H3ABioNet

Pan African Bioinformatics Network for H3Africa

16SrRNA Intermediate Bioinformatics Online Course: Int_BT_2019

16S rRNA analysis pipeline

Taxonomic classification and alignment using the dada2 pipeline

Module 5: 16S rRNA Analysis Pipeline

- Session 1:

QC and ASV picking using the dada2 pipeline

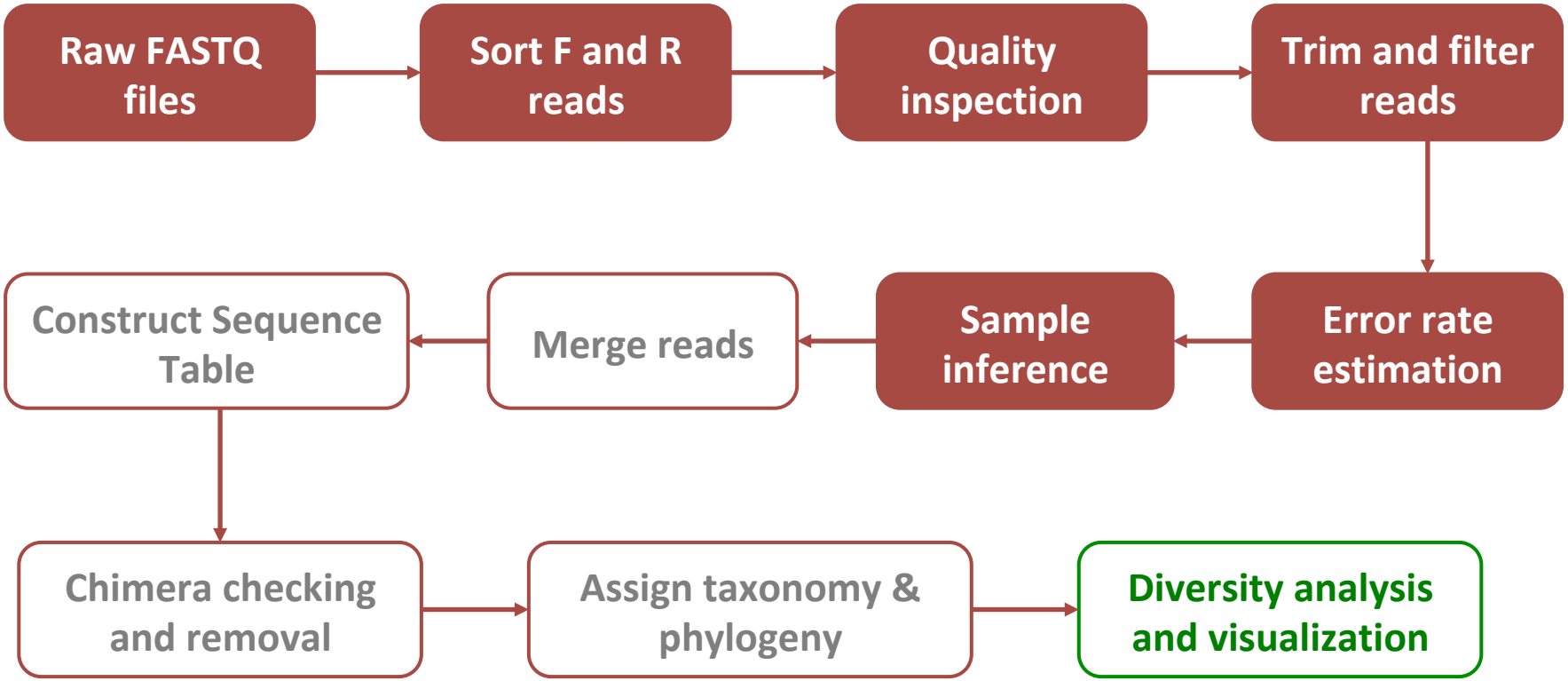
- Session 2:

