MAP-RSeq User Guide, version 1.2.1

Division of Biomedical Statistics and Informatics, Mayo Clinic

May 2014

**Contents**

1. Introduction
2. Quick Start Virtual Machine
3. Standalone System Requirements
4. Installation, Setup, and Testing
	1. System Software Pre-requisites
	2. MAP-RSeq Install
	3. Post-Install Configuration
	4. Install validation and test run with example data set
5. Step-by-step instructions to run MAP-RSeq on a user’s sample(s)
6. Using alternate reference sequences
7. Included Bioinformatics Software
8. Limitations to the workflow
9. Post-MAP-RSeq Differential Expression Analysis
10. Contact Information / Support
11. **Introduction**

MAP-RSeq, Mayo Analysis Pipeline for RNA Sequencing offers an end-to-end solution to analyze and interpret next generation RNA sequencing data. MAP-RSeq:

* Conducts quality analysis using FastQC & RSeQC.
* Aligns reads using Tophat2/Bowtie.
* Performs downstream analyses such as gene count, exon count, SNP calling. and fusion detection.
* Provides a comprehensive HTML report of all samples.

MAP-Rseq, provides two modes of execution, standalone single machine and parallel Sun Grid Engine cluster version.

Source code, executable tools and reference files are all available to download via:

<http://bioinformaticstools.mayo.edu/research/maprseq/>

1. **Quick Start Virtual Machine**

A virtual machine image is available for download at

<http://bioinformaticstools.mayo.edu/research/maprseq/>

This includes a sample dataset, references (limited to Chromosome 22), and the complete MAP-RSeq pipeline pre-installed. Please make certain that the host system meets the following system requirements:

* Oracle Virtual Box software ( free for Windows, Mac, and Linux at <https://www.virtualbox.org/wiki/Downloads> )
* At least 4GB of physical memory
* At least 10GB of available disk.

Although our sample data is on Human Chromosome 22, this virtual machine can be extended for all chromosomes and species. But this requires allocating more memory (~16GB) than may be available on a typical desktop system and building the index references files for the species of interest. If you have questions about expanding the VM please contact us for assistance.

Bockol, Matthew A. Bockol.Matthew@mayo.edu

Most recent desktops will have virtualization extensions enabled by default. Once VirtualBox is installed and the virtual machine image is downloaded you can launch the software by clicking on the MAPRSeqVM.ova file:



Click on the “Import” button to load the virtual machine:



It will appear in the list of available Virtual Machines. Clicking the green start arrow will launch the system:



Once launched the virtual machine will present instructions for starting the workflow.

1. **Standalone System Requirements**

To use MAP-RSeq you will need:

1. Linux (64-bit) workstation. We currently do not support any Windows environments. We recommend four cores with 16GB ram to get optimal performance.
2. Approximately 8GB of storage space for source, tools and reference file installation.
3. A high speed internet connection to download large reference files.
4. All of the pre-requisites outlined in the Software Requirements section below.
5. Additional storage space for analyzing input data and writing output files is recommended.
6. **Installation, Setup, and Testing**
7. **System Software Pre-requisites**

Linux distributions come with different sets of default packages installed. Your environment may be customized even further. We have tested the MAP-RSeq pipeline with Ubuntu 13.10 and Centos 6.5. The prerequisites for Centos are quite involved, but full details are outlined below. Other distributions/versions should work as well, but the packages required to satisfy the pipeline’s dependencies will differ. To begin, any distribution should include:

* + JAVA version 1.6.0\_17 or higher
	+ Perl version 5.10.0 or higher
	+ Python version 2.7 or higher

On **Ubuntu 13.10**, you must include the following packages:

* + python-dev
	+ cython
	+ python-numpy
	+ python-scipy
	+ gcc
	+ g++
	+ zlib1g-dev
	+ libncurses5-dev
	+ r-base
	+ libgd2-xpm-dev
	+ libgd-gd2-perl
	+ bsd-mailx

This list assumes an existing minimal desktop install. A full list of system packages is available in the docs/Required\_Packages\_For\_Ubuntu.txt file . Installing these packages will require root access and can be done via:

sudo apt-get install *package-name package-name package-name …*

On **Centos 6.5**, you must include the following packages:

