BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 6 2011, pages 863–864
doi:10.1093/bioinformatics/btr026

Sequence analysis

Advance Access publication January 28, 2011

Quality control and preprocessing of metagenomic datasets

Robert Schmieder[1,2,*] and Robert Edwards[1,3,*]

[1]Department of Computer Science, [2]Computational Science Research Center, San Diego State University, San Diego, CA 92182 and [3]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: Here, we present PRINSEQ for easy and rapid quality control and data preprocessing of genomic and metagenomic datasets. Summary statistics of FASTA (and QUAL) or FASTQ files are generated in tabular and graphical form and sequences can be filtered, reformatted and trimmed by a variety of options to improve downstream analysis.

Availability and Implementation: This open-source application was implemented in Perl and can be used as a stand alone version or accessed online through a user-friendly web interface. The source code, user help and additional information are available at http://prinseq.sourceforge.net/.

Contact: rschmied@sciences.sdsu.edu; redwards@cs.sdsu.edu

Received on November 8, 2010; revised on January 11, 2011; accepted on January 12, 2011

1 INTRODUCTION

High-throughput sequencing has revolutionized microbiology and accelerated genomic and metagenomic analyses; however, downstream sequence analysis is compromised by low-quality sequences, sequence artifacts and sequence contamination, eventually leading to misassembly and erroneous conclusions. These problems necessitate better tools for quality control and preprocessing of all sequence datasets.

For most next-generation sequence datasets, the quality control should include the investigation of length, GC content, quality score and sequence complexity distributions; sequence duplication; contamination; artifacts; and number of ambiguous bases. In the preprocessing step, the sequence ends should be trimmed and unwanted sequences should be filtered.

Here, we describe an application able to provide graphical guidance and to perform filtering, reformatting and trimming on FASTA (and QUAL) or FASTQ files. The program is publicly available through a user-friendly web interface and as a stand alone version. The web interface allows online analysis and data export for subsequent analysis.

2 METHODS

2.1 Sequence complexity

The sequence complexity is evaluated as the mean of complexity values using a window of size 64 and a step size of 32. There are two types of sequence complexity measures implemented in PRINSEQ. Both use overlapping nucleotide triplets as words and are scaled to a maximum value of 100. The first is an adaptation of the DUST algorithm (Morgulis *et al.*, 2006) used as BLAST search preprocessing for masking low complexity regions:

$$\mathrm{CD} = \sum_{i=1}^{k} \frac{n_i(n_i-1)(w-2)s}{2(l-1)l} \quad (1)$$

where $k = 4^3$ is the alphabet size, w is the window size, n_i is the number of words i in a window, $l \leq 62$ is the number of possible words in a window of size 64 and $s = 100/31$ is the scaling factor.

The second method evaluates the block-entropies of words using the Shannon–Wiener method:

$$\mathrm{CE} = -\sum_{i=1}^{k} \left(\frac{n_i}{l}\right) \log_k \left(\frac{n_i}{l}\right) \quad (2)$$

where n_i is the number of words i in a window of size w, l is the number of possible words in a window and k is the alphabet size. For windows of size $w < 66$, $k = l$ and otherwise $k = 4^3$.

2.2 Dinucleotide odds ratio

The basic version of the dinucleotide odds ratio calculation (Burge *et al.*, 1992) is used without taking into account the occurrence of ambiguous characters such as N. In addition, the commonly used version that accounts for the complementary antiparallel structure of double-stranded DNA introduces an additional dinucleotide by simply concatenating the sequence with its reverse complement. To account for this, the odds ratios are calculated using the number n_X of nucleotide X and the number n_{XY} of dinucleotide XY only for nucleotides A, C, G and T on the forward strand:

$$\rho_{XY} = \frac{n_{XY} + n_{Y'X'}}{(n_X + n'_X)(n_Y + n'_Y)} \frac{2m^2}{d} \quad (3)$$

where X′ is the complement of nucleotide X, m is the number of valid nucleotides and d is the number of valid dinucleotides in the sequence.

2.3 Tag sequence probability

Tag sequences are artifacts at the sequence ends such as adapter or barcode sequence. A k-mer approach is used to calculate the probability of a tag sequence at the 5′- or 3′-end. The k-mers are aligned and shifted before calculating the frequencies as described in (Schmieder *et al.*, 2010) to account for sequencing limitations.

2.4 Sequence duplication

Sequence replication can occur during different steps of the sequencing protocol, and can therefore generate artificial duplicates (Gomez-Alvarez *et al.*, 2009). Here, duplicates are categorized into the following groups: (i) exact duplicate, (ii) 5′ duplicate (sequence matches the 5′-end of a longer sequence), (iii) 3′ duplicate, (iv) exact duplicate with the reverse complement of another sequence and (v) 5′/3′ duplicate with the reverse complement of another sequence. The duplicates are identified independently by sorting and prefix/suffix matching of the sequences.