Taxonomic classification and alignment using the dada2 pipeline

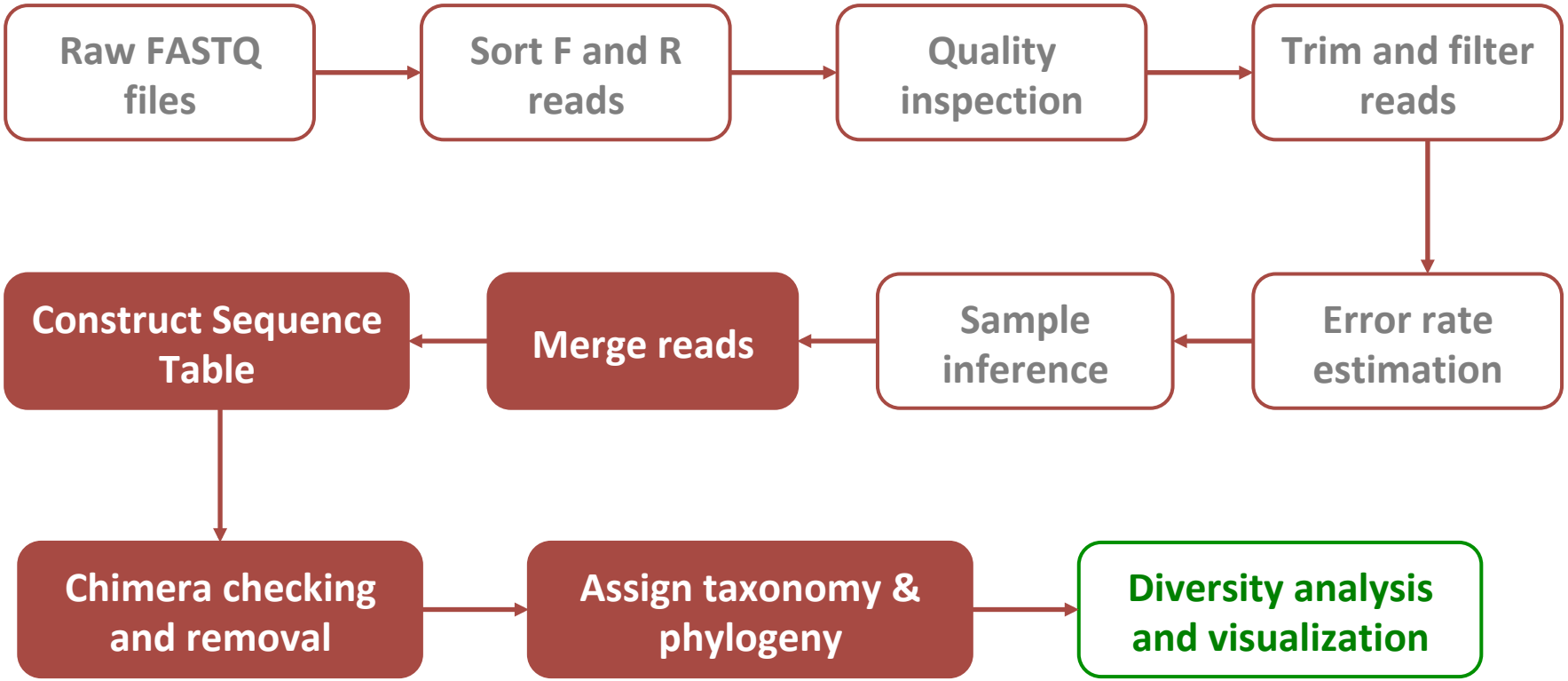
Outline

- Quality Control
- DADA2 background
- **DADA2 workflow**

DADA2 workflow



DADA2 workflow



DADA2 workflow

Merge Reads

```
merge.reads <- mergePairs(dadaF, filt.dataF, dadaR, filt.dataR, verbose=TRUE)
```

```
## 76306 paired-reads (in 1153 unique pairings) successfully merged out of 82378 (in 2526 pairings) input.
```

```
## 54948 paired-reads (in 720 unique pairings) successfully merged out of 60139 (in 1811 pairings) input.
```

```
## 83013 paired-reads (in 843 unique pairings) successfully merged out of 89310 (in 2175 pairings) input.
```

```
## 74533 paired-reads (in 1093 unique pairings) successfully merged out of 81135 (in 2674 pairings) input.
```

```
## 67364 paired-reads (in 855 unique pairings) successfully merged out of 71053 (in 1743 pairings) input.
```

```
## 69262 paired-reads (in 1194 unique pairings) successfully merged out of 76778 (in 2891 pairings) input.
```

mergePairs merges reads only if they exactly overlap.