* atlas
* atlas-devel
* cairo
* cairo-devel
* cairomm
* cairomm-devel
* cpp
* gcc
* gcc-c++
* gcc-gfortran
* gd
* gd-devel
* glib2-devel
* lapack
* lapack-devel
* bzip2-devel
* libpng
* libpng-devel
* libsigc++20
* libsigc++20-devel
* libtiff-devel
* libX11-devel
* libXext-devel
* libXft-devel
* libXt-devel
* ncurses-devel
* openssl-devel
* pango-devel
* perl-Clone
* perl-GD
* perl-HTML-Parser
* perl-Time-HiRes
* perl-IO-String
* readline-devel
* tcl
* tcl-devel
* tk
* tk-devel
* xorg-x11-server-Xvfb
* zlib-devel

This list assumes an existing minimal desktop install. A full list of system packages is available in the docs/Required\_Packages\_For\_Centos.txt file . Installing these packages will require root access and can be done via:

sudo yum install *package-name package-name package-name …*

A default Centos 6.5 install includes Python v2.6.6, but the RSeqC package requires v2.7 or higher. To satisfy this dependency, you will need to install a parallel version of Python to your system if it’s not already available.

We recommend installing Python v2.7.6 available here:

<http://www.python.org/ftp/python/2.7.6/Python-2.7.6.tgz>

To build Python, unpack the archive, configure, make, and install:

tar xfz Python-2.7.6.tgz

cd Python-2.7.6

./configure –prefix=/path/to/your/python-2.7.6

make

make install

MAP-RSeq uses a number of Python extensions, and each of these will need to be installed to your local Python as well:

Cython available at

<http://cython.org/release/Cython-0.20.1.tar.gz>

tar xfz Cython-0.20.1.tgz

cd Cython-0.20.1

/path/to/your/python-2.7.6/bin/python setup.py install

NumPy available at <http://downloads.sourceforge.net/project/numpy/NumPy/1.8.0/numpy-1.8.0.tar.gz>

tar xfz numpy-1.8.0.tar.gz

cd numpy-1.8.0

/path/to/your/python-2.7.6/bin/python setup.py install

SciPy available at

<http://downloads.sourceforge.net/project/scipy/scipy/0.13.3/scipy-0.13.3.tar.gz>

tar xfz scipy-0.13.3.tar.gz

cd scipy-0.13.3

/path/to/your/python-2.7.6/bin/python setup.py install

MAP-RSeq relies on the R statistical computing package. This is not included in Centos 6.5 and will need to be installed manually. R is available at:

<http://cran.us.r-project.org/src/base/R-3/R-3.0.2.tar.gz>

tar xfz R-3.0.2.tar.gz

cd R-3.0.2

./configure –prefix=/path/to/your/R-3.0.2

make

make install

Once these packages have been installed you will need to configure your account to use them. The simplest way to do this is to update your account’s PATH variable. You can do this by appending the following lines to your ~/.bashrc file:

PATH=/path/to/your/python-2.7.6/bin:/path/to/your/R-3.0.2/bin:$PATH

export PATH

Log out of your account, and when you re-connect the new version of Python and R should be available.

1. **MAP-Rseq Install**

The standalone MAP-RSeq package contains an install.pl script which unpacks and builds an included copy of all the required bioinformatics tools that the pipeline relies upon. The list and sources for these tools are detailed in the “Included Bioinformatics Software” section below. The install.pl script also pre-configures the pipeline to execute a run against the included sample dataset.

To install the workflow on an existing server or cluster environment, download

<https://s3-us-west-2.amazonaws.com/mayo-bic-tools/maprseq/maprseq-1.2.1.tgz>

Steps to run the installer:

1. Unpack the file, it will create a *MAPRSeq\_VERSION* directory

tar -zxvf MAPRSeq\_VERSION.tar.gz
2. Change to *MAPRSeq\_VERSION* directory
3. Execute install script

./install.pl --prefix=/PATH/TO/INSTALL\_DIR

Note. Be sure *INSTALL\_DIR* exists before running *install.pl*.
4. Install script will perform following tasks:
	1. Unpack and install *src* directory
	2. Unpack and install *lib* directory
	3. Unpack and install *bin* directory
	4. Unpack and install *sample\_data* directory
	5. Unpack and install *config* directory
	6. Unpack and install *references* directory
	7. *logs* directory contains *stderr* and *stdout* for each tool installed
5. **Post-Install Configuration**

The *config* directory in your *INSTALL\_DIR* will contain the following files:

memory\_info.txt

run\_info.txt

sample\_info.txt

tool\_info.txt

These will be pre-configured for your environment, but in some cases you will need to manually update them prior to a run.

