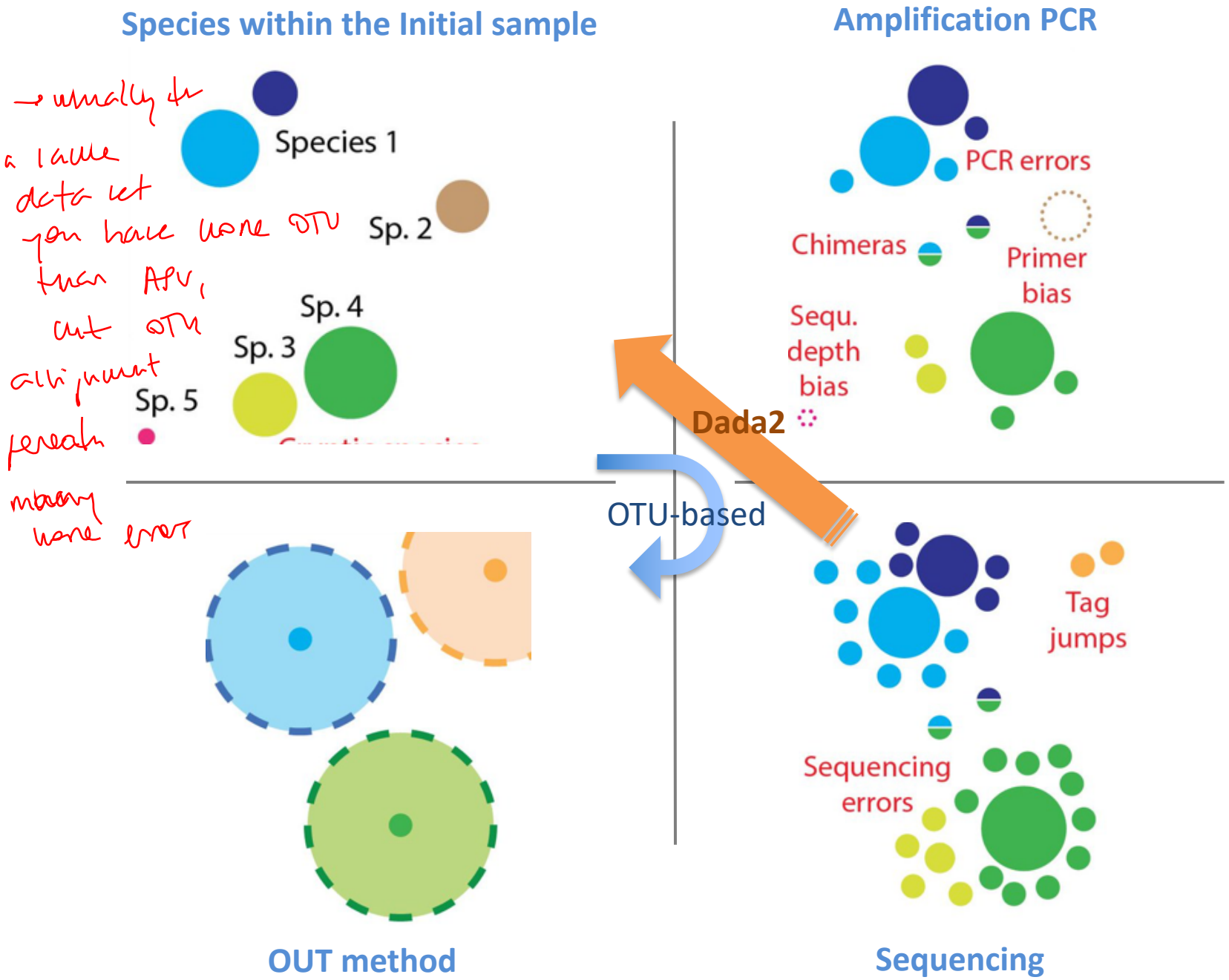
very important!! → you data need to be good quality, fit the model



