NGS TOOLBOX

release 2

- Documentation -

What is NGS TOOLBOX

The NGS TOOLBOX is a collection of simple open source Perl scripts that perform basic analyses and processing steps using next generation sequencing (NGS) datasets. Each tool is designed to ensure convenient and intuitive usage. Installation and usage does not require any bioinformatics skills. All scripts work out-of-the-box. Advanced users may use the command line based Perl scripts to build automated sequence analyses/processing pipelines.

Prerequisites

Running Perl scripts on your computer requires installation of a Perl interpreter. A Perl interpreter is commonly preinstalled on Linux/Unix and Mac computers. If you are using a Windows computer you can download and install one of the following freely available Perl distributions:

- Strawberry Perl (at http://strawberryperl.com/)
- ActivePerl (at http://www.activestate.com/activeperl/downloads)

Quick start

Download and copy NGS TOOLBOX into the same directory as your NGS datasets. If you downloaded the zip-compressed folder, extract the zip-folder here. The Perl scripts must be started from the command line (Windows: *command prompt*, Linux/Unix and Mac: *terminal*) so you have to open a *command prompt* or *terminal* window and navigate to the folder that contains your NGS datasets and the NGS TOOLBOX files. You can get detailed information how to start and use each script by using e.g the following command.

```
perl TB2.1_basic-analyses.pl -help
```

Scripts that accept both FASTA and FASTQ input format will automatically check the format of the input file.

Tools (for detailed information start the scripts with the option -help)

Name	Description	Input format
basic-analyses	Counts the number of sequence reads and non-identical sequences. Calculates the total nucleotide composition and GC content. Calculates the sequence length distribution and positional nucleotide composition.	FASTA/FASTQ
clip	Removes specified adapter sequences. Is a very customizable tool since it applies a simple RegEx-like search function.	FASTA/FASTQ
collapse	Removes identical sequences from your dataset. Information on sequence read counts for identical sequences will be output in the FASTA/FASTQ header line.	FASTA/FASTQ
concatenate	Concatenates all files from one directory with one or more specified file extensions.	FASTA/FASTQ
duster	Removes low-complexity sequences from your dataset.	FASTA/FASTQ
fastq2fasta	Converts FASTQ formatted files to FASTA formatted files.	FASTQ
length-filter	Filters your sequence reads according to a specified minimum and maximum sequence length.	FASTA/FASTQ
pingpong	Screens map files for a so-called ping-pong signature (10 nt 5’ overlap of mapped sequences) and calculates ping-pong z-scores. A ping-pong signature is a hallmark of secondary piRNA biogenesis. Use map files produced by	MAP files from sRNAmapper.pl

	SeqMap (Jiang and Wong 2008) or sRNAmapper (small RNA mapping tool that comes along with proTRAC).	
q-check	Performs a quality check based on Phred scores. Calculates the total average Phred score and the average Phred score for each position. Calculates the total average sequence accuracy (probability to contain 0 miscalled bases) and outputs the total distribution of Phred scores.	FASTQ
q-filter	Performs quality filtering based on Phred scores. Can apply three different cutoff types: i) Minimum average Phred score of a sequence read ii) Minimum quality of the worst called base within one sequence read iii) Minimum accuracy of a sequence read (probability to contain 0 miscalled bases).	FASTQ
rev-comp	Creates reverse complementary sequences (or reverse/complementary only)	FASTA/FASTQ
split	Splits large sequence files into smaller parts specified by i) sequence counts, ii) file size or iii) fixed number of output files. Does not disrupt FASTA or FASTQ format.	FASTA/FASTQ

Citation

If you use software from the NGS toolbox for your publication, please cite the following paper:
Rosenkranz D, Chung-Ting H, Roovers EF, Zischler H, Ketting RF. Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. *Genomics Data* 2015 5:309-313.

Contact

If you have any questions or comments or found any bugs in the software please do not hesitate to contact.

David Rosenkranz
Institute of Anthropology, small RNA group
Johannes Gutenberg University Mainz, Germany
Email: rosenkranz@uni-mainz.de
Web: http://www.smallRNAgroup-mainz.de