We provide the check\_install script to confirm that all the tools are properly available. It reads the tool\_info.txt file and tests. From the *INSTALL\_DIR*  you can execute it with:

./check\_install –toolinfo config/tool\_info.txt

When check\_install finds a dependency it cannot execute properly it will prompt you to supply the correct path to the particular program or library. Once complete it will update the tool\_info.txt file with the new values.

Troubleshooting failed installs can be difficult. Please see the *INSTALL\_DIR/logs* directory for possible causes. If you have issues, please feel free to contact us at bockol.matthew@mayo.edu for help diagnosing the problem.

1. **Install validation and test run with example data set**

To test your install as a standalone single box run, execute the MAP-RSeq workflow with the test data provided:

/INSTALL\_DIR/src/mrna.pl -r=/INSTALL\_DIR/config/run\_info.txt

If you have access to SGE, and would like to run MAP-RSeq as a cluster job, edit the following parameters to match your SGE environment:

*tool\_info.txt* under *INSTALL\_DIR/config* set:

 STANDALONE=NO

 QUEUE=*SGE queue you have access to submit jobs*

 GATK\_QUEUE=*SGE queue you have acces to submit jobs*

**Navigate results**

Upon successful completion of the test run, you will receive an email notification stating that the workflow has completed and results are ready. The results from the test run are stored in following folder structure:

 INSTALL\_DIR/sample\_output/USERNAME/mrnaseq/test

 | \_ alignment

 | \_ tophat\_SAMPLENAME

 | \_ accepted\_hits.bam

 | \_ unmapped.bam

 | \_ SAMPLENAME\_sorted.bam

 | \_ SAMPLENAME.samtools.flagstat

 | \_ SAMPLENAME.flagstat

 | \_ prep\_reads.info

 | \_ fusion

 | \_ tophat\_fusion\_report.txt

 | \_ circos\_fusion\_all.png

 | \_ result.html

 | \_ potential\_fusion.txt

 | \_ Reports

 | \_ GeneCount.tsv

 | \_ ExonCount.tsv

 | \_ SampleStatistics.tsv

 | \_ Main\_Document.html

The Main\_Document.html contains the summary information of the results and various links to more details about the samples and the analysis. The other files in the output directory serve as supplemental content.

1. **Step-by-Step instruction to run MAP-RSeq on a user’s sample(s)**

The MAP-RSeq workflow processes sequencing data from the Illumina sequencing platform. The workflow accepts two different types of input files:

1. fastq files (extension “.fastq”)
2. compressed fastq files (extension “.fastq.gz”)

To run MAP-RSeq on user sample(s), four configuration files need to be modified. Copy all four skeleton configuration files from INSTALL\_DIR/config/skeleton to desired location.

* Edit

run\_info.txt configuration file

* + **PAIRED=**1
	*Indicate whether the samples are paired end or not. A value of 1 means yes and a value of 0 means no.*
	+ **READLENGTH=**<read length of input sample>
	*It is important to identify read length of your input sample. If read lengths are variable use average read length.* ***Number must be a whole number and if the read length is something other than 50 or 100, you will have to change the SEGMENT\_SIZE parameter.***
	+ **PI=**<username>
	*Username of who is executing the workflow, this value is used to create distinct project output folders.*
	+ **MEMORY\_INFO=**<path of memory\_info.txt config file>
	*Complete path of where memory\_info.txt file is located.*
	+ **SAMPLE\_INFO=**<path of sample\_info.txt config file>
	*Complete path of where sample\_info.txt file is located.*
	+ **TOOL\_INFO=**<path of tool\_info.txt config file>
	*Complete path of where tool\_info.txt file is located.*
	+ **INPUT\_DIR=**<path of input fastq.qz files>
	*Complete path of where input samples files are.* ***Files must be in fastq.gz or fastq format.***
	+ **BASE\_OUTPUT\_DIR=**<output dir location>
	*Complete path of output folder.*
	+ **OUTPUT\_FOLDER=**<output dir name>
	*Name of output folder.*
	+ **SAMPLENAMES=**<sampleName>:[sampleName2]:…:[sampleNameN]
	*Sample names to be processed. It’s a colon (:) delimited list. Sample names must be the exact same as listed in sample\_info.txt.*
	+ **LABINDEXES=**<lab index>:[lab index2]:…:[lab indexN]
	*Adaptor metadata for each sample. If not available use dashes (-) for each sample. Also a colon (:) delimited list.*
	+ **LANEINDEX=**<lane>:[lane2]:…:[lane3]
	*Lane metadata indication which lane each sample was sequenced. If not available use dashes (-) for each sample. Also a colon (:) delimited list.*
	+ **FASTQC=**<Yes/No>
	Indicate if FASTQC module of the workflow should be executed.
	+ **USE\_SUBREAD\_FEATURECOUNTS=**No