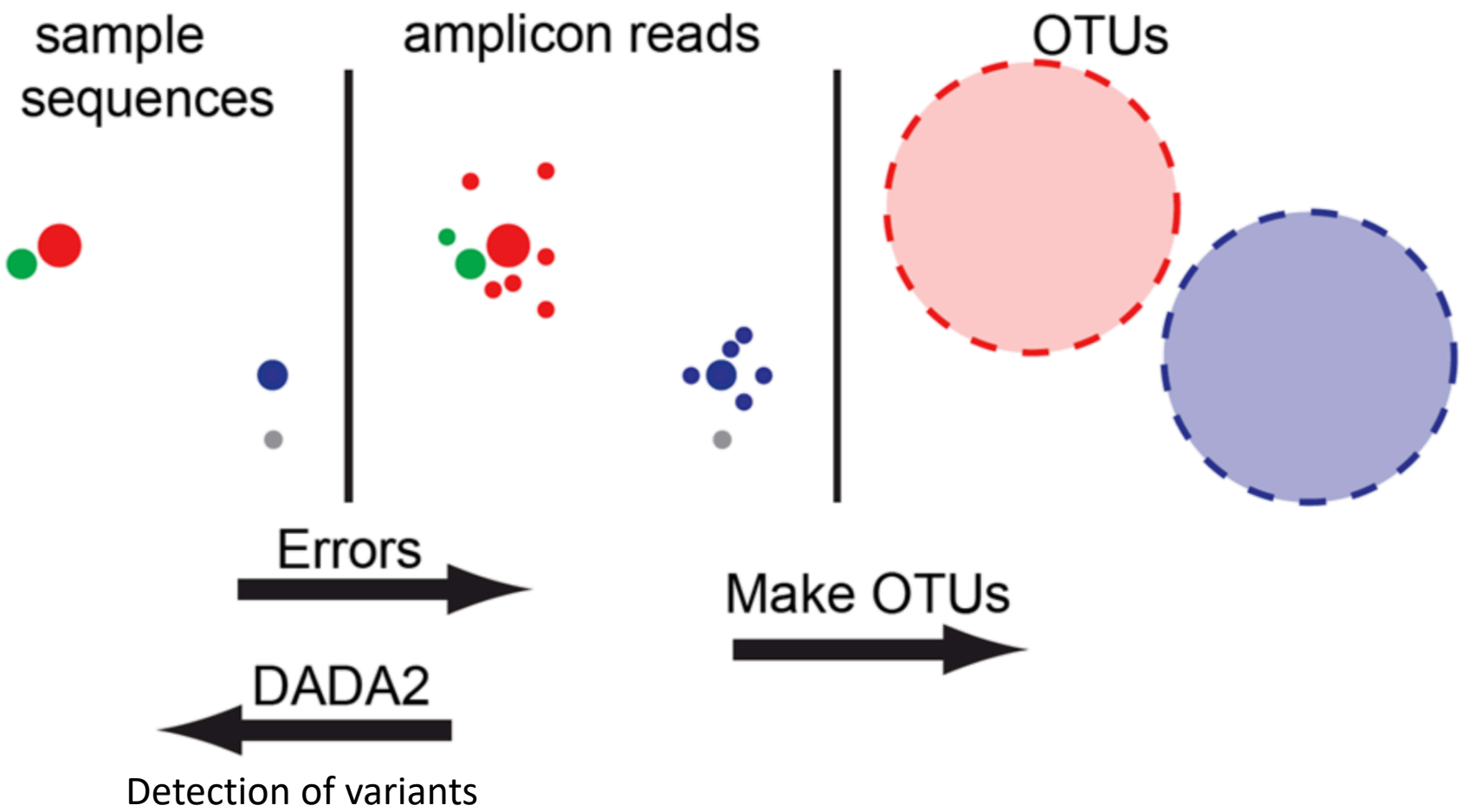
very bad data!!!