The length of your overlap, by default is 20 nt for DADA2, you can lower it by using this parameter **minOverlap**.

DADA2 workflow

Merge Reads

```
head(merge.reads[[1]])
```

```
##
sequence
## 1  TGTGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGC
TCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGG
CGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGAAACCCGAGTAGTCCGGCTGACTGACTCGAGA
## 2  TGTGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGC
TCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGG
CGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGAAACCCTAGTAGTCCGGCTGACTGACTCGAGA
## 3  TGTGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGC
TCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGG
CGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGAAACCCCAGTAGTCCGGCTGACTGACTCGAGA
## 4 TTGTGTGCCAGCAGCCGCGGTAATACGTAGGGGGCTAGCGTTATCCGGATTTACTGGGCGTAAAGGGTGCGTAGGCGGTCTTTCAAGTCAGGAGTTAAAGGCTACGGC
TCAACCGTAGTAAGCTCCTGATACTGTCTGACTTGAGTGCAGGAGAGGAAAGCGGAATTCCCAGTGTAGCGGTGAAATGCGTAGATATTGGGAGGAACACCAGTAGCGAAGGC
GGCTTTCTGGACTGTAACTGACGCTGAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGAAACCCTAGTAGTCCGGCTGACTGACTCGAGAG
## 5  TGTGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGC
TCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGG
CGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGAAACCCTTGTAGTCCGGCTGACTGACTCGAGA
## 6  TGTGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGC
TCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGG
CGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGAAACCCTGGTAGTCCGGCTGACTGACTCGAGA
##   abundance forward reverse nmatch nmismatch nindel prefer accept
## 1       460       2       1    253          0      0      2   TRUE
## 2       456       1       1    253          0      0      2   TRUE
## 3       421       5       1    253          0      0      2   TRUE
## 4       414       7       2    252          0      0      2   TRUE
## 5       401       6       1    253          0      0      2   TRUE
## 6       400       4       1    253          0      0      2   TRUE
```

DADA2 workflow

Construct Amplicon Sequence Variant (ASV) Table

```
seqtab <- makeSequenceTable(merge.reads)
dim(seqtab)
```

```
## [1]    15 13527
```

```
table(nchar(getSequences(seqtab)))
```

```
##
##  311  312  313  315
##  107 9136 4283    1
```

DADA2 workflow

Construct Amplicon Sequence Variant (ASV) Table

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		TTGTGTGCCA	TTGTGTGCCA	TGTGTGCCA	TGTGTGCCA	TTGTGTGCCA	TTGTGTGCCA	TGTGTGCCA	TGTGTGCCA	TTGTGTGCCA	TTGTGTGCCA	TGTGTGCCA	TTGTGTGCCA
2	Dog1	0	0	242	205	0	0	0	223	0	0	195	0
3	Dog10	0	0	0	0	0	0	0	0	0	0	0	0
4	Dog15	0	0	0	0	0	0	0	0	0	0	0	0
5	Dog16	0	0	0	0	0	0	0	0	0	0	0	0
6	Dog17	0	0	0	0	0	0	0	0	0	0	0	0
7	Dog2	373	0	0	0	276	0	0	0	283	277	0	322
8	Dog22	1926	0	0	0	1516	0	0	0	1453	1459	0	1366
9	Dog23	0	955	0	0	0	805	0	0	0	0	0	0
10	Dog24	0	0	0	0	0	0	1747	0	0	0	0	0
11	Dog29	0	0	0	0	0	0	0	0	0	0	0	0
12	Dog3	0	921	0	0	0	944	0	0	0	0	0	0
13	Dog30	0	0	0	0	0	0	0	0	0	0	0	0
14	Dog31	0	0	1596	1625	0	0	0	1523	0	0	1536	0
15	Dog8	0	0	0	0	0	0	0	0	0	0	0	0
16	Dog9	0	0	0	0	0	0	0	0	0	0	0	0
17													
18													

