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 exists 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.

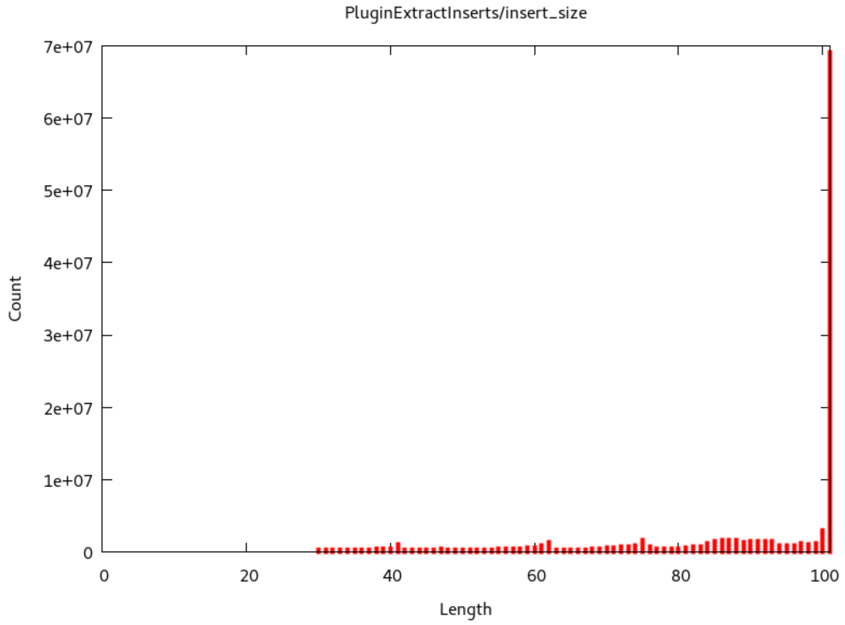

Figure 1: Size distribution of the output sequences. Short sequences (< min_insert_size_trimmed) were removed. [PluginExtractInserts_insert_size.png]

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	134258074
	Smallest read (bp)	101
	Largest read (bp)	101
	Mode (bp)	0
	Mean (bp)	0.0
Output results:	total	5392165
	Rejected	14965534
	Low complexity reads	81579
	Mode (bp)	91
	Mean (bp)	90.2
	Output paired reads	113818796
	Total output reads	119210961
Linkers:		

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

Vectors	sequences
Cloning vector pKOHPRT complete sequence.	647151
Cloning vector pAAV-MCS, complete sequence.	645587
Enterobacteria phage lambda	173399
Cloning vector pVLH/hsp	57527
Cloning vector pWormgate2, complete sequence.	52833

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

Adapters	sequences
ABI_Solid3_Adapter_A	353704
TruSeq_Universal_Adapter	346755
ABI_Solid3_GAPDH_Reverse_Primer	297956
Illumina_Single_End_Adapter_1	276505
ABI_Solid3_GAPDH_Forward_Primer	259178

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

Contaminants	sequences
rRNA_small_subunit_Metazoa_Myotis	160507
rRNA_long_subunit_Metazoa_Dasypus	138008
Saccharomyces cerevisiae S288c chromosome XII, complete sequence	131251
Schizosaccharomyces pombe 972h- chromosome III, complete sequence	97226
rRNA_small_subunit_Metazoa_Equus	82926

Table 4: Summary of nucleotides removed in every plugin.

Plugin	Nucleotides	Percent	Warnings
Low Quality	2054940225	Inf %	OK
Low Complexity	86417525	Inf %	OK
Poly T	40988164	Inf %	OK
Poly A	47730806	Inf %	OK
Contaminants	89606753	Inf %	OK
Adapters	118154180	Inf %	OK
Vectors	73567226	Inf %	OK
Indeterminations	125311	Inf %	OK
Inserts	10762219825	Inf %	iW1

iW1 Warning!, only Inf % of nucleotides are useful

4 Rejected reads

Input sequences	134258074
Output sequences	5392165
Rejected sequences	14965534
Output paired sequences	113818796
Total output sequences	119210961
Low complexity sequences	81579

Table 5: Summary of reads removed in every plugin.

Case	Number of sequences	Percent	Warnings
	14965534	11.147 %	OK
Short inserts	9544141	7.109 %	OK
Empty Inserts	4440010	3.307 %	rdW3
Contaminants	952341	0.709 %	OK
No Valid Inserts	19982	0.015 %	OK
Low Complexity	8242	0.006 %	OK
Indeterminations	708	0.001 %	OK
Unexpected Vector	110	0.000 %	OK
Total rejected	14965534	11.147 %	OK

rdW3 Warning!, a 3.307 % of your sequences are empty (without an insert)

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