MAYO CLINIC

CAP-miRSeq User Guide

A comprehensive analysis pipeline for deep microRNA sequencing

3/9/2014

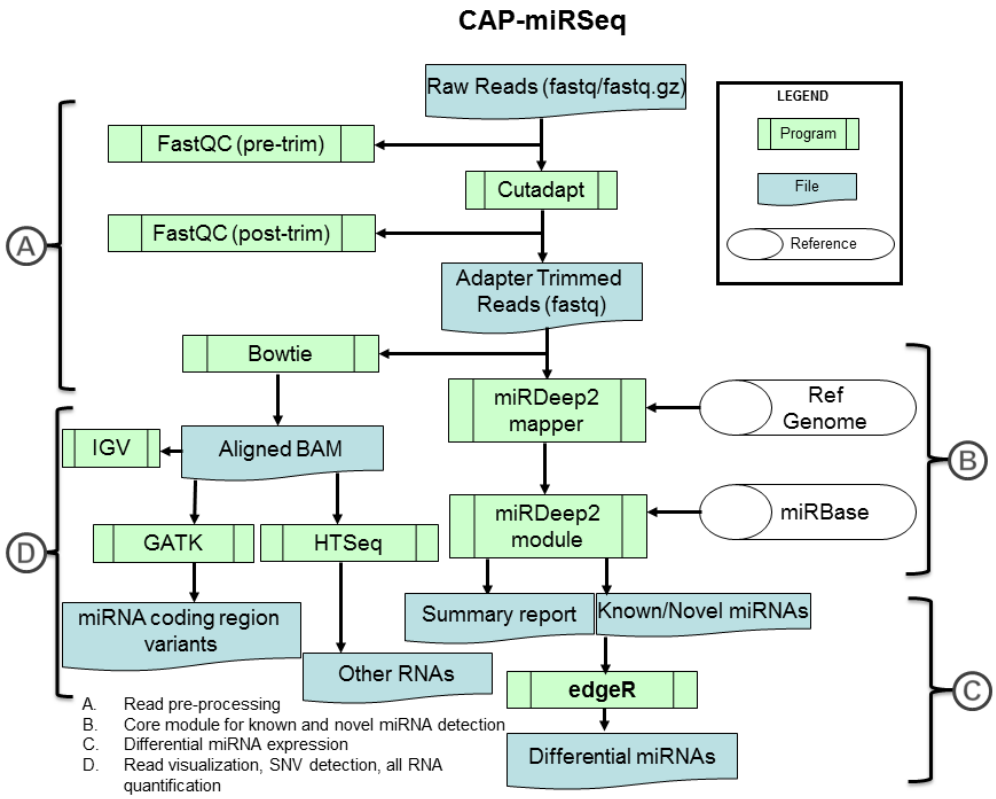
Table of Contents

1. Introduction

miRNAs play a key role in normal physiology and various diseases such as cancer. Hybridization based microarray technology has been used for miRNA profiling, but is hindered by its narrow detection range, more susceptibility to technical variation, and lack of ability to characterize novel miRNAs and sequence variation. miRNA profiling through next generation sequencing overcomes those limitations and provides a new avenue for biomarker discovery and clinical applications. However, analyzing miRNA sequencing data is challenging. Significant amount of computational resources and bioinformatics expertise are needed. Several analytical tools have been developed over the past few years; however most of these tools are web-based and can only process one or a pair of samples at time, which is not suitable for a large scale study with tens or even hundreds of samples. Lack of flexibility and reliability of the web service (such as outdated references, unknown parameters used, server down, and slow performance) are also common issues. Although some tools provide differential miRNA analysis, they either limit to a pair of samples or use a model not suitable to a study design. Moreover, miRNA SNVs or mutations become increasingly important but none of the tools provide SNV/mutation detection. Herein, we present a comprehensive analysis pipeline for deep microRNA sequencing (CAP-miRSeq) that integrates read preprocessing, alignment, mature/precursor/novel miRNA qualification, variant detection in miRNA coding region, and flexible differential expression between experimental conditions. According to computational infrastructures, users can run samples sequentially or in parallel for fast processing. In either a case, summary and expression reports for all samples are generated for easier quality assessment and downstream analyses. Using well characterized data, we demonstrated the pipeline's superior performances, flexibilities, and practical use in research and biomarker discovery.

