

SeqTrimNext

Statistics of pre-processing

Plataforma Andaluza de Bioinformática
Universidad de Málaga

April 23, 2020

1 Output Files

SeqTrimNext provides several files, the most interesting ones are in the following directories:

- `output_files`
 - `output.less`, containing an extensive information about the trimming of each sequence. It can be visualised on terminal using the command `less -R`.
 - `used_params.txt`, containing the complete set of parameters used for execution of SeqTrimNext with your data
 - `rejected.txt`, containing a list of rejected sequences together with the reason for their removal.
 - `initial_stats.json`, containing statistics for raw sequences.
 - `stats.json`, containing the statistics of the cleaning process.
 - There is a collection of `folders` that gather sequences with the same MID; each folder contains a `sequences` file (in FASTQ format) with useful reads. There may also exist a file with reads containing low complexity regions. If you want to reconstruct a SFF with the useful segment of each pre-processed read, use `sff_info` file in combination with the original SFF file for the `sfffile` tool.
- `graphs`
 - `size_stats.png`, a graph with the distribution of read lengths in raw data (see Fig. ??).
 - `qualities.png`, a graph to inspect read qualities in raw data (see Fig. ??).
 - `PluginExtractInserts_insert_size.png`, a graph with the distribution of read lengths after SeqTrimNext pre-processing (see Fig. 1).
 - There are other graphs (mostly bar plots) that illustrate the quality of pre-processed reads. All are in PNG format.
- `latex`
 - It is provided as a compressed file `latex.zip` containing all “.tex” files required to compile this document. Graphs are taken from the `graph` folder

2 Relevant parameters

In this section, the relevant parameters used in your experiment are shown. Full information about the parameters can be obtained from file `used_params.txt`

2.1 General

Plugins applied to every sequence, separated by commas. Order is important

1. PluginIndeterminations
2. PluginFindPolyAt
3. PluginAbAdapters
4. PluginUserContaminants
5. PluginContaminants
6. PluginVectors
7. PluginLowQuality
8. PluginLowComplexity
9. PluginExtractInserts

Remove duplicated (clonal) sequences (using CD-HIT 454)

```
remove_clonality: false
```

Minimum insert size for every trimmed sequence

```
min_insert_size_trimmed: 30
```

Minimum insert size for each end of paired-end reads; true paired-ends have both single-ends longer than this value

```
min_insert_size_paired: 40
```

Seqtrim version

```
seqtrim_version: 2.0.67
```

```
min_sequence_size_raw:
```

2.2 Quality

Minimum quality value for every nucleotide

```
min_quality: 20
```

```
window_width:
```

2.3 Contaminants

Blast E-value used as cut-off when searching for contaminations

```
blast_evalue_contaminants: 1.0e-10
```

Minimum required identity (%) for a reliable contamination

```
blast_percent_contaminants: 85
```

Minimum hit size (nt) for considering a true contamination

```
min_contam_seq_presence: 40
```

Genus of input data: contaminations belonging to this genus will be ignored

```
genus:
```

Is a contamination considered a source of sequence rejection? (setting to false will only trim contaminated sequences instead of rejecting the complete read)

```
contaminants_reject: true
```

Path for contaminants database

```
contaminants.fasta
```

```
cont_ribosome.fasta
```

3 Pre-processing statistics

Next figure is equivalent to Figure ?? but using output reads (useful sequences). The mode is expected to decrease but the shape of the plot should be similar.

