SeqSense Analysis Toolkit

Tutorial Guide

Version 1.0

BIO-RAD

SeqSense Analysis Toolkit

Tutorial Guide

BIO-RAD

Legal Notices

Technical Support

The Bio-Rad Technical Support department in the U.S. is open Monday through Friday, 5:00 AM to 5:00 PM, Pacific time.

Phone: 1-800-424-6723, option 2

Email: Support@bio-rad.com (U.S./Canada Only)

For technical assistance outside the U.S. and Canada, contact your local technical support office or click the Contact us link at www.bio-rad.com.

Table of Contents

Table of Contents

Chapter 1 Introduction

The Bio-Rad SeqSense Analysis Toolkit is a Docker container with command line scripts and libraries that process FASTQ files as input for secondary analysis, and produces BAM files, count matrices, and reports as output for tertiary analysis.

This tutorial illustrates how the SeqSense Analysis Toolkit is used with the SeqSense Complete Stranded RNA Library Prep Kit, and provides the necessary information, scripts, and libraries to analyze the SeqSense Complete RNA data.

Note: Instructions for obtaining human, rat, and mouse reference genomes for analysis are provided in Appendix A, Downloading the Reference Genome.

This tutorial is presented in an Ubuntu Terminal interface, but you can use the commands in any environment that supports UNIX commands.

Requirements

The SeqSense Analysis Toolkit is packed into a a Docker container. Therefore, to use the Toolkit you must install the free Community Edition of Docker from the Docker website:

https://www.docker.com/get-started

This tutorial assumes that Docker is installed and running. You do not need advanced knowledge of Docker to use the Toolkit, but an optional tutorial is available on the Docker website.

Table 1 specifies the requirements for installing and running Docker and the SeqSense Analysis Toolkit.

Table 1. System requirements

Component	Minimum	Recommended
Operating system	Ubuntu OS 16.04 or higher	Ubuntu OS 16.04 or higher
Docker version	Docker v18.08.7 or higher	Docker v18.08.7 or higher
CPU cores	16	24 or greater
Memory	RAM 32 GB	RAM 64 GB or greater
Available disk space	500 GB	1 TB

Important: If you are running a system with higher than minimum requirements, you must add the following command line arguments to fully utilize its capabilities:

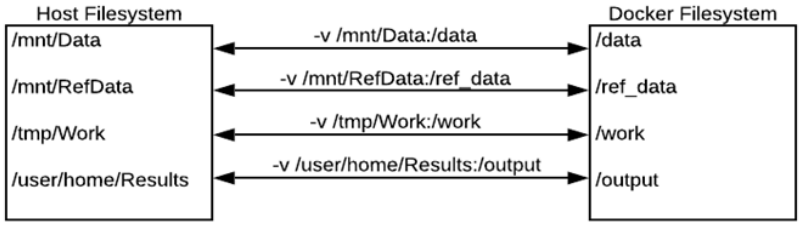
- --max_cpus
- --max_memory

Chapter 2 Using the Toolkit

The SeqSense Analysis Toolkit is designed to process one sample index at a time, where each sample is represented by a set of paired FASTQ files. These files represent the entry point into the command workflow. To view an illustration of the workflow, see Understanding the Output Step Workflow on page 11.

Container Structure

When you run the Docker command, Docker launches the SeqSense Analysis Toolkit and mounts the required directories to pass input data and receive output data. The following graphic shows the sample directory structure for the raw FASTQ files from one sample that are used in this tutorial:



All analysis will proceed from this directory structure. Each directory is briefly described in Table 2.

Table 2. Container directories

Directory	Description
/data	Input directory, which contains your FASTQ files **-v /local/path/to/fastqdir:/data**
/ref_data	reference data directory, where the local copy of the reference genome is stored **-v /local/path/to/ref_data:ref_data**
/work	working directory, where all intermediary work is stored **-v /local/path/toi/workdir:/work**
/output	Output directory, where formal outputs of the pipeline are written **-v /local/path/to/outputdir:/output**

Mounting the Directories

Complete the steps below to mount the required directories and launch the SeqSense Analysis Toolkit container.