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

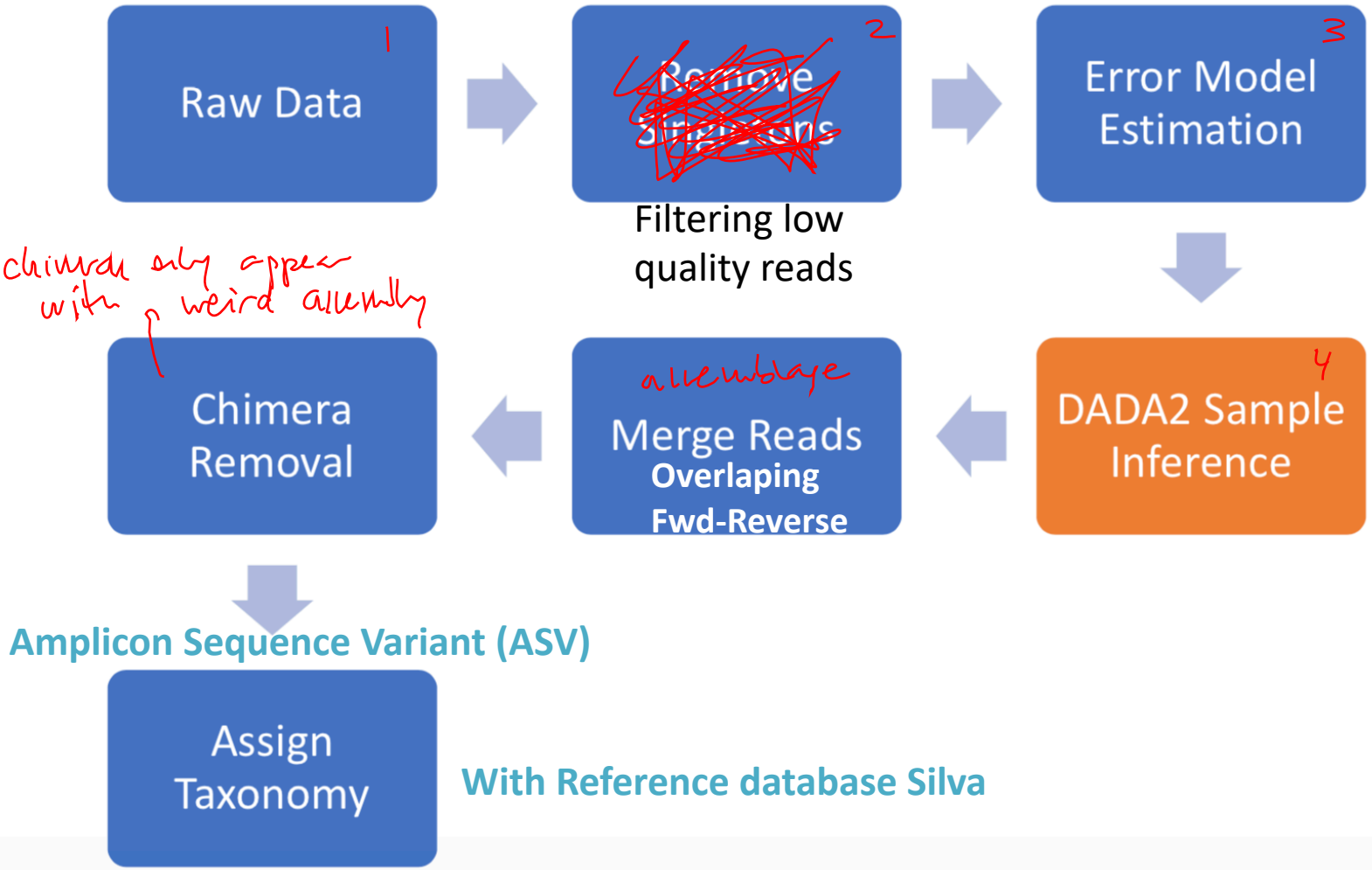
ASV is actually below species, it goes down to the strain level!!!



