SeqyClean User Manual

Ilya Y. Zhbannikov[1,3], Samuel S. Hunter[1,3], Matthew L. Settles[1,3]

[1]Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844-3051,
[2]Department of Bioinformatics and Computational Biology, University of Idaho, Moscow, ID 83844-3051,

Keywords:
Bioinfomatics, Genomics

Correspondence:
Ilya Y. Zhbannikov
Department of Bioinformatics and Computational Biology
University of Idaho
Moscow, ID 83844-30521
Email: zhba3458@vandals.uidaho.edu

Running Title:
SeqyClean User Manual

1 Abstract

Raw data from sequencing machines is not usually completely prepared for analysis since sequences often contain remnants of nucleotides that have been used during DNA library preparation. Such nucleotides can potentially case hurdles during data analysis and hereby need to be removed. In additional, the noise, which is presented as low-quality bases, has serious impact on genome assembly and mapping. We propose SeqyClean, a specialized cleaning pipeline that alleviates these issues.

2 Introduction

2.1 Importance of the quality of sequence for both genome assembly and genome mapping applications

With newly developed next-generation sequencing technologies and increasing interest in gene discovery, DNA mapping, functional genomics and genome annotations, the amount of sequencing has exponential growth curve and doubles every month. Sequence data from automatic sequencing machines should not be considered as "ready-to-use" data for analysis due to contaminant remnants and adapters used during DNA sequencing. Poly A/T tails and low quality of sequenced data due to base-call errors make the library even worse. One can say that cleaning next-generations sequence data is not worthwhile because of large number of reads within a DNA library (1 million for Roche 454 and 10 millions for Illumina on average). However, we would argue with this and in order to support our claim we present several examples that show the importance of sequence cleaning. As an example, adapters left within a read along with base-call errors can potentially lead to unexpected results after genome assembly having large amount of small contigs, which indicates that genome was not assembled properly. Another example shows how important to trim pieces of adapters within a library of interest. Illumina technology employs specific adapters named TruSeq

in order to immobilize DNA pieces on the surface. This situation is described in Figure XX. In this figure, two large contigs could not join together because of abundance of adapters at the end of reads (marked yellow). The assembler stopped building the left contig and begun another one. Another example shows impact of contaminant screening on genome assembly when results lead to the wrong conclusion as the evidence of contaminants within a library of interest. However, after performing a BLAST search in order to find similar genomic sequences, it turned out that in addition to anolis data, the published genome contains genomic sequences of PhiX virus, which cannot be presented in anolis genome. Apparently, there was no contaminant screening provided before genome assembly. That is why it is very important to filter contaminants that may potentially be in a library. An additional cleaning step is highly desirable. This step is intended to filter adapters, contaminants, vector particles along with trimming of low-quality regions in order to prepare sequences for further analysis.

2.2 Sequence cleaning software

2.2.1 Individual cleaning aspects

We divide the cleaning tasks into the following categories:

- Adapter trimming

  Adapter trimming should be one of the major cleaning task. Adapters are short pieces of nucleotides used during sequencing. Different sequencing technologies employ different types of adapters. Cleaning software must be flexible enough to detect such artifacts within reads.

- Quality trimming

  Quality trimming is important part of cleaning pipeline intended to trim low-quality bases caused by base-call errors.

- Poly A/T tails trimming

  Trimming of poly A and poly T tails is used mostly for Roche 454 technology.

- Contaminants screening

  Removes those reads that can be contaminants.

- Vector trimming

  Removes vector pieces within reads. Necessary step if bacteria vector was used during the dna library preparation.

2.2.2 Existing cleaning software

Today's sequence cleaning software are mostly tools that perform individual cleaning aspects (e.g. TrimEST (removes poly A/T tails from sequence), VectorStrip and VecScreen (both trim bacteria vector particles), ESTPrep(Scheetz et al., 2003), Figaro(White, Roberts, Yorke, & Pop, 2008), Lucy(Chou & Holmes, 2001)), and AlienTrimmer (Criscuolo & Brisse, 2013) there are a few programs that offer cleaning pipelines. Also most of these tools perform cleaning of Sanger sequences, which is not NGS. A few tools perform several cleaning aspects such as adapters removal, contaminants/vector screening, poly-A/T and quality trimming, in single cleaning pipeline for next-generation sequencing.

Lucy (Chou & Holmes, 2001) is a cleaning software that offers quality, poly A/T and vector trimming. It operates both Sanger and Next-Generation sequences. Its quality trimming algorithm tries to find the minimum-sized window of acceptable quality within a read and if success it uses found window as a seed in order to further extend it from both ends. Similarly with its poly A/T trimming algorithm. The vector trimming algorithm offered by this tool requires a special vector splice site file in order to find vector coordinates. Major disadvantages with Lucy is that it does not provide adapter trimming and unnable to work with SFF and FastQ formats.

AlienTrimmer is another cleaning software which was recently announced (Criscuolo & Brisse, 2013). It provides adaper and vector trimming along with contaminants screening. It also accepts FastQ formats. The algoritm utilizes hash tables to be able to find these artifacts within a read. However, AlienTrimmer still does not accept SFF files (which means the user has to convert his SFF-formatted files into FastQ with loosing valuable flowgram information) and does not offer quality and poly A/T trimming.

For these reasons there is a need for a next-generation sequence cleaning tool that combines all cleaning tasks into one single pipeline. We present SeqyClean – a bioinformatics software tool for next-generation sequence cleaning. The first purpose of SeqyClean is to incorporate various cleaning methods, such as adapter, contaminant, poly A/T and quality trimming into a single software pipeline; the second is an ability to clean second-generation sequencing data such as Roche 454 and Illumina sequences. SeqyClean can recognize and remove both technological components (adapters, primers, mid tags) along with contaminants and vector. The LUCY© quality trimming algorithm was incorporated to SeqyClean in order to remove low-quality and poly-A/T erroneous data. Below we conducted a survey which describes several sequence cleaning applications (Table 1). None of these application, except SeqyClean, offers the full cleaning pipeline and the all aspects for complete next-generation sequence cleaning. In addition, SeqyClean has also advanced features, such as overlap and duplicates removal, which is helpful for genome assembly and mapping.

*It considers adapter as a contaminant

SeqyClean has been intensively used in our lab with constant improving and we believe that research community will benefit from using it.

Table 1: Comparison of existing cleaning tools

Application	SFF	FastQ	Contaminants	Vector	Adapter	Quality	Poly A/T	Overlap	Duplicates
SeqyClean	X	X	X	X	X	X	X	X	X
AlienTrimmer		X	X	X	X*				
Lucy				X		X	X		
SeqClean			X			X			
TrimEST							X		
VectorStrip				X					
VectorScreen				X					
Figaro				X					

3 Materials and Methods

3.1 Downloading and running seqyclean

3.1.1 How to download

Follow the link to download: https://bitbucket.org/izhbannikov/seqyclean/get/stable.zip. Save the file under some name you wish.

Then compile:

```
cd path_to_SeqyClean_directory
make clean
make
```

3.1.2 Usage

Roche 454

```
./seqyclean [options] -454 input_filename -o output_prefix
```

Main arguments

`-454 input_filename`	The filename of library to be cleaned. Can be in SFF or FASTQ formats. "-454" tells the program to clean Roche 454 reads
`-o output_prefix`	The files produced will start with the output_prefix followed by some formatted ending (see section "Output files: naming convention")

Roche 454 options

Options are shown in Table 2

Illumina paired- and single-end sequences

Paired-end sequences

```
./seqyclean [options] -1 input_filename_R1 -2 input_filename_R2 -o output_prefix
```

The descriptions of main arguments for paired-end libraries are shown in Table 3.

Single-end sequences

```
./seqyclean [options] -U input_filename -o output_prefix
```

The descriptions of main arguments for single-end libraries are shown in Table 4.

The descriptions of `[options]` for Illumina libraries are shown in Table 5.

Help

For help please use: `seqyclean -?  or --help`

Quick examples of usage

Example for 454 reads: `./seqyclean -v test_data/vectors.fasta -qual 30 25 -454 in.sff -o cleaned_data/Small454Test_cleaned`

Example for Illumina reads: `./seqyclean -v test_data/vectors.fasta -c test_data/contaminants.fasta -qual -1 test_data/R1.fastq.gz -2 test_data/R2.fastq.gz -o cleaned_data/cleaned`

Table 2: Roche 454 options

-v vector_file	This option is used for vector trimming. If you choose this option, the program assumes the reference genome provided in vector_file. This file must be given in FASTA format. Note: vector reference genome(s) must be provided! Example: ./seqyclean -v vectors.fa -454 in.sff -o Test
-c file_of_contaminants	This option is used for contaminants screening. If you choose this option, the program assumes the reference genome provided in file_of_contaminants. This file must be given in FASTA format. When SeqyClean recognizes contaminants in the sequence, the whole sequence gets discarded. Note: contaminant reference genome(s) must be provided! Example: ./seqyclean -c contaminants.fa -454 in.sff -o Test
-m file_of_RL_MIDS	This option works in 454 mode only. Use this option to provide your own RLMIDs. SeqyClean will use them and will not use those provided by default. Example: ./seqyclean -m file_of_custom_RL_MIDS -454 in.sff -o Test
-k k_mer_size	Use this option in order to specify a size of k-mer. Default size is 15 bases. Example: ./seqyclean -k 18 -454 in.sff -o Test
-kc k_mer_size	Special k-mer size for contaminant screening. Use this option only if you want to have different k-mer sizes for contaminant dictionary. Sometimes this option is useful because it prevents false detection of contaminants when program discards too many reads. Example: ./seqyclean -kc 25 -454 in.sff -o Test
-f overlap	For Roche 454 only. This option is intended to impose an overlap between two consecutive kmers. By default it is set to 1 bp. Refer to Fig. 1 Example: ./seqyclean -f 10 -454 in.sff -o Test
-t number_of_threads	Specifies a number of threads in order to take advantage from using a multicore system. Example: ./seqyclean -t 16 -454 in.sff -o Test
-qual max_avg_error max_error_at_ends	LUCY parameters for quality trimming. if "-qual" is set that means you have to provide max_avg_error and max_error_at_ends. Otherwise default values [20 20] will be used. Examples: ./seqyclean -qual -454 in.sff -o Test ./seqyclean -qual 25 30 -454 in.sff -o Test
--qual_only	Use --qual_only parameter if you want to do only quality trimming. Example: ./seqyclean --qual_only -qual -454 in.sff -o Test
--fastq	If input is given in SFF format, by default the output will be also in SFF format. Use this option if you want to have FASTQ format on the output in addition to SFF. Example: ./seqyclean --fastq -454 in.sff -o Test
--keep_fastq	Use this option only if you want to keep original FASTQ file from your input SFF Example: ./seqyclean --keep_fastq -454 in.sff -o Test
-minimum_read_length value	Use this option in order to define the minimum number of base pairs when read is still considered as acceptable. If after the cleaning process the read has a length which is less than minimum_read_length parameter, such the read will be discarded. By default, the minimum_read_length is set to 50 base pairs. Example: ./seqyclean -minimum_read_length 100 -454 in.sff -o Test
-polyat [cdna] [cerr] [crng]	This option provides trimming of poly A/T tails from nucleotide sequences. cdna – tail length (10 by default); cerr – maximum number of errors per tail (3 by default); crng – range to search poly A/T tails (50 by default) Examples: ./seqyclean -polyat -454 in.sff -o Test ./seqyclean -polyat 12 5 67 poly_test.fastq.gz -o Test_polyAT

Table 3: Paired-end main arguments

`-1 input_filename_R1`
 	
`-2 input_filename_R2`
	The filenames of the library to be cleaned. Must be in FASTQ formats only (the program also accepts zipped (.gz) FASTQ files).

