SEQuoia Express Analysis Toolkit

User Guide

Version 1.0

BIO-RAD

SEQuoia Express Analysis Toolkit

User Guide

Version 1.0

BIO-RAD

Bio-Rad Technical Support Department

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.

Legal Notices

Revision History

Document	Date	Description of Change
SEQuoia Express Analysis Toolkit, User Guide DIR No. 10000154645 Ver A	June 2022	Create new document (software version 1.0)

Table of Contents

Chapter 1 Introduction

The Bio-Rad SEQuoia Express Analysis Toolkit is a Linux command line tool that processes FASTQ files as input for secondary analysis, and then produces BAM files count matrices, and reports as downstream output for tertiary analysis.

This document describes how to use the SEQuoia Express Analysis Toolkit with the SEQuoia Express Stranded RNA Library Prep Kit, and provides the necessary information and commands to analyze SEQuoia Express Stranded RNA-Seq data.

Note: For information on obtaining references for future analysis, see Appendix 1, Downloading the Reference Genome.

System Requirements

Table 1 specifies the requirements for installing and running the SEQuoia Express Analysis Toolkit.

Table 1. System requirements

Component	Minimum	Recommended
Operating system	Ubuntu OS 16.04 or higher	Ubuntu OS 16.04 or higher
Processors	8 cores	16 cores 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 run the `--max_cpus` and `--max_memory` commands to fully utilize its capabilities. For information, see Executing the Pipeline on page 9.

Software Requirements

To set up and run the SEQuoia Express Analysis Toolkit. you must install the software specified in Table 2.

Table 2. Software requirements

Software application	Minimum version	Recommended version
Docker, Community Edition https://www.docker.com/get-started	v18.08.7 or higher	v18.08.7 or higher
Nextflow https:/www.nextflow.io	v20.10.0 or higher	v20.10.0.5430 or higher

Note: The software for the SEQuoia Express Analysis Toolkit is packed into a Docker container, so Docker must be installed and running.

Chapter 2 Using the Toolkit

The SEQuoia Express Analysis Toolkit is designed to use FASTQ files to process samples through the pipeline using a directory of one or more sequencing files. To view an illustration of the workflow, see Understanding the Output Step Workflow on page 15.

Executing the Pipeline

Use the information and commands in this section to launch the SEQuoia Express Analysis Toolkit pipeline.

Note: This document assumes that the Docker container application and Nextflow pipeline workflow application are installed and running. Using the Ubuntu Terminal interface is recommended, but you can use any command line that supports UNIX or Linux.

To access user assistance for Nextflow, run the following command:

```
nextflow run ~/Sequoia_express_toolkit/main.nf --help
```

Important: If your system was set up with higher than minimum requirements, you must run the command parameters specified in Table 3, with appropriate corresponding values, to fully utilize your system capabilities.

Table 3. System parameters

Parameter	Description
`--max_cpus`	Enter the number of local system cores to be used.
`--max_memory`	Enter the total local system RAM to use for the analysis.

To launch the toolkit pipeline

1. Open the command terminal interface and run the following command:

   ```
   nextflow run
   ```

2. Enter commands to launch the pipeline toolkit and specify genome and storage parameters.

   Command line code strings should be similar to the following example, which shows a basic run with parameters hightlighted:

   ```
   nextflow run Sequoia_express_toolkit/main.nf --outDir ./output/ --reads
   '~/read/express/' --genome hg38 --genomes_base ./genomes/
   ```

Although additional default pipeline parameters are available to get started, and can be removed as needed, the parameters specified in Table 4 are required.

Table 4. Required parameters and values

Parameter	Description	Default or example value
`--reads`	Path to the directory containing the FASTQ files to be read.	`~/read/express`
`--genome`	Enter the genome to use for the analysis .	Choose from `hg38`, `mm10`, `rnor6`, `tair10`, `sacCer3`, `dm6`, `danRer11`, or `ce11`:
`--genome_base`	Path to the genomes directory; this is the parent directory.	`~/genomes/` **Note:** The genomes directory contains the downloaded and unpacked genomes from Dropbox.

Use the parameters in Table 5 to configure outputs as needed.

Table 5. Advanced parameters and values

Parameter	Description	Default value
`--w`	The directory where Nextflow should store temporary files for the pipeline.	`./work`
`--fivePrimeQualCutoff`	Enter the read quality below which bases will be trimmed on the 5' end.	`0` through `42`

