Quality control

Sequencing technologies are not perfect and the quality control (QC) is an essential step to ensure that the data used for downstream analysis is not compromised of low-quality sequences, sequence artifacts, or sequence contamination that might lead to erroneous conclusions. The easiest way of QC is looking at summary statistics of the data. There are different programs that can produce those statistics. Web applications allow users to easily share and discuss the results with other people without transferring large data files. The following QC steps are implemented in and all graphics generated by PRINSEQ (http://prinseq.sourceforge.net).

Content:

Necessary resources

Hardware

Computer connected to the Internet

Software

Up-to-date Web browser (Firefox, Safari, Chrome, Internet Explorer, …)

Files

FASTA file with sequence data

QUAL file with quality scores (if available)

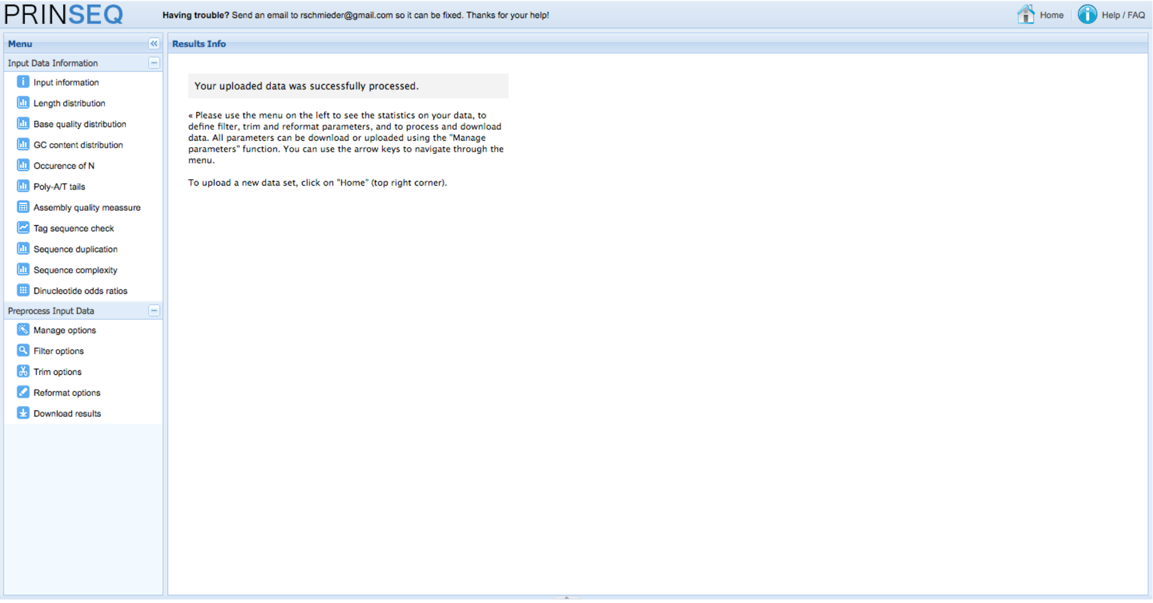
FASTQ file (as alternative format)

Uploading data to the PRINSEQ web version

To upload a new dataset in FASTA and QUAL format (or FASTQ format) to PRINSEQ, follow these steps:

1. Go to http://prinseq.sourceforge.net
2. Click on "Use PRINSEQ" in the top menu on the right (the latest PRINSEQ web version should load)
3. Click on "Upload new data"
4. Select your FASTA and QUAL files or your FASTQ file and click "Submit"

After the data is parsed and processed successfully, the user interface will show a menu on the left and a message in the main panel as shown below.



Having trouble? Send an email to rschmieder@gmail.com so it can be fixed. Thanks for your help!

Your uploaded data was successfully processed.

« Please use the menu on the left to see the statistics on your data, to define filter, trim and reformat parameters, and to process and download data. All parameters can be download or uploaded using the "Manage parameters" function. You can use the arrow keys to navigate through the menu.

To upload a new data set, click on "Home" (top right corner).

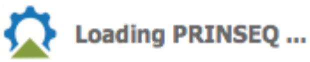
Notes

After clicking the submit button, a status bar (not progress bar) will be displayed until the file upload is completed. During the data processing, several progress bars will show the progress of the data parsing and statistics calculation steps.

Possible problems

1. The PRINSEQ web interface does not load / is not visible.

   You only see this and nothing else happens:

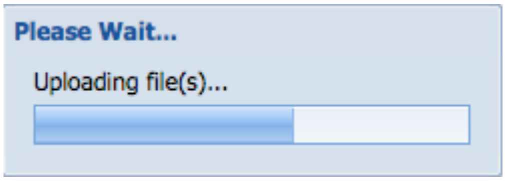
   

   Loading PRINSEQ ...

   Solution: Make sure that you have JavaScript activated in your browser, as this is required to load and use PRINSEQ's web interface.

2. The upload status bar does not disappear.

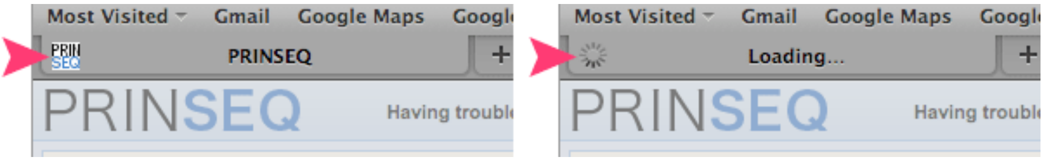
   After clicking on the submit button you see this and it does not disappear:

   Please Wait...

   Uploading file(s)...

   Solution: The first thing to check is if the file is still uploading. The easiest way to do this is by checking the loading icon in your browser.