Whether to use subread’s featureCount tool instead of HTSeq-count (default)

* + **CENTER=**<meta data>
	+ **PLATFORM=**<meta data>
	+ **SAMPLEINFORMATION=**<metadata>
* Create

 sample\_info.txt configuration file

* + Sample alias/short name followed by “=” (equal to) sign followed by R1 and R2 separated by tab. R1 and R2 must be exact file name of the files specified in **INPUT\_DIR** of run\_info.txt file.
	+ Each line must contain single read pair.

A two sampled paired end fastq analysis:

sampleA=NameOf\_SampleA\_Read1.fastq NameOf\_SampleA\_Read2.fastq

sampleB=NameOf\_SampleB\_Read1.fastq NameOf\_SampleB\_Read2.fastq

* Edit

tool\_info.txt configuration file

* + The tool\_info.txt file has been created based on the installation parameters passed when install.pl was executed.
	+ For the most part this file does not need any changes and can be simply copied to desired location.
* Edit

memory\_info.txt configuration file

* + The memory\_info.txt file has been created based on over one hundred runs to extract optimal performance from the workflow. If the execution system does not meet recommended hardware specifications, you may need to edit memory\_info.txt to adjust Java Vitual Memory request respectively.
	+ For the most part this file does not need any changes and can be simply copied to desired location.

**Description of the identifiers in run\_info.txt configuration file**

|  |  |  |
| --- | --- | --- |
| **Identifier**  | **Format** | **Description** |
| **TOOL** | **MAPRSeq** | Name of the tool. |
| **VERSION** | **1.2** | Version number. |
| **TYPE** | **RNA** | To create output folder structure. |
| **ALIGNER** | **Tophat** | Type of aligner (only one supported currently). |
| **ANALYSIS** | **All** | Run complete or part of workflow. |
| **PAIRED** | **1** | Only paired end is supported at the moment. |
| **READLENGTH** | 100 | Input number of bases of each sequence from FASTQ. If not uniform provide average length. |
| **DISEASE** | Cancer | Provide metadata about the samples. |
| **PI** | Username | Username/unique id to keep results organized by PI/Study. |
| **MEMORY\_INFO** | /path/to/file | Full path of memory\_info.txt file. |
| **TOOL\_INFO** | /path/to/file | Full path of tool\_info.txt file. |
| **SAMPLE\_INFO** | /path/to/file | Full path of sample\_info.txt file. |
| **INPUT\_DIR** | /path/to/input data | Location of all input FASTQ files. Must be a single directory. |
| **BASE\_OUTPUT\_DIR** | /path/to/output | Base location where output will be stored. |
| **OUTPUT\_FOLDER** | Output folder name | Output folder name.  |
| **SAMPLENAMES** | sampleA:sampleB  | Sample aliases delimited by colon (:) as indicated in sample\_info.txt file. |
| **LANEINDEX** | 1:2 | Metadata for each sample. One per sample use dash (-) if not available. List is colon (:) delimited. |
| **LABINDEX** | ABC:XYZ | Metadata for each sample. One per sample use dash (-) if not available. List is colon (:) delimited. |
| **CHRINDEX** | **1:2:3:4:..:X:Y:M** | All chr values. |
| **FASTQC** | **Yes/No** | Indicate whether to run FASTQC module of the workflow. |
| **USE\_SUBREAD\_FEATURECOUNTS** | **Yes/No** | Indicate whether to use subread for gene feature counts (Yes) or the default HTSeq-count (No) |
| **CENTER** | Mayo | Provide Metadata. |
| **PLATFORM** | Illumina | Provide Metadata |
| **GENOMBUILD** | hg19 | Provide Metadata |
| **SAMPLEINFORMATION** | Sample meta data | Provide Metadata |
|  |  |  |