Tip: To view additional options or other help information, run the following command:

```
docker run -t bioraddbg/sequoia_analysis_toolkit --help
```

To mount the directories and launch the container

1. Run the `docker run -t` command to launch the container.
2. Docker creates the directories comprising the container structure.

   -v /local/path/to/workdir:/work
   -v /local/path/to/ref_data:/ref_data
   -v /local/path/to/outputdir:/output
   -v /local/path/to/fastqdir:/data

3. Name the container using the following syntax:

   `bioraddbg/sequoia_analysis_toolkit`

4. Use the following settings to specify storage locations in the container:
   - `--reads '/data/myreads_*R{1,2}*.fastq.gz'` for FASTQ files
   - `--outDir /output/myreads` for output files
   - `--genomes_base /ref_data` for the reference genome
   - `-w /work` for the working directory
5. Use `profile indocker` for the context of this tutorial.
6. Use `--genome {hg38,mm10,rnor6}` to specify the reference genome.

The complete invocation to set up the pipeline is shown below:

```
docker run --rm -t -v /local/path/to/workdir:/work \
    -v /local/path/to/ref_data:/ref_data \
    -v /local/path/to/outputdir:/output \
    -v /local/path/to/fastqdir:/data \
    bioraddbg/sequoia_analysis_toolkit \
    --reads '/data/myreads_*R{1,2}*.fastq.gz \
    --outDir /output/myreads \
    --genomes_base /ref_data \
    -w /work \
    -profile indocker \
    --genome hg38
```

Inputs

Following are examples of g-zipped FASTQ input files in a data directory:

/data/mm10/A23-276048775/

├── IndexA23_S23_L001_R1_001.fastq.gz

└── IndexA23_S23_L001_R2_001.fastq.gz

Note: When a sample is run across multiple lanes, a FASTQ file is generated for each lane. Before running the toolkit, merge the files together using the following commands:

cat /local/data/samplename*L*_R1_*.fastq.gz > /local/data/samplename_R1.fastq.gz

cat /local/data/samplename*L*_R2_*.fastq.gz > /local/data/samplename_R2.fastq.gz

Outputs

The output structure of the SeqSense Analysis Toolkit is listed alphabetically, as shown below:

/mnt/toolkit_test/output/IndexA23/

├── **calcRPKMTPM**

│ └── gene_counts_rpkmtpm.txt

├── **cutAdapt**

│ ├── trimlog.log

│ └── trimmed_R1.fastq.gz

├── **debarcode**

│ ├── debarcode_stats.txt

│ └── IndexA23_S23_L001_debarcoded_R1.fastq.gz

├── **dedup**

│ ├── Aligned.sortedByCoord.deduplicated.out.bam

│ ├── Aligned.sortedByCoord.deduplicated.out.bam.bai

│ └── dedup.log

├── **fastqc**

│ ├── IndexA23_S23_L001_R1_001_fastqc.html

│ ├── IndexA23_S23_L001_R2_001_fastqc.html

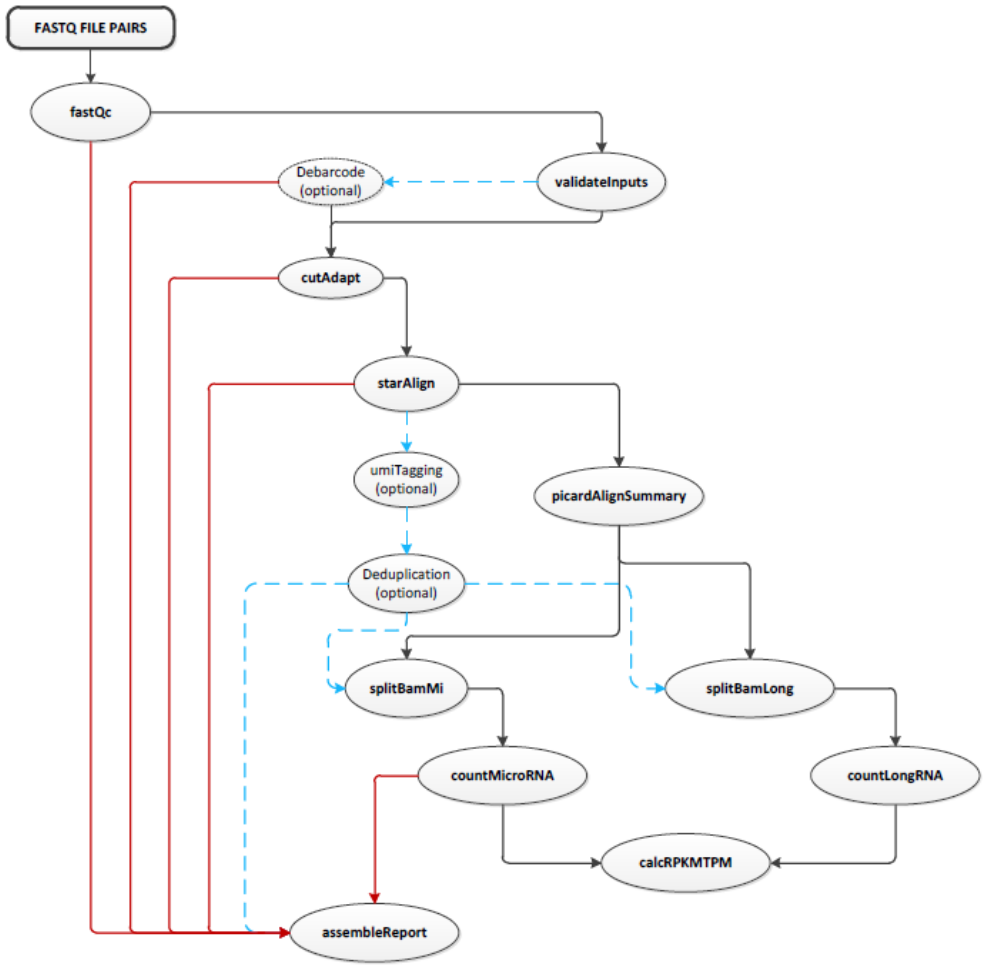
│ └── **zips**

│ ├── IndexA23_S23_L001_R1_001_fastqc.zip

```
│ └── IndexA23_S23_L001_R2_001_fastqc.zip
├── longRNACounts
│ ├── gene_counts_longRNA
│ └── gene_counts_longRNA.summary
├── microRNACounts
│ ├── gene_counts_miRNA
│ └── gene_counts_miRNA.summary
├── picardAlignSummary
│ └── rna_metrics.txt
├── pipeline_info
│ ├── execution_report.html
│ ├── execution_timeline.html
│ ├── execution_trace.txt
│ └── pipeline_dag.dot
├── report
│ ├── htmlReport.html
│ └── pdfReport.pdf
├── splitBamLong
│ └── out.longRNAs.bam
├── splitBamMi
│ └── out.miRNAs.bam
├── star
│ ├── Aligned.sortedByCoord.out.bam
│ ├── Aligned.sortedByCoord.out.bam.bai
│ └── Log.final.out
└── umiTagging
├── Aligned.sortedByCoord.tagged.bam
└── Aligned.sortedByCoord.tagged.bam.bai
```

Understanding the Output Step Workflow

The following graphic illustrates the output directory structure in order of step execution. Table 3 on page 12 describes each step output.



LINE COLOR LEGEND

■	Standard execution order
■	Path to report output
■	Optional outputs

Table 3. Outputs

Output Directory	Description
fastqc	Holds the HTML reports for each of the FASTQ files in the input directory. **(https://www.bioinformatics.babraham.ac.uk/projects/fastqc)**
debarcode (optional)	Contains the output of the debarcode step, which removes the UMI barcode from R2, and inserts it into the name of the R1 read. **Tip:** To skip deduplication (if running with only R1, or with R1 and R2), invoke the `--skipUMI` command.