   If you see the loading icon (right) instead of the PRINSEQ icon (left), your file is still uploading and you should give it more time. If you see the PRINSEQ icon instead of the loading icon, your file did not upload completely and this caused an error. If you have a slow connection to the Internet or try to upload large files, the connection to the web server can time out before the upload was completed. If you did not upload compressed files, try to compress your files with any of the supported compression algorithms (ZIP, GZIP, ...).

   In rare cases, the issue can also be caused by certain Firefox plugins or extensions. If possible, use an alternative browser to test if this was the case. If the browser caused the problem, updating Firefox and the plugins/extension to the latest version might solve the problem.

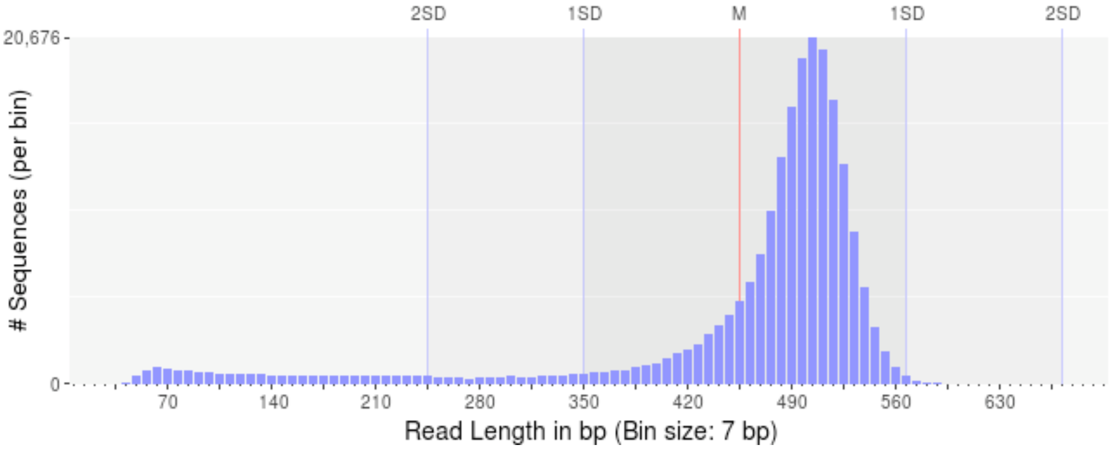
Number and length of sequences

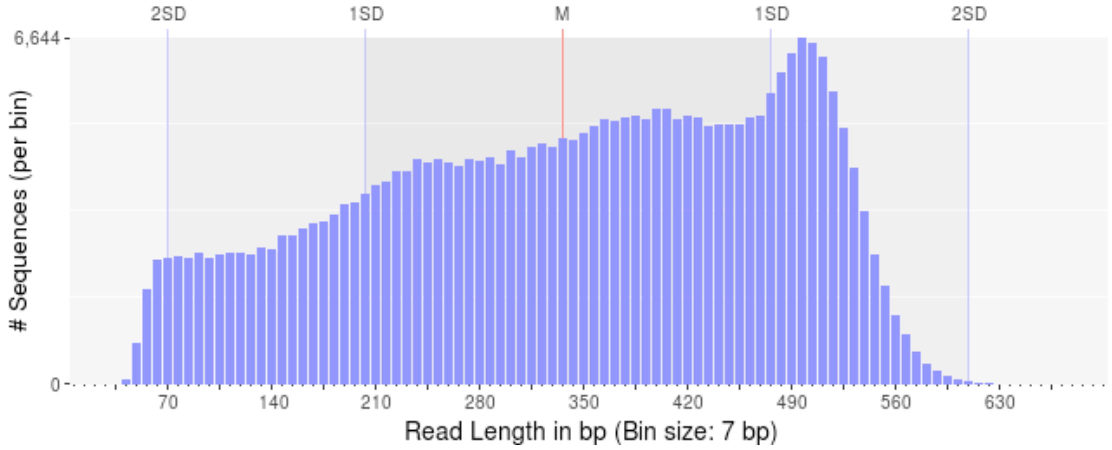
Check those numbers to make sure it matches approximately the manufacturer estimates. If your numbers are off too much, check the raw data and filter statistics in "454BaseCallerMetrics" and "454QualityFilterMetrics".

Length distribution

The length distribution can be used as quality measure for the sequencing run. You would expect a normal distribution for the best result. However, most sequencing results show a slowly increasing and then a steep falling distribution. The plots in PRINSEQ mark the mean length (M) and the length for one and two standard deviations (1SD and 2SD), which can help to decide where to set length thresholds for the data preprocessing. If any of the sequences is longer than 100 bp, the lengths will be binned in the plots generated by PRINSEQ. The number of sequences for each bin is then shown instead of the number of sequences for a single length (values might therefore be bigger than shown in the table for non-binned lengths).

The following two datasets have approximately the same number of sequences, however the length distributions look different.





Both distributions have the highest number of sequences around 500 bp, but for the first dataset the mean of the sequence lengths is higher and the standard deviation is lower. A certain number of shorter reads might be expected, but if the sample contained mainly longer fragments, this number should be low.

Assuming that both samples contained enough fragments of at least 500 bp and all fragments were sequenced with the same number of cycles (sequencing flows), we would expect that the majority of the sequences would have approximately the same length. The higher amount of shorter reads in the second dataset suggests that those reads might have been of lower quality and were trimmed during the signal processing. If the sample contained many short fragments, the shorter reads might be from those fragments and not of lower quality.

Minimum and maximum read length

Sequences in the SFF files can be as short as 40 bp (shorter sequences are filtered during signal processing). For multiplexed samples, the MID trimmed sequences can be as short at 28 bp (assuming a 12 bp MID tag). Such short sequences can cause problems during, for example, database searches to find similar sequences. Short sequences are more likely to match at a random position by chance than longer sequences and may therefore result in false positive functional or taxonomical assignments. Furthermore, short sequences are likely to be quality trimmed during the signal-processing step and of lower quality with possible sequencing errors.

