DADA2

High-resolution sample inference from Illumina amplicon data

Presented by: Guy Bivas

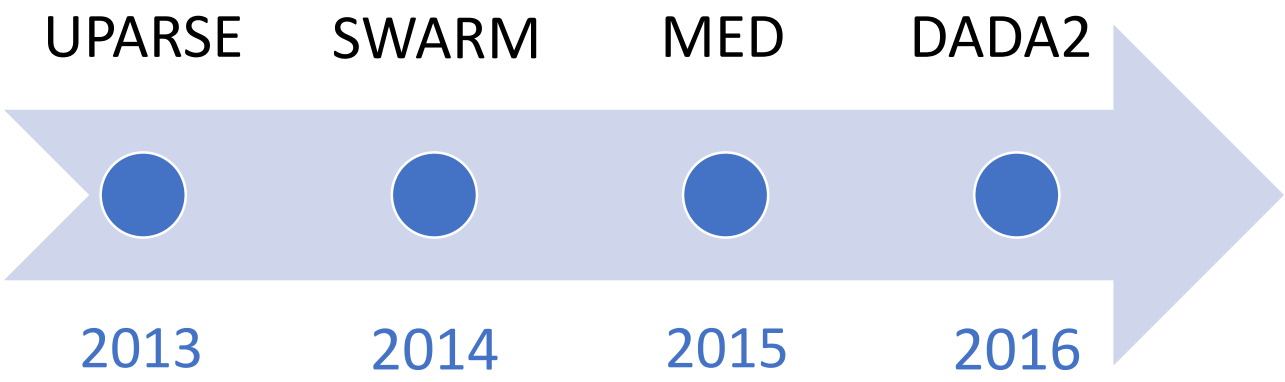
Course Instructor: Prof Elhanan Borenstein

TAU CS 0368-3116: Seminar in Computational

Methods in Metagenomics and Microbiome Research

ביה"ס למדעי המחשב ע"ש בלווטניק
הפקולטה למדעים מדויקים
ע"ש ריימונד ובברלי סאקלר
אוניברסיטת תל אביב

Timeline



Outline

Background

True Variation vs. Sequencing Errors

DADA2 Core Algorithm

DADA2 Full Workflow

Some Results

Summary

Outline

Background

True Variation vs. Sequencing Errors

DADA2 Core Algorithm

DADA2 Full Workflow

Some Results

Summary

Background


SUGAR

PCR

Polymerase Chain Reaction

PCR Machine

Target DNA

Primers

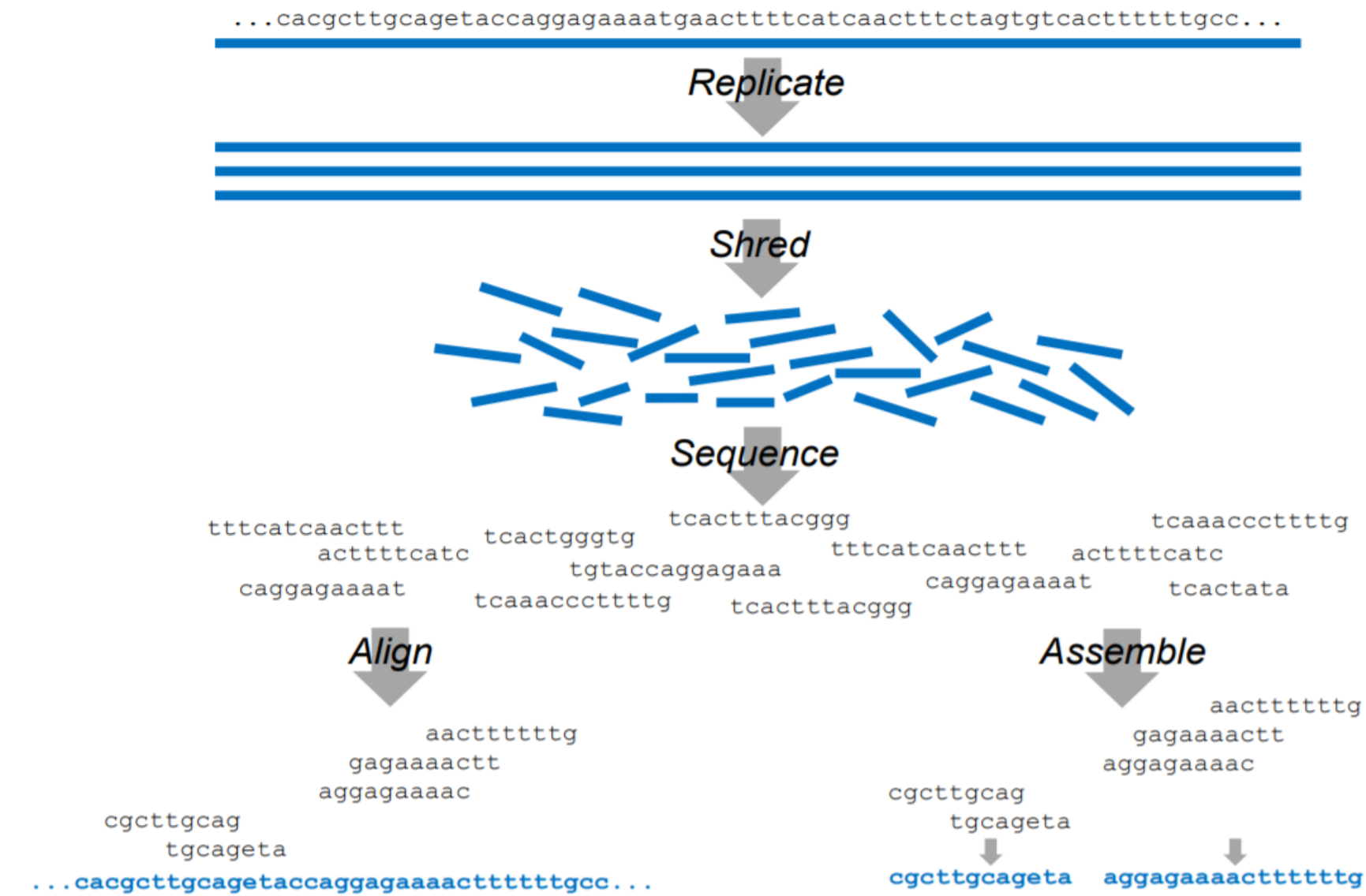
Nucleotides

DNA polymerase

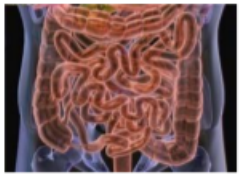
Buffer
(creates the optimum pH for the enzyme)