`-o output_prefix`	The files produced will start with the output_prefix followed by some formatted ending (see section "Output files: naming convention")

Table 4: Single-end main arguments

`-U input_filename`	The filenames of the library to be cleaned. Can be in FASTQ formats only (the program also accepts .gz files).
`-o output_prefix`	The files produced will start with the output_prefix followed by some formatted ending (see section "Output files: naming convention")

3.2 Output files: naming conventions

Depending on the given parameters and the cleaning strategy, the name of output file can be different and has the formats described below.

3.2.1 Roche 454

After processing Roche 454 reads, SeqyClean outputs a cleaned file by default in Standard Flowgam Format (SFF) and (if option `--fastq` was chosen) in FASTQ format. Also two report files: Prefix_SummaryStatistics.txt (which contains information about how many reads were processed, trimmed, discarded and some other information) and Prefix_Report.csv file which holds the detailed statistics for every read.

- Output_prefix.sff , .fastq (optionally)
- Output_prefix_Report.tsv
- Prefix_SummaryStatistics.txt
- Prefix_SummaryStatistics.tsv

3.2.2 Illumina

After processing Illumina reads, SeqyClean generates two (shuffled file and file with single-end reads) or three (PE1 and PE2 files that contain paired-end reads and one file with single-end reads) output files in FASTQ format. Optionally, if the *–overlap* flag was used, SeqyClean generates a file that contains single-end reads combined from paired-end sequences by overlap.

- Output_prefix_PE1.fastq
- Output_prefix_PE2.fastq
- Output_prefix_shuffled.fastq (if `--shuffle` flag was set)
- Output_prefix_SE.fastq
- Output_prefix_SEOLP.fastq (if `--overlap` flag was set)
- Output_prefix_PE1_Report.tsv
- Output_prefix_PE2_Report.tsv
- Prefix_SummaryStatistics.txt
- Prefix_SummaryStatistics.tsv

3.2.3 Supported RLMIDs

The set of supported Roche 454 RL MIDs is shown in Table 6.

3.3 Removal of non-interested sequences

The general workflow diagram is shown in 1 and described below. The workflow consists of several atomic steps: (1) Input data pre-processing; (2) Trimming poly A/T tails; (3) Vector and contaminants trimming; (4) Adapter trimming; (5) Quality trimming; (6) Extension by overlap; (7) PCR

duplicates removal; (8) Establishing clip points; (9) Generating output files and summary statistics. Stages 2, 3, 4, 5, 6, 7 are optional depending on chosen cleaning strategy.

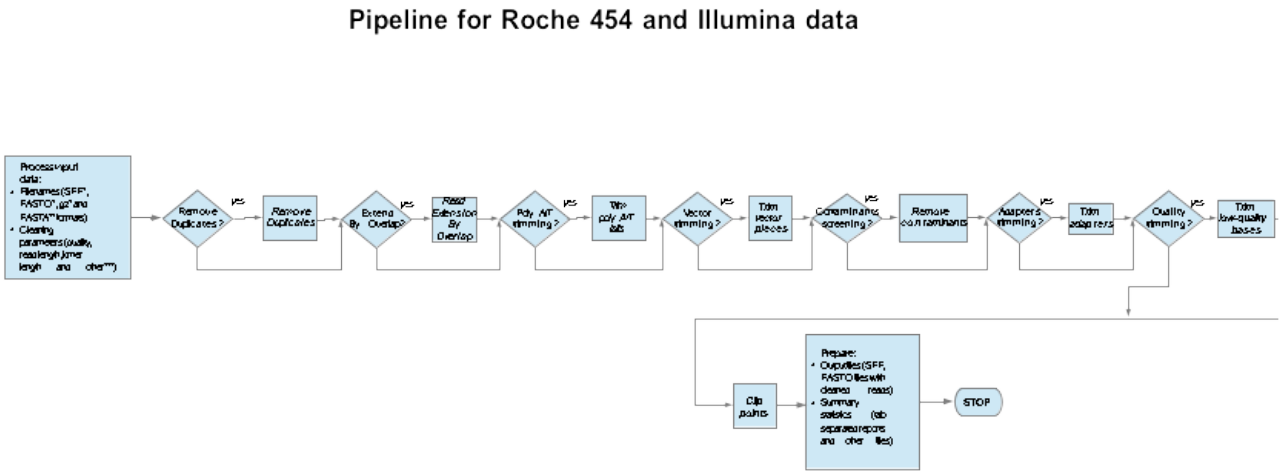

Figure 1: The workflow diagram for SeqyClean.

3.3.1 Input data pre-processing step

The first processing of the input data is performed on the preprocessing stage. Preprocessing stage analyzes user's inputs and depending on chosen strategy, calls appropriate cleaning methods. SeqyClean accepts FastQ, SFF and FastA-formatted files the following types:

- Raw DNA library files (format can be SFF or FastQ) that contain sequenced data. Use *-454*

in order to tell the program to enter to the 454 mode for cleaning Roche 454 or Ion Torrent libraries; or *-1*, *-2* or *-U* to clean Illumina paired- or single-end libraries.

- References genomes in FastaA-formatted file. Reference genomes are vector of contaminants genomes. SeqyClean employs two separate FastA file: one for contaminants and another one for vector. Such FastaA files can contain single or multiple reference genomes. Use flag *-c* to tell the program to use contaminant screening and flag *-v* for vector trimming.

3.3.2 Removing PCR duplicates

Removing duplicates can be considered to mitigate the effects of PCR amplification bias introduced during the library construction. We define duplicates as reads with identical sequence. Removal of duplicates is dependent on an error rate in the library. Removing duplicates also reduces the pool of reads before mapping and thereby reduces alignment time. The algorithm considers bases within the position from 10 to 25 in the read and does not look at the whole sequence (Figure 2).

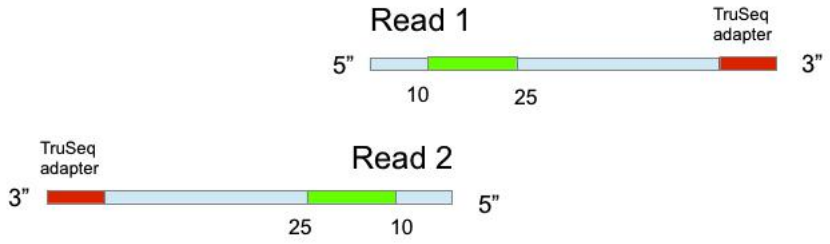

Figure 2: Paired-end Illumina sequences. Green regions are to be checked for duplicates.

3.3.3 Contaminants and vector detection and trimming

In majority of cases DNA library contains foreign DNA at the same time. Contaminants in DNA library are one of the main obstacles for sequence assembly and mapping. If vector is used (Figure 3.3.3), it is highly possible that reads contain remnants of vector genome, which is also a contaminant. Contaminant genomes are reference genomes and read as a query sequence. It is up to user

to decide which types of contaminants are likely to be in a library and therefore provide reference genomes. Algorithm uses exact kmer matching and works as follows. (1) Reference genomes along with reverse complements are sampled into consecutive kmers of n bases (15 bases by default) long and then each kmer is stored in a hash table named "dictionary". Each dictionary item stores a key, which is a kmer string and a value that represents an array of 2-tuples: pos, id where pos represents the position within a reference of each occurrence of each kmer and id represents a record name in provided FastA file. Searching for a contaminant within a query sequence is performed by dictionary lookup of each kmer from the query sequence. Refer to Figure 3.3.3. There are two cases: (1) Vector trimming when vector sequence occupies only a part of a read. When vector coordinates are approximately found, then exact coordinates are obtained by applying a pairwise alignment of a vector site in the query sequence and corresponding area in the reference vector genome. The read is discarded if a vector is more than 80% of the read. (2) In case of contaminant screening the whole read considered as contaminant and discarded as soon as it meets several consecutive successful lookups.

3.3.4 Adapter trimming

Roche 454 and Illumina Next-Generation Sequencing technologies employ short synthetic nucleotides called adapters or primers to immobilize DNA pieces on a surface. Depending on technology, adapters may be different types (Roche 454 and Illumina adapters) and attached to the ends of the DNA particle. Adapters can potentially have a significant impact on genome assembly as it is shown in 3.3.3. In this case due to adapter on the end, assembler does not extend the growing contig but stops and initializes another one. Two different algorithms are used for adapter trimming: (1) Banded Pairwise alignment (Fickett, 1984) is used for Roche 454 adapters. This is a variation of Smith-Waterman algorithm with reduced omputational complexity in order to increase the alignment performance. (2) SSAHA kmer matching algorithm (Ning, Cox, & Mullikin, 2001),

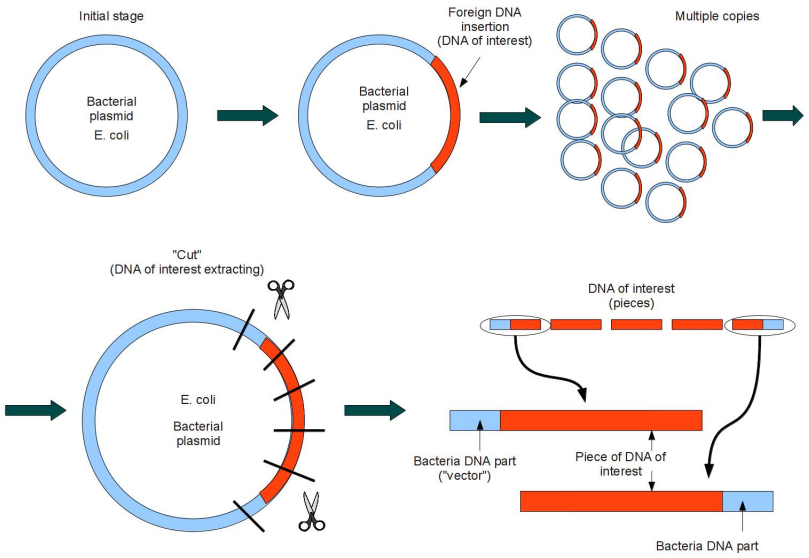

Figure 3: Bacteria vector cloning technology.