*To whom correspondence should be addressed.

3 FEATURES

3.1 Quality control

The summary statistics provided include the number of sequences and number of bases in the FASTA or FASTQ file, tables with minimum, maximum, range, mean, standard deviation and mode for read length and GC content, charts for read length distribution, GC content distribution, quality scores, sequence complexity, sequence duplicates, occurrence of Ns and poly-A/T tails. Additionally, the base frequencies at the sequence ends and the probability of tag sequences are provides to the user. The dinucleotide odds ratios can be used to identify possibly contamination (Willner *et al*., 2009) and the dinucleotide relative abundance profile can be used to compare the user metagenome to other microbial or viral metagenomes using principal component plots. The assembly measures such as N50 or N90 are helpful for datasets containing contigs.

3.2 Sequence filtering

Sequences can be filtered by their length, quality scores, GC content, number or percentage of ambiguous base N, non-IUPAC characters for nucleic acids, number of sequences, sequence duplicates, sequence complexity (for example, to remove simple repeat sequences such as ATATATATAT), and custom filters defined by the user given a predefined grammar.

3.3 Sequence trimming

The trimming options allow users to trim sequences to a specific length, trim bases from the 5′- and 3′-end, trim poly-A/T tails and trim by quality scores with user-defined options. The trimming of sequences can generate new sequence duplicates and therefore, trimming is performed before most filtering steps.

3.4 Sequence formatting

The sequences can be modified to change them to upper or lower case (for example, to remove soft-masking), convert between RNA and DNA sequences, change the line width in FASTA and QUAL files, remove sequence headers or rename sequence identifiers. Additionally, FASTQ inputs can be converted into FASTA and QUAL format, and vice versa.

3.5 Web interface

The web version includes sample datasets to compare and test the program. All graphics are generated using the Cairo graphics library (http://cairographics.org/). The web interface allows the submission of compressed FASTA (and QUAL) or FASTQ files to reduce the time of data upload. Currently, ZIP, GZIP and BZIP2 compression algorithms are supported allowing direct processing of compressed data from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra). The filter, trim and reformat options can be exported and imported for similar processing of different datasets. Additionally, the web interface provides predefined option sets to perform different types of preprocessing. Data uploaded using the web interface can be shared or accessed at a later point using unique data identifiers.

4 BRIEF SURVEY OF ALTERNATIVE PROGRAMS

There are different applications that provide quality control and preprocessing features for sequence datasets. PRINSEQ was compared with three other available programs, each offering various additional features and functions. Although the programs have been designed to process short read data, they are able to process longer read sequences. SolexaQA (Cox *et al*., 2010) is software written in Perl that allows investigation and trimming of sequences by their base quality scores. The software does not provide additional summary statistics or preprocessing features and requires a working installation of R and Perl modules such as GD to produce graphical outputs. FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) is software written in Java that provides summary statistics for FASTQ files. In its current version, FastQC does not provide data preprocessing features. The FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) is a collection of command line tools that provide preprocessing features and summaries for quality scores and nucleotide distributions. The tools were recently integrated into the Galaxy platform (Blankenberg *et al*., 2010). All of these programs are still in active development and new functions will undoubtedly be added over time.

5 CONCLUSION

PRINSEQ allows scientists to efficiently check and prepare their datasets prior to downstream analysis. The web interface is simple and user-friendly, and the stand alone version allows offline analysis and integration into existing data processing pipelines. The results reveal whether the sequencing experiment has succeeded, whether the correct sample was sequenced and whether the sample contains any contamination from DNA preparation or host. The tool provides a computational resource able to handle the amount of data that next-generation sequencers are capable of generating and can place the process more within reach of the average research lab.

ACKNOWLEDGEMENT

We thank the PRINSEQ users for comments and suggestions.

Funding: Advances in Bioinformatics from the National Science Foundation (grant DBI 0850356).

Conflict of Interest: none declared.

REFERENCES

Blankenberg,D. *et al*. (2010) Manipulation of FASTQ data with Galaxy. *Bioinformatics*, **26**, 1783–1785.

Burge,C. *et al*. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.

Cox,M.P. *et al*. (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, 485.

Gomez-Alvarez,V. *et al*. (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J.*, **3**, 1314–1317.

Morgulis,A. *et al*. (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.*, **13**, 1028.

Schmieder,R. *et al*. (2010) TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics*, **11**, 341.

Willner,D. *et al*. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Envir. Microbiol.*, **11**, 1752–1766.