$MgCl_2$
(needed for the enzyme to work)

Sequencing, Alignment, Assembly (≈1 course)



Key Challenge 1



Microbiome sample

Extract DNA

Targeted sequencing (16S)

Shotgun sequencing

AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AGGAGGTGCA-A
AGGAGGTGCA-A
AAA--CTGCTCA
AGGAGCT--TCA

AGGAGCTGCTCA AGCGCGATCTGAA GAACCTGGTGA
ATTGAGAACCCGGCTG CCATTGCCGAGC TGAAGGCAATCA
TTCTACCCC TATGATGTGACCGT GCCCATTGCCGAG ACCACCAAG
TTCGCCCAGCTGAAG ACCGTCGATGTCC TGATCAAC

Cluster into species or OTUs

AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA

AGGAGGTGCA-A
AGGAGGTGCA-A
AGGAGGTGCA-A

AAA--CTGCTCA

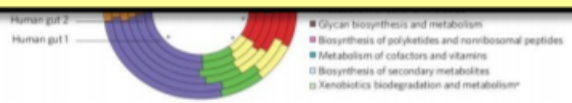
AGGAGCT--TCA

Species/Taxonomic composition

Who's there?

From 16S Sequences to Taxa Composition

- Clear clusters are not always feasible
- True variation vs. sequencing errors?
- Low resolution
- A fixed threshold doesn't always reflect the same phylogenetic closeness
  - Different species clustered together
  - Strains clustered separately



What are they doing?

Divisive Amplicon Denoising Algorithm

Our goal is not to find the best clusters

We want to determine if a sequence read came from True Variation or Sequencing Error