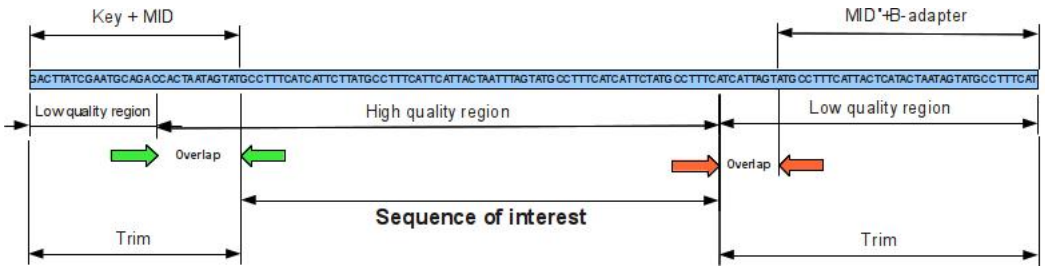

Figure 4: Artifacts within a read (Roche 454).

which is used to detect Illumina TruSeq adapters.

3.3.5 Quality trimming

Errors in reads are unavoidable and assumption that reads are error-free is often misleading. In fact, majority of reads in DNA library have base-call errors and it is important to obtain as less errors as possible. The quality of nucleotide bases decreases toward the end of the read 3.3.5. To obtain nucleotide sequence and corresponding quality scores, a program known as a base caller

converts the raw output from DNA sequencer into nucleotide bases, each of which is typically assigned a quality score that estimates the probability that the base has been sequenced correctly. Quality scores are needed to trim the low quality ends of a read (Roche 454) and for determination a consensus sequence. SeqyClean utilizes quality trimming approach described in (Chou & Holmes, 2001) and employs the scoring model with an integer scale derived from the error probability p (where p is the probability that a particular base has been read correctly) via the formula $Q = -10log_{10}p$, a scoring scheme named Phred scheme. Nucleotides and corresponding Phred quality scores with other metadata are stored in ascii – delimited text FastQ format which Illumina employs (files are usually compressed with gzip) and binary Sff format which is used by Roche 454. Having the raw data, algorithm discards low-quality regions that may exist within a read by computing average quality of a window. Windows with quality lower than pre-defined threshold are trimmed. The minimum read length to keep is by default 50 nucleotide bases. In order to be able to process short Illumina reads with length about 30 bases, the size of the window can be adjusted.

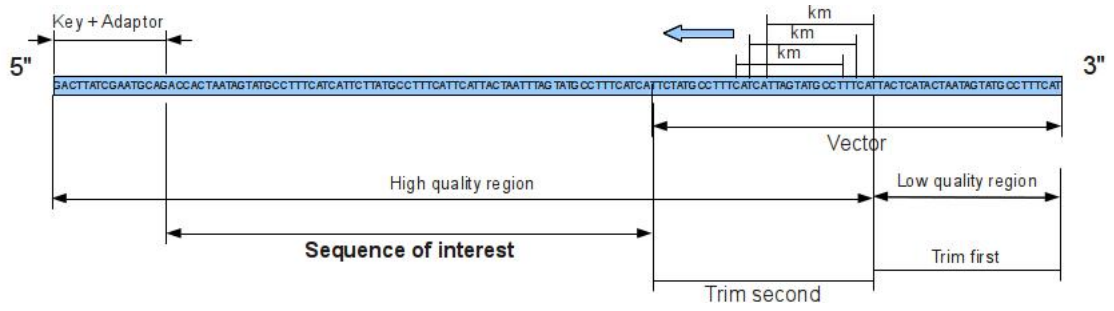

Figure 5: Quality trimming.

3.3.6 Extending Illimina reads by overlaps between paired-end sequences

Illumina paired-end sequencing technology produces a very large amount of short reads. In spite of high quality and significant amount (up to 100 million) paired-end reads, their length (100-250 nucleotides) still makes genome assembly a difficult and challenging task. There is a potential

benefit to take advantage of paired-end paradigm to extend the reads that may potentially overlap into one single-end sequence with length up to almost twice as the length of a pair. This is can be considered as a library pre-processing before conducting the downstream analysis (Magoč & Salzberg, 2011). We included such length extension algorithm to SeqyClean. The core of this algorithm is to take advantage of paired-end sequences by merging pairs together into long, single-end sequence. The algorithm works as follows:

1. Align the pair of reads so that they overlap completely by the full length of the shorter read.
2. Repeat while the overlap is longer than minoverlap:
    a Calculate the overlap length.
    b Calculate the score for the overlap as the ratio between the number of mismatches and the overlap length.
    c If the score of the overlap is equal or bigger that the pre-defined mismatch threshold:
        - Calculate the new quality values of all bases within the overlap by taking maximum quality scores if bases are equal. Otherwise assign the new quality value as a subtraction of quality scores.
        - Save the overlap and progress toward the next pair.
    d Otherwise slide the reads apart by one base, reducing the overlap by one.
3. If the score is less than the mismatch threshold, report that no good overlap was found.

This promises the potential improvement of detection of short remnants of TruSeq adapters (9-15 bases) that may still be left within reads. We applied algorithm that increases the sensitivity of TruSeq (Illumina) adapter trimming. Our assumption is that if an adapter is found on one of the end of the read, than the corresponding adapter should be found at the end of another read. In this

case, by applying overlap alignment is possible to have a situation where adapters will be on the both end of the consensus sequence and it is a trivial problem to trim them. The algorithm is shown in Figure 6

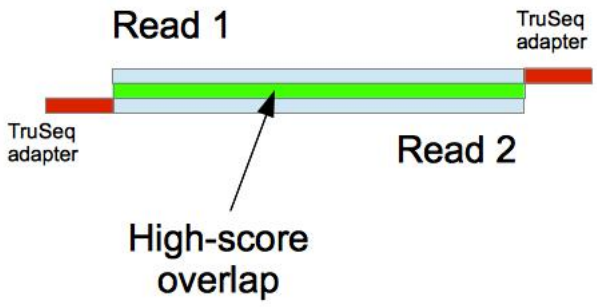

Figure 6: Adapter trimming by overlap. Green region has high overlap scores and this makes us to assume that the short remnats of TruSeq adapters exist on the both of the ends.

3.3.7 Poly A/T tails trimming

Expressed sequence tags (ESTs), are single-stranded DNA sequencing reads made from complementary DNA (cDNA) clone libraries derived from a known source (cells, tissues, or organs from different organisms). Sequencing a large number of these clones from such a library allows one to sample the set of expressed genes, or transcripts, in that particular tissue and experimental state, thus providing a snapshot of the active genes under those defined conditions (Nagaraj, Gasser, & Ranganathan, 2007).

The raw DNA sequences obtained from an EST library contains poly-A/T tags that have to be removed before conducting the downstream analysis (see Figure 7). SecyClean provides this service by searching for the first minimum occurrence of 10 bases or longer poly-T fragment within the first initial search range of 50 bases inside the good region, then attempts to extend from this initial poly-T seed toward the center of the sequence, allowing no more than maximum error (three by default) mismatches between every min span (10) consecutive T bases in the search region.

GenBank: DY008075.1

GenBank FASTA

IDENTIFIERS

dbEST Id: **35691756**
EST name: 19ACACYS_UP_022_A11_29OCT2004_095
GenBank Acc: DY008075
GenBank gi: 119423037

CLONE INFO
Clone Id: (5')
DNA type: cDNA

PRIMERS
PolyA Tail: Unknown

SEQUENCE

TGGTACGGTCAGATGCTTGCTAAAGGAGAAATAAATAGAGACATGGGTGATAGTATAAGC
GGAAAGGGAATGATTCAGGGTGTTTCTGCAGTGGGAGCGTTTTACCAACTGCTTAGTCAG
TCCAGCCTAAGTATATTGCATTCTGAAGAGAAGAAACCTGTGGCTCCGGTTGAATCATGT
CCTATTTTGAAAACACTCTACAAGATACTCATCACAAGAGAACAATCAACACAAGCGATT
CTGCAAGCATTAAGGGATGAAACACTGAATGACCCAAGAGACAGGATTGAGATTGCACAG
AGCCATGCATTCTACAGGCCTTCCCTTCTAGATCAGCCTTGATTAGTCTGTCATGGCTCA
TAATCCGAACTTCTAAGATCTTACTTGTGCAAACTGCAGATTCTGCTATGTTAAACATCA
TGTCTTAAAATTGATTGTTGTTCAGCCAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAT
GTC

Entry Created: Dec 15 2006
Last Updated: Dec 15 2006

Figure 7: Poly-A tails occurs within a sequenced tag.

This can be done quickly since it is a linear time algorithm. The same algorithm is employed to the poly-A tail trimming at the other end of the sequence.

3.3.8 Establishing clip points

Clip point is a position within a read where all bases are discarded after this position. SeqyClean establishes the most conservative clip points, in other words it chooses the right-most clip point from the left and left-most clip point from the right side of the read. This implies, for example, that if quality clip points are located inside of vector clip points, SeqyClean uses the first ones as the actual trim points of the read (Figure 8).

3.3.9 Output files and summary statistics

The output files are FastQ or SFF-formatted files containing either trimmed reads (FastQ-file) or reads with new established clip points (SFF-file). In addition, detailed statistics for each read and summary statistics are generated both in human- (TXT) and machine-readable (TSV) formats.

3.4 Quality and adapter trimming impacts on genome assembly and mapping

We evaluated SeqyClean from two sides: genome assembly and mapping. Our interest was (1) to explore impacts of quality and adapter trimming on genome assembly and (2) genome mapping. We also determined the optimal quality parameters for sequence cleaning. Finally we cleaned human genome data with found optimal quality parameters for SeqyClean.

We tested the impact of quality and adapter trimming on two NGS libraries: *Escherichia coli* which is Roche 454 pyrosequencing data, 621,578 reads, 327,471,374 bases, and *Saccharomyces cerevisiae* (*Yeast*) W303 which is Illumina MiSeq data, 2x 3,875,453 reads, 2x 972,738,703 bases. As a reference we used *E.coli* str. K-12 substr. MG1655 strain obtained from GenBank [Accession number is NC_000913.2]. In case with *Yeast* we used *Saccharomyces cerevisiae* W303-K60001 strain (Ralser et al., 2012). We cleaned both *E.coli* and *Yeast* libraries with SeqyClean by applying quality trimming parameter pairs maximum_average_error, maximum_error_at_ends (Chou

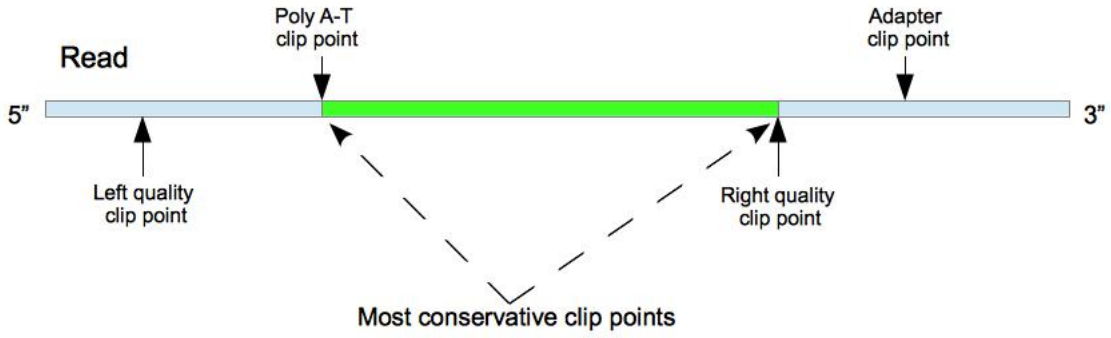

Figure 8: Establishing the most-conservative clip points. Green region is to be kept.