→ usually the a same data set you have more OTU than ASV, but OTU alignment perform mostly more error


sample sequences
amplicon reads
OTUs
Errors
Make OTUs
DADA2
Detection of variants

DADA2 Workflow



Raw Data: Fastaq Format

Line1 : Identifiant
Line2 : Sequence
Line3 : +
Line4 : Base Quality (ASCII)

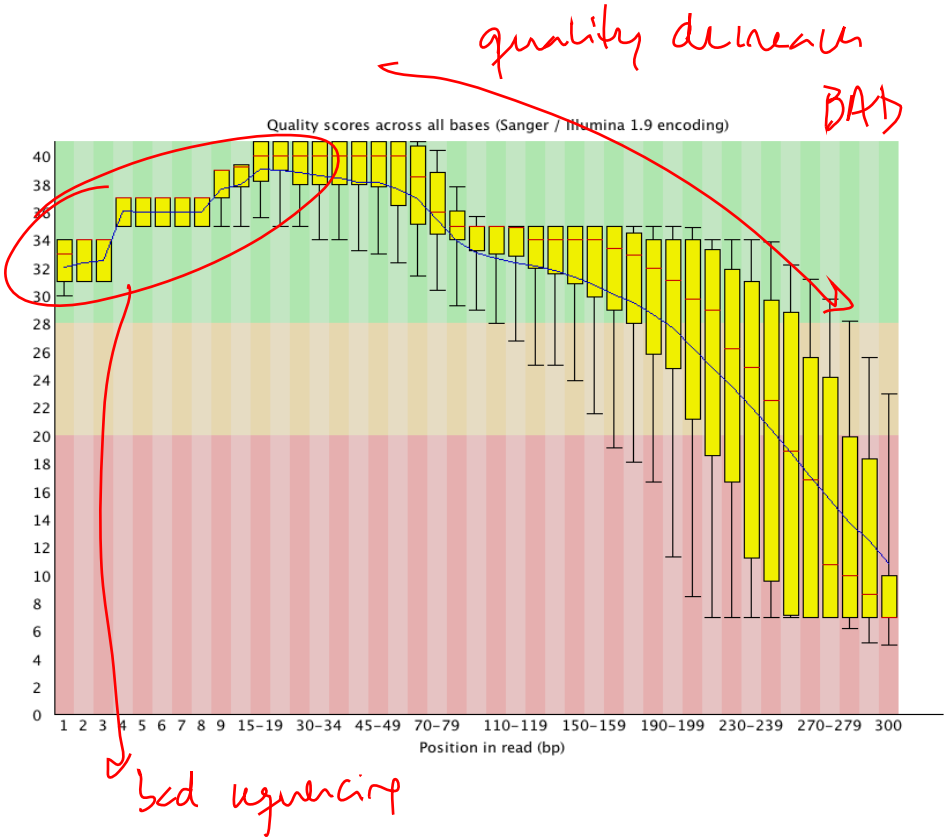
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

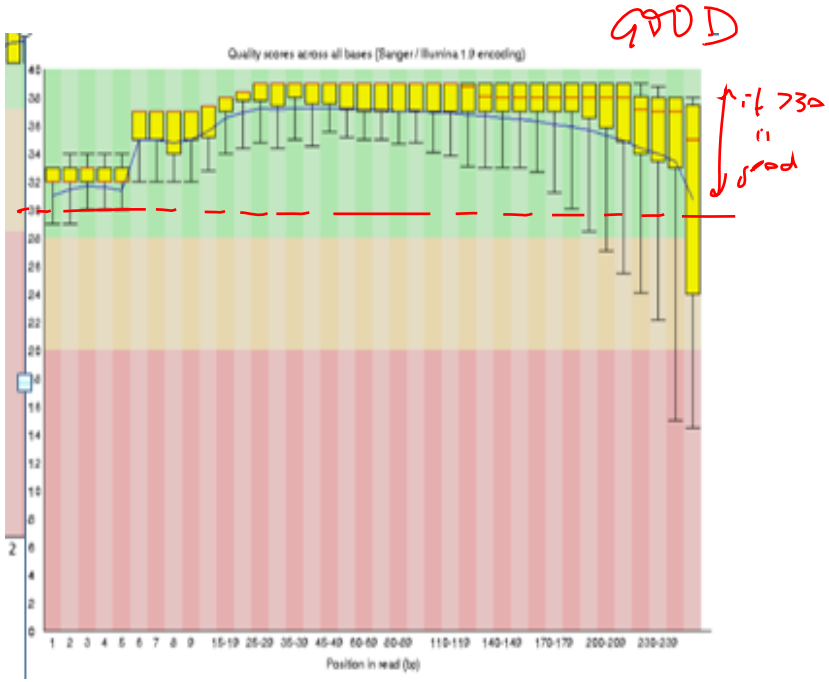
ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