Outline

Background

True Variation vs. Sequencing Errors

DADA2 Core Algorithm

DADA2 Full Workflow

Some Results

Summary

True Variation vs. Sequencing Errors

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

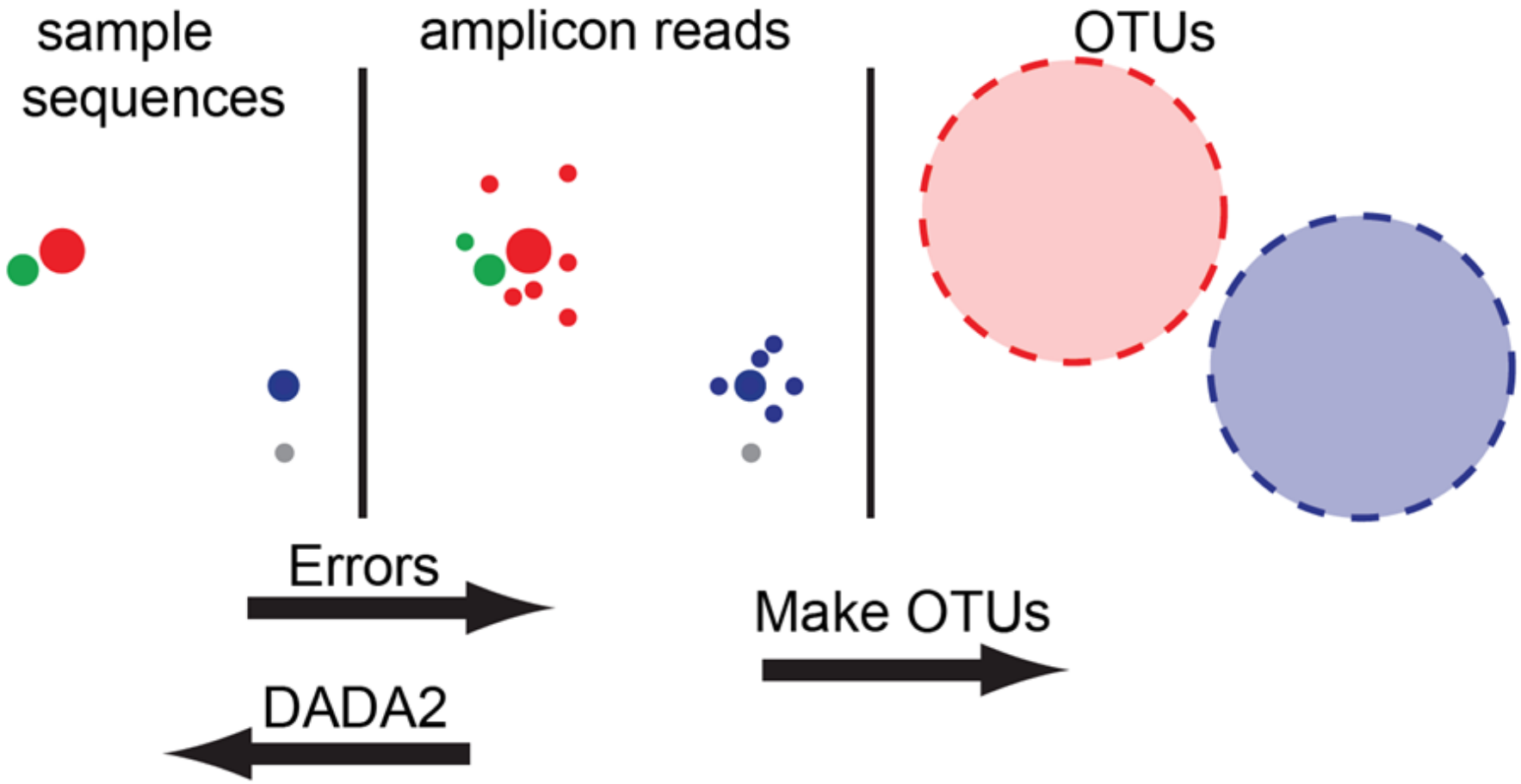
True Variation vs. Sequencing Errors