& Holmes, 2001) presented in Phred scores. We varied both of them from 1 to 40 (inclusive), where 1 denotes 20.57% of chance that the base was read correctly and 40 implies 99.99% that a base was read correctly. Values of these parameters are presented in Table 7. We call the pair maximum_average_error, maximum_error_at_ends as quality parameters. Parameters for adapter trimming were default.

To explore impacts of quality and adapter trimming on assembly we performed assemblies both cleaned *E.coli* and *Yeast* libraries with GS De Novo Assembler (Newbler) 2.6, a Roche 454 assembler (454 Life Science, n.d.). We conducted five assemblies for each of the quality parameter pair in order to exclude impacts of assembler's variations.

Roche 454 sequencing offers its own clip points stored alongside with the nucleotide data and other information in SFF-formatted file [CITE]. Such "native" clip points can be potentially used as an alternative to SeqyClean's clip ponts. In order to evaluate the impact of Roche's native clip points we performed assembly of *E.coli* data trimmed with Roche clip points only.

We also conducted the assembly of the raw, non-cleaned sequences for both *E.coli* and *Yeast* libraries.

In order to explore effects of quality and adapter trimming on the assembly we estimated the following assembly parameters: (1) $N50$; (2) total number of contigs; (3) average read length; (4) read coverage; (5) insertions/deletions (indels); (6) total length of the assembly.

We also evaluated impacts of quality and adapter trimming on mapping.

In this case we used the same *E.coli* and *Yeast* libraries which were previously cleaned with SeqyClean. For those libraries we estimated the number of single-nucleotide polymorphisms and insertions/deletions and then compared them with raw data and data which was trimmed with Roche clip points. For mapping we used *bowtie2* (Langmead & Salzberg, 2012) aligner along with *samtools* (Li et al., 2009) in order to extract SNPs and indels.

We estimate the coverage with the following equation:

$$\text{Coverage} = \frac{\text{Total number of bases}}{\text{Total number of reads}}$$

This equation provides a rough coverage, but still can be used for evaluation.

4 Results

4.0.1 Application to Escherichia coli

The assembly results are shown in Figure 9. As quality parameters { *maximum_average_error, maxim_error_at_ends* } increase (Figure 9) and while they are lower 12, the value of $N50$ of assembly composed from cleaned data is slightly higher than $N50$ of assembly made on Roche-trimmed data. The maximum $N50$ is achieved when parameters are within 17...22 Phred. These values are considered to be optimal. Further, the value of $N50$ drops toward 0. This happens because, while quality parameter values increase and trimming becomes more stringent, SeqyClean trims and discards larger number of reads, and the rest of the library data does not supply enough coverage for sufficient assembly.

As it is seen from the Figure 10, the program does not trim significantly when parameters are within the range of 1...14. When parameters are within 15...25, the average read length decreases slightly and significantly goes down to 0 (when Phred is 40) after that. In this case all reads in the library are discarded. From Figure 10 we see that until 11, the coverage does not change significantly and does not have large difference between coverage of raw data. After 25 and more, the coverage significantly decreases to zero.

We estimated the number of indels per 100kb using QUAST analytic tool (Gurevich, Saveliev, Vyahhi, & Tesler, 2013). Our assumption is that removal low-quality bases and regions will decrease false-positive indels and SNPs. It is clear to see that the number of indels is lower in comparison to the raw data. Obviously, it is not possible to trust the data with base-call error. On the other side, the coverage should be good enough in order to obtain proper SNP-call. As it is seen from the Figure 11, the number of indels is slightly lower or the same in comparison to Roche-trimmed

data.

Total number of contigs is shown in Figure 12. Because of complexity of DNA assembly problem it is not easy to obtain original sequence on the output. However, it is possible to obtain small number of contigs that still represent the whole genome.

The important assembly metrics is the total length of the assembly which shows how the assembly output (which is a set of contigs) close to the reference sequence (if one exists). In Figure 13 the total assembly lengths are shown for different types of assembly. Green line shows the reference

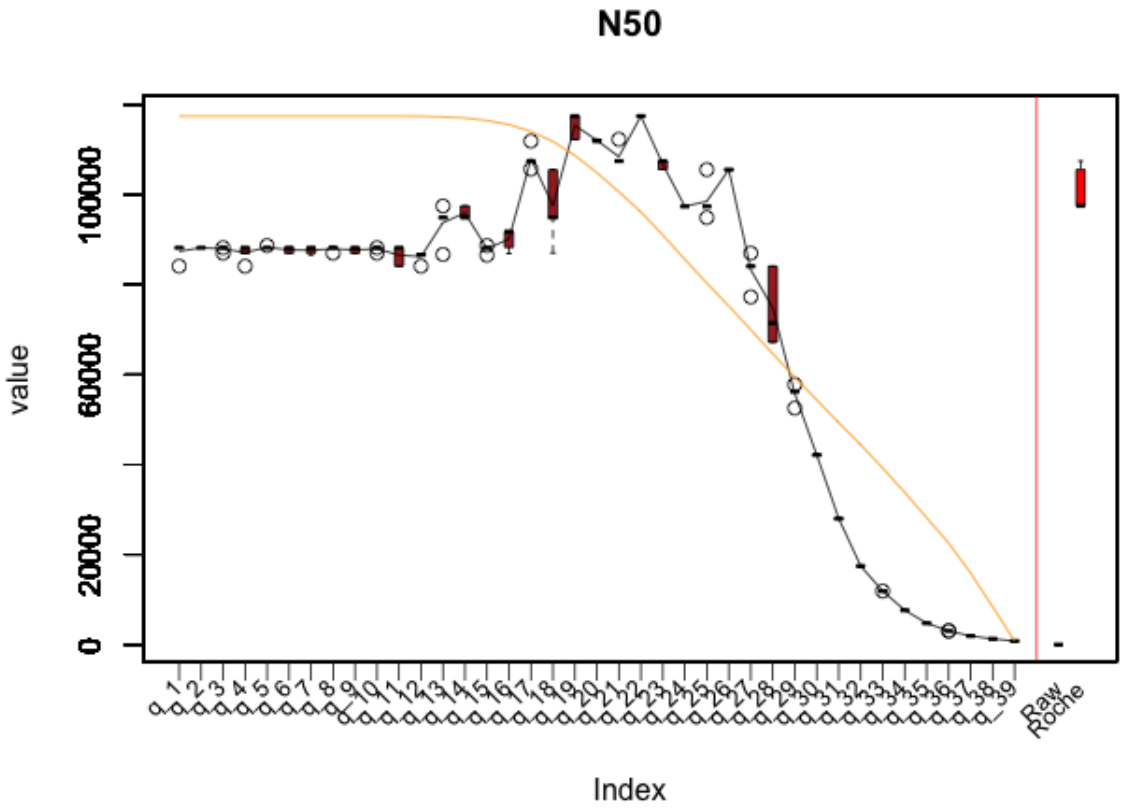

Figure 9: The left side of the plot shows N50 of assemblies made of library cleaned with SeqyClean with different quality thresholds. The right side shows N50 of assemblies built from raw data (blue boxplot) and Roche data (red boxplot).

length which serves as an indicator to compare different assemblies. Total lengths of assemblies made of libraries cleaned with SeqyClean are shown on the left of the Figure 13. Those ones made out of raw and Roche data are shown on the right side of this figure.

We guess that Roche quality clip points are made base on Phred scores of 20. Due to low coverage when Phred scores > 30, the number of indels raises quite fast toward the Phred crores equal to 40.

After mapping (we employed bowtie2 and samtools), results are shown in Figure 14. The number of insertions/deletions is significantly larger than number of SNPs. This was expected

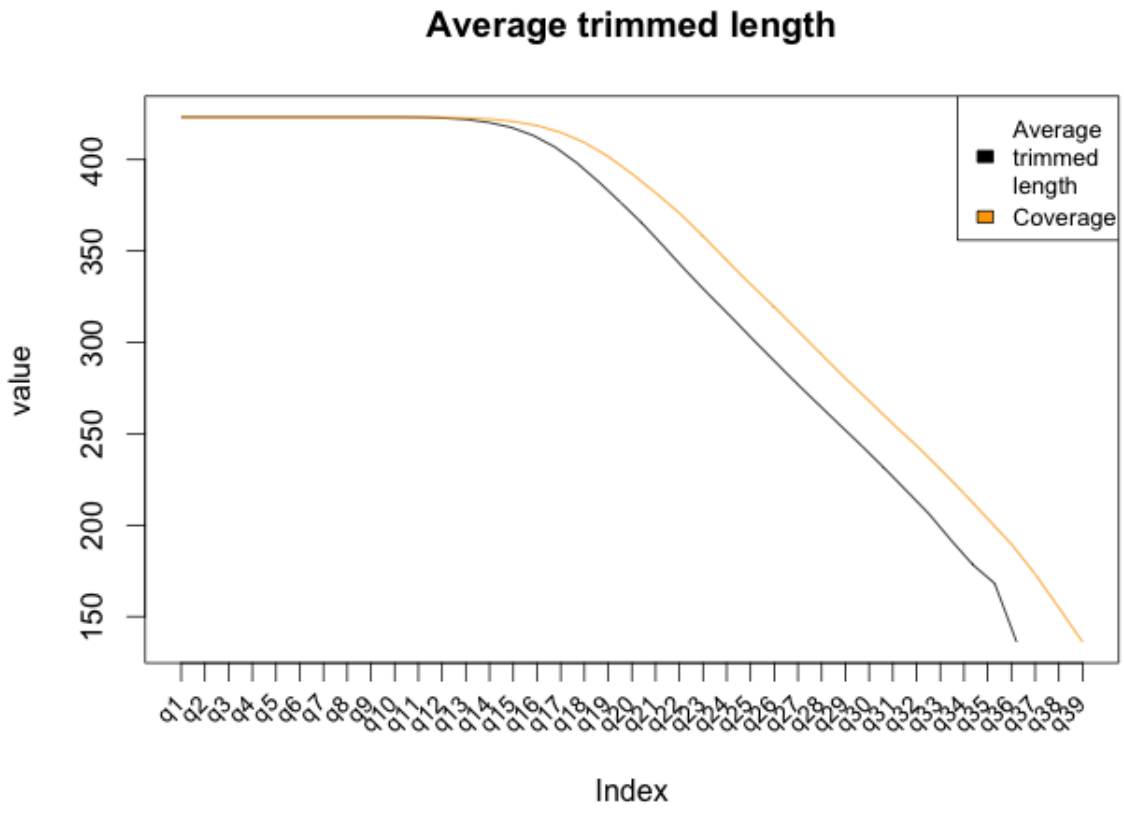

Figure 10: Average trimmed length - the average read length within the library after trimming (black line). The coverage is shown with orange line.

because Roche 454 sequences contain large number of homopolymers. The number of indels and SNPs decreases significantly after 25 because the depth is too low to call SNP/indels. However, for those parameter values less than 25 we assume we have proper variance call because the coverage (depth) is high enough. We can not say the same for those values larger than 35 because of low coverage.