DADA2 workflow

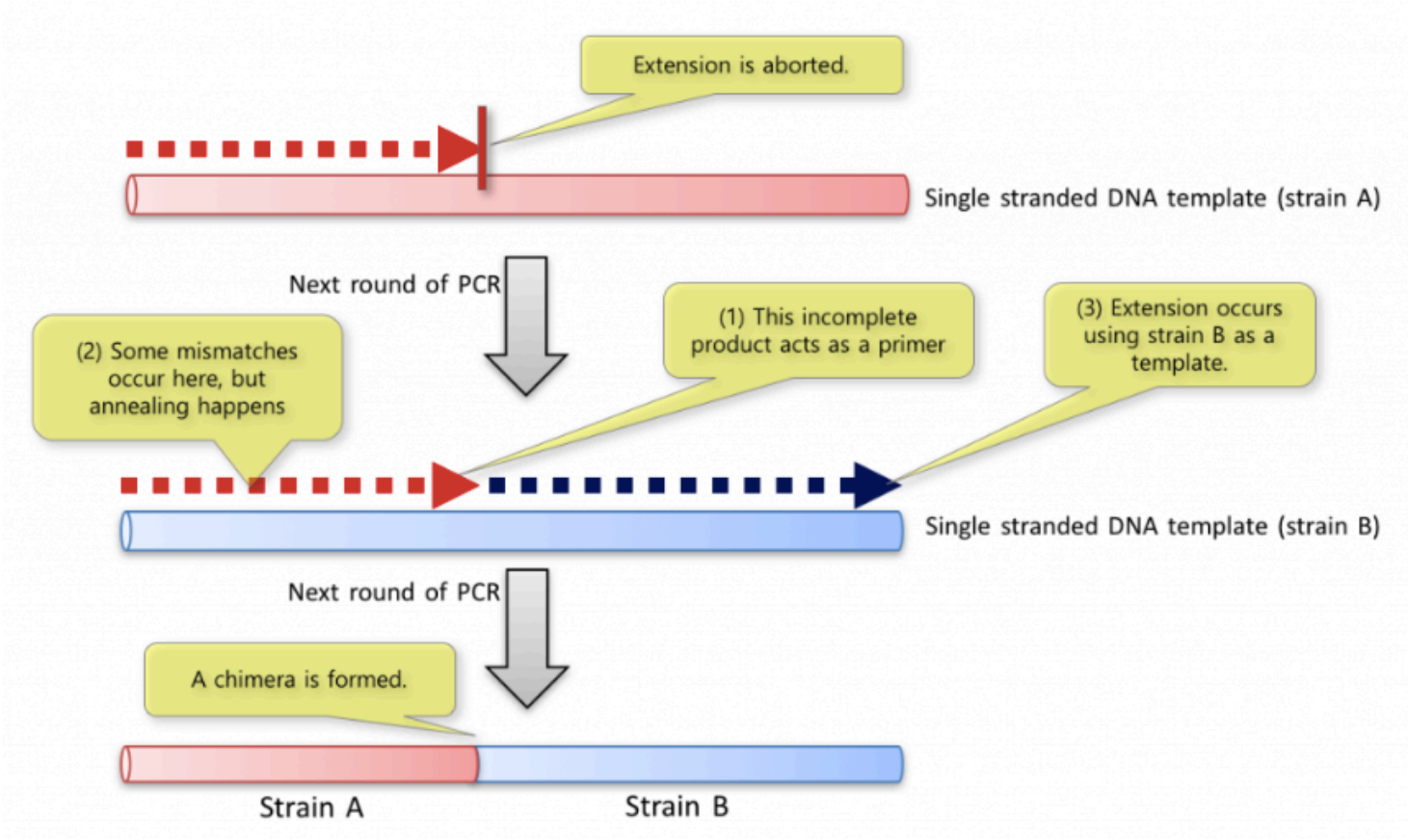
H3ABioNet

Chimera Sequence

- Chimeras are sequences formed from two or more biological sequences joined together.
- Amplicons with chimeric sequences can be formed during PCR.
- Chimeras are rare with shotgun sequencing but are common in amplicon sequencing when closely related sequences are amplified.