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

Visualization



Outline

Background

True Variation vs. Sequencing Errors

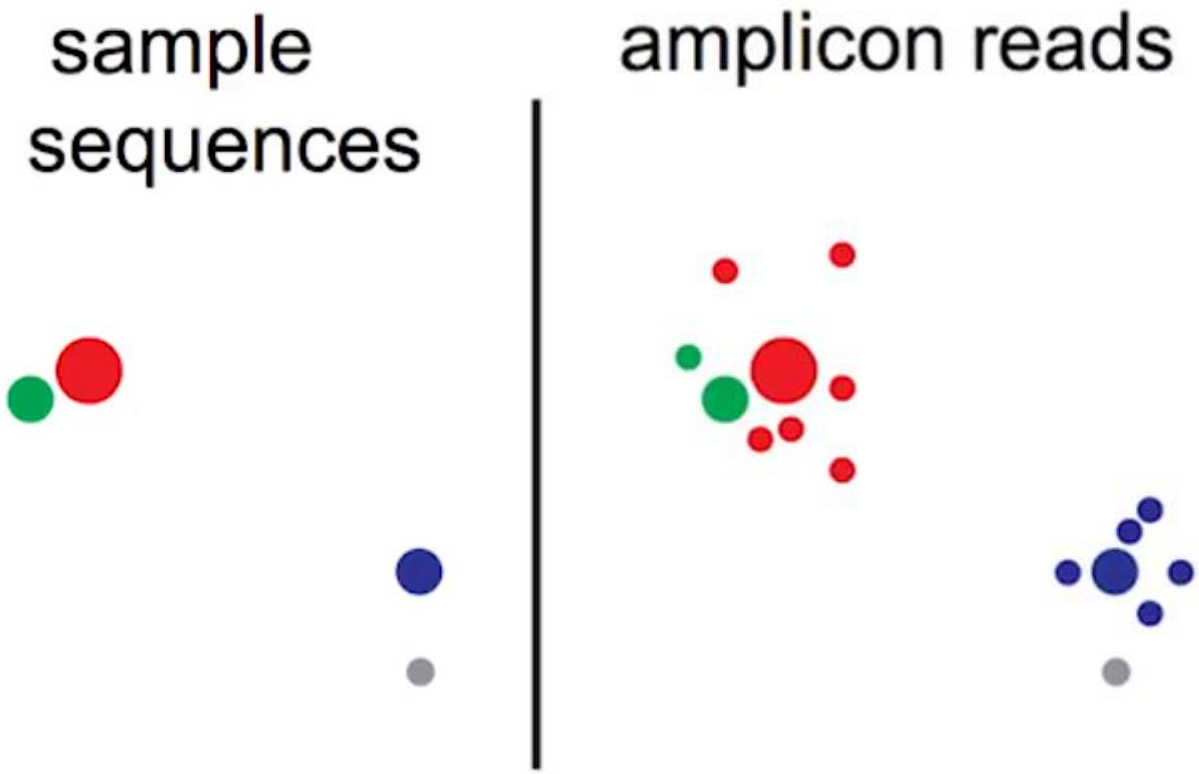
DADA2 Core Algorithm

DADA2 Full Workflow

Some Results

Summary

DADA2



Outline

Background