cutAdapt	Contains the output of the cutAdapt step, which trims the poly-A tails and first base from reads, and allows for trimming from the 5' or 3' end based on quality score of the following passed in options:
	`--fivePrimeQualCutoff`
	`--threePrimeQualCutoff`
	(https://cutadapt.readthedocs.io/en/stable/)

starAlign	Contains the output (aligned BAM file and STAR log file) of the starAlign step, which aligns the reads to the reference genome selected. **Note:** STAR aligner (https://github.com/alexdobin/STAR) is used as a single pass alignment that aligns both long and short RNA at the same time.
picardAlignSummary	Contains the output (alignment QC stats) of the picard step, when run on the aligned BAM file. **(https://broadinstitute.github.io/picard/)** **Note:** The output directory contains a metrics file that is the result of the `CollectRnaSeqMetrics` command. **(https://broadinstitute.github.io/picard/command-line-overview.html #CollectRnaSeqMetrics).**
umiTagging (optional)	Contains the output of an Intermediary step, which adds an XU tag indicating the UMI to each read in the aligned BAM file. **Important:** Applicable only if both R1 and R2 are present and `--skipUMI` has not been set.

Table 3. Outputs, continued

Output Directory	Description
deduplication (optional)	Contains the result of PCR deduplication (deduplicated BAM file and umi_tools log file) based on the UMIs. Deduplication is performed using umi_tools with the `--method=unique` setting. **(https://github.com/CGATOxford/UMI-tools)** Applicable only if both and R1 and R2 are present, and `--skipUMI` has not been set.
splitBamMi	Holds the BAM file containing all reads that align entirely within an annotated miRNA. Overlapping reads result from intersecting the aligned BAM file with the annotated BED file containing known small RNA. Bedtools is used for the intersection. **(https://bedtools.readthedocs.io/en/latest/index.html)**
splitBamLong	Holds the BAM file containing all reads that do not intersect a known small RNA. Bedtools is used for the intersection. **(https://bedtools.readthedocs.io/en/latest/index.html)**
countMicroRNA	Holds the result (counts file and summary) of running featureCounts on the small RNA BAM file with the small RNA annotation set. **(http://subread.sourceforge.net/)**
countLongRNA	Holds the result (counts file and summary) of running featureCounts on the long RNA BAM file with the long RNA annotation set. **(http://subread.sourceforge.net/)**
calcRPKMTPM	Holds the result of the aggregation and normalization of the combined long RNA and small RNA counts.
assembleReport	Holds both PDF and HTML versions of the assembled report.
pipeline_info	Holds graphs and reports on the runtime of each of the steps.

Logging

The stderr command prompts the SeqSense Analysis Toolkit to output its status while running. The Toolkit also writes to a log file (.nextflow.log) in the directory that is mounted to /work. This log file captures the steps run and the command line options set.

Appendix A Downloading the Reference Genome

To download the reference genome, you must install the awscli tools per the instructions at the following link:

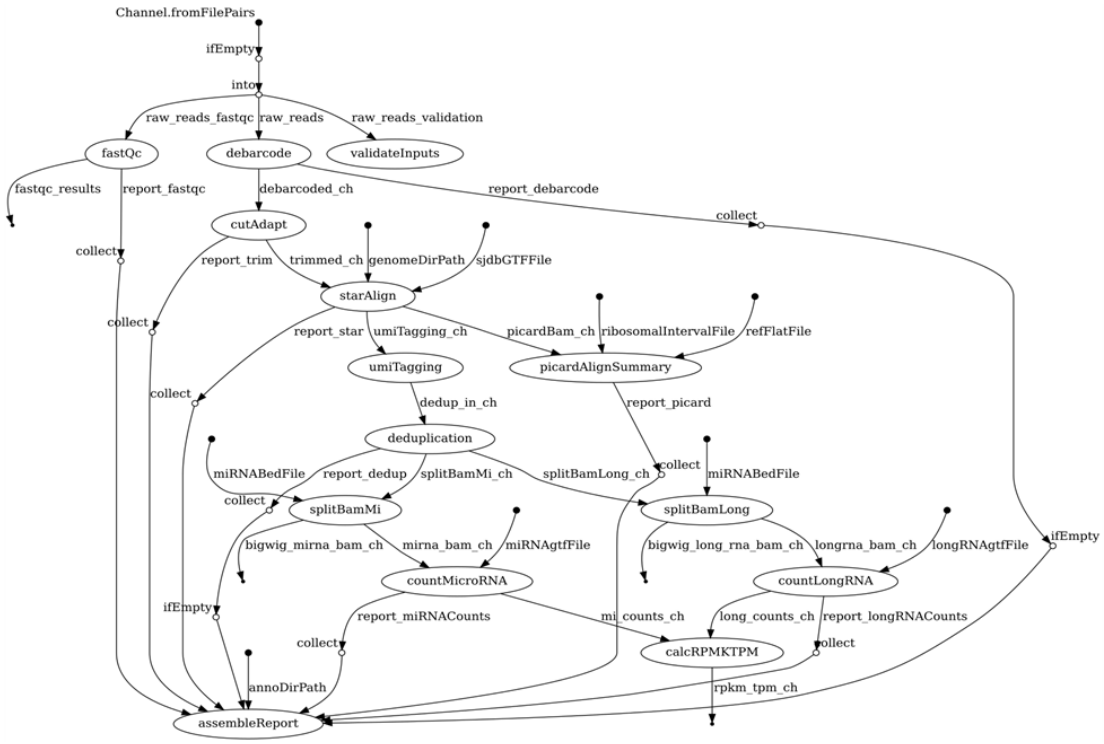
https://docs.aws.amazon.com/cli/latest/userguide/cli-chap-install.html

After you have installed the awscli tools, execute the following commands to download the genome:

```
mkdir ref_data

cd ref_data

aws s3 cp --recursive s3://dbg-cloudpipeline-data-us-west-2-prod/ref_
data/sequoia_analysis/latest/hg38.tar.gz ./

aws s3 cp --recursive s3://dbg-cloudpipeline-data-us-west-2-prod/ref_
data/sequoia_analysis/latest/mm10.tar.gz ./

aws s3 cp --recursive s3://dbg-cloudpipeline-data-us-west-2-prod/ref_
data/sequoia_analysis/latest/rnor6.tar.gz ./

tar xvzf hg38.tar.gz

tar xvzf mm10.tar.gz

tar xvzf rnor6.tar.gz

md5sum -c ./*/*.chk
```

Appendix B Full Process Example

Refer to the following illustration to see all Toolkit steps and commands for processing the FASTQ files into analysis data.



Bio-Rad
Laboratories, Inc.

Life Science
Group

Web site bio-rad.com **USA** 1 800 424 6723 **Australia** 61 2 9914 2800 **Austria** 43 1 877 89 01 177 **Belgium** 32 (0)3 710 53 00 **Brazil** 55 11 3065 7550
Canada 1 905 364 3435 **China** 86 21 6169 8500 **Czech Republic** 420 241 430 532 **Denmark** 45 44 52 10 00 **Finland** 358 09 804 22 00
France 33 01 47 95 69 65 **Germany** 49 89 31 884 0 **Hong Kong** 852 2789 3300 **Hungary** 36 1 459 6100 **India** 91 124 4029300
Israel 972 03 963 6050 **Italy** 39 02 216091 **Japan** 81 3 6361 7000 **Korea** 82 2 3473 4460 **Mexico** 52 555 488 7670 **The Netherlands** 31 (0)318 540 666
New Zealand 64 9 415 2280 **Norway** 47 23 38 41 30 **Poland** 48 22 331 99 99 **Portugal** 351 21 472 7700 **Russia** 7 495 721 14 04
Singapore 65 6415 3188 **South Africa** 27 (0) 861 246 723 **Spain** 34 91 590 5200 **Sweden** 46 08 555 12700 **Switzerland** 41 026 674 55 05
Taiwan 886 2 2578 7189 **Thailand** 66 2 651 8311 **United Arab Emirates** 971 4 8187300 **United Kingdom** 44 020 8328 2000

10000124694 Ver A US/EG Sig 0120