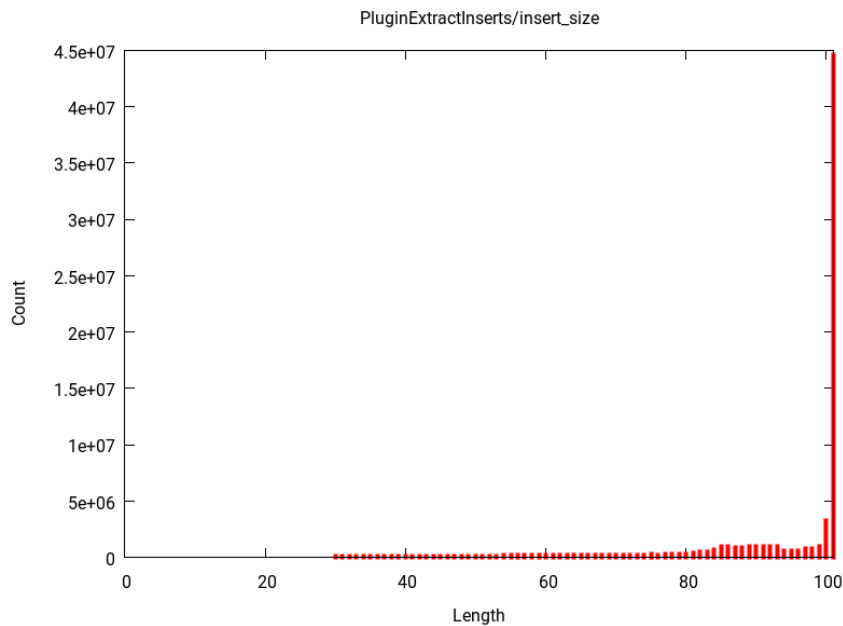


Figure 1: Size distribution of the output sequences. Short sequences ($< \text{min_insert_size_trimmed}$) were removed. [PluginExtractInserts_insert_size.png]

Summary statistics of the SeqTrimNext analysis. Be careful and read all warnings that are indicating concerns about your data. In the files `initial_stats.json` and `stats.json` can be found a full statistics of your data and SeqTrimNext pre-processing

Input reads:	total	75686398
	Smallest read (bp)	101
	Largest read (bp)	101
	Mode (bp)	0
	Mean (bp)	0.0
Output results:	total	1790525
	Rejected	4836935
	Low complexity reads	39758
	Mode (bp)	91
	Mean (bp)	93.9
	Output paired reads	69019180
	Total output reads	70809705
Linkers:		

Table 1: List of the most frequent Vectors found among your reads

Vectors	sequences
Cloning vector pAAV-MCS, complete sequence.	402703
Cloning vector pKOHPRT complete sequence.	385216
Enterobacteria phage lambda	123065
Cloning vector pVLH/hsp	40331
Cloning vector pWormgate2, complete sequence.	33631

Table 2: List of the most frequent Adapters found among your reads

Adapters	sequences
TruSeq_Universal_Adapter	659189
Illumina_Multiplexing_Read2_Sequencing_Primer	252926
ABI_Solid3_Adapter_A	216184
ABI_Solid3_GAPDH_Reverse_Primer	176682
TruSeq_Adapter_Index_4	163357

Table 3: List of the most frequent Contaminants found among your reads

Contaminants	sequences
rRNA_small_subunit_Metazoa_Myotis	114078
Saccharomyces cerevisiae S288c chromosome XII, complete sequence	80420
rRNA_long_subunit_Metazoa_Dasytus	66105
Schizosaccharomyces pombe 972h- chromosome III, complete sequence	62640
rRNA_small_subunit_Metazoa_Equus	58795

Table 4: Summary of nucleotides removed in every plugin.

Plugin	Nucleotides	Percent	Warnings
Low Quality	605975290	Inf %	OK
Low Complexity	60728749	Inf %	OK
Poly T	23714750	Inf %	OK
Poly A	25729764	Inf %	OK
Contaminants	56109868	Inf %	OK
Adapters	93701327	Inf %	OK
Vectors	47095876	Inf %	OK
Indeterminations	234766	Inf %	OK
Inserts	6652473042	Inf %	iW1

iW1 Warning!, only Inf % of nucleotides are useful

4 Rejected reads

Input sequences	75686398
Output sequences	1790525
Rejected sequences	4836935
Output paired sequences	69019180
Total output sequences	70809705
Low complexity sequences	39758

Table 5: Summary of reads removed in every plugin.

Case	Number of sequences	Percent	Warnings
	4836935	6.391 %	OK
Short inserts	2943422	3.889 %	OK
Empty Inserts	1279025	1.690 %	rdW3
Contaminants	594945	0.786 %	rdW4
No Valid Inserts	12951	0.017 %	OK
Low Complexity	5407	0.007 %	OK
Indeterminations	1062	0.001 %	OK
Unexpected Vector	123	0.000 %	OK
Total rejected	4836935	6.391 %	OK

rdW3 Warning!, a 1.690 % of your sequences are empty (without an insert)

rdW4 Warning!, a 0.786 % of your sequences are from a contaminant organism or from organelles

References

- [1] Falgueras et al. SeqTrim: a high-throughput pipeline for preprocessing any type of sequence reads. *BMC Bioinformatics* 11:38 (2010)
- [2] Weizhong Li & Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* (2006) 22:1658-9

Thanks you for use SeqTrimNext! Send us any comment to scbi support