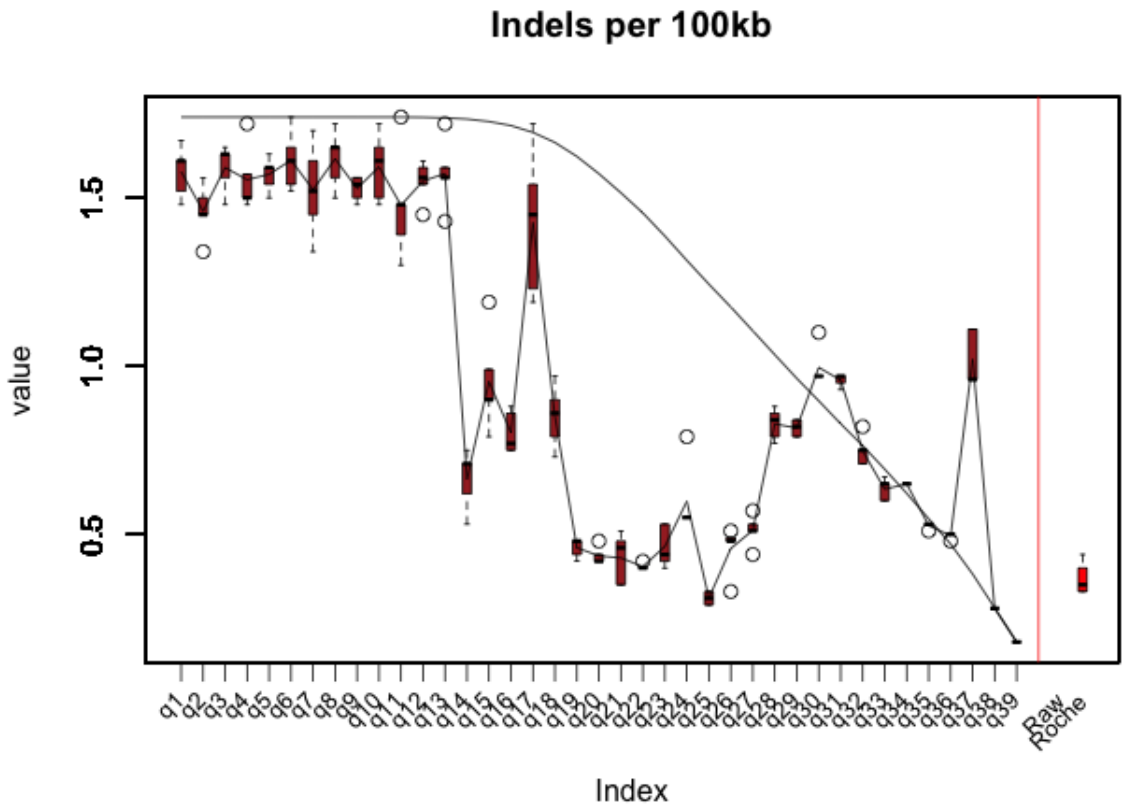

Figure 11: Indels per 100 bp (brown boxplots) and coverage (orange line). The left side of the plot shows assembly results made of data cleaned with SeqyClean. The right side of the plot shows assembly results made of raw (uncleaned data, blue boxplot) and data cleaned with Roche's clip poitns (red boxplot)

4.0.2 Application to Yeast

The same N50 pattern we see for the *Yeast* Illumina data (Figure 15), except the optimal Phred scores are shifted left [not clear to me].

In Figure 16, trimmed lengths for PE1 and PE2 does not changes until Phred scores of 18. This happens because of high quality of Illumina data. After Phred scores of 18 it decreases and then when Phred scores achieve 36 it drops to zero.

From the Figure 16, expected coverage started to decrease when Phred scores achieve the value

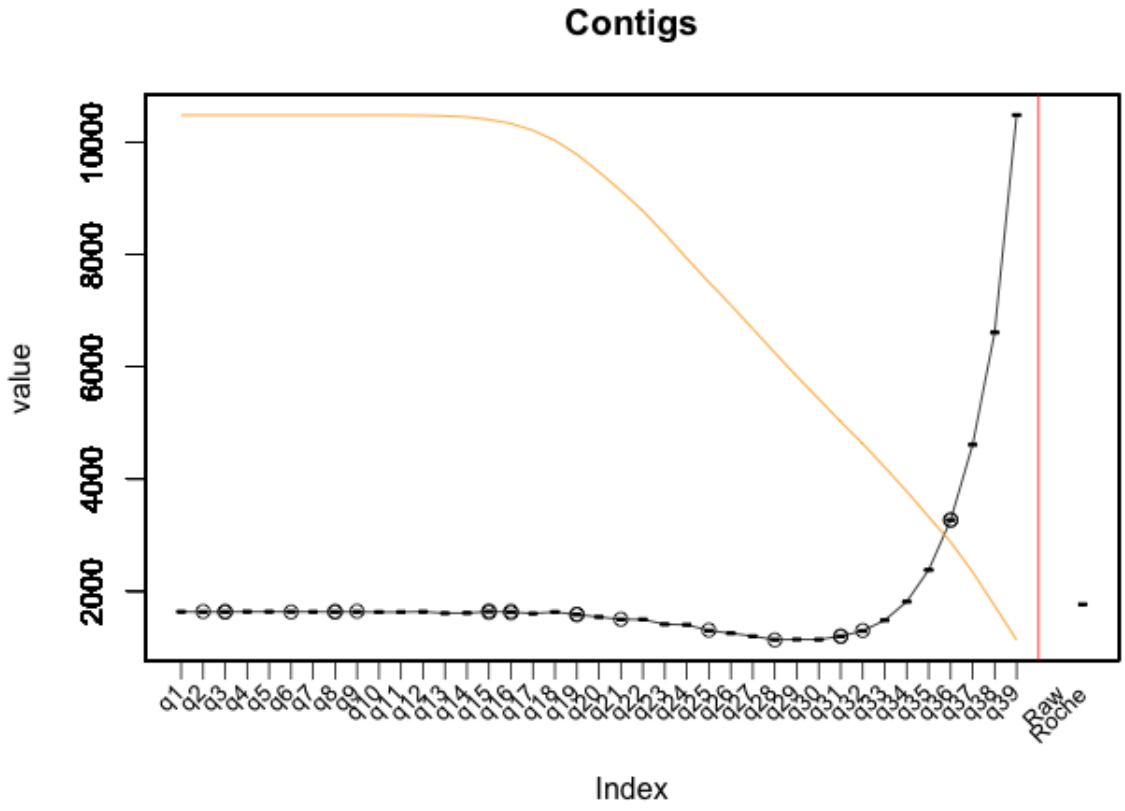

Figure 12: Total number of contigs. Raw and Roche result are shown with blue and red boxplots on the right and SeqyClean results are on the left. Orange line represents the coverage.

of 16 and does not change until this value. This is because of high quality if Illumina data.

Number of indels for *Yeast* data (Figure 17), are lower in comparison to Raw data, except of short region when some indels are slightly higher (perhaps the number of reads should be larger than 600kb). There is the opposite (to *E.coli*) pattern for number of indels when Phred scores $>$ 30. (needs explanations)

Total length (or total number of bases) in an assembly is the sum of all of its contig lengths (Figure 18). In the beginnig, when Phred scores are still less 26, the total number of bases in the assembly remains roughly constant (but concordant to the reference). Then it raises and becomes

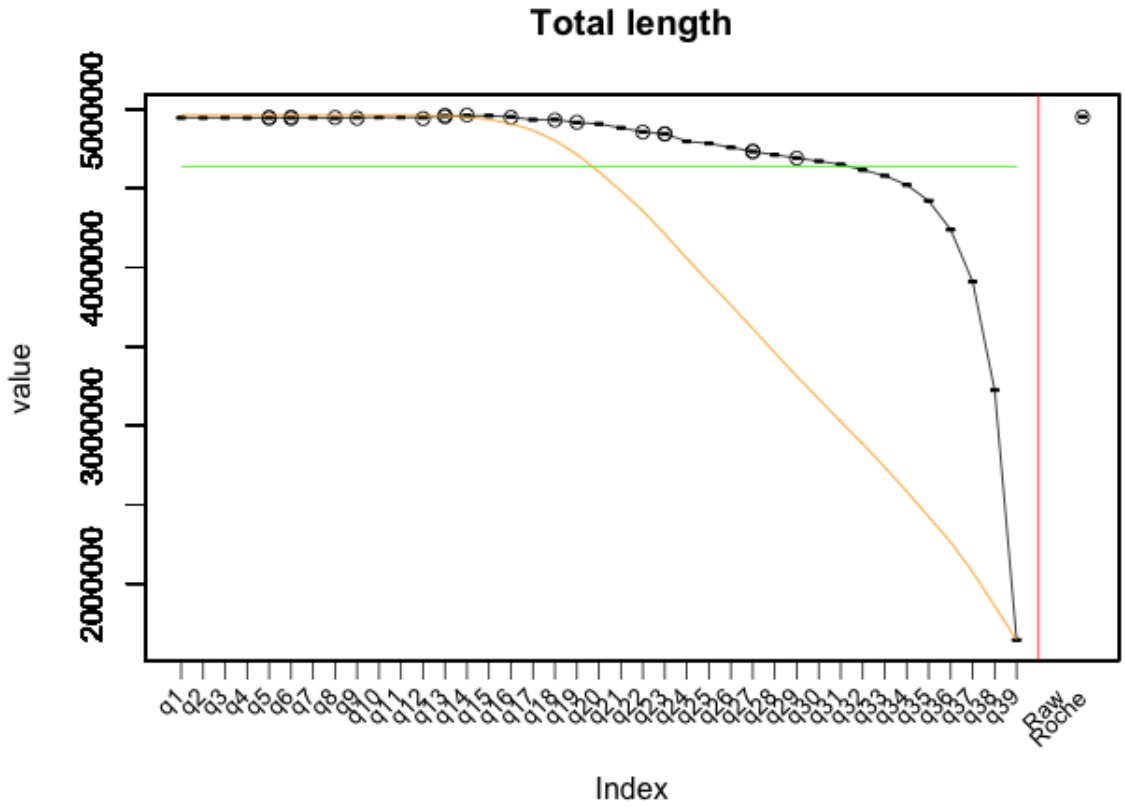

Figure 13: Total number of contigs. Raw and Roche result are shown with blue and red boxplots on the right and SeqyClean results are on the left. Orange line represents the coverage.

(almost) equal to the reference and when Phred scores continue to raise, it significantly drops down to zero (no coverage). Within the range of Phred scores of 18...25 we have the optimum.

