QuasR: Quantify and Annotate Short Reads in R

Anita Lerch, Dimos Gaidatzis, Florian Hahne and Michael Stadler

Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Maulbeerestrasse 66, CH-4056 Basel
Novartis Institute for Biomedical Research, Klybeckstrasse 141, CH-4057 Basel

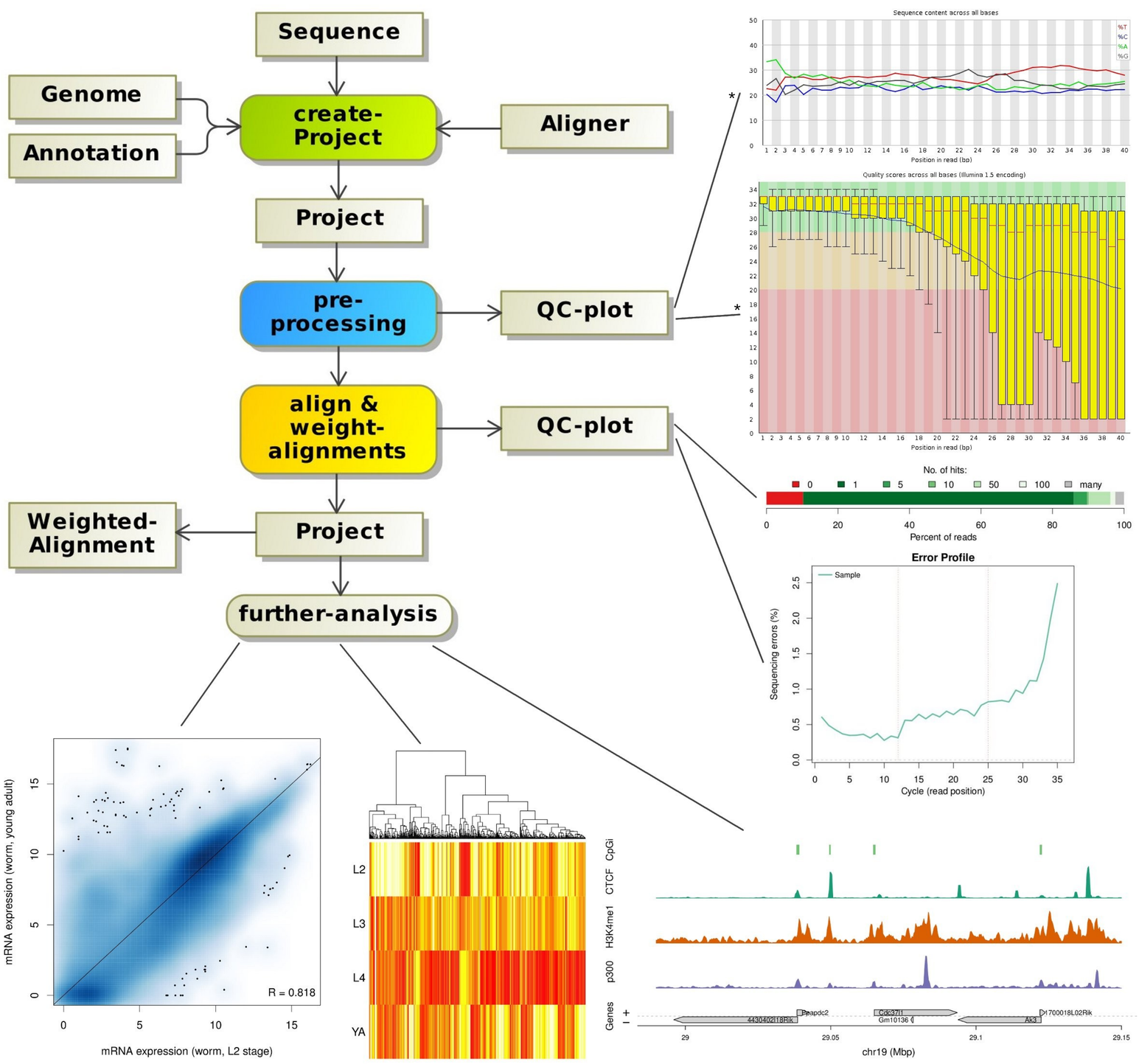
Deep sequencing technology, due to its high throughput and low cost, has become a powerful research tool in a wide range of applications, such as RNA-seq and ChIP-seq. In the last years there have been many efforts in the bioinformatics community to provide software in R/Bioconductor to simplify the processing and the biological interpretation of such large data sets. However until now, there is no integrated start-to-end analysis solution within R that sufficiently abstracts the technical details and would be suitable for use by biologists. In particular, alignments need to be performed outside of R and genome annotation information must be manually incorporated. Here we outline the deep sequencing analysis package QuasR, a further development of the FMI deep sequencing pipeline, built to make efficient use of available hardware resources and to simplify analysis of next generation sequencing data.