Table 5. Advanced parameters and values, continued

Parameter	Description	Default value
`--minBp`	Enter the value at which reads with fewer base pairs will be rejected.	`0` through `500` The default value is `15`.
`--minGeneCutoff`	Enter the cutoff double value to indicate the minimum number of reads required for a gene to be counted.	Value depends on selection for `minGeneType`.
`--minGeneType`	Enter a metric for quantifying gene expression and filtering the output of the Reads counts for downstream usage.	Choose from `None, reads, RPKM, or TPM`.
`--minMapqToCount`	Minimum MapQ score for an aligned read to count toward a feature count.	`0` through `255`
`--noTrim`	Indicates whether or not trimming is skipped on the reads	`True` or `False` Default value is `False`.
`--outDir`	Indicates the results directory folder as the output directory where results are written.	`./results`
`--reverseStrand`	Indicates that your library is reverse stranded.	`True` or `False` Default value is `False`.
`--seqType`	Sequencing method used.	`SE` (single-end) or `PE` (paired-end)
`--skipUmi`	If `True`, deduplication of reads will not occur. If `False` (default) UMIs are tagged and deduplication occurs. **Note:** Deduplication is available for PE (paired-end) only.	`True` or `False` Default value is `False`.
`--spikeType`	The type of spike-in samples; none (default) or ercc.	`none` (default) or `ercc`
`--threePrimeQualCutoff`	Read quality below which bases are trimmed on the 3' end	`0` throught `42`

Table 5. Advanced parameters and values, continued

Parameter	Description	Default value
`--validateInputs`	Ensures that input meets the standards and is below 500 million reads.	`True` or `False` Default value is `True`.

Inputs

Following are examples of g-zipped (.gz) FASTQ input files in a data directory:

```
test_set/
├── NS4_S4_L001_5M_R1_001.fastq.gz
└── NS4_S4_L001_5M_R2_001.fastq.gz
```

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 concatenation commands in the concat directory:

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

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

  Important: Input file names must contain R1 or R2 to indicate the specified file.

Note: This step is not required when you are using the SeqSense Analysis Solution web application. After the file upload, reads that contain different lanes, but the same sample name and read number (R1/R2), are automatically merged to a SampleName_L00C_R*.fastq.gz file.

Outputs

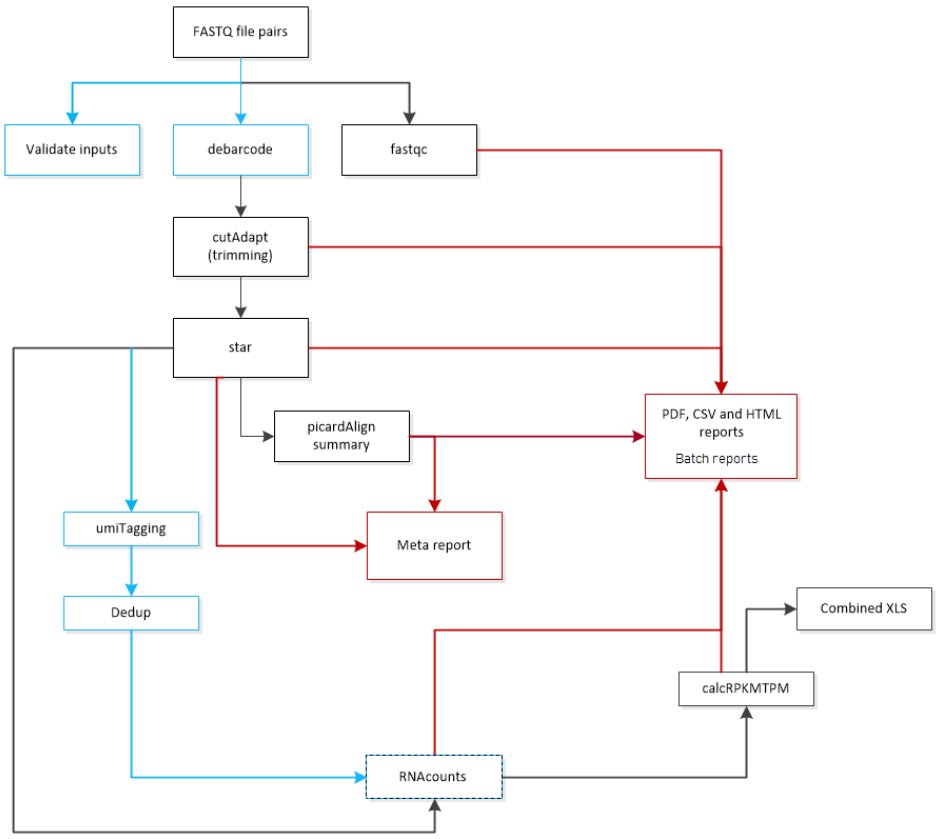
The output structure of the SEQuoia Express Analysis Toolkit is listed alphabetically, as shown below:

```
├── pipeline_info
│   ├── execution_report.html
│   ├── execution_timeline.html
│   ├── execution_trace.txt
│   └── pipeline_dag.dot
├── report
│   ├── batch_summary.csv
│   ├── batch_summary.html
│   ├── batch_summary.pdf
│   ├── NS4_S4_L001_5M_csvReport.csv
│   ├── NS4_S4_L001_5M_htmlReport.html
│   └── NS4_S4_L001_5M_pdfReport.pdf
└── Sample_Files
    └── NS4_S4_L001_5M
        ├── calcRPKMTPM
        │   ├── gene_counts_rpkmtpm.txt
        │   └── readcount_report.xlsx
        ├── cutAdapt
        │   ├── trimlog.log.NS4_S4_L001_5M
        │   ├── trimmed_R1.fastq.gz
        │   └── trimmed_R2.fastq.gz
        ├── debarcode
        │   ├── NS4_S4_L001_5M_R1_001_debarcoded.fastq.gz
        │   ├── NS4_S4_L001_5M_R1_barcode_stats.tsv
        │   └── NS4_S4_L001_5M_R2_001_debarcoded.fastq.gz
        ├── dedup
        │   ├── dedup.log.NS4_S4_L001_5M
        │   ├── rumi_dedup.sort.bam
        │   └── rumi_dedup.sort.bam.bai
```

```
├── fastqc
│   ├── NS4_S4_L001_5M_R1_001_fastqc.html
│   ├── NS4_S4_L001_5M_R2_001_fastqc.html
│   └── zips
│       ├── NS4_S4_L001_5M_R1_001_fastqc.zip
│       └── NS4_S4_L001_5M_R2_001_fastqc.zip
├── picardAlignSummary
│   └── rna_metrics.txt.NS4_S4_L001_5M
├── RNACounts
│   ├── gene_counts_longRNA.NS4_S4_L001_5M
│   ├── gene_counts_longRNA.summary.NS4_S4_L001_5M
│   └── rumi_dedup.sort.bam.featureCounts.bam
└── star
    ├── Aligned.sortedByCoord.out.bam
    ├── Aligned.sortedByCoord.out.bam.bai
    ├── Log.final.out.NS4_S4_L001_5M
    ├── Unmapped.out.mate1
    └── Unmapped.out.mate2
```