**Description of the identifies in tool\_info.txt configuration file**

|  |  |  |
| --- | --- | --- |
| **Identifier**  | **Format** | **Description** |
| **STANDALONE** | Yes/No | Indicate if running work in grid environment or on a single machine. |
| **QUEUE** | q-name | Queue name to submit jobs if running in SGE grid environment. |
| **GATK\_QUEUE** | q-name | Queue name to submit jobs if running in SGE grid environment. Can be same as QUEUE. |
| **NOTIFICATION\_OPTION** | Queue options | Notification options flag for SGE queue master. |
| **FUSION** | BLANK/non-human | If input sample is human leave this blank. Otherwise indicate non-human. Be sure to change reference files if input samples are non-human. |
| **SEGMENT\_SIZE** | 25 | General rule of thumb is to keep this value about half the input read length if read length <= 50 or 25 if read length >= 100. This is an important value to be set for Tophat to run successfully. |
| **MAX\_HITS** | 20 | Default Tophat parameter values.  |
| **INSERT\_SIZE** | 50 | Default Tophat parameter values. *Will be updated at run time.* |
| **MATE\_SD** | 20 | Default Tophat parameter values. *Will be updated at run time.* |
| **FUSION\_MIN\_DIST** | 50000 | Default Tophat parameter values.  |
| **GATK\_UG\_PARAM** | Undefined | Unified Genotyper Options |
| **GATK\_VQSR\_FEATURES** | ReadPosRankSum:FS | Annotations to use |
| **GATK\_VQSR\_GAUSSIANS** | 4 | The maximum number of Gaussians to try during variational Bayes algorithm |
| **GATK\_VQSR\_PCT\_BAD\_VARIANTS** | 0.05 | What percentage of the worst scoring variants to use when building the Gaussian mixture model of bad variants. 0.07 means bottom 7 percent. |
| **GATK\_VQSR\_TRENCH** | 99.0 | The truth sensitivity level at which to start filtering, used here to indicate filtered variants in the model reporting plots |
| **GATK\_HARD\_FILTERS\_EXP** | "FS > 20.0":"ED > 5":"ReadPosRankSum < -8.0":"ReadPosRankSum >-8.0" | Filter expressions |
| **GATK\_HARD\_FILTERS\_NAMES** | FSFilter:EDFilter:RPRSFilter:RPRSFilter | Filter names |

**In built QC in the workflow**

* For each step of the workflow we validate the input file and if there is a discrepancy the user will get an email and the workflow is paused at that stage. When the workflow is paused there is a file which is created with extension **\*.err** within the error directory.
* User should fix the output file and then delete the **\*.err** file and the workflow will resume.
* Example email:

Subject: Error in MAPRSeq workflow executing Scripture

There is an error executing Scripture.

Following file seem to be missing

/complete/path/to/file.name

Logs are available here

Once you have fixed the error delete the file

/complete/path/to/error/filename.err

…

1. **Using alternate reference sequences**

It's possible to use alternate references with the MAP-RSeq pipeline. You will need to provide a number of alternate files and modify the tool\_info.txt file to point at them. Below is a list of the settings to change and sources for those files.

**REF\_GENOME**

The fasta file for your reference file.

**REF\_BOWTIE**

Indexes for the REF\_GENOME fasta file generated by bowtie-build (\*.ebwt files)

**TRANSCRIPTOME\_HG19\_INDEX**

Path to the transcriptome fasta file and bowtie-build generated indexes This is labeled "HG19" but can be pointed at alternate references.

**ENSGENE**

Available from http://tophat.cbcb.umd.edu/fusion\_tutorial.html

**REF\_GENE\_BED**

Generated via UCSC Table Browser

http://genome.ucsc.edu/cgi-bin/hgTables?command=start

**REF\_SAMPLE\_BED**

A filtered version of refGene.txt provided to speed analysis with the sample data provided. Replace with the value in REF\_GENE\_BED for new analyses.