We also conducted mapping. To be written

In Figure XX there is the opposite situation: the number of variance (SNPs) is higher than

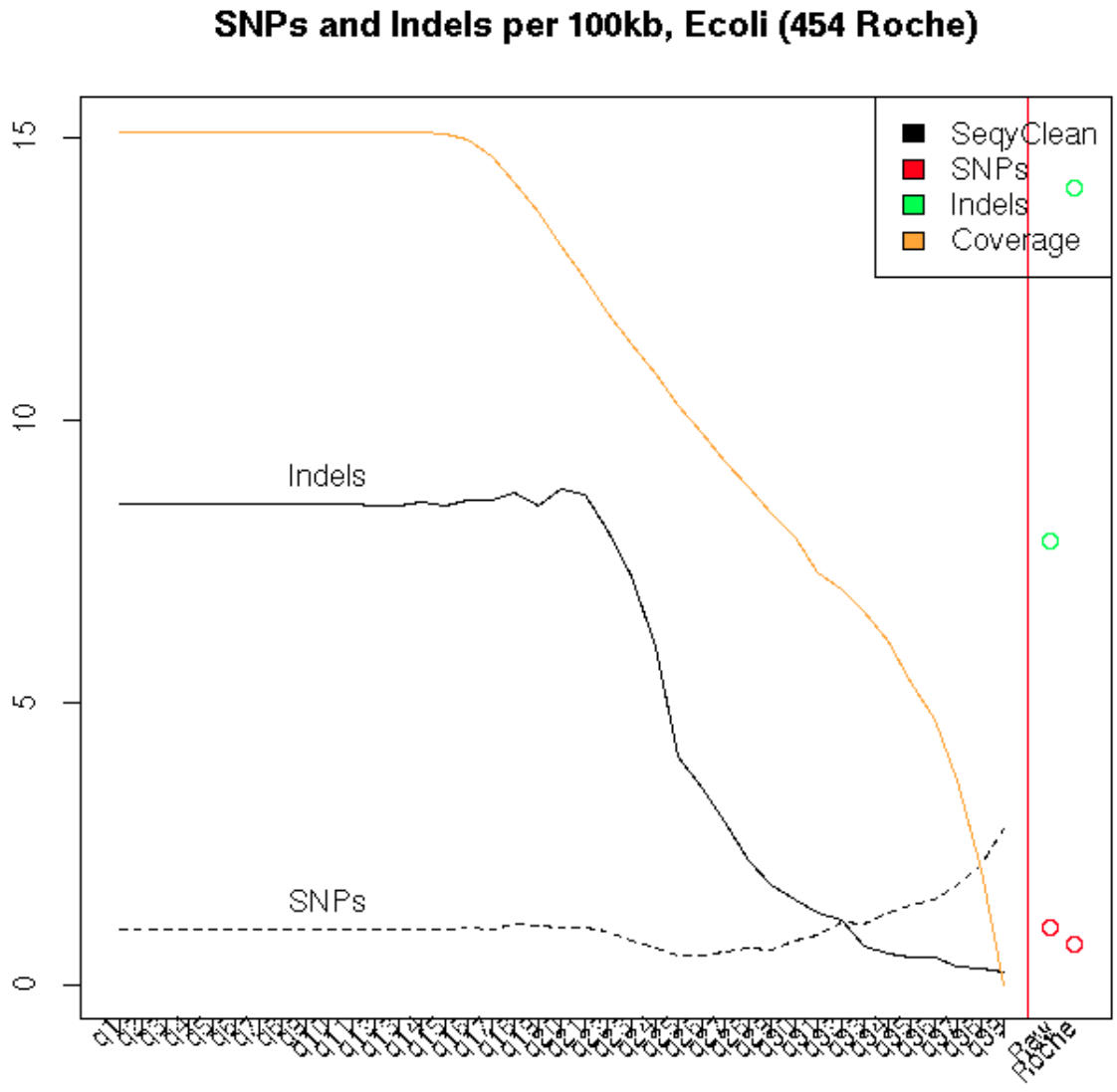

Figure 14: Mapping results: Single-Nucleotide Polymorphisms (SNPs) and indertions/deletions (Indels) for libraries cleaned with SeqyClean (black), untrimmed (raw) and trimmed with Roche clip points. Raw and Roche result are shown on the right. Orange line represents the coverage.

number of indels. This was also expected, due to lower abundance of erroneous homopolymers in comparison to Roche 454. And these results are concordant to E.*E.coli* and prove our hypothesis as well.

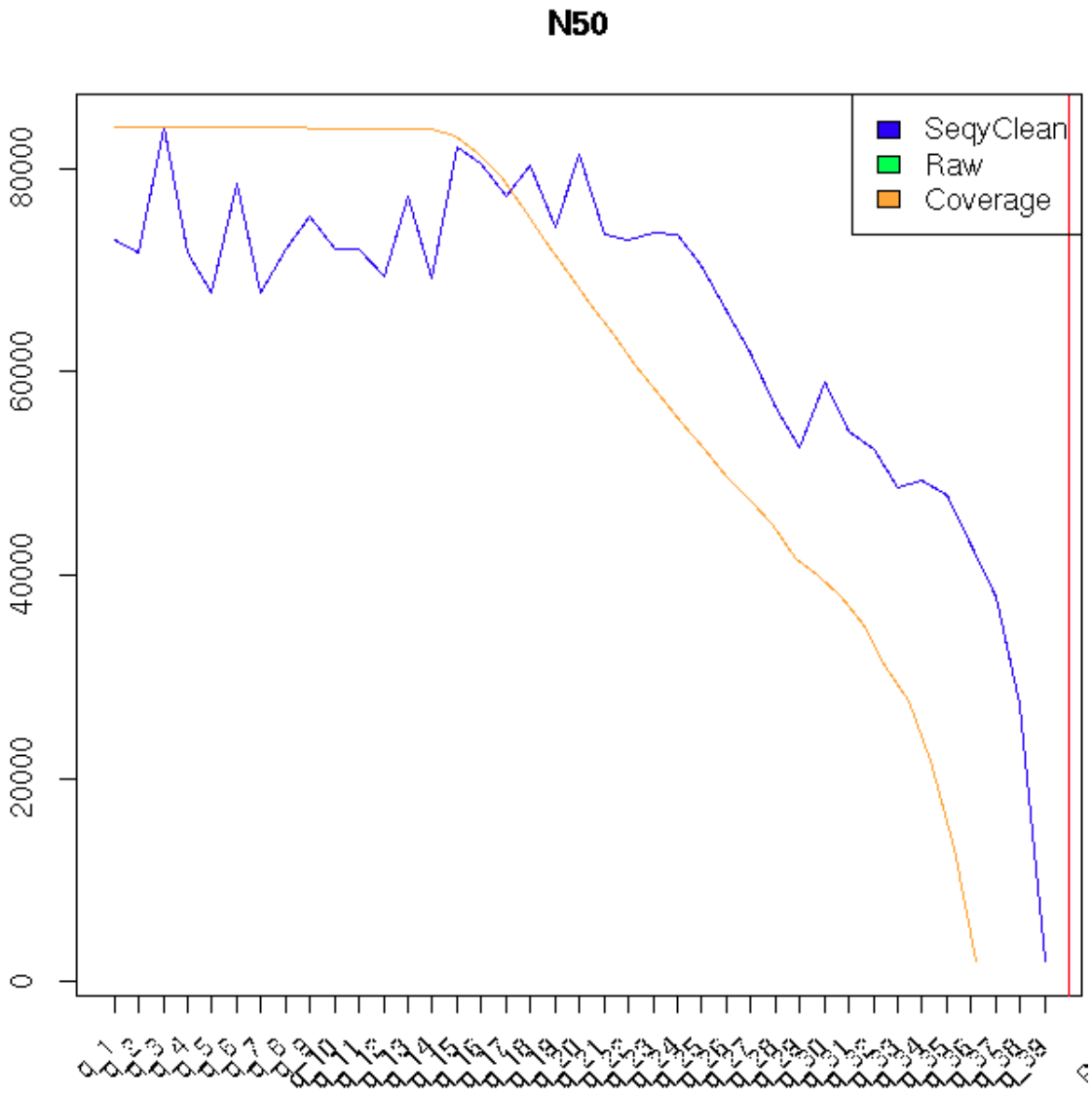

Figure 15: Indels per 100 bp (blue line) and coverage (black line). The left side of the plot shows assembly results made of data cleaned with SeqyClean. The right side of the plot shows assembly results made of raw (uncleaned data).

5 Discussion

We applied sequence cleaning with SeqyClean for two different sequence libraries: *E.coli* and *Yeast*. Libraries were the next-generation sequence libraries: Roche 454 (*E.coli*) and Illumina (*Yeast*). SeqyClean is a bioinformatic cleaning pipeline. The advantages of having such pipeline

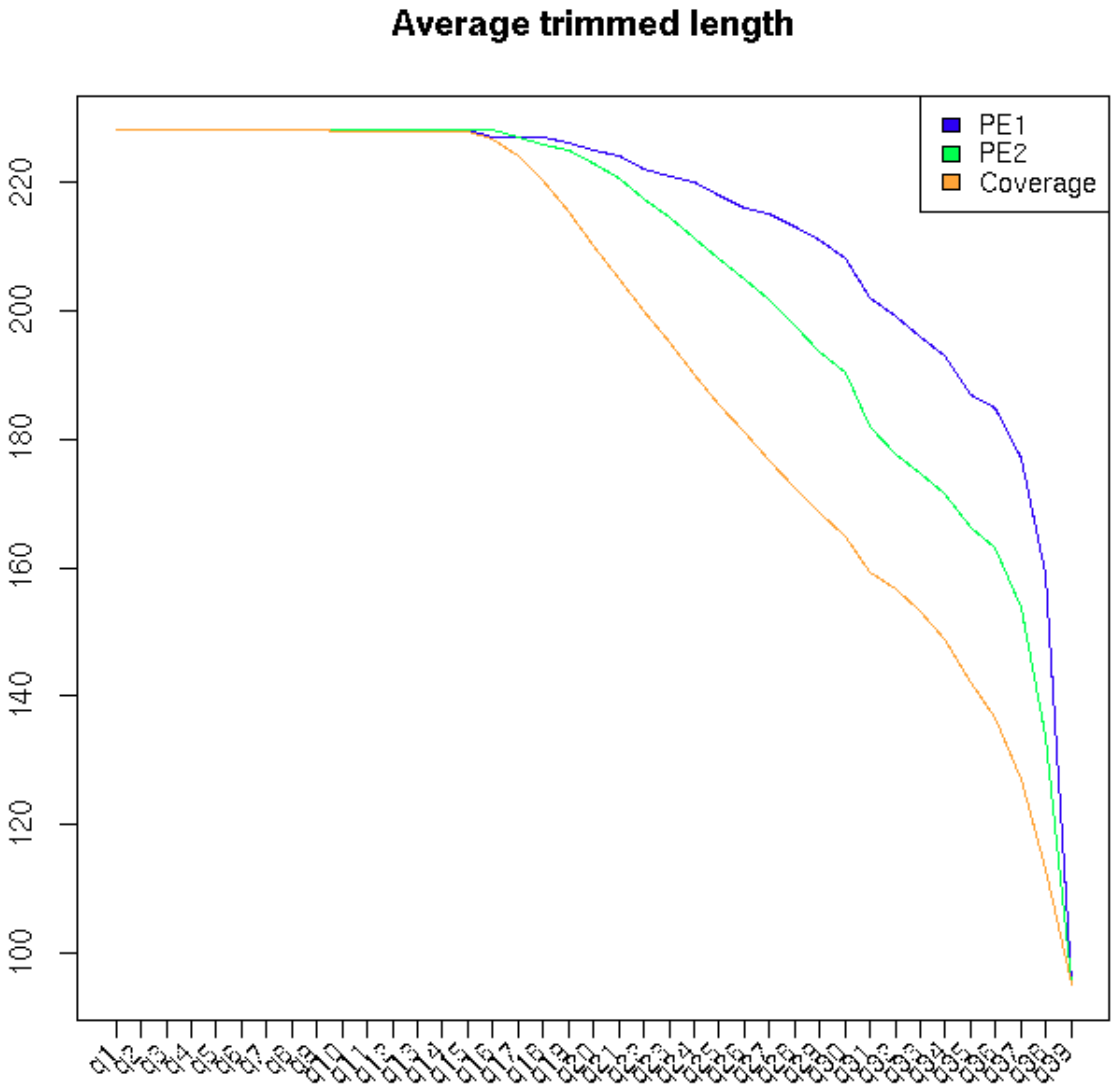

Figure 16: Average trimmed length - the average read lengths after cleaning with SeqyClean (blue line for PE1 and green for PE2 line). The coverage is shown with orange line.