The package can be downloaded from the CAP-miRSeq website:
http://bioinformaticstools.mayo.edu/research/cap-mirseq/

The CAP-miRSeq analysis workflow

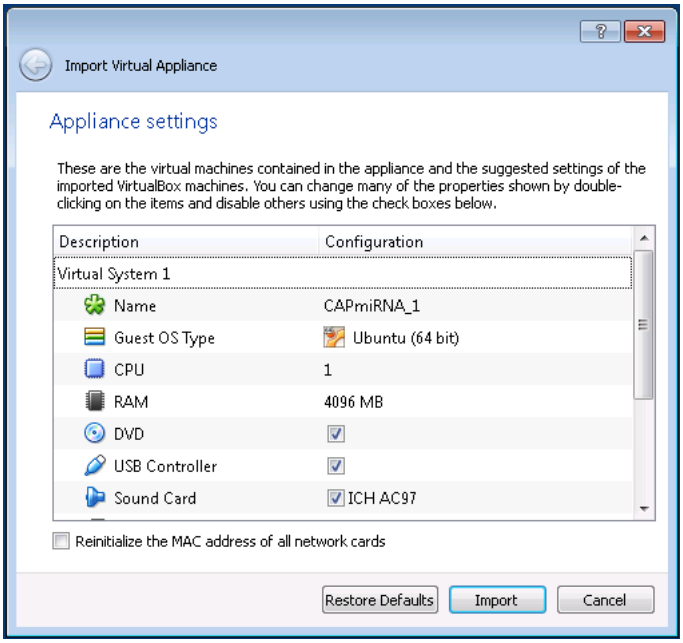


2. Quick Start Virtual Machine

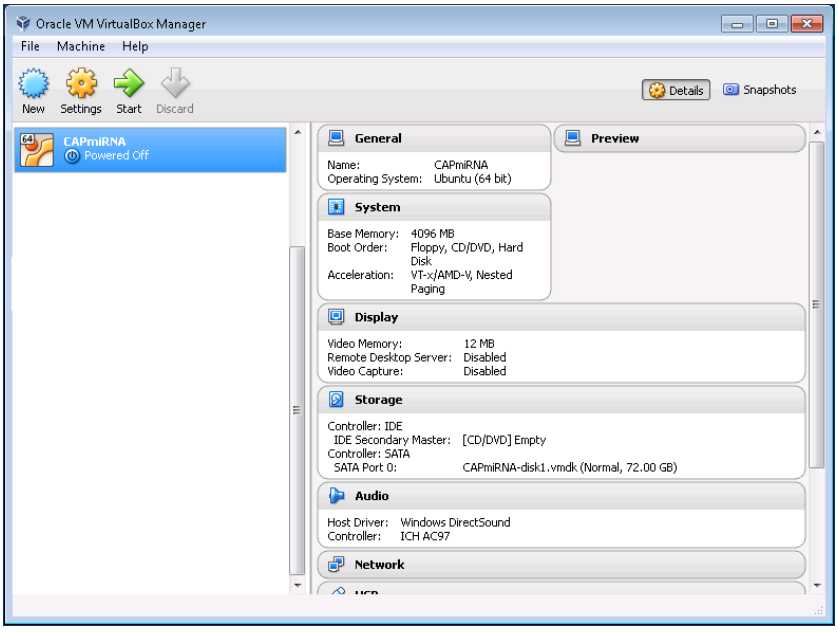
Along with the full CAP-miRSeq local installation package, we also provide a Virtual Machine (VM) image to allow users the opportunity to quickly test the workflow or run a limited number of samples without a need or environment for parallelization. . The VM image even includes some test samples and reference files. The VM version can be run on a Windows, Mac, or Linux machine with at least 8GB of RAM and 10GB of free disk space. Follow these steps to run the VM version of CAP-miRSeq:

1. Download and install the free tool VirtualBox which will allow you to run the VM image: https://www.virtualbox.org/
2. Download the CAP-miRSeq VM image: https://s3-us-west-2.amazonaws.com/mayo-bic-tools/cap-mirseq/CAP-miRNA-VM.ova
3. Double click on the OVA file and VirtualBox will launch.

4. Import the VM (popping up a few windows that you can click next through)



5. Click the green arrow at the top to start and the VM will launch into the desktop with brief instructions on how to start CAP-miRSeq. Note: you may need to reconfigure the network interface NAT to start the VM properly.



Once the Virtual Machine has started you will be presented with brief instructions for launching the workflow against the included sample data. You can use this system to run real sample data, but this requires updating the reference files as appropriate (only Chromosome 1 is used in the sample data) and you may need to increase the available memory. The included test dataset takes approximately 4 hours to process on a 3GHz Intel i5 desktop system.

3. Local Installation to your system

Please see the "README.txt" file included with the CAP-miRSeq source distribution (CAP-miRSEQ.tgz) for a step by step guide to installing all the software required to run the pipeline. This installation allows you to run the pipeline more efficiently and take advantage of parallel computing on an SGE cluster with multiple samples running simultaneously. After successful installation and test-run of the included example data, you may follow the instructions below to start analyzing your own samples.

