**The PARMA toolkit - useful tools for NGS data analysis**

**Overview**

The PARMA toolkit provides tools for the analysis of NGS data, especially of (PAR-)CLIP sequence reads. The most important tool is the mapping tool which embeds the PARMA algorithm for read alignment. The following tools are available in Version 0.5 alpha:

Table : Overview of tools accessible through the PARMA toolkit.

|  |  |
| --- | --- |
| **Tool** | **Description** |
| map |  |
| comb |  |
| benchmark |  |
| error |  |
| clust |  |

**Requirements**

The PARMA toolkit (hyperlink) and the PARMA algorithm, embedded in the open source aligner BWA (hyperlink), which still provides the indexing with the basic BWA algorithm (so far, only BWA version 0.7.8 is supported), are required. If the BWT-index is created with another version than BWA 0.7.8, the PARMA algorithm could produce unforeseen errors during the alignment process. After installing the PARMA algorithm it should be included in the PATH environment, otherwise the PARMA toolkit is not able to find the algorithm. Alternatively, you can create a file called “parma.properties” in the PARMA toolkit directory and set the path to PARMA there, like in the following:

PARMA\_LOCATION=myPathToPARMA/PARMA-master/

The basic command for executing the PARMA toolkit is as follows:

java -jar parma.jar *MODE*  [options]

whereas *MODE* is one of the tools from Table 1. To print an overview of the available tools, just execute the jar-file without any further options. A more detailed description of every tool can be printed by executing the basic mode, like the following for the mapping tool:

java –jar parma.jar map

which will print the instructions for the mapping tool.

**Workflow for mapping**

First, an index for a reference genome sequence has to be created using the index function of the BWA algorithm as follows:

bwa index *REFERENCE*

Afterwards, the PARMA mapping tool can be executed as follows:

java -jar parma.jar map -q *INPUT* -r *REFERENCE* -p *THREADS* -o *OUTPUT* --refine

To allow mapping against multiple databases, the command looks as follows:

java -jar parma.jar map -q *INPUT* -r *REFERENCE* -p *THREADS* -o *OUTPUT* -t *TRANSCRIPT\_REFERENCE* --refine

where *TRANSCRIPT\_REFERENCE* is a multiple fasta file containing sequences of known transcripts for a given organism. For this multiple fasta file, a BWT-index has to be created in a first step, too.

**Combine tool**

We additionally provide our implementation of combining the results of a genomic reference mapping and the results of a transcriptomic reference mapping in the PARMA toolkit. Therefore, the two alignment files must be stored in a BAM-format and are used as input for the tool, as follows:

java -jar parma.jar comb *GENOMIC\_MAPPING TRANSCRIPTOMIC\_MAPPING OUTPUT*

The result is saved in the *OUTPUT* file in a BAM-format.

**Error profile tool**

The calculation of the error profile for a given sequence read dataset is possible using the error profile tool of the PARMA toolkit. Therefore, a reference-based read alignment has to be calculated (and stored in a BAM-file) and can be used as input for the error profile tool:

java -jar parma.jar error *MAPPING REFERENCE MAX\_READ\_LENGTH*

**Clustering tool**