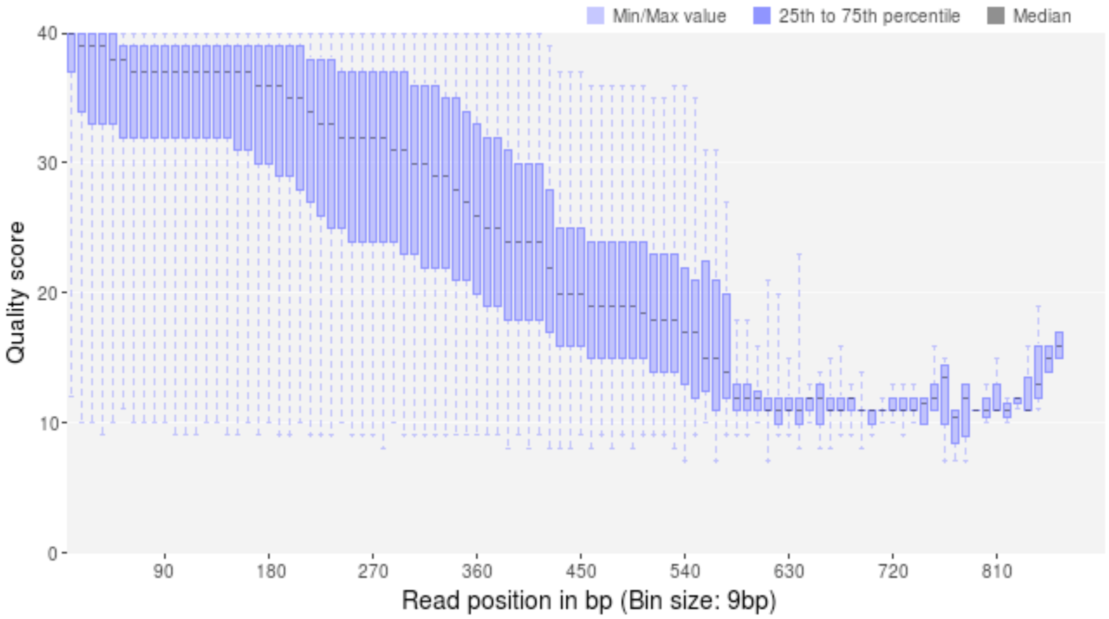
In some cases, sequences can be much longer than several standard deviations above the mean length (e.g. 1,500+ bp for a 500 bp mean length with a 100 bp standard deviation). Those sequences should be used with caution as they likely contain long stretches of homopolymer runs as in the following example. Homopolymers are a known issue of pyrosequencing technologies such as 454/Roche [1].

```
aactttaaccttttaaaacccccttaaaaaaactttaaaccccgtaaaccccccgggttt
ttttttaaaaaaccgttttttacggggtttaccccgttttaccggggttttgggggttt
taaaaaaaacgggttttaaacgggttaacccccgggttttccgggggtttaaaaagtttt
tttaaacgggggttttcccgtaaaaaaaaaaccccgtttaaaaaaaggggttaaaaaaaa
aaggggttaacccccccggggtttaaaaaaaaccttttttttttttaaaaaaaacgttttt
tttttttaaaaggggttttttttacgggggtaaacgggggggttaaaaaaaaacccccc
cgggggggtttaaaaaaaaaaccccccggttttaaaaaaccccgttttaacccctttaaaa
aaaaaacggggggggttttaaaaaaaaaaggggtttttttttttaaaaacccgtttta
aaaaccccccgtttttaacccgggttaaacccccccgggggggtaaaaccccccccccc
ggggtaacccccttttttaaaacccccccccgtttttacccgggggttttacccccg
gggggggtaaaaaaacgggggttttttttttttaaaaccggggtttttttttttttaaa
ccccggttttaaaaaccggttttaccccgggggtttttacccccgggggggggttttt
aaaaccccccggtttaaaactttaaaaacccgggtaaccccggggttttaaaaaaaaaaaa
aaaacccccccccgttaaaaaaaaaaaacccgtttttttttttaaaaaaaacccccccccgg
ttttaaaacccccccccggggggttttaccccggggttttaaaaaaaaacccgtttaaaaaa
accgggtttttaaaggggttttaaacccccccccc
```

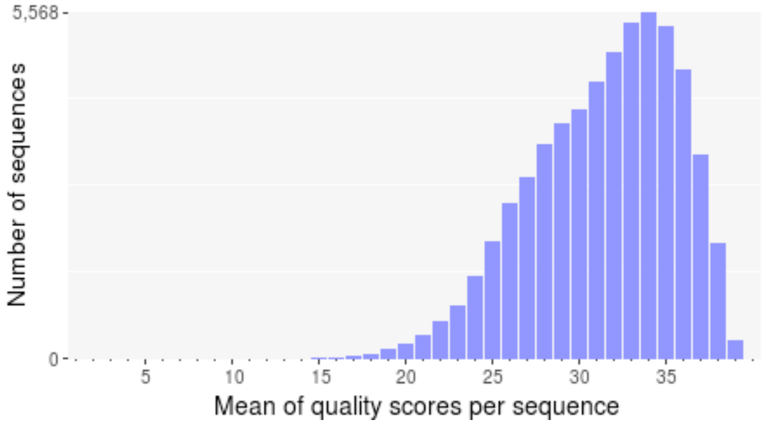
Base qualities

As for Sanger sequencing, next-generation sequencers produce data with linearly degrading quality across the read. The quality scores for 454/Roche sequencers are Phred-based since version 1.1.03, ranging from 0 to 40. Phred values are log-scaled, where a quality score of 10 represents a 1 in 10 chance of an incorrect base call and a quality score of 20 represents a 1 in 100 chance of an incorrect base call.