DADA2 workflow

Chimera Sequence



https://help.ezbiocloud.net/

DADA2 workflow

Chimera Checking and Removal

```
seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE, verbose=TRUE)
```

```
## Identified 9112 bimeras out of 13527 input sequences.
```

```
dim(seqtab.nochim)
```

```
## [1]   15 4415
```

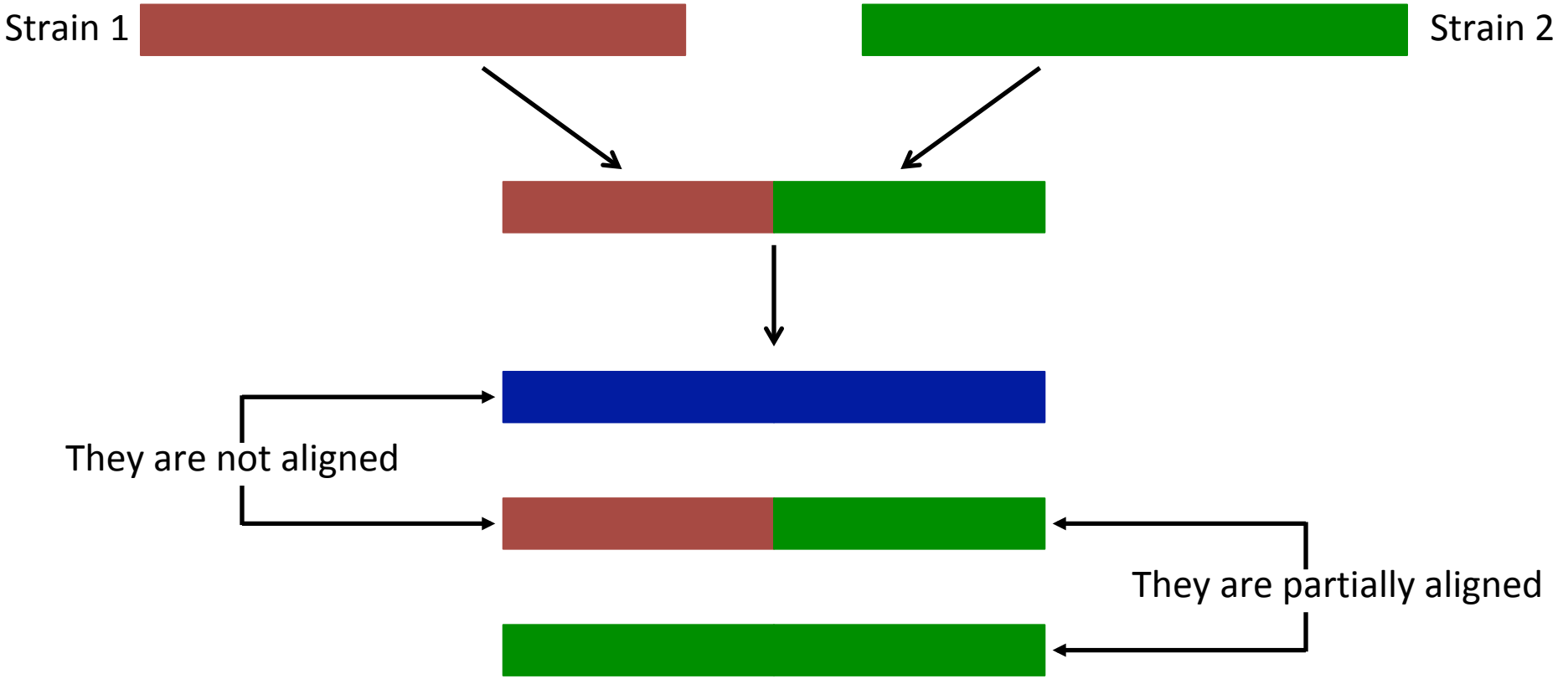
```
sum(seqtab.nochim)/sum(seqtab)
```

```
## [1] 0.5094968
```

- It uses *de novo* to check for two parent chimeras.
- Chimeric sequences are identified if they can be exactly reconstructed by combining a left-segment and a right-segment from two more abundant "parent" sequences.