4. Reference Files

In order to run CAP-miRSeq, a local copy of the following genome and miRNA reference files are required.

1. Reference genome assembly sequences in FASTA format (.fai, .dict, and bowtie index files are also needed, but if they don't exist then CAP-miRSeq will automatically generate the indexes)
2. miRBase known miRNA references:
    a. hairpin precursor miRNA sequences in FASTA format
    b. mature miRNA sequences in FASTA format
    c. miRNA GFF3 file
3. Gencode annotation GTF file

Here are some links to where these references can be downloaded for commonly used human and mouse species:

hg19 References	**Download Link**
Reference genome assembly	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz
miRBase precursor miRNA sequences	ftp://mirbase.org/pub/mirbase/20/hairpin.fa.gz
miRBase mature miRNA sequences	ftp://mirbase.org/pub/mirbase/20/mature.fa.gz
miRBase GFF3 file	ftp://mirbase.org/pub/mirbase/20/genomes/hsa.gff3
Gencode annotation file	ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz

mm10 References	Download Link
Reference genome assembly	http://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/chromFa.tar.gz
miRBase precursor miRNA sequences	ftp://mirbase.org/pub/mirbase/20/hairpin.fa.gz
miRBase mature miRNA sequences	ftp://mirbase.org/pub/mirbase/20/mature.fa.gz
miRBase GFF3 file	ftp://mirbase.org/pub/mirbase/20/genomes/mmu.gff3
Gencode annotation file	ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_mouse/release_M2/gencode.vM2.annotation.gtf.gz

After downloading the above reference files, unzip the compressed files. The reference assembly files will be split into separate files for each chromosome. Before running CAP-miRSeq please concatenate the chromosome files into a single FASTA file:

```
cat chr1.fa chr2.fa chr3.fa chr… > hg19.fa
```

The hairpin.fa and mature.fa files downloaded from miRBase will contain all species; however, miRDeep2 requires that they only contain a single species and the RNA sequences be converted to DNA. miRDeep2 (see next section for download link) includes a couple of Perl scripts to easily convert these files. For example, with hg19 references, the following miRDeep2 commands would be run:

```
perl extract_miRNAs.pl hairpin.fa hsa > hairpin.hsa.fa
perl extract_miRNAs.pl mature.fa has mature > mature.hsa.fa
perl rna2dna.pl hairpin.hsa.fa > hairpin.hsa.dna.fa
perl rna2dna.pl mature.hsa.fa > mature.hsa.dna.fa
```

5. Running CAP-miRSeq

To run CAP-miRSeq, 3 config files must be created (described in detail below):

- run_info.txt
- sample_info.txt
- tool_info.txt

Once these 3 files are created, CAP-miRSeq can be started using the following command:

```
./CAP-miRseq.sh run_info.txt
```

If the workflow is being run on a single machine then CAP-miRseq.sh will continue running until the entire workflow is finished. If it is being run on an SGE cluster then CAP-miRSeq.sh will quickly run and submit all of the SGE jobs. In either case an email will be sent to the user upon completion.

5.1 run_info.txt

The run_info.txt file contains various parameters and information specific to the current run of CAP-miRSeq. Here is an example of a run_info.txt file:

```
TOOL=CAPmiRSeq
VERSION=1.1
PI=Firstname_Lastname
FLOWCELL=100627_R0174436_0079
GENOME_BUILD=hg19
MIRBASE_VERSION=19
EMAIL=your.name@institution.edu
SAMPLENAMES=SampleName1:SampleName2:SampleName3:SampleName4
LANEINDEX=1:2:3:4
TRIM_ADAPTER=YES
CALL_SNVS=YES
DIFF_EXPRESSION=YES
DIFF_EXPRESSION_ANALYSES=dicer_knockdown
INPUT_DIR=/path/to/fastqs/
OUTPUT_DIR=/path/to/capmirseq/output/
TOOL_INFO=/path/to/config_files/tool_info.txt
SAMPLE_INFO=/path/to/config_files/sample_info.txt
DELIVERY_FOLDER=/path/to/where/results/will/be/permanently/stored/
USE_SGE=1
```