is that it offers almost complete adapter, vector and low-quality trimming, contaminants screening and trimming poly A/T tails. SeqyClean has an ability to output cleaned data in SFF files, preserving the flowgram and other information about reads needed for Roche 454 assembler. Majority of today's cleaning software tools offer only single cleaning task: for example adapter or quality or vector trimming trimming only. In opposite to, SeqyClean offers a full cleaning pipeline which

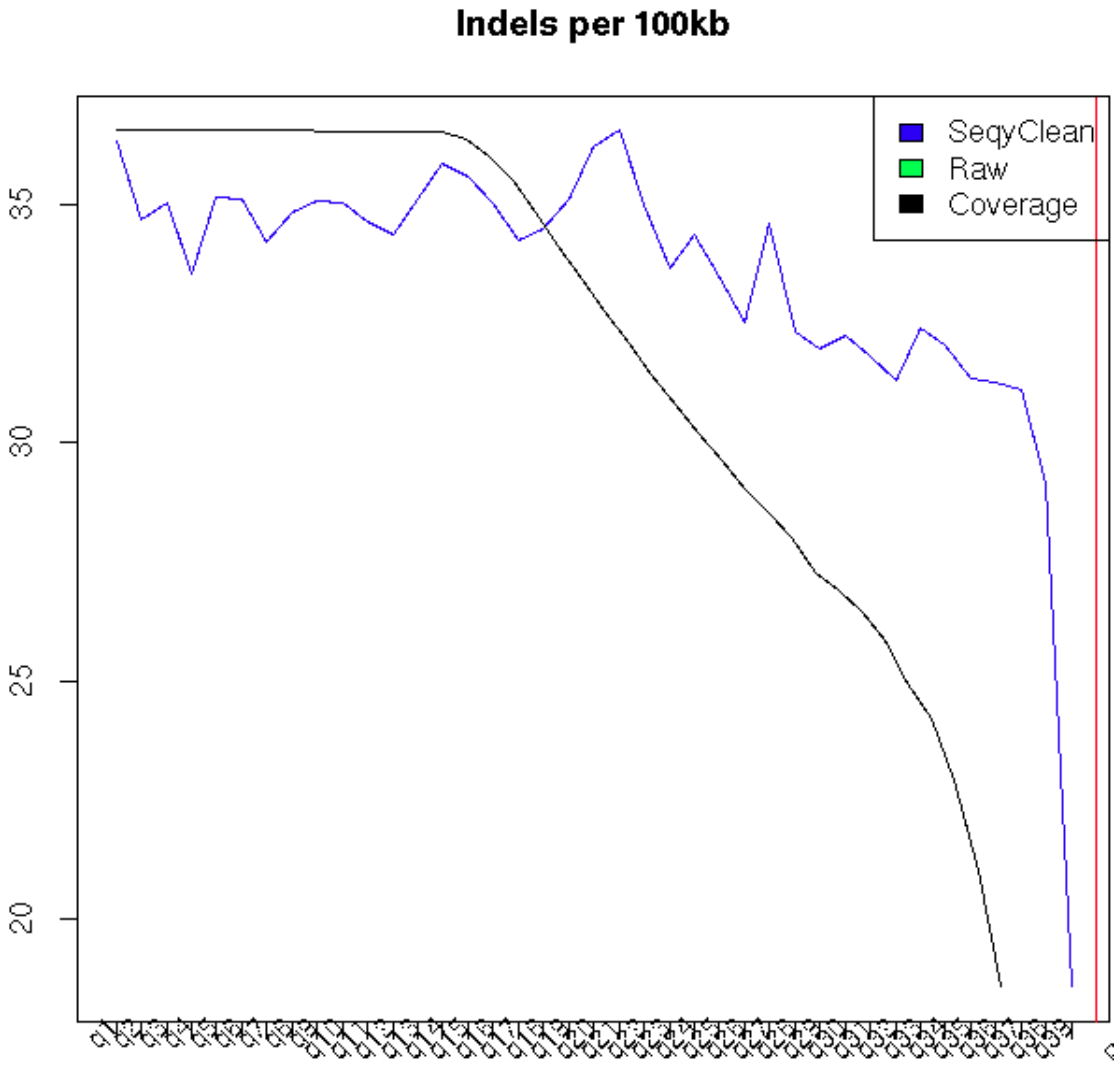

Figure 17: Average trimmed length - the average read lengths after cleaning with SeqyClean (blue line for PE1 and green for PE2 line). The coverage is shown with orange line.

incorporates several cleaning tasks. We varied cleaning parameter values needed for the quality trimming from 1 to 40. It turned out that for those values within a range of 1 to 12, quality trimming does not increase significantly and only adapter trimming made significant impact on assembly and mapping in this case. Perhaps it depends on library. We have also tried perform the assembly on

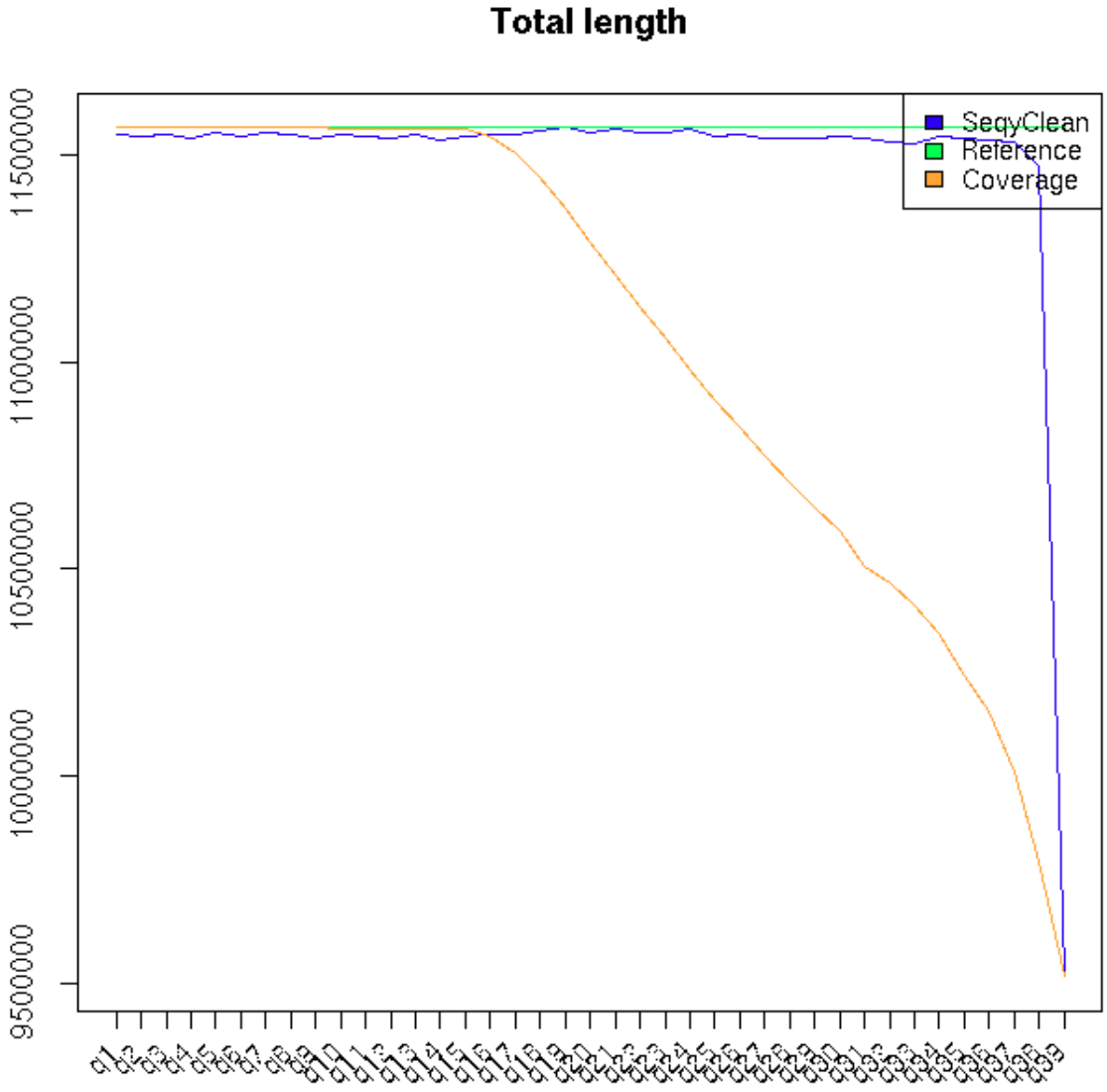

Figure 18: Total length of the assembly (blue line) build of data cleaned with SeqyClean. The coverage is shown with orange line and the reference length is shown with the green line.

raw, untrimmed data. Unfortunately, after two weeks of work, the assembler (we used Roche 454 Newbler assembler for this task) was not able to finish while assemblies of cleaning data were able to finish less than a day. This happens because by trimming adapters and low-quality nucleotides, SeqyClean reduces complexity of data for further genome assembly.. In case of mapping we obtained significant reduction of single-nucleotide polymorphisms and indels. By trimming adapters and low-quality bases, SequClean reduces the potential chance of false-positive SNPs and indels. Such false-positive variations can be mistakenly treated as true-positive and to avoid this situation is extremely important when the genome of interest was not sequenced before (which means it was not annotated) and related genome was used as a reference.