In PRINSEQ, the quality scores are plotted across the reads using box plots. The x-axis indicates the absolute position if all reads are no longer than 100 bp and the relative position (in % of read length) if any read is longer than 100 bp. For datasets with any read longer than 100 bp, a second plot shows binned quality values to keep its absolute positions. This plot is helpful to identify quality scores at the end of longer reads, which would otherwise be grouped with the ends of the shorter reads. The following example shows the quality scores across the read length for fragments sequenced with GS FLX using the Titanium kit. The sequences with low quality scores at the ends should be trimmed during data preprocessing.



In addition to the decrease in quality across the read, regions with homopolymer stretches will tend to have lower quality scores. Huse *et al.* [1] found that sequences with an average score below 25 had more errors than those with higher averages. Therefore, it is helpful to take a look at the average (or mean) quality scores. PRINSEQ provides a plot that shows the distribution of sequence mean quality scores of a dataset, as shown below. The majority of the sequences should have high mean quality scores.



Low quality sequences can cause problems during downstream analysis. Most assemblers or aligners do not take into account quality scores when processing the data. The errors in the reads can complicate the assembly process and might cause misassemblies or make an assembly impossible.

GC content

The GC content distribution of most samples should follow a normal distribution. In some cases, a bi-modal distribution can be observed, especially for metagenomic data sets. The GC content plot in PRINSEQ marks the mean GC content (M) and the GC content for one and two standard deviations (1SD and 2SD). This can help to decide where to set the GC content thresholds, if a GC content filter will be applied. The plot can also be used to find the thresholds or range to select sequences from a bi-modal distribution.

Poly-A/T tails

Poly-A/T tails are considered repeats of As or Ts at the sequence ends. In PRINSEQ, the minimum length of a tail is 5 bp and sequences that contain only As or Ts are counted for both ends. A small number of tails can occur even after trimming poly-A/T tails. For example, a sequence that ends with AAAAATTTTT and that has been trimmed for the poly-T will still contain the poly-A.

Trimming poly-A/T tails can reduce the number of false positives during database searches, as long tails tend to align well to sequences with low complexity or sequences with tails (e.g. viral sequences) in the database.

Ambiguous bases

Sequences can contain the ambiguous base N for positions that could not be identified as a particular base. A high number of Ns can be a sign for a low quality sequence or even dataset. If no quality scores are available, the sequence quality can be inferred from the percent of Ns found in a sequence or dataset. Huse *et al.* [1] found that the presence of any ambiguous base calls was a sign for overall poor sequence quality.

Ambiguous bases can cause problems during downstream analysis. Assemblers such as Velvet and aligners such as SHAHA2 or BWA use a 2-bit encoding system to represent nucleotides, as it offers a space efficient way to store sequences. For example, the nucleotides A, C, G and T might be 2-bit encoded as 00, 01, 10 and 11. The 2-bit encoding, however, only allows to store the four nucleotides and any additional ambiguous base cannot be represented. The different programs deal with the problem in different ways. Some programs replace ambiguous bases with a random base (e.g. BWA [2]) and others with a fixed base (e.g. SHAHA2 and Velvet replace Ns with As [3]). This can result in misassemblies or false mapping of sequences to a reference sequence and therefore, sequences with a high number of Ns should be removed before downstream analysis.

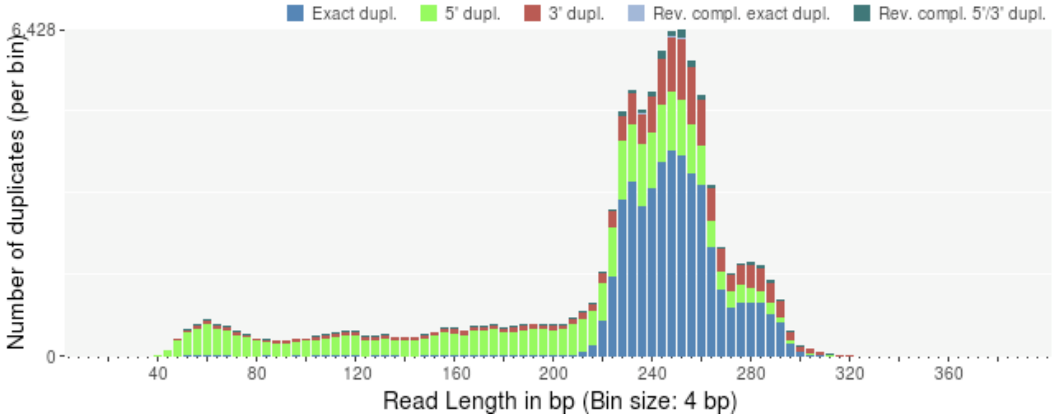
Sequence duplications

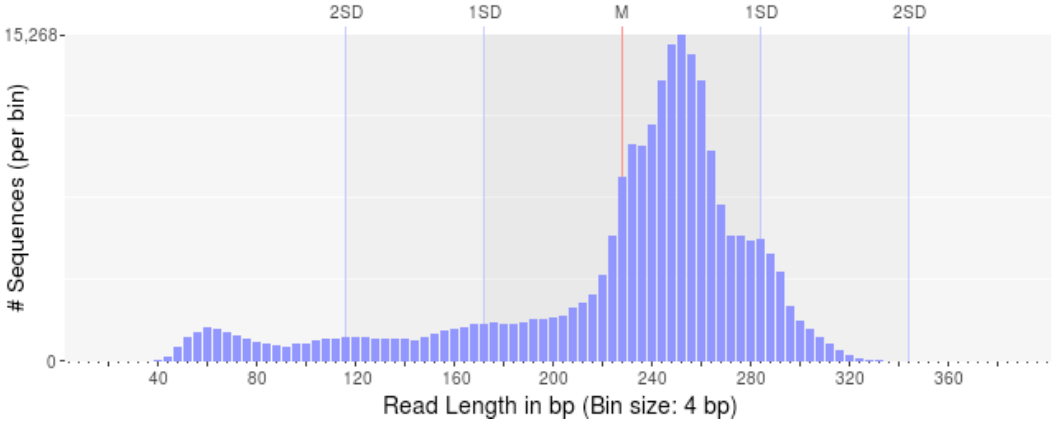
Real or artificial? Assuming a random sampling of the genomic material in an environment such as in metagenomic studies, reads should not start at the same position and have the same errors (at least not in the numbers that they have been observed in most metagenomes). Gomez-Alvarez *et al.* [5] investigated the problem in more detail and did not find a specific pattern or location on the sequencing plate that could explain the duplications.