Details of run_info.txt parameters:

run_info.txt Keys	Values	Description
TOOL	CAPmiRSeq	The name of the tool being run.
VERSION	1.1	Version # of the tool being run.
PI	Name	Name of the Principle Investigator or owner of the data.
FLOWCELL	Flowcell ID	ID of the flowcell or run that the samples were sequenced on.
GENOME_BUILD	hg19/mm10/other	Name of the reference genome being used.
MIRBASE_VERSION	20	The version of miRBase being used.
EMAIL	name@email.com	Email address of individual running CAP-miRSeq
SAMPLENAMES	S1:S2:S3	A colon separated list of the sample names
LANEINDEX	1:2:3	A colon separated list of lane numbers for each sample
TRIM_ADAPTER	YES/NO	Whether or not adapter sequences should be trimmed from fastqs
CALL_SNVS	YES/NO	Whether or not SNVs should be called using GATK's UnifiedGenotyper
DIFF_EXPRESSION	YES/NO	Whether or not differential expression should be performed using EdgeR
DIFF_EXPRESSION_ANALYSES	Name	Name of the specific differential expression analysis to perform. Can be colon separated to run multiple differential expression

		analyses
INPUT_DIR	/folder/path/	Path to location of input FASTQ files
OUTPUT_DIR	/folder/path/	Path to where output and intermediate CAP-miRSeq files will be written
TOOL_INFO	/path/tool_info.txt	Path to the tool_info.txt file for this run
SAMPLE_INFO	/path/sample_info.txt	Path to the sample_info.txt file for this run
DELIVERY_FOLDER	/folder/path/	Path to where the results should be stored
USE_SGE	1/0	Whether or not to run the Sun Grid Engine version of CAP-miRSeq. 1=YES, 0=NO

5.2 sample_info.txt

The sample_info.txt config file contains sample-specific information about the samples being analyzed by CAP-miRSeq. Here is an example of a sample_info.txt file:

```
SampleName1=SRR326279.sra.fastq
SampleName2=SRR326280.sra.fastq
SampleName3=SRR326281.sra.fastq
SampleName4=SRR326282.sra.fastq

dicer_knockdown:SAMPLES=SampleName1:SampleName2:SampleName3:SampleName4
dicer_knockdown:GROUPS=control:control:knockdown:knockdown
dicer_knockdown:PAIRS=cytoplasmic:total:cytoplasmic:total
```

Details of sample_info.txt parameters:

sample_info.txt Keys	**Values**	**Description**
SampleName	FASTQ file name	The sample name should be the exact sample name used in the run_info.txt file and the FASTQ file name is the name of the FASTQ found in the INPUT_DIR specified in the run_info.txt. This row should be repeated for as many samples in your analysis. FASTQ files can be in either .fastq or .fastq.gz formats.
name:SAMPLES	S1:S2:S3:S4	A colon separated list of the samples being used in this differential expression. The differential expression name is defined before the :SAMPLES.
name:GROUPS	G1:G1:G2:G2	A colon separated list of the groups that the above samples belong to. The order is very important. In this example S1 and S2 both belong to the first group and S3 and S4 belong to the second

		group. There should only be 2 groups defined.
name:PAIRS	P1:P2:P1:P2	This is an optional parameter that can be defined if samples across the groups are paired. For example, this parameter could be used if S1 and S3 are the same sample, but at different timepoints.