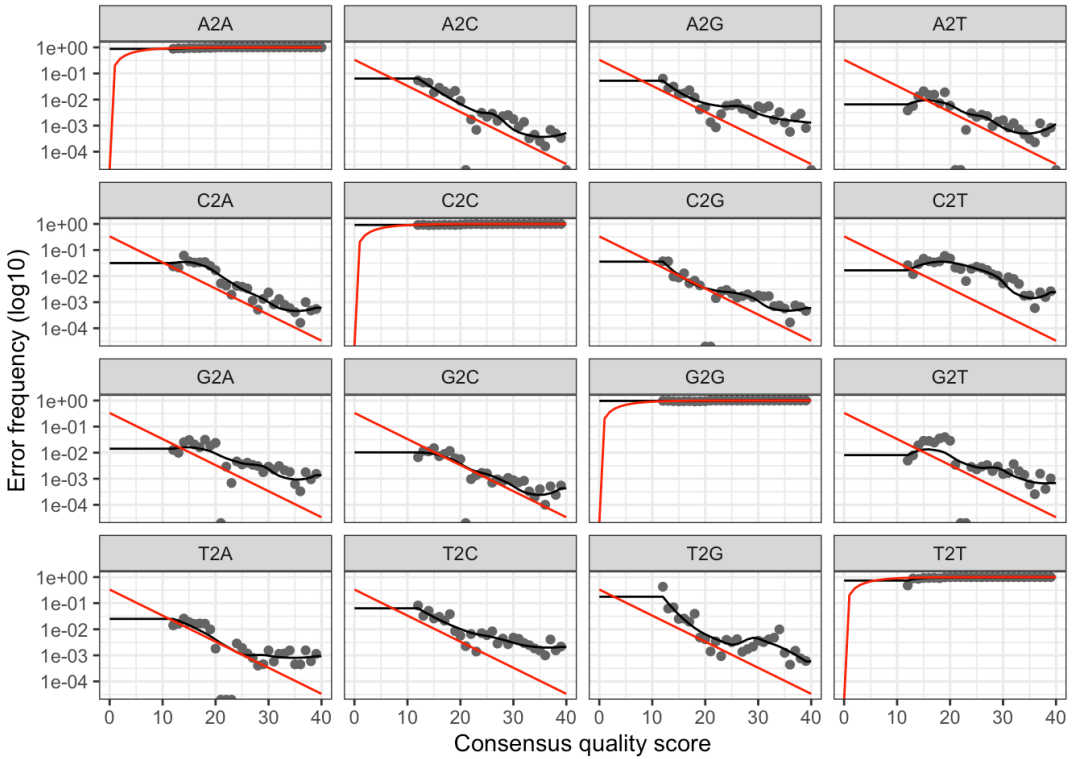
Remove Singleton….

Filtering bad reads....