True Variation vs. Sequencing Errors

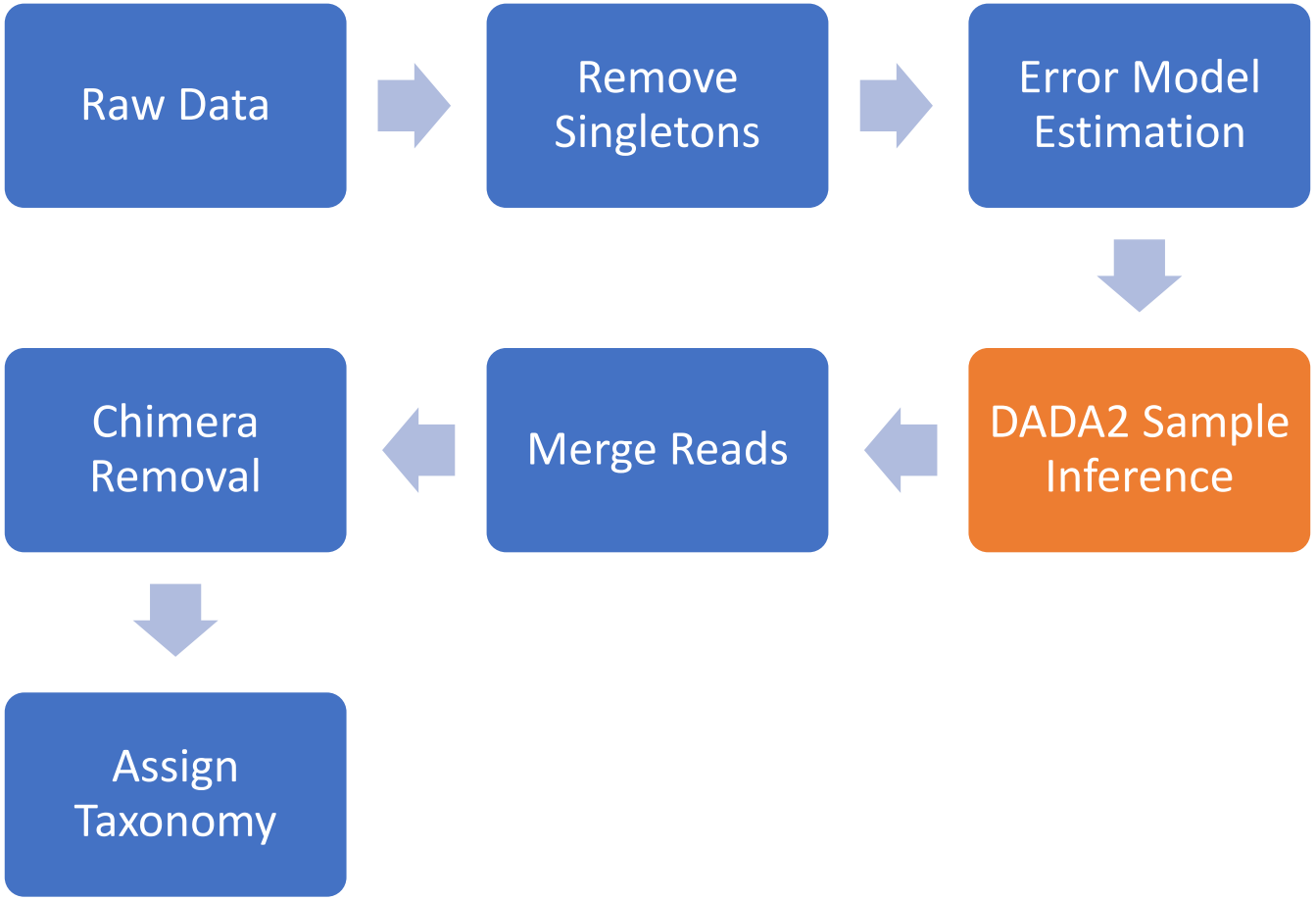
DADA2 Core Algorithm

DADA2 Full Workflow

Some Results

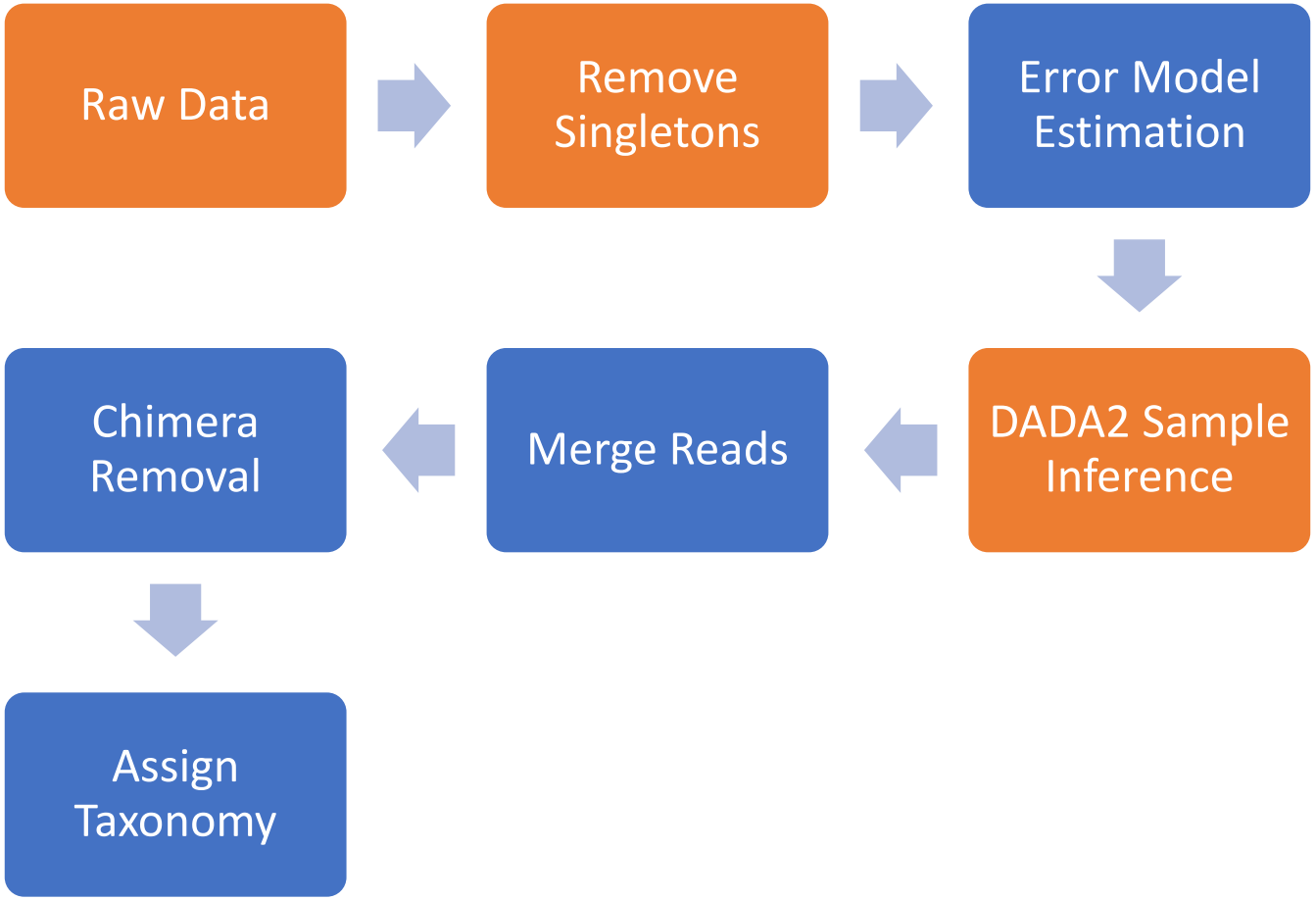
Summary

DADA2 Workflow

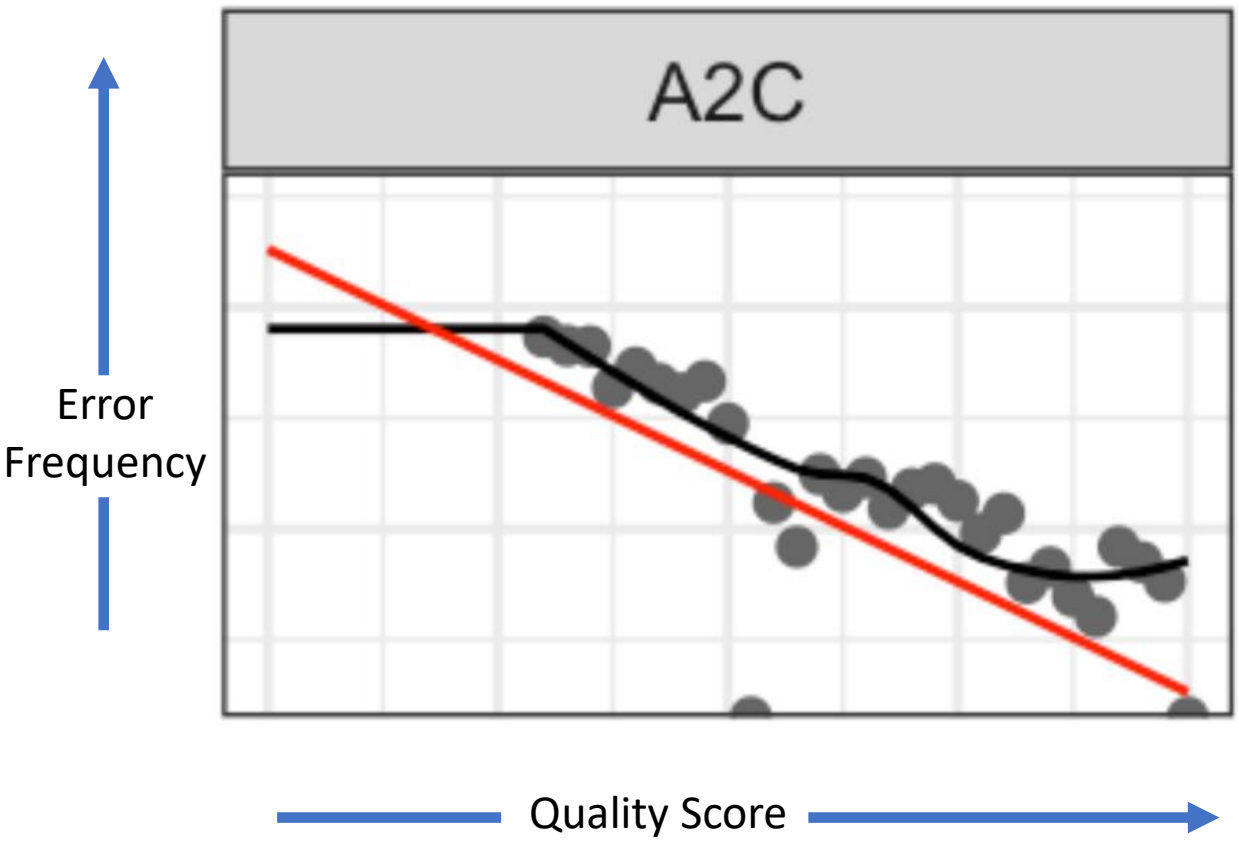


Remove Singletons