5.3 tool_info.txt

The tool_info.txt config file defines the paths to tools and reference files used by CAP-miRSeq, as well as tool parameters and memory usage information. The check_install script can be supplied with a template tool_info.txt file and will guide you through populating the values. Here is an example of a tool_info.txt file:

```
## Reference Files
REF_GENOME=/path/to/hg19.fa
BOWTIE_REF=/path/to/hg19
MIRBASE_HAIRPIN=/ path/to/mirbase/v19/hairpin.hsa.dna.fa
MIRBASE_MATURE=/path/to/mirbase/v19/mature.hsa.dna.fa
MIRBASE_GFF=/path/to/mirbase/v19/hsa.gff3
GENCODE_GTF=/path/to/gencode.v18.annotation.gtf

## Tool Paths
SCRIPT_PATH=/path/to/capmirseq/code/
MIRDEEP2_PATH=/path/to/mirdeep/2.0.0.5/
BOWTIE_PATH=/path/to/bowtie/0.12.7/
RANDFOLD_PATH=/path/to/mirdeep/2.0.0.5/essentials/randfold-2.0/
SQUID_PATH=/path/to/mirdeep/2.0.0.5/essentials/squid-1.9g/
VIENNA_PATH=/path/to/mirdeep/2.0.0.5/essentials/ViennaRNA-
1.8.4/install_dir/bin/
PDFAPI2_PM_PATH=/path/to/mirdeep/2.0.0.5/lib/lib/perl5/
JAVA_PATH=/usr/java/latest/bin/
PICARD_PATH=/path/to/picard/1.77/
FASTQC_PATH=/path/to/FastQC
CUTADAPT_PATH=/path/to/cutadapt/0.9.5/
SAMTOOLS_PATH=/path/to/samtools/samtools-0.1.18
BEDTOOLS_PATH=/path/to/BEDTools/2.15.0/bin/
GATK_JAR=/path/to/GenomeAnalysisTK/2.7-2-g6bda569/GenomeAnalysisTK.jar
VCFTOOLS_PATH=/path/to/vcftools/0.1.9/bin/
VCFTOOLS_PERLLIB=/path/to/vcftools/0.1.9/perl/
HTSEQ_PATH=/path/to/htseq/0.5.3p9/bin/
HTSEQ_LIB_PATH=/path/to/htseq/0.5.3p9/lib/python2.7/site-packages/
PYTHON_PATH=/path/to/python/2.7.3/bin/

## Tool Parameters
CUTADAPT_PARAMS=-b AATCTCGTATGCCGTCTTCTGCTTGC -O 3 -m 17 -f fastq
MAPPER_PARAMS=-e -h -q -m -r 5 -u -v -o 4
MIRDEEP2_PARAMS=-P -t Human
MIRDEEP2_CLOSE_SPECIES=none
QUANTIFIER_PARAMS=-P -W
BOWTIE_PARAMS=-p 4 -S -q -n 1 -e 80 -l 30 -a -m 5 --best --strata
ADDORREPLACEREADGROUPS_PARAMS=MAX_RECORDS_IN_RAM=1800000
VALIDATION_STRINGENCY=SILENT RGLB=hg19 RGCN=Mayo RGPL=Illumina
SORTSAM_PARAMS=MAX_RECORDS_IN_RAM=1800000 VALIDATION_STRINGENCY=SILENT
```

```
UNIFIEDGENOTYPER_PARAMS=-glm SNP -dcov 1000
HTSEQ_PARAMS=-m intersection-nonempty -q -t exon -s no
QUEUE=1-day

## Memory Parameters
# QSUB
REFERENCE_INDEXES_MEM=-l h_vmem=3G -l h_stack=10M
CUTADAPT_MEM=-l h_vmem=2G -l h_stack=10M
FASTQC_MEM=-l h_vmem=3G -l h_stack=10M
BAMS_MEM=-l h_vmem=2G -l h_stack=10M
MIRDEEP2_MAPPER_MEM=-l h_vmem=1G -l h_stack=10M
MIRDEEP2_MEM=-l h_vmem=2G -l h_stack=10M
VARIANTS_MEM=-l h_vmem=3G -l h_stack=10M
EXPRESSION_REPORTS_MEM=-l h_vmem=1G -l h_stack=10M
DIFF_EXPRESSION_MEM=-l h_vmem=2G -l h_stack=10M
GENCODE_CLASSIFICATION_MEM=-l h_vmem=3G -l h_stack=10M
SAMPLE_SUMMARY_MEM=-l h_vmem=1G -l h_stack=10M
MAIN_DOC_MEM=-l h_vmem=1G -l h_stack=10M
# JVM
ADDORREPLACEREADGROUPS_JVM_MEM=-Xmx512m -Xms512m
SORTSAM_JVM_MEM=-Xmx1g -Xms512m
UNIFIEDGENOTYPER_JVM_MEM=-Xmx512m -Xms512m
CREATEDICTIONARY_JVM_MEM=-Xmx512m -Xms512m
```

To use the check_install script, either create a new tool_info file from the template above or use one of the examples provided in the sample_config directory. Run:

scripts/check_install –t tool_info.csv

The script will example each of the tool_info properties and confirm that the binaries, libraries, and reference files are all in place as specified. When not found, it will give you the opportunity to set the correct value and peform a check on what you provide. Once complete, an updated tool_info file will overwrite the one you specified and a summary of any missing configuration options will be displayed. You can re-run check_install on the same file repeatedly until all requirements are satisfied.

Details of tool_info.txt parameters:

tool_info.txt Keys	Default Values	Description
REF_GENOME		Full path to reference sequence assembly in FASTA format
BOWTIE_REF		Full path to the reference assembly used by BOWTIE. This can be the same as the REF_GENOME path, with the .fa extension removed as required by BOWTIE
MIRBASE_HAIRPIN		A FASTA file containing the miRBase DNA precursor miRNA sequences for a single species (see the Reference Files section of this manual for instructions to create this file)