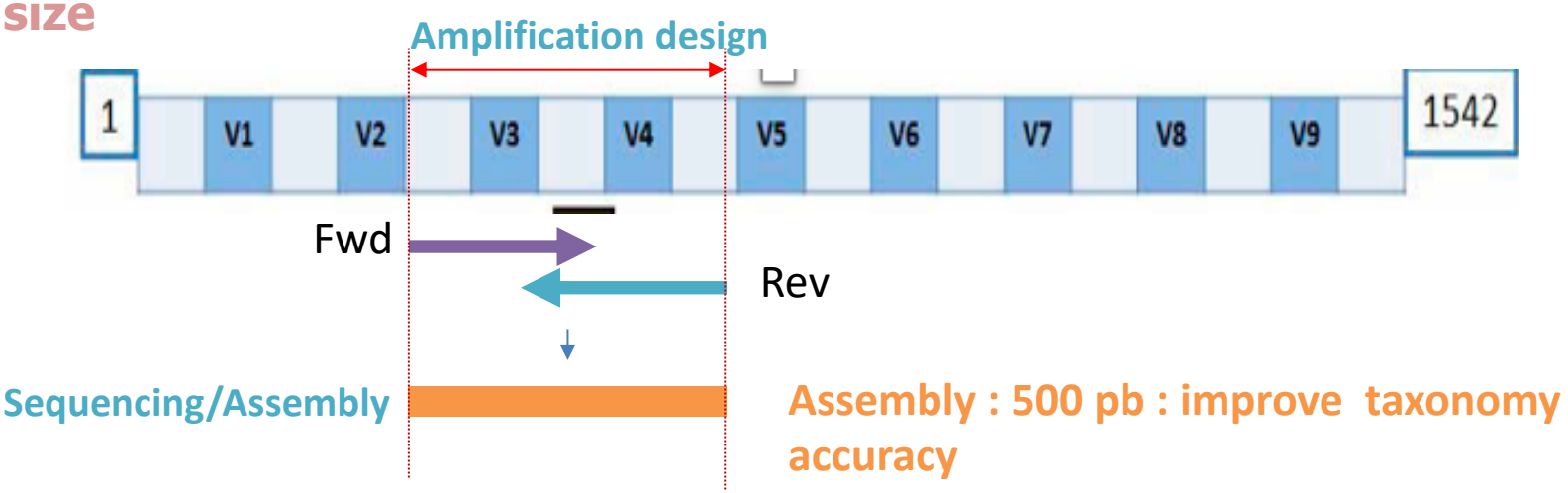


Error model Estimation



Merge : Assembly of Fwd & Reverse

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



- **Remove Chimera**

Assign Taxonomy for Amplicon Sequence Variant

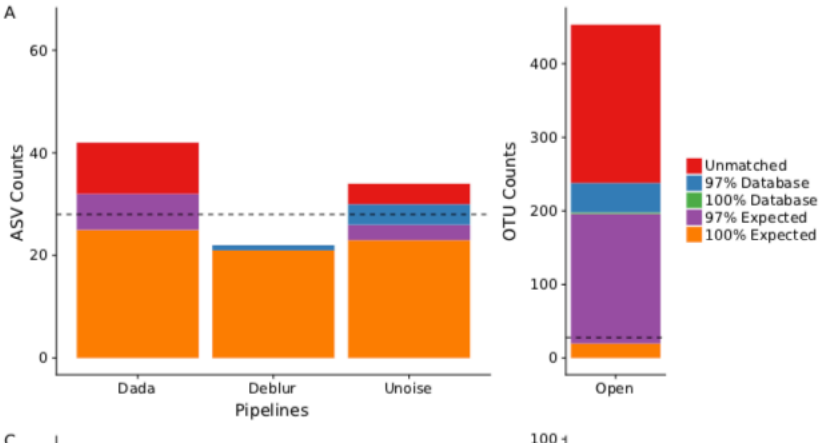
- Formatted training fasta files for **RDP**, **Greengenes** and **Silva** reference databases are maintained
- Silva is probably the more complete database

What the literature says ….

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			
HMP	Forward	DADA2	95.1	151	23	112			
		UPARSE	96.7	161	20	123			
		MED	80.9	83	23	2			
		Mothur	95.4	849	20	177			
		QIIME	97.4	1,375	20	177			
	Merged	DADA2	92.3	67	23	40			
		UPARSE	67.7	94	20	59			
		MED	64.8	32	23	3			
		Mothur	62.1	121	20	82			
		QIIME	67.6	290	20	71			
Extreme	Forward	DADA2	99.5	68	26	35			
		UPARSE	99.5	74	21	40			
		MED	86.4	95	16	0			
		Mothur	–	–	–	–			
		QIIME	99.5	3,237	20	44			
	Merged	DADA2	97.6	25	24	1			
		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



Nearing et al. 2018

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