Duplicates can arise when there are too few fragments present at any stage prior to sequencing, especially during any PCR step. Furthermore, the theoretical idea of one micro-reactor containing one bead for 454/Roche sequencing does not always translate into practice where many beads can be found in a single micro-reactor. Unfortunately, artificial duplicates are difficult to distinguish from exactly overlapping reads that naturally occur within deep sequence samples.

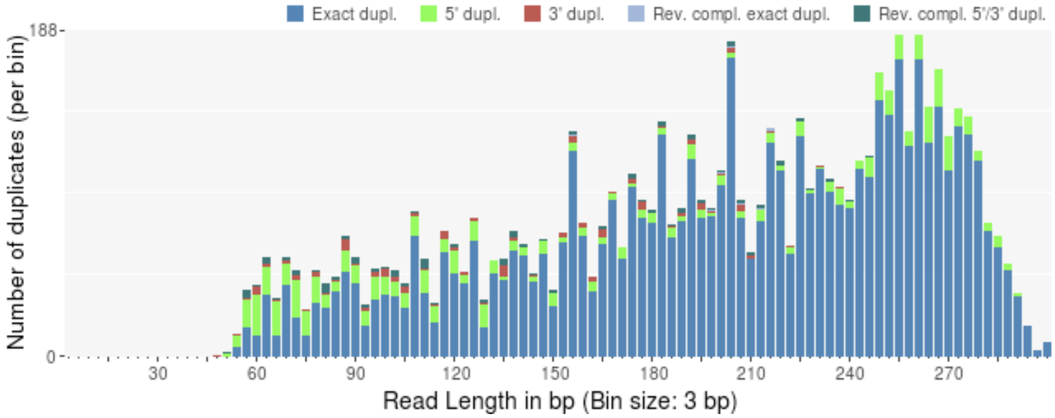
The number of expected sequence duplicates highly depends on the depth of the library, the type of library being sequenced (whole genome, transcriptome, 16S, metagenome, ...), and the sequencing technology used. The sequence duplicates can be defined using different methods. *Exact duplicates* are identical sequence copies, whereas *5' or 3' duplicates* are sequences that are identical with the 5' or 3' end of a longer sequence. Considering the double-stranded nature of DNA, duplicates could also be considered sequences that are identical with the *reverse complement* of another sequence.

The different plots in PRINSEQ can be helpful to investigate the degree of sequence duplications in a dataset. The following plot shows the number of sequence duplicates for different lengths. The distribution of duplicates should be similar to the length distribution of the dataset. The number of 5' duplicates is higher for shorter sequences (as observed in the example below), suggesting that exact sequence duplicates may have been trimmed during signal processing.





The number of exact duplicates is often higher than the number of 5' and 3' duplicates as in the following example.



PRINSEQ offers additional plots to investigate the sequence duplicates from different points of view. The plot showing the sequence duplication levels (with number of sequences with one duplicate, two duplicates, three duplicates, …) can be used to identify the distribution of duplicates (e.g. do many sequences have only a few duplicates). The plot showing the highest number of duplicates for a single sequence (top 100) can help to indentify if only a few sequences have many duplicates (e.g. as a result of specific PCR amplification) and what the highest duplication numbers are.

Depending on the dataset and downstream analysis, it should be considered to filter sequence duplicates. The main purpose of removing duplicates is to mitigate the effects of PCR amplification bias introduced during library construction. In addition, removing duplicates can result in computational benefits by reducing the number of sequences that need to be processed and by lowering the memory requirements. Sequence duplicates can also impact abundance or expression measures and can result in false variant (SNP) calling. The example below shows the alignment of sequences against a reference sequence (gray). The sequence duplicates (starting at the same position) suggest a possibly false frequency of base C at the position marked in bold.

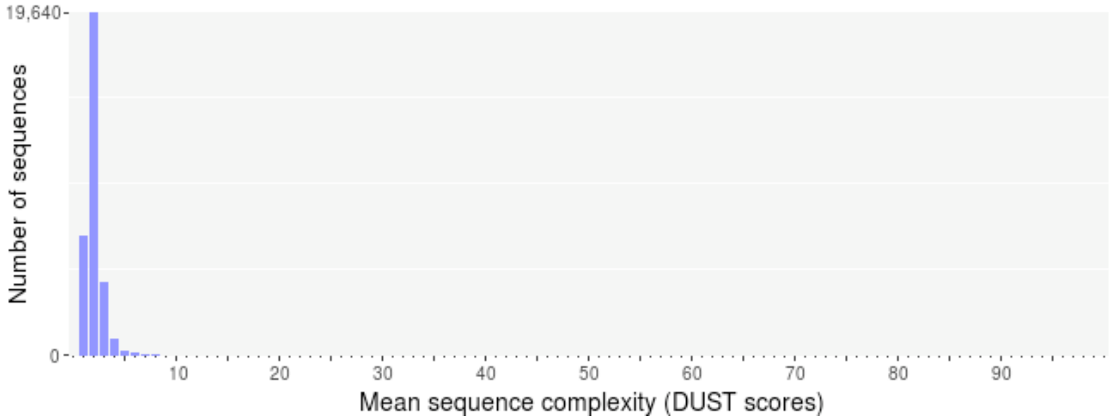
```
...ACCACACGTGTTGTGTACATGAACACAGTATATGAGCATACAGAT...
          GTGTTGTGTACATGAACACAGTATATGAGCATACAGAT...
               GTGTACATGAACACAGTATATGAGCATACAGAT...
                      TGAACACAGTCTATGAGCATACAGAT...
                      TGAACACAGTCTATGAGCATACAGAT...
                      TGAACACAGTCTATGAGCATACAGAT...
                      TGAACACAGTCTATGAGCATACAGAT...
                      TGAACACAGTCTATGAGCATACAGAT...
```