MIRBASE_MATURE		A FASTA file containing the miRBase DNA mature miRNA sequences for a single species (see the Reference Files section of this manual for instructions to create this file)
MIRBASE_GFF		The miRBase GFF3 file for a single species
GENCODE_GTF		Gencode GTF file
SCRIPT_PATH		Full path to the CAP-miRSeq code directory
MIRDEEP2_PATH		Full path to miRDeep2 code directory
BOWTIE_PATH		Full path to directory containing the BOWTIE binaries
RANDFOLD_PATH		Full path to directory containing randfold binary
SQUID_PATH		Full path to squid code directory
VIENNA_PATH		Full path to vienna binary directory
PDFAPI2_PM_PATH		Full path to the location of the PDFAPI2 Perl module
JAVA_PATH		Full path to directory containing Java binary
PICARD_PATH		Full path to directory containing the Picard Jar files
FASTQC_PATH		Full path to directory containing fastqc script
CUTADAPT_PATH		Full path to directory containing the cutadapt binary
SAMTOOLS_PATH		Full path to directory containing the samtools binary
BEDTOOLS_PATH		Full path to directory containing the BEDTools binaries
GATK_JAR		Full path of the GATK Jar file
VCFTOOLS_PATH		Full path to the VCFtools binaries
VCFTOOLS_PERLLIB		Full path to the VCFtools Perl library directory
HTSEQ_PATH		Full path to the directory containing the HTSeq binaries
HTSEQ_LIB_PATH		Full path to the HTSeq python packages
IGVSESSION_SERVER		If applicable, an ftp server that the bams will be moved to. The IGV session will contain this server path, otherwise just leave this parameter blank.
PYTHON_PATH		Full path to the directory containing the python binary
CUTADAPT_PARAMS	-b AATCTCGTATGCCG TCTTCTGCTTGC -O 3 -m 17 -f fastq	Parameter for Cutadapt which performs adapter trimming. The adapter sequence (-b) should be match the adapter used during library preparation. Multiple adapters can be passed in using additional -b arguments if necessary.
MAPPER_PARAMS	-e -h -q -m -r 5 -u -v -o 4	miRDeep2 mapper parameters
MIRDEEP2_PARAMS	-P -t Human	miRDeep2 parameters

MIRDEEP2_CLOSE_SPECIES	none	Optional close species to be used by miRDeep2
QUANTIFIER_PARAMS	-P -W	miRDeep2 quantifier parameters
BOWTIE_PARAMS	-p 4 -S -q -n 1 -e 80 -l 30 -a -m 5 --best --strata	Parameters for the BOWTIE alignment (not within miRDeep2) to be used for variant calling, IGV visualization, and QC.
ADDORREPLACEREADGROUPS_PARAMS	MAX_RECORDS_IN_RAM=1800000 VALIDATION_STRINGENCY=SILENT RGLB=hg19 RGCN=Mayo RGPL=Illumina	Parameters for Picard's AddOrReplaceReadGroups tool which is used to sort and add read group information to the bowtie BAM files.
SORTSAM_PARAMS	MAX_RECORDS_IN_RAM=1800000 VALIDATION_STRINGENCY=SILENT	Parameters for Picard's SortSam tool.
UNIFIEDGENOTYPER_PARAMS	-glm SNP -dcov 1000	Parameters for GATK's UnifiedGenotyper which is used to call SNVs jointly across all samples.
HTSEQ_PARAMS	-m intersection-nonempty -q -t exon -s no	HTSeq parameters
QUEUE		If running CAP-miRSeq on an SGE cluster, this parameter defines the name of the queue that the jobs should be submitted to.
REFERENCE_INDEXES_MEM	-l h_vmem=3G -l h_stack=10M	Memory allocation parameters for the Reference Indexes SGE job
CUTADAPT_MEM	-l h_vmem=2G -l h_stack=10M	Memory allocation parameters for the Cutadapt SGE job
FASTQC_MEM	-l h_vmem=3G -l h_stack=10M	Memory allocation parameters for the Fastqc SGE job
BAMS_MEM	-l h_vmem=2G -l h_stack=10M	Memory allocation parameters for the Bams SGE job
MIRDEEP2_MAPPER_MEM	-l h_vmem=1G -l h_stack=10M	Memory allocation parameters for the miRDeep2 Mapper SGE job

MIRDEEP2_MEM	-l h_vmem=2G -l h_stack=10M	Memory allocation parameters for the miRdeep2 SGE job
VARIANTS_MEM	-l h_vmem=3G -l h_stack=10M	Memory allocation parameters for the Variants SGE job
EXPRESSION_REPORTS_MEM	-l h_vmem=1G -l h_stack=10M	Memory allocation parameters for the Expression Reports SGE job
DIFF_EXPRESSION_MEM	-l h_vmem=2G -l h_stack=10M	Memory allocation parameters for the Diff Expression SGE job
GENCODE_CLASSIFICATION_MEM	-l h_vmem=3G -l h_stack=10M	Memory allocation parameters for the Gencode Classification SGE job
SAMPLE_SUMMARY_MEM	-l h_vmem=1G -l h_stack=10M	Memory allocation parameters for the Sample Summary SGE job
MAIN_DOC_MEM	-l h_vmem=1G -l h_stack=10M	Memory allocation parameters for the Main Doc SGE job
ADDORREPLACEREADGROUPS_JVM_MEM	-Xmx512g -Xms512m	Java memory allocations for Picard's AddOrReplaceReadGroups
SORTSAM_JVM_MEM	-Xmx1g -Xms512m	Java memory allocations for Picard's SortSam
UNIFIEDGENOTYPER_JVM_MEM	-Xmx512m –Xms512m	Java memory allocations for GATK's UnifiedGenotyper
CREATEDICTIONARY_JVM_MEM	-Xmx512g -Xms512m	Java memory allocations for Picard's CreateDictionary