There are many ways of taking advantage of parallel processing. SeqyClean combines itself several sequential cleaning methods. In the same time, DNA libraries, especially paired-end ones have sizes up to several decades of gigabytes of sequences meaning that it is not efficient to process reads sequentially. There is a plenty of room to parallelize the cleaning pipeline by executing several cleaning steps in the same time as well as parallelization of individual read. We have improved sensitivity of adapter detection within Illumina paired-end reads. SeqyClean detects remnants of Illumina TruSeq adapters less then 15 bases.

Compressed output FastQ files. NGS libraries are quite large: typical Illumina library occupies several gigabytes of data. Modern hard discs can have very large storage space but it still costly to maintain giga- and terabytes of data. Compression of cleaned libraries offers significant reduction of storage space and cost of maintenance.

6 Acknowledgements

This work was supported by IBEST COBRE, grant NIH/NCRR P20RR16448 and the University Research Office at the University of Idaho.

References

454 Life Science, A. R. C. (n.d.). *Point-and-click tools for assembly, mapping and amplicon variant analysis.* Retrieved from http://454.com/products/analysis-software/index.asp

Chou, H., & Holmes, M. H. (2001). DNA sequence quality trimming and vector removal. *Bioinformatics/computer Applications in The Biosciences*, *17*, 1093–1104. doi: 10.1093/bioinformatics/17.12.1093

Criscuolo, A., & Brisse, S. (2013). Alientrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*(0), -. Retrieved from http://www.sciencedirect.com/science/article/pii/S0888754313001481 doi: http://dx.doi.org/10.1016/j.ygeno.2013.07.011

Fickett, J. W. (1984). Fast optimal alignment. *Nucleic Acids Research*, *12*, 175–179. doi: 10.1093/nar/12.1Part1.175

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). Quast: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075. Retrieved from http://bioinformatics.oxfordjournals.org/content/29/8/1072.abstract doi: 10.1093/bioinformatics/btt086

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat Meth*, *9*(4). Retrieved from http://dx.doi.org/10.1038/nmeth.1923 doi: 10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Subgroup, . G. P. D. P. (2009). The sequence alignment/map (sam) format and samtools. *Bioinformatics*. Retrieved from http://bioinformatics.oxfordjournals.org/content/early/2009/06/08/bioinformatics.btp352.abstract doi: 10.1093/bioinformatics/btp352

Magoč, T., & Salzberg, S. L. (2011). Flash: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957-2963. Retrieved from http://bioinformatics.oxfordjournals.org/content/27/21/2957.abstract doi: 10.1093/bioinformatics/btr507

Nagaraj, S. H., Gasser, R. B., & Ranganathan, S. (2007). A hitchhiker's guide to expressed sequence tag (est) analysis. *Briefings in Bioinformatics*, *8*(1), 6-21. Retrieved from http://bib.oxfordjournals.org/content/8/1/6.abstract doi: 10.1093/bib/bbl015

Ning, Z., Cox, A. J., & Mullikin, J. C. (2001). SSAHA: A Fast Search Method for Large DNA Databases. *Genome Research*, *11*, 1725–1729. doi: 10.1101/gr.194201

Ralser, M., Kuhl, H., Ralser, M., Werber, M., Lehrach, H., Breitenbach, M., & Timmermann, B. (2012). The saccharomyces cerevisiae w303-k6001 cross-platform genome sequence: insights into ancestry and physiology of a laboratory mutt. *Open Biology*, *2*(8). Retrieved from http://rsob.royalsocietypublishing.org/content/2/8/120093.abstract doi: 10.1098/rsob.120093

Scheetz, T. E., Trivedi, N., Roberts, C. A., Kucaba, T., Berger, B., Robinson, N. L., ... Casavant, T. L. (2003). ESTprep: Preprocessing CDNA Sequence Reads. *Bioinformatics/computer Applications in The Biosciences*, *19*, 1318–1324. doi: 10.1093/bioinformatics/btg159

White, J. R., Roberts, M., Yorke, J. A., & Pop, M. (2008). Figaro: a novel statistical method for vector sequence removal. *Bioinformatics/computer Applications in The Biosciences*, *24*, 462–467. doi: 10.1093/bioinformatics/btm632

Table 5: Illumina options

--shuffle	With this option SeqyClean will combine output paired-end libraries into one single file named <output_prefix>_shuffled.fastq. However, SeqyClean still does keep single-end reads (reads without corresponding pairs) in <output_prefix>_SE.fastq file. Example: ./seqyclean --shuffle -1 R1.fastq -2 R2.fastq -o Test
-v vector_file	This option is used for vector trimming. If you choose this option, the program assumes the reference genome provided in vector_file . This file must be given in FASTA format. Example: ./seqyclean -v vectors.fa -1 R1.fastq -2 R2.fastq -o Test
-c file_of_contaminants	This option is used for contaminants screening. If you choose this option, the program assumes the reference genome provided in file_of_contaminants . This file must be given in FASTA format. When SeqyClean recognizes contaminants in the sequence, the whole sequence gets discarded. Example: ./seqyclean -c contaminants.fa -1 R1.fastq -2 R2.fastq -o Test
-k k_mer_size	Use this option in order to specify a size of k-mer. Default size is 15 bases. In Illumina mode this option defines a size of kmer that will be used as a dictionary word size. Example: ./seqyclean -k 14 -1 R1.fastq.gz -2 R2.fastq.gz -o Test
-kc k_mer_size	Special k-mer size for contaminant screening. Use this option only if you want to have different k-mer sizes for contaminant dictionary. Sometimes this option is useful because it prevents false detection of contaminants when program discards too many reads. Example: ./seqyclean -kc 31 -1 R1.fastq.gz -2 R2.fastq.gz -o Test
-qual max_avg_error max_error_at_ends	LUCY parameters for quality trimming. if "-qual" is set that means you have to provide max_avg_error and max_error_at_ends. Otherwise default values [20 20] will be used. Examples: ./seqyclean -qual -1 R1.fastq.gz -2 -1 R2.fastq.gz -o Test ./seqyclean -qual 30 25 -1 R1.fastq.gz -2 R2.fastq.gz -o Test
--qual_only	Use --qual_only parameter if you want to do only quality trimming. Example: ./seqyclean --qual_only -qual -1 R2.fastq.gz -2 R2.fastq.gz -o Test
-minimum_read_length value	Use this option in order to define the minimum number of base pairs when read is still considered as acceptable. If after cleaning process the read has length which is less than minimum_read_length parameter such read will be discarded. By default, the minimum_read_length is set to 50 base pairs. Note: in this case no adapter/vector/contaminants cleaning is performed. Example: ./seqyclean -minimum_read_length 100 -1 R1.fastq.gz -2 R2.fastq.gz -o Test
-polyat [cdna] [cerr] [crng]	This option provides trimming of poly A/T tails from nucleotide sequences. Parameters: cdna – tail length (10 by default) cerr – maximum number of errors per tail (3 by default) crng – range to search poly A/T tails (50 by default) Examples: ./seqyclean -polyat -454 in.sff -o Test ./seqyclean -polyat 15 4 55 poly_test.fastq.gz -o Test_polyAT
-overlap <value>	This option provides overlap of read 1 and read 2. The consensus single-end sequence is stored under the name ¡output_prefix¿_SEOLP.fastq. Parameter: <value¿ – the minumum length of overlap in bases. By default it is set to 10 bases.
-ot value	This option sets the similarity threshold for adaprer trimming by overlap. By default its value is set to 0.9.
-adapter_length value	This option sets the maximum adapter length for adaprer trimming by overlap. By default its value is set to 40 bases.
--ow	This option allows SeqyClean overwrite output files.

Table 6: Supported RLMIDs by default

#	Left MID	Right MID	#	Left MID	Right MID
RL1	ACACGACGACT	AGTCGTGGTGT	RL19	ATAGTATACGT	ACGTATAGTAT
RL2	ACACGTAGTAT	ATACTAGGTGT	RL20	CAGTACGTACT	AGTACGTGCTG
RL3	ACACTACTCGT	ACGAGTGGTGT	RL21	CGACGACGCGT	ACGCGTGGTCG
RL4	ACGACACGTAT	ATACGTGGCGT	RL22	CGACGAGTACT	AGTACTGGTCG
RL5	ACGAGTAGACT	AGTCTACGCGT	RL23	CGATACTACGT	ACGTAGTGTCG
RL6	ACGCGTCTAGT	ACTAGAGGCGT	RL24	CGTACGTCGAT	ATCGACGGACG
RL7	ACGTACACACT	AGTGTGTGCGT	RL25	CTACTCGTAGT	ACTACGGGTAG
RL8	ACGTACTGTGT	ACACAGTGCGT	RL26	GTACAGTACGT	ACGTACGGTAC
RL9	ACGTAGATCGT	ACGATCTGCGT	RL27	GTCGTACGTAT	ATACGTAGGAC
RL10	ACTACGTCTCT	AGAGACGGAGT	RL28	GTGTACGACGT	ACGTCGTGCAC
RL11	ACTATACGAGT	ACTCGTAGAGT	RL29	ACACAGTGAGT	ACTCACGGTGT
RL12	ACTCGCGTCGT	ACGACGGGAGT	RL30	ACACTCATACT	AGTATGGGTGT
RL13	AGACTCGACGT	ACGTCGGGTCT	RL31	ACAGACAGCGT	ACGCTGTGTGT
RL14	AGTACGAGAGT	ACTCTCGGACT	RL32	ACAGACTATAT	ATATAGTGTGT
RL15	AGTACTACTAT	ATAGTAGGACT	RL33	ACAGAGACTCT	AGAGTCTGTGT
RL16	AGTAGACGTCT	AGACGTCGACT	RL34	ACAGCTCGTGT	ACACGAGGTGT
RL17	AGTCGTACACT	AGTGTAGGACT	RL35	ACAGTGTCGAT	ATCGACAGTGT
RL18	AGTGTAGTAGT	ACTACTAGACT	RL36	ACGAGCGCGCT	AGCGCGCGCGT

Table 7: Quality parameter values along with corresponding real error

Phred value for the pair {maximum_average_error, maxim_error_at_ends}	Corresponding error probability	Phred value for the pair {maximum_average_error, maxim_error_at_ends}	Corresponding error probability
1,1	0.2057	21,21	0.9921
2,2	0.3690	22,22	0.9937
3,3	0.4988	23,23	0.9950
4,4	0.6019	24,24	0.9960
5,5	0.6838	25,25	0.9968
6,6	0.7488	26,26	0.9975
7,7	0.8005	27,27	0.9980
8,8	0.8415	28,28	0.9984
9,9	0.8741	29,29	0.9987
10,10	0.9000	30,30	0.9990
11,11	0.9206	31,31	0.9992
12,12	0.9369	32,32	0.9994
13,13	0.9499	33,33	0.9995
14,14	0.9602	34,34	0.9996
15,15	0.9684	35,35	0.9997
16,16	0.9749	36,36	0.9997
17,17	0.9800	37,37	0.9998
18,18	0.9842	38,38	0.9998
19,19	0.9874	39,39	0.9999
20,20	0.9900	40,40	0.9999