Keep in mind that the number of sequence duplicates also depends on the experiment. For short-read datasets with high coverage such as in ultra-deep sequencing or genome re-sequencing datasets, eliminating *singletons* can present an easy way of dramatically reducing the number of error-prone reads.

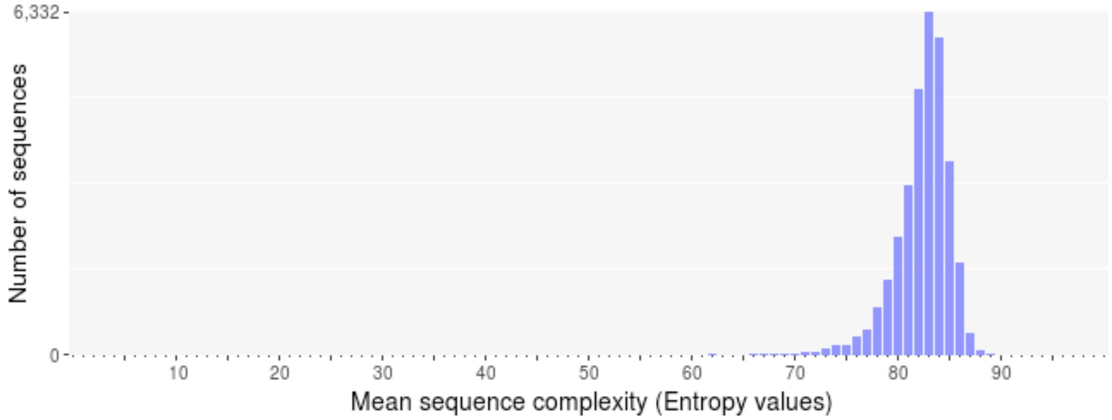
Sequence complexity

Genome sequences can exhibit intervals with low-complexity, which may be part of the sequence dataset when using random sampling techniques. Low-complexity sequences are defined as having commonly found stretches of nucleotides with limited information content (e.g. the dinucleotide repeat CACACACACA). Such sequences can produce a large number of high-scoring but biologically insignificant results in database searches. The complexity of a sequence can be estimated using many different approaches. PRINSEQ calculates the sequence complexity using the DUST and Entropy approaches as they present two commonly used examples.

The *DUST* approach is adapted from the algorithm used to mask low-complexity regions during BLAST search preprocessing [6]. The scores are computed based on how often different trinucleotides occur and are scaled from 0 to 100. Higher scores imply lower complexity and complexity scores above 7 can be considered low-complexity. A sequence of homopolymer repeats (e.g. TTTTTTTTT) has a score of 100, of dinucleotide repeats (e.g. TATATATATA) has a score around 49, and of trinucleotide repeats (e.g. TAGTAGTAGTAG) has a score around 32.

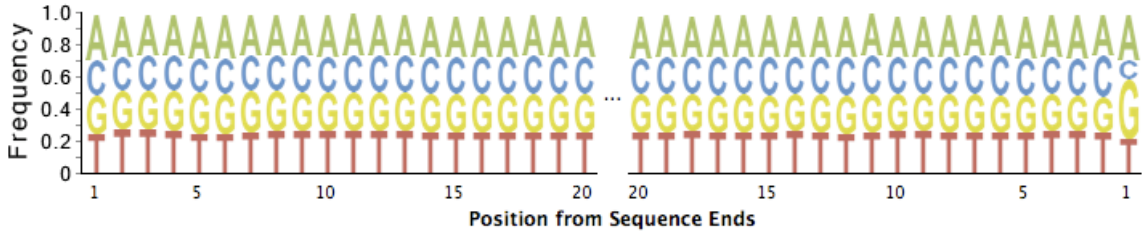


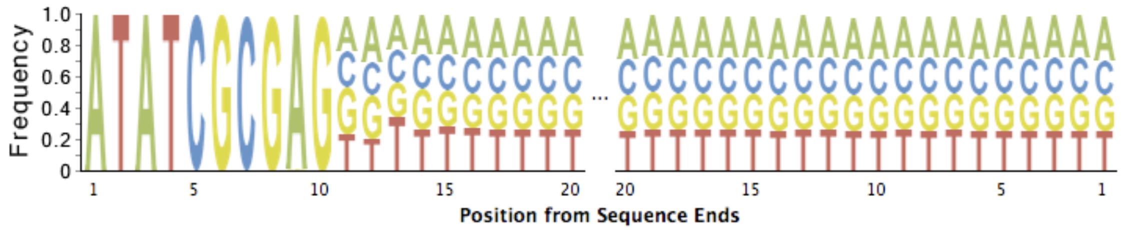
The *Entropy* approach evaluates the entropy of trinucleotides in a sequence. The entropy values are scaled from 0 to 100 and lower entropy values imply lower complexity. A sequence of homopolymer repeats (e.g. TTTTTTTTT) has an entropy value of 0, of dinucleotide repeats (e.g. TATATATATA) has a value around 16, and of trinucleotide repeats (e.g. TAGTAGTAGTAG) has a value around 26. Sequences with an entropy value below 70 can be considered low-complexity.