5.4 Completion

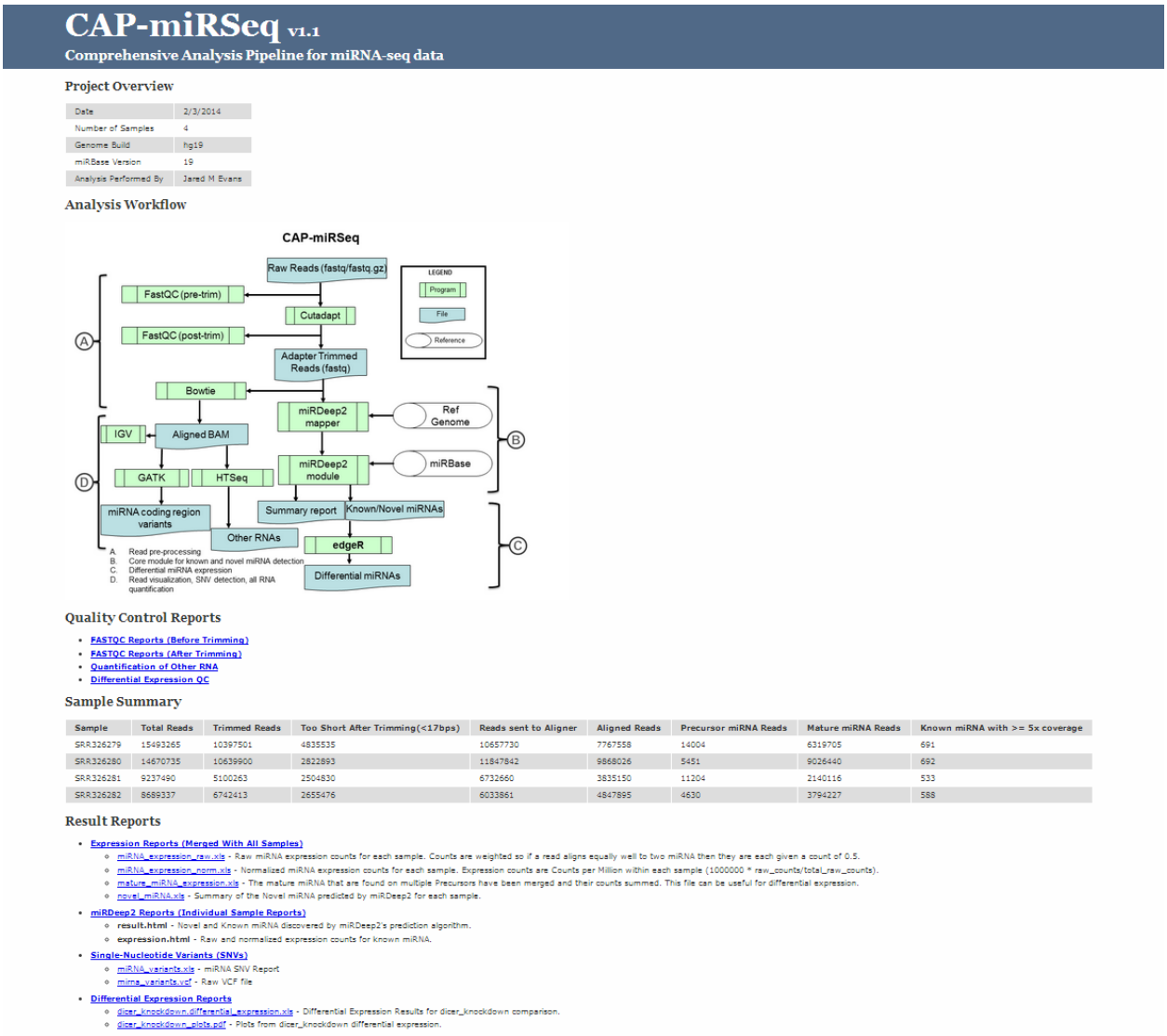
Once CAP-miRSeq finishes running, an email will be sent to the user indicating that the workflow has completed. CAP-miRSeq has built-in error checking that checks the integrity of intermediate files as well as the final reports. In the completion email, CAP-miRSeq will let the user know whether the run was a SUCCESS or ERROR run. If an unsuccessful run occurred, the user can check 2 text files in the CAP-miRSeq output directory which will give more detail about the problem:

- errorlog: Messages in the errorlog will show up if intermediate files or final reports didn't get generated correctly. It is important to fix any error listed in this file because they usually point to results missing from the workflow output.
- warninglog: Messages in the warninglog are notifications that something may be wrong, but could be ok depending on the quality of the samples.