Understanding the Output Step Workflow

The following graphic illustrates the output directory structure in order of step execution. Table 6 on page 16 describes each step output.



COLOR LEGEND

- Standard execution order
- Path to report output
- Optional steps and processes

Table 6. 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), you must 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/)

star	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 6. Outputs, continued

Output directory	Description
dedup (optional)	Contains the result of PCR deduplication (deduplicated BAM file and rumi log file) based on the UMIs. Deduplication is performed using rumi with `--is_paired` and `--umi_tag` XU parameters. **(https://github.com/ sstadick/rumi)** **Note:** Applicable only if both and R1 and R2 are present, and `--skipUMI` has not been set.
RNAcounts	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.
report	Holds PDF, CSV, and HTMLversions of the assembled report, as well as batch reports.
pipeline_info	Holds graphs and reports on the runtime of each of the steps.

Logging

The `stderr` command prompts the SEQuoia Express Analysis Toolkit to output its status while running.

The Toolkit also writes to a log file (.nextflow.log) in the /work directory.

Appendix A Downloading the Reference Genome

To download the reference genome, you must use the link provided from Dropbox and name the directories appropriately. Use the following command line example in Nextflow to create the directory and download the prepared reference genome. If applicable, replace `hg38` with the genome you are using.

```
mkdir ./ref_data/genome-annotations

cd ./ref_data/genome-annotations

wget -O hg38.tar.gz
https://www.dropbox.com/s/hm6kyp70dtbqovr/hg38.tar.gz?dl=0

tar xvzf hg38.tar.gz
```

Note: After the genome is downloaded, use the `cd ~/` command to return to your default directory.

To run the analysis afterward, use the following options:

`--genome hg38` and `--genome_base /ref_data/genome-annotations/`

Bio-Rad Laboratories, Inc.

Life Science Group

Website *bio-rad.com* ***USA*** *1 800 424 6723* ***Australia*** *61 2 9914 2800* ***Austria*** *00 800 00 24 67 23* ***Belgium*** *00 800 00 24 67 23* ***Brazil*** *4003 0399*
Canada *1 905 364 3435* ***China*** *86 21 6169 8500* ***Czech Republic*** *00 800 00 24 67 23* ***Denmark*** *00 800 00 24 67 23* ***Finland*** *00 800 00 24 67 23*
France *00 800 00 24 67 23* ***Germany*** *00 800 00 24 67 23* ***Hong Kong*** *852 2789 3300* ***Hungary*** *00 800 00 24 67 23* ***India*** *91 124 4029300* ***Israel*** *0 3 9636050*
Italy *00 800 00 24 67 23* ***Japan*** *81 3 6361 7000* ***Korea*** *82 2 3473 4460* ***Luxembourg*** *00 800 00 24 67 23* ***Mexico*** *52 555 488 7670*
The Netherlands *00 800 00 24 67 23* ***New Zealand*** *64 9 415 2280* ***Norway*** *00 800 00 24 67 23* ***Poland*** *00 800 00 24 67 23* ***Portugal*** *00 800 00 24 67 23*
Russian Federation *00 800 00 24 67 23* ***Singapore*** *65 6415 3188* ***South Africa*** *00 800 00 24 67 23* ***Spain*** *00 800 00 24 67 23* ***Sweden*** *00 800 00 24 67 23*
Switzerland *00 800 00 24 67 23* ***Taiwan*** *886 2 2578 7189* ***Thailand*** *66 2 651 8311* ***United Arab Emirates*** *36 1 459 6150* ***United Kingdom*** *00 800 00 24 67 23*

10000154645 Ver A

Sig 0122