Circ-Seq User Guide

A comprehensive bioinformatics workflow for circular RNA detection from transcriptome sequencing data

02/03/2016

Table of Contents

Introduction

Circular RNAs (circRNAs) are recently discovered members of the noncoding RNA family that range in length from a few hundred to thousands of nucleotides. In contrast to linear RNA transcripts, which are normally spliced tail-to-head, circRNAs are formed by the covalent bonding of their 3´ and 5´ (head-to-tail) ends. The lack of open sites at the 5´ and 3´ ends exempts circRNAs from endonuclease degradation, making them stable in cells. Additionally, studies have shown remarkable capabilities of circRNAs to sequester several miRNAs away from messenger RNA targets using shared miRNA binding sites (MRE – miRNA response elements). Depending on the number of MRE sites available, circRNAs can compete with messenger RNAs for a common pool of miRNAs, thereby regulating gene expression. Such networks of complex interactions between coding and non-coding RNAs within the cell are termed as competing endogenous RNA (ceRNA) networks. Hence, these unique features of circRNAs to remain stable and act as competing endogenous RNAs (ceRNAs) make