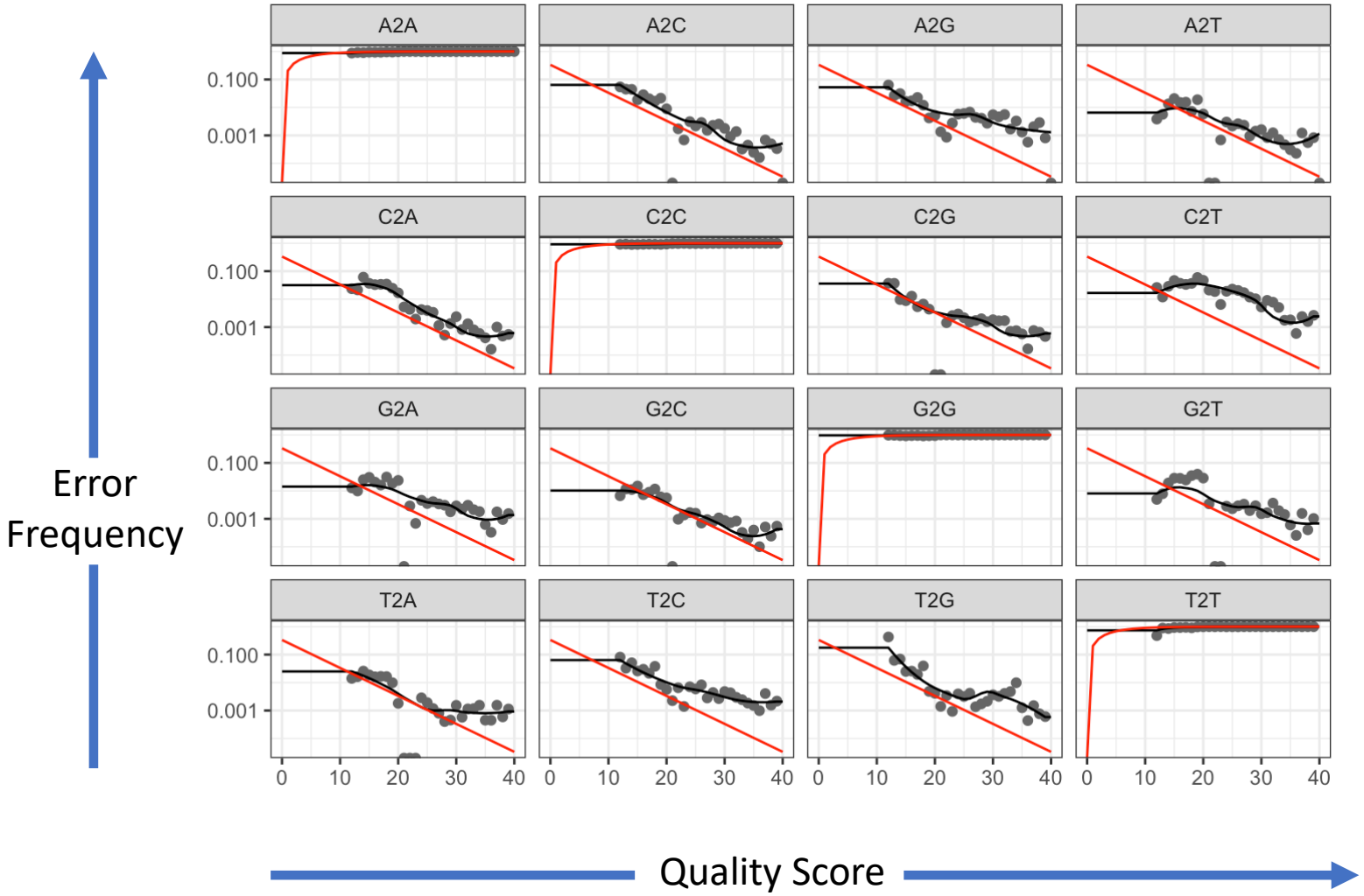
DADA2 Workflow



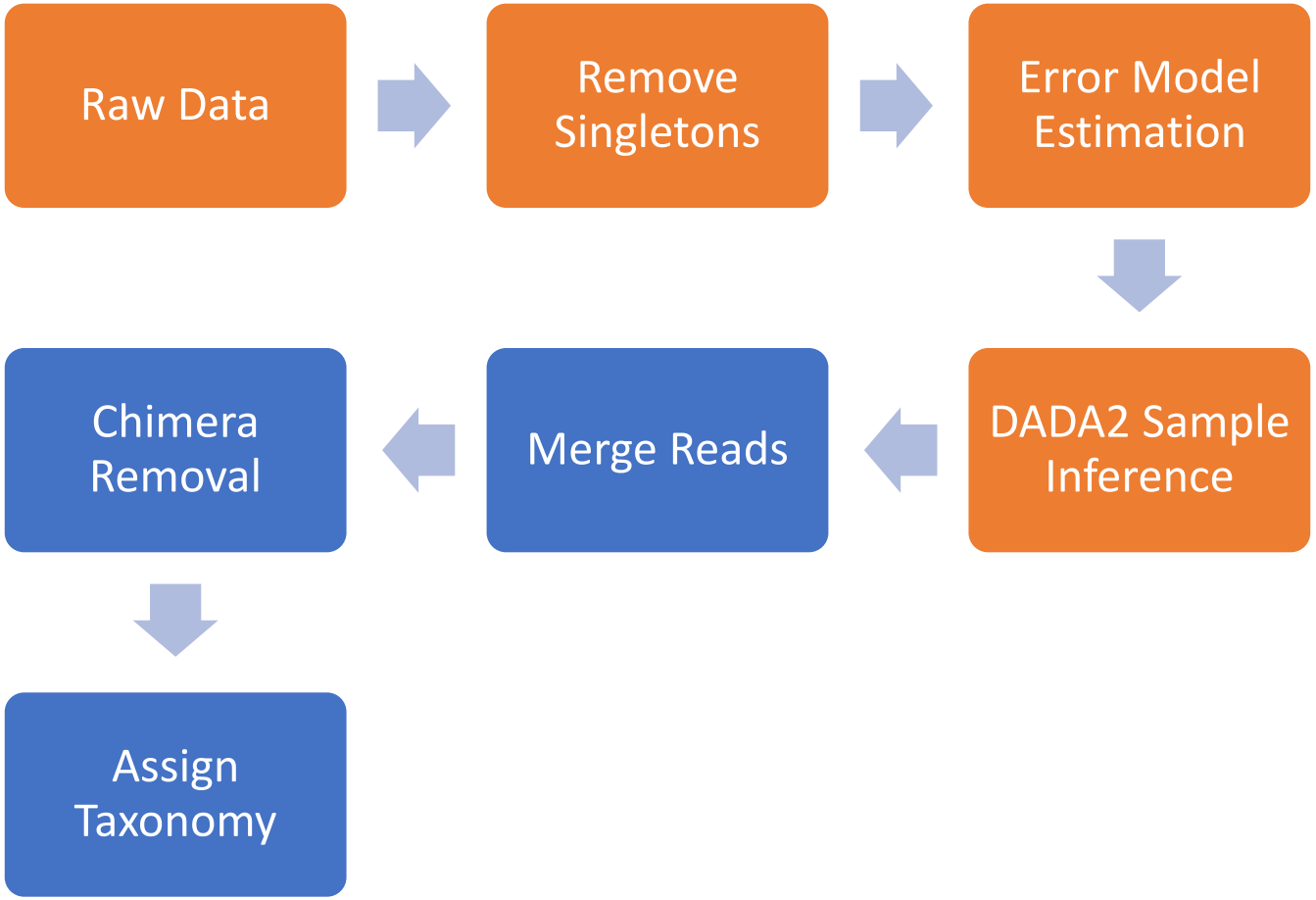
Error model



Error model

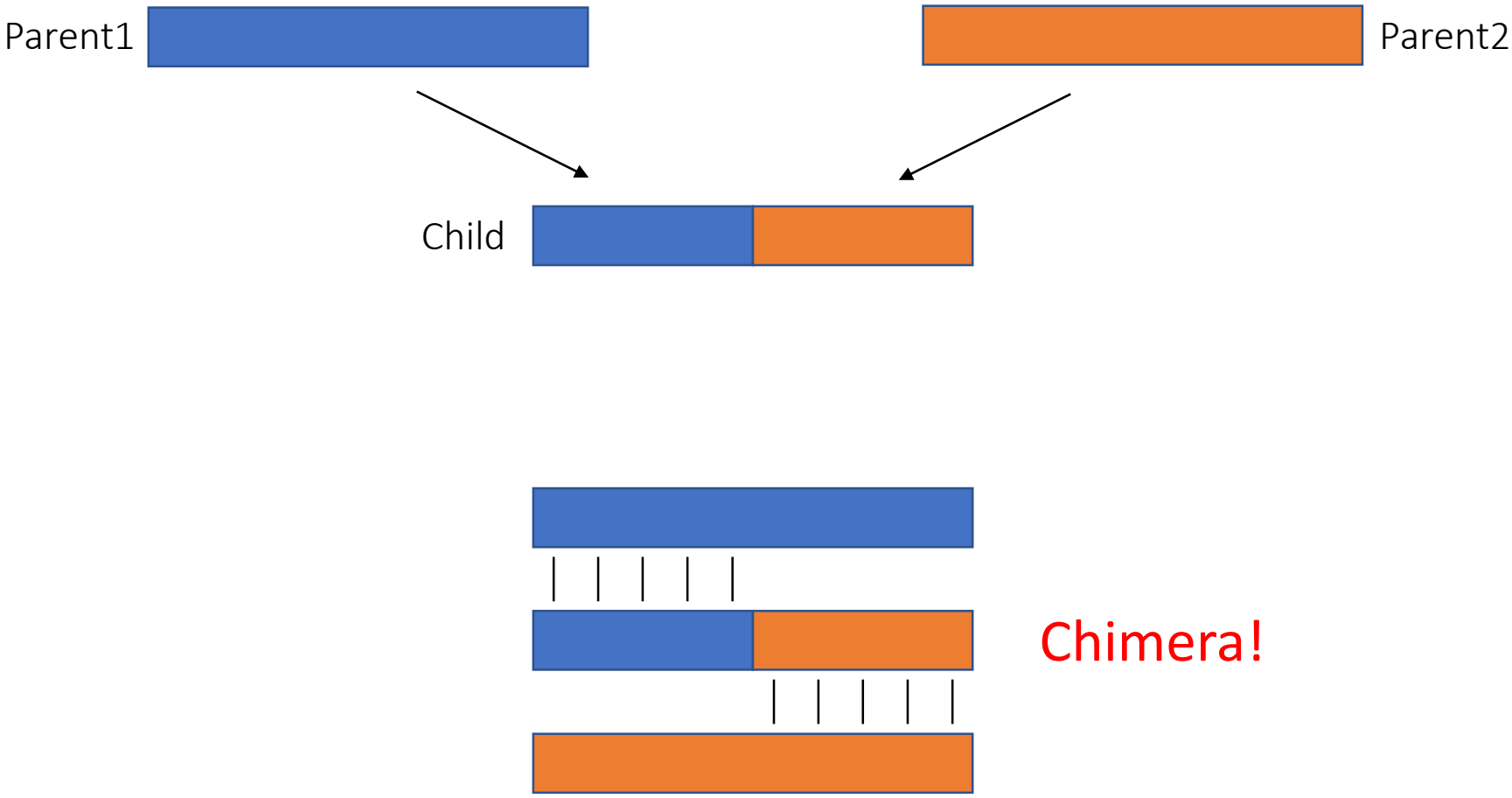

A2A
A2C
A2G
A2T
C2A
C2C
C2G
C2T
G2A
G2C
G2G
G2T
T2A
T2C
T2G
T2T
0.100
0.001
0
10
20
30
40
Error Frequency
Quality Score