**MCL**

Gene publication references. Can be left unchanged.

**KARYOTYPE**

Provided with the Circos software. See:

 INSTALL\_DIR/bin/circos/0.64/data/karyotype/

**MASTER\_GENE\_FILE**

Bed file containing genes to include in variant calling. Generated via UCSC Table Browser

http://genome.ucsc.edu/cgi-bin/hgTables?command=start

The following files are available from TopHat at

 http://tophat.cbcb.umd.edu/igenomes.shtml

**CHROMSIZE**

A list of each chromosome in the reference and the number of bases in each.

See ChromInfo.txt in the TopHat iGenomes download.

**FEATURES**

Reference annotation GTF file.

See genes.gtf in the TopHat iGenomes download.

**REF\_FLAT**

Gene prediction table.

See refFlat.txt in the TopHat iGenomes download.

**REFGENE**

See refGene.txt in the TopHat iGenomes download.

The following files are available as part of the GATK bundle at:

 ftp://ftp.broadinstitute.org/bundle

 username: gsapubftp-anonymous

**HAPMAP\_VCF**

A list of high quality SNPs used as a filter in GATK

**OMNI\_VCF**

A list of high quality SNPs used as a filter in GATK

1. **Included Bioinformatics Software**

MAP-RSeq relies on the bioinformatics tools listed below. If your environment already has these installed you can modify the tool\_info.txt file to point at your existing copies provided the versions are compatible. Versions differing from those tested may not execute successfully.

Please note, that the RseQC and Tophat packages have been patched for use in the pipeline. Using the standard versions will cause it to fail. Details of the modification are available in the patches/ directory of the standalone distribution.

BEDTools v2.17.10

<http://code.google.com/p/bedtools/downloads/detail?name=BEDTools.v2.17.0.tar.gz>

UCSC Blat’s faToTwoBit and wigToBigWig

<http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/faToTwoBit>

<http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/wigToBigWig>

Bowtie v0.12.9.0

<http://downloads.sourceforge.net/project/bowtie-bio/bowtie/0.12.9/bowtie-0.12.9-linux-x86_64.zip>

Circos v0.64

<http://circos.ca/distribution/circos-0.64.tgz>

FastQC v0.10.1

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.10.1.zip>

GATK v1.6.9

ftp://ftp.broadinstitute.org/distribution/gsa/GenomeAnalysisTK/GenomeAnalysisTK-1.6-9-g47df7bb.tar.bz2

HTSeq v0.5.3p9

<https://pypi.python.org/packages/source/H/HTSeq/HTSeq-0.5.3p9.tar.gz>

SubRead 1.4.4

<http://downloads.sourceforge.net/project/subread/subread-1.4.4/subread-1.4.4-source.tar.gz>

Picard Tools v1.92

<http://downloads.sourceforge.net/project/picard/picard-tools/1.92/picard-tools-1.92.zip>

RSeQC v2.3.7 (customized)

<http://downloads.sourceforge.net/project/rseqc/RSeQC-2.3.7.tar.gz>

We have applied patches to this version. Please use the files included with the installer.

Samtools v0.1.19

<http://downloads.sourceforge.net/project/samtools/samtools/0.1.19/samtools-0.1.19.tar.bz2>

TopHat v2.0.6

<http://tophat.cbcb.umd.edu/downloads/tophat-2.0.6.Linux_x86_64.tar.gz>

These packages are included pre-built in the MAP-RSeq install, but you can point the workflow at your own versions via the tool\_info.txt file as needed.

1. **Limitations to the workflow**
* Sample names cannot start with a number or a special character. For example, characters such as “( ){ }[ ] . , $-” are not permitted.
* The workflow does not run in any Windows environment.
1. **Post-MAP-RSeq Differential Expression Analysis**

Included in the contrib/ directory is a set of scripts you can use to perform

differential expression analysis against two groups of samples processed with

the MAP-RSeq pipeline. Please see the documentation at

 contrib/Differential\_Expression/README.txt

for full details on using the scripts.

1. **Contact Information / Support**

If you have questions or need assistance using the MAP-RSeq workflow, please feel free to contact bockol.matthew@mayo.edu .