Aim

- To simplify analysis of next generation sequencing data due to the integration of common alignment programs e.g. Bowtie, BWA and QC tools such as FastQC.
- Efficient use of CPU and memory, use of multiple cores if possible and data streaming to loading limited amount of data into the memory (default max. 2GB).
- Provide a scalable solution that runs on a laptop, big server or a cluster.
- The pipeline should be as flexible as possible for future extension with new analysis strategies.
- To build a R/Bioconductor package, which makes extensive use of existing core Bioconductor libraries.
- Operating system independent (running on Linux, Windows or MacOS).
- Easy to install and use for a biologist (comparable to the analysis of a microarray experiment).

Scaffold of the Deep Sequencing Pipeline

The standard procedure for analyzing sequences (short reads) from a RNA-seq or Chip-seq experiment is first to filter out sequences with low quality and complexity (e.g. shorter than 14 nt or more than two N bases). Barcoded samples have to be demultiplexed and the adapter has to be removed from the sequence. Afterwards the alignment is performed against the genome assembly and optionally against additional sequences, such as an exon-exon junction database or E. coli or other genomes to check for contamination. Finally the coverage of each nucleotide or transcript is calculated.

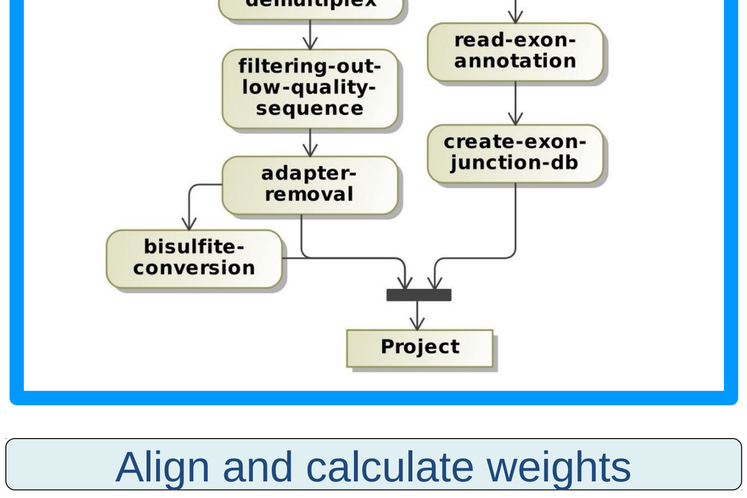


The minimal required steps are shown in the activity diagram of the deep sequencing pipeline scaffold. The selected picture illustrate possible visualizations of the data at different steps of the analysis.
* Source: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Example

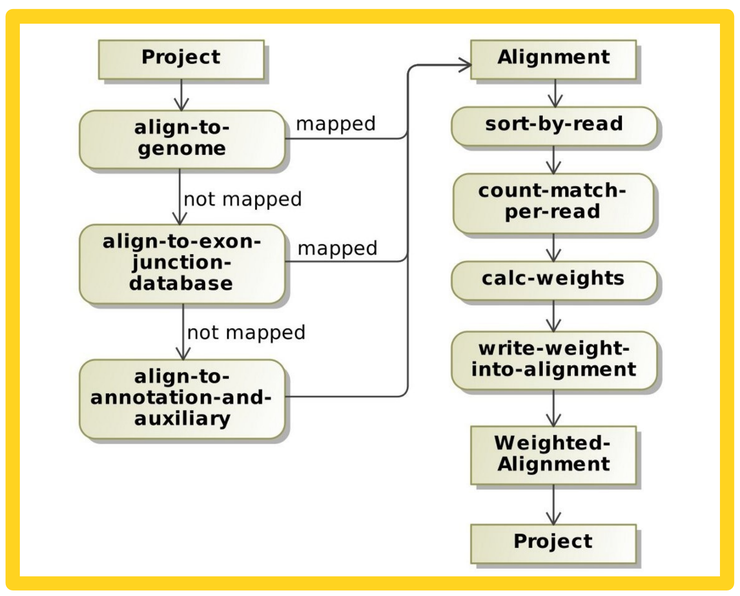
```
library("QuasR")
sampleFile <- "samples.txt"
annotationFile <- "annotations.txt"
genomeName <- "dm3"
project <- ProjectInfo(sampleFile, genomeName, annotationFile)
project <- preprocessing(project)
project <- align(project)
dat <- calcLevels(project, "mRNA")
saveProjectInfo(project, "project.rds")
```

Create Project Object



Preprocessing

Align and calculate weights



Information about the Library

Implementation/Technologies:

The library will be implemented in R, with computing intensive processes implemented in C for performance reasons.

Availability:

Currently the software is still in the primary phase. The source code is available from the Bioconductor subversion repository https://hedgehog.fhcrc.org/bioconductor/trunk/madman/Rpacks/QuasR as development version. In future it will be part of the R/Bioconductor software packages and can be installed as follows:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("QuasR")
```

Road map:

1. Construction of the Scaffold of the Deep Sequencing Pipeline:
   - Implementation of the constructor of the 'Project' Object, which contains all information needed for processing.
   - Implementation of the indexing, aligning and weighting function.
2. Implementation of the preprocessing and the building of the exons junction database.
3. Implementation of further analysis algorithms and connectors to other libraries e.g. Genome Browser, DEseq, edgeR, rtracklayer, headmap, BayesPeak

Questions and suggestions are welcome! anita.lerch@fmi.ch