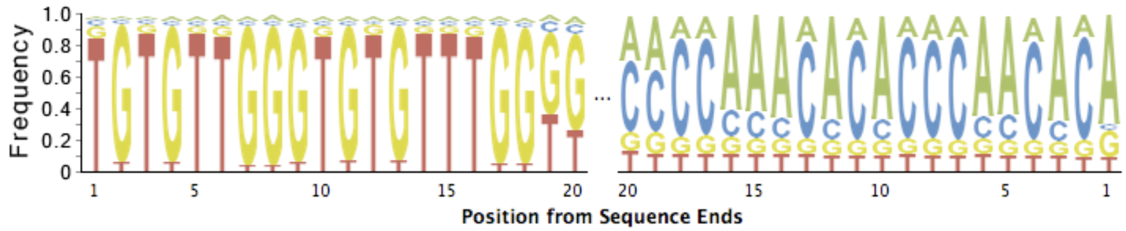


Tag sequences

Tag sequences are artifacts at the ends of sequence reads such as multiplex identifiers, adapters, and primer sequences that were introduced during pre-amplification with primer-based methods. The base frequencies across the reads present an easy way to check for tag sequences. If the distribution seems uneven (high frequencies for certain bases over several positions), it could indicate some residual tag sequences. The following three examples show the base frequencies of datasets with no tag sequence, multiplex identifier (MID) tag sequence, and whole transcriptome amplified (WTA) tag sequence.





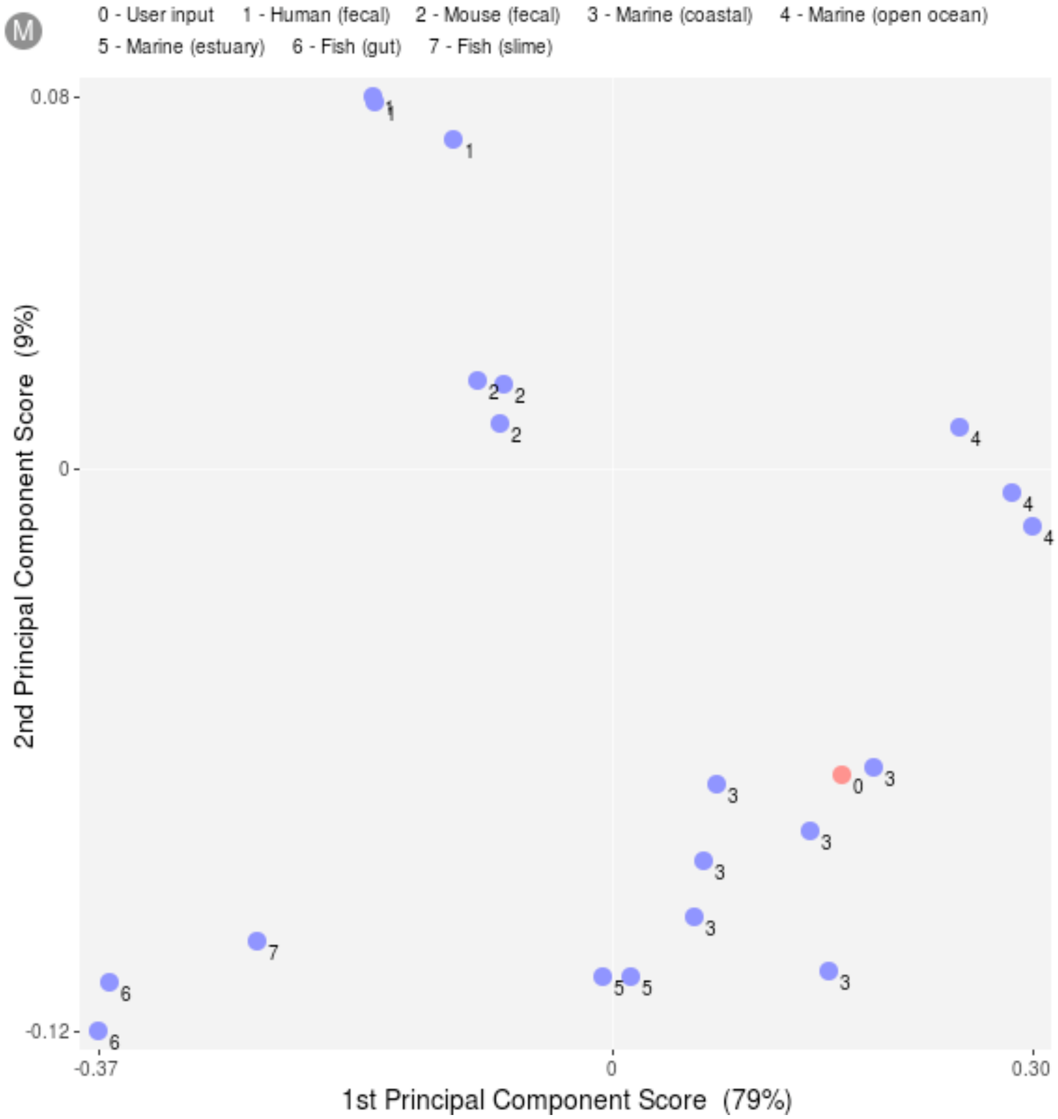


Those tag sequence should be trimmed using a program such as TagCleaner (http://tagcleaner.sourceforge.net) [4]. The input to any such trimming program should be untrimmed reads (e.g. not quality trimmed), as this will allow easier and more accurate identification of tag sequences. PRINSEQ can be used after tag sequence trimming to check if the tags were removed sufficiently.