DADA2 Workflow

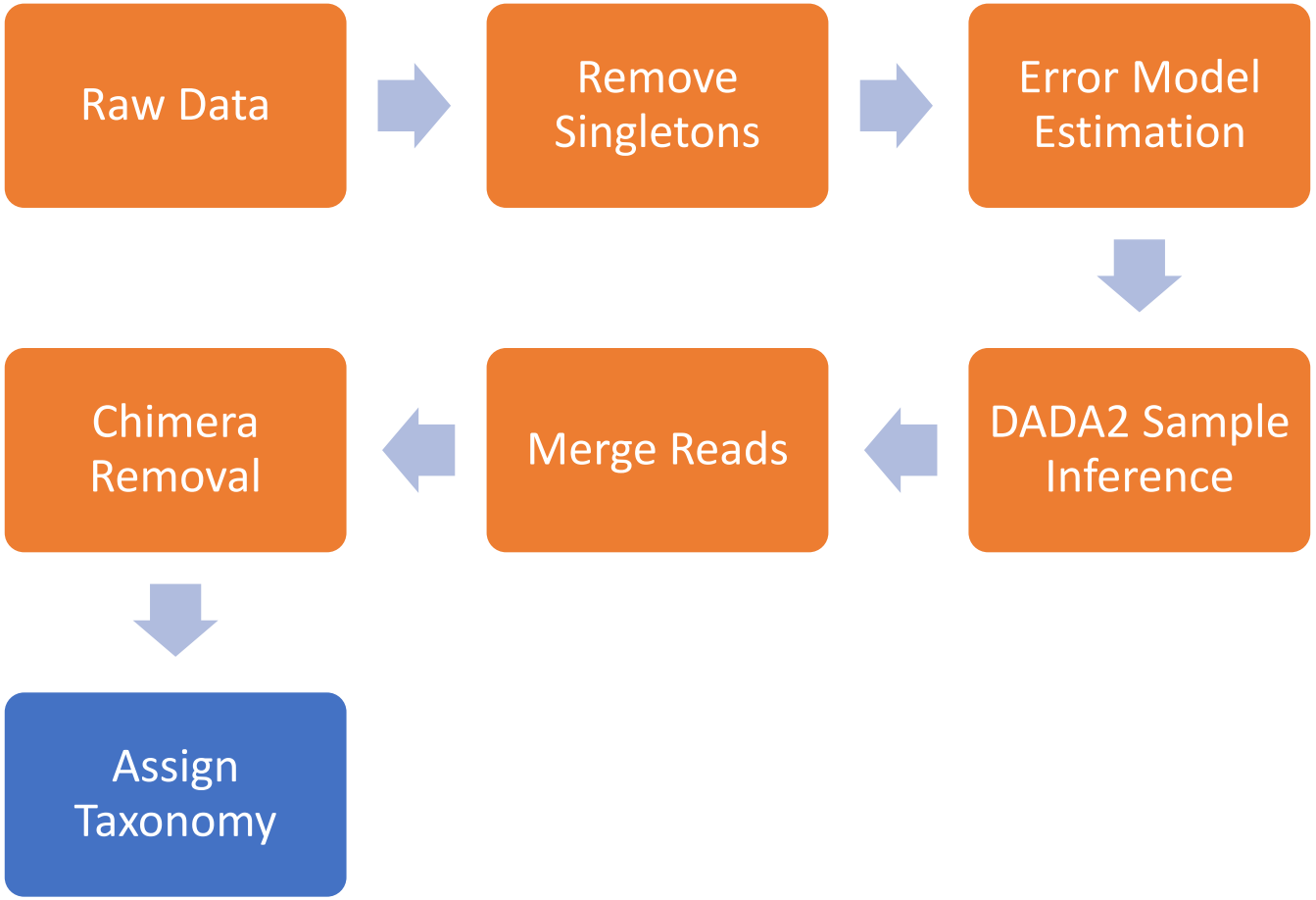


What is a Chimera?

Chimera Detection



DADA2 Workflow



Benefits to DADA2

- Open Source
- Parallelizable
- Customizable
- Compatible with all amplicon types (not only 16S)
- Provide single-nucleotide resolution
- Lower false-positive rates

Outline

Background

True Variation vs. Sequencing Errors

DADA2 Core Algorithm

DADA2 Full Workflow

Some Results

Summary

Results

Compared on 3 mock community datasets against 4 algorithms:

- UPARSE
- MED
- Mothur
- QIIME

Results

- UPARSE 9s
- QIIME 17s
- DADA2 21s
- Mothur 2m 26s
- MED 2m 34s

Results

On real communities DADA2 revealed a diversity of previously undetected variants!


SUMMARY

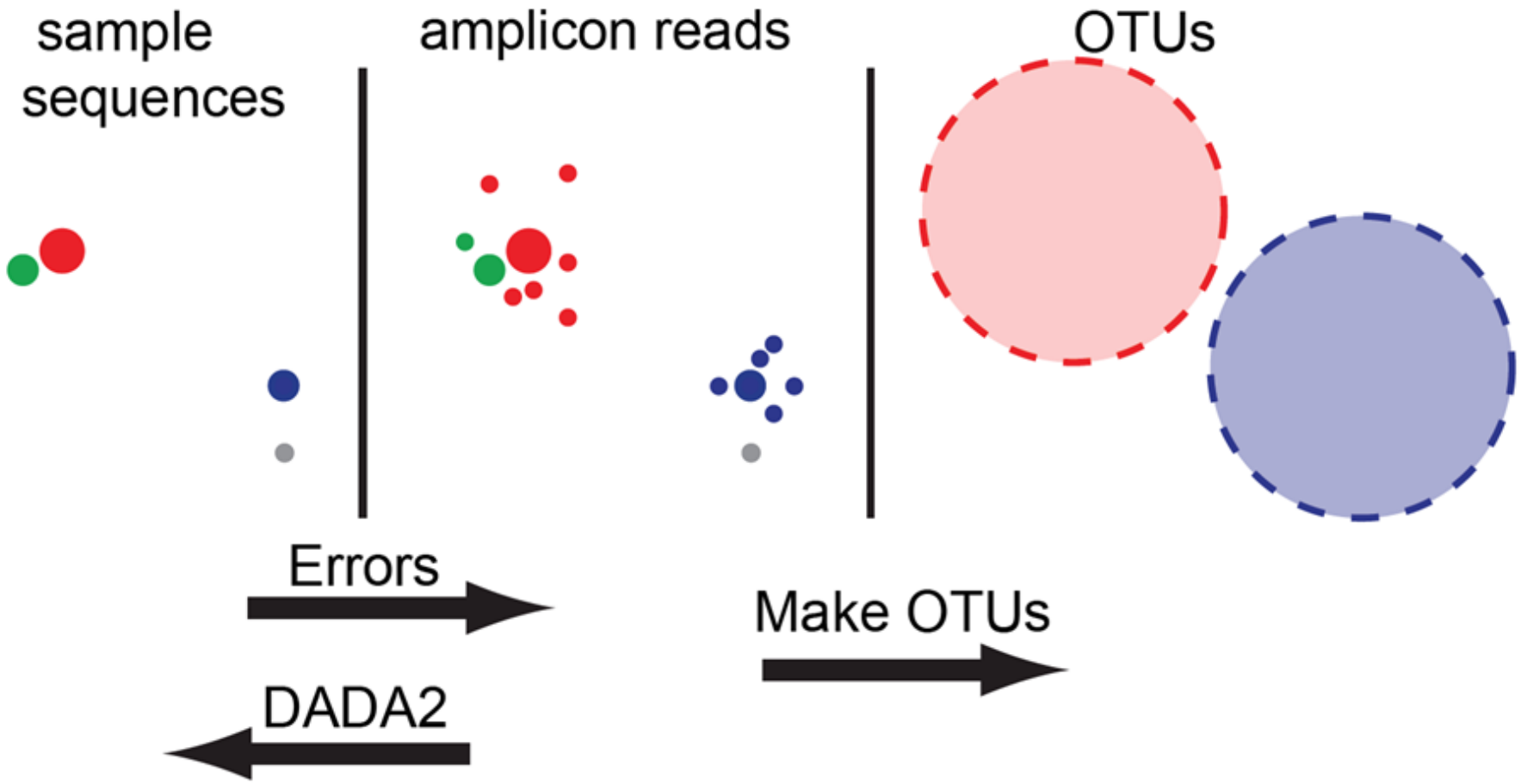
True Variation vs. Sequencing Error

Sequence Read 1: acttcatgataccacatgatacg

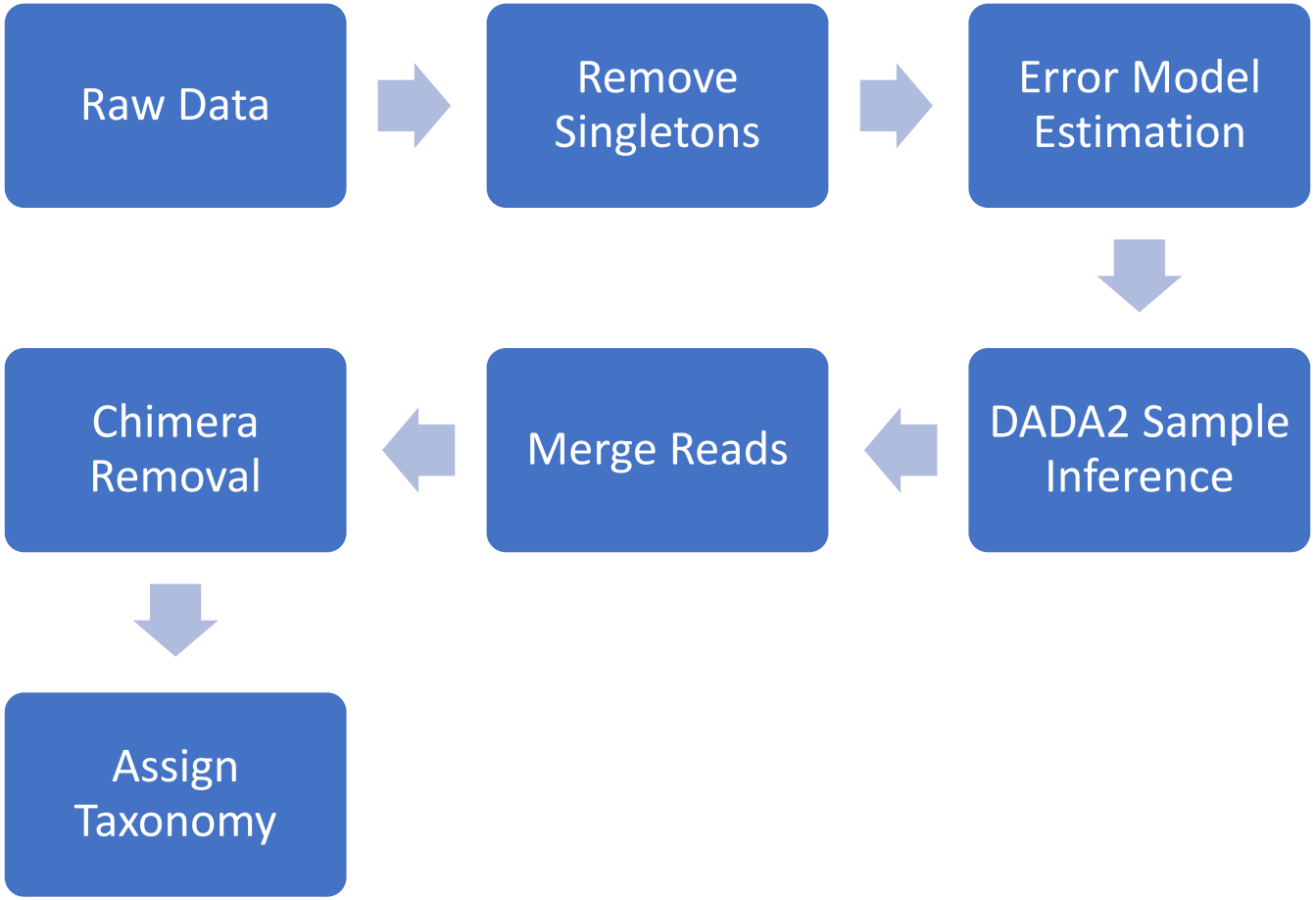
Sequence Read 2: acttcatgctaccacatgatacg

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

Visualization



DADA2 Workflow



For More Information…

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan[1], Paul J McMurdie[2], Michael J Rosen[3], Andrew W Han[2], Amy Jo A Johnson[2] & Susan P Holmes[1]

RESEARCH ARTICLE

Bioconductor workflow for microbiome data analysis: from raw reads to community analyses [version 1; referees: 3 approved]

Ben J. Callahan[1], Kris Sankaran[1], Julia A. Fukuyama[1], Paul J. McMurdie[2], Susan P. Holmes[1]

[1]Statistics Department, Stanford University, Stanford, CA, 94305, USA
[2]Whole Biome Inc., San Francisco, CA, 94107, USA

Discussion Points

- How didn’t we think about it before?
- Do you think we can improve?

Thanks for Listening