In addition to the summarized error and warning logs, CAP-miRSeq stores all of the stdout and sdterr logs generated by SGE in the logs folder of the output directory.

6. CAP-miRSeq Results

When CAP-miRSeq finishes running, a MainDocument.html file is produced which should be referred to for a summary of the run as well as links and descriptions to all of the QC and expression reports.



CAP-miRSeq v1.1

Comprehensive Analysis Pipeline for miRNA-seq data

Project Overview

Date	2/3/2014
Number of Samples	4
Genome Build	hg19
miRBase Version	19
Analysis Performed By	Jared M Evans

Analysis Workflow

Quality Control Reports

- FASTQC Reports (Before Trimming)
- FASTQC Reports (After Trimming)
- Quantification of Other RNA
- Differential Expression QC

Sample Summary

Sample	Total Reads	Trimmed Reads	Too Short After Trimming(<17bps)	Reads sent to Aligner	Aligned Reads	Precursor miRNA Reads	Mature miRNA Reads	Known miRNA with >= 5x coverage
SRR326279	15493265	10397501	4835535	10657730	7767558	14004	6319705	691
SRR326280	14670735	10639900	2822893	11847842	9868026	5451	9026440	692
SRR326281	9237490	5100263	2504830	6732660	3835150	11204	2140116	533
SRR326282	8689337	6742413	2655476	6033861	4847895	4630	3794227	588

Result Reports

- Expression Reports (Merged With All Samples)
  - miRNA_expression_raw.xls - Raw miRNA expression counts for each sample. Counts are weighted so if a read aligns equally well to two miRNA then they are each given a count of 0.5.
  - miRNA_expression_norm.xls - Normalized miRNA expression counts for each sample. Expression counts are Counts per Million within each sample (1000000 * raw_counts/total_raw_counts).
  - mature_miRNA_expression.xls - The mature miRNA that are found on multiple Precursors have been merged and their counts summed. This file can be useful for differential expression.
  - novel_miRNA.xls - Summary of the Novel miRNA predicted by miRDeep2 for each sample.
- miRDeep2 Reports (Individual Sample Reports)
  - **result.html** - Novel and Known miRNA discovered by miRDeep2's prediction algorithm.
  - **expression.html** - Raw and normalized expression counts for known miRNA.
- Single-Nucleotide Variants (SNVs)
  - miRNA_variants.xls - miRNA SNV Report
  - mirna_variants.vcf - Raw VCF file
- Differential Expression Reports
  - dicer_knockdown_differential_expression.xls - Differential Expression Results for dicer_knockdown comparison.
  - dicer_knockdown_plots.pdf - Plots from dicer_knockdown differential expression.

Figure 1: An example of a MainDocument.html report generated by CAP-miRSeq

6.1 Output Structure

The output directory of CAP-miRSeq will be organized in the following structure:

```
bams/
     Sample1.bam
     Sample1.bam.bai
config/
     run_info.txt
     sample_info.txt
     tool_info.txt
differential_expression/
```

```
     Analysis1.differential_expression.xls
     Analysis1_plots.pdf
     expression_boxplots.pdf
expression/
     mature_miRNA_expression.xls
     miRNA_expression_norm.xls
     miRNA_expression_raw.xls
     novel_miRNA.xls
fastqs/
     Sample1.cutadapt.fastq
     Sample1.tooshort.fastq
igv/
     igv_session.xml
     IGV_Setup.doc
logs/
     (SGE stdout and stderr logs for all jobs)
mirdeep2/
     Sample1/
          (miRDeep2 output files for this sample)
qc/
     fastqc_posttrim/
          (FastQC reports for each sample after adapter trimming)
     fastqc_pretrim/
          (FastQC reports for each sample before adapter trimming)
     other_rna/
          Sample1_gencode_counts.txt
          Sample1_gencode_piecharts.pdf
variants/
     mirna_variants.vcf
     miRNA_variants.xls
CAP-miRSeq_workflow.png
Errorlog
MainDocument.html
SampleSummary.xls
Warninglog
```

7. Contact Information

Please visit the CAP-miRSeq website for the latest updates:
http://bioinformaticstools.mayo.edu/research/cap-mirseq/

Feel free to contact Jared Evans (evans.jared@mayo.edu) for technical questions, Matt Bockol (bockol.matthew@mayo.edu) for installation, or Dr. Zhifu Sun (sun.zhifu@mayo.edu) for general questions about CAP-miRSeq and miRNA-seq analysis.