In addition to the frequency plots, PRINSEQ estimates if the dataset contains tag sequences. The probabilities for a tag sequence at the 5'- or 3'-end require a certain number of sequences (10,000 should be sufficient). A percentage below 40% does not always suggest a tag sequence, especially if it cannot be observed from the base frequencies. The estimation does not work for sequence datasets that target a single loci (e.g. 16S) and should only be used for randomly sequenced samples such as metagenomes.

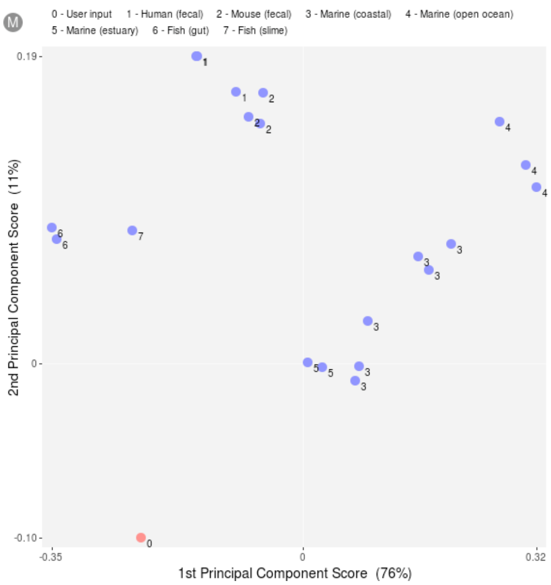
Sequence contamination

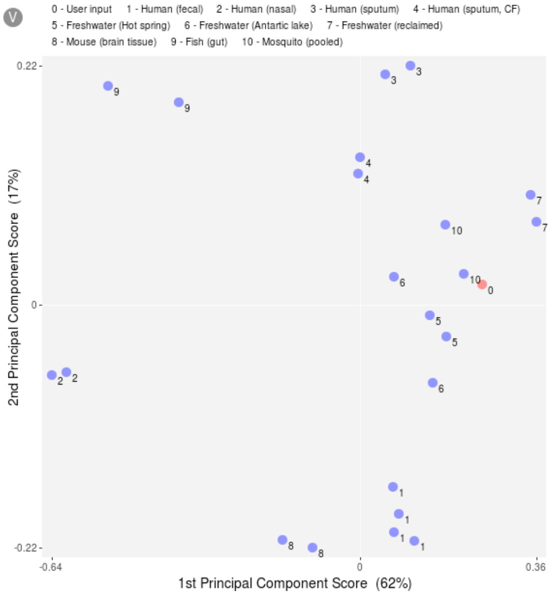
Sequences obtained from impure nucleic acid preparations may contain DNA from sources other than the sample. Those sequence contaminations are a serious concern to the quality of the data used for downstream analysis, possibly causing erroneous conclusions. The dinucleotide odds ratios as calculated by PRINSEQ use the information content in the sequences of a dataset and can be used to identify possibly contamination [7]. Furthermore, dinucleotide abundances have been shown to capture the majority of variation in genome signatures and can be used to compare a metagenome to other microbial or viral metagenomes. PRINSEQ uses principal component analysis (PCA) to group metagenomes from similar environments based on dinucleotide abundances. This can help to investigate if the correct sample was sequenced, as viral and microbial metagenomes show distinct patterns. As samples might be processed using different protocols or sequenced using different techniques, this feature should be used with caution.



The PCA plots in PRINSEQ show how the user metagenome (represented by a red dot) groups with other metagenomes (blue dots). Since the plots are generated for microbial and viral metagenomes separately, they are marked with an M or V (top left corner). The percentages in parenthesis show the explained variation in the first and second principal component. The plots are generated using preprocessed data from published metagenomes that were sequenced using the 454/Roche sequencing platform. If sequences contain tag sequences or are targeted to a certain loci (e.g. 16S), this approach will not be able to group the user data to metagenomes from the same environment. The plot above shows how a microbial metagenome might be related to other microbial metagenomes. (This plot suggest that the metagenome is likely a marine metagenome sampled in a coastal region.)

The following plots show how a viral metagenome does not group with the microbial metagenomes (left) but closely with other mosquito metagenomes (right).





PRINSEQ additional lists the dinucleotide relative abundance odds ratios for the uploaded dataset. Anomalies in the odds ratios can be used to identify discrepancies in metagenomes such as human DNA contamination (depression of the CG dinucleotide frequency).

Assembly quality measures

The Nxx contig size is a weighted median that is defined as the length of the smallest contig *C* in the sorted list of all contigs where the cumulative length from the largest contig to contig *C* is at least xx% of the total length (sum of contig lengths). Replace xx by the preferred value such as 90 to get the N90 contig size. The higher the Nxx value, the higher the rate of longer contigs and the better the dataset. If the dataset does not contain contigs or scaffolds, this information can be ignored.

References

1. Huse S, Huber J, Morrison H, Sogin M, Welch D: **Accuracy and quality of massively parallel DNA pyrosequencing**. *Genome Biology* 2007, **8**:R143.

2. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**:1754-1760.

3. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing**. *Brief Bioinform* 2010.

4. Schmieder R, Lim YW, Rohwer F, Edwards R: **TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets**. *BMC Bioinformatics* 2010, **11**:341.

5. Gomez-Alvarez V, Teal TK, Schmidt TM: **Systematic artifacts in metagenomes from complex microbial communities**. *ISME J* 2009, **3**:1314-1317.

6. Morgulis A, Gertz EM, Schäffer AA, Agarwala R: **A fast and symmetric DUST implementation to mask low-complexity DNA sequences**. *J. Comput. Biol* 2006, **13**:1028-1040.

7. Willner D, Thurber RV, Rohwer F: **Metagenomic signatures of 86 microbial and viral metagenomes**. *Environ. Microbiol* 2009.