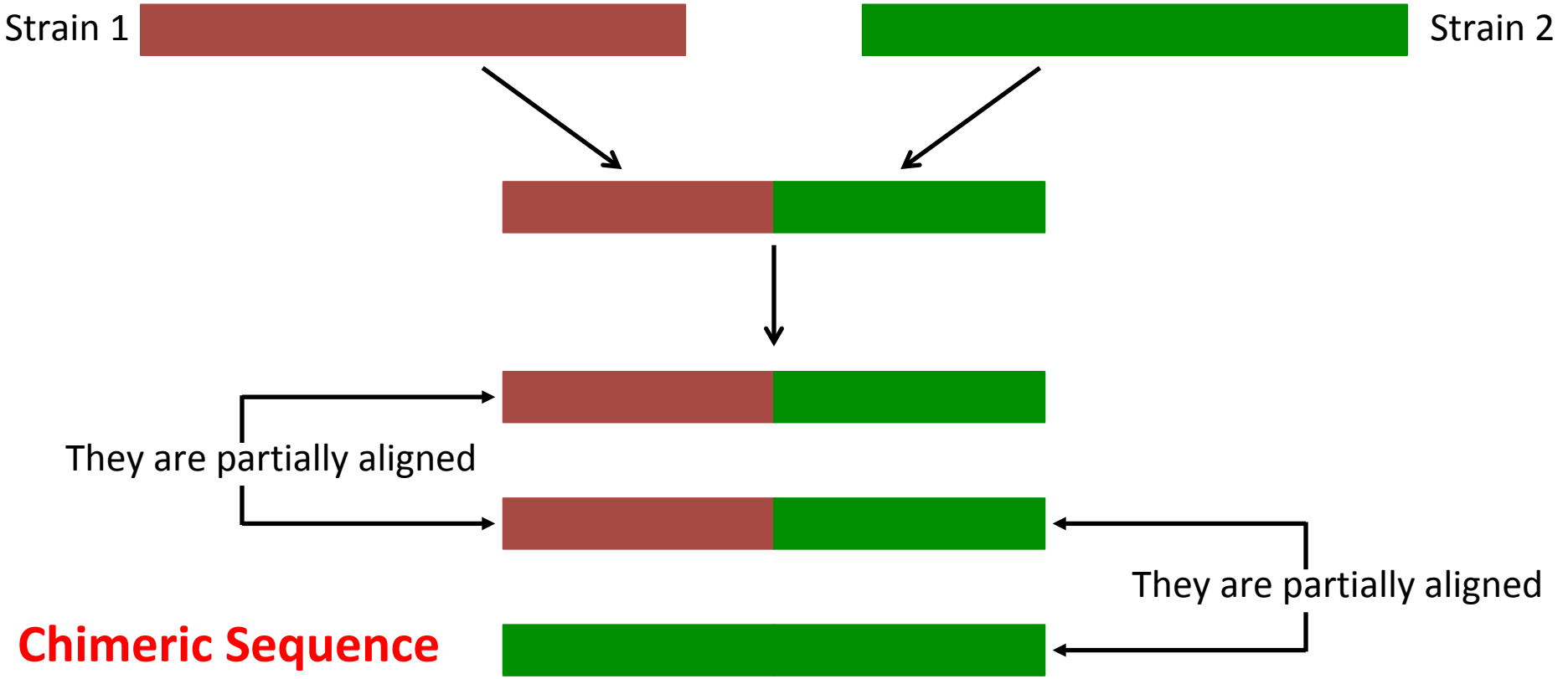
DADA2 workflow

Chimera Checking and Removal



DADA2 workflow

Chimera Checking and Removal



DADA2 workflow

Chimera Checking and Removal

```
seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE, verbose=TRUE)
```

```
## Identified 9112 bimeras out of 13527 input sequences.
```

```
dim(seqtab.nochim)
```

```
## [1]   15 4415
```

```
sum(seqtab.nochim)/sum(seqtab)
```

```
## [1] 0.5094968
```

- It uses *de novo* to check for two parent chimeras.
- Chimeric sequences are identified if they can be exactly reconstructed by combining a left-segment and a right-segment from two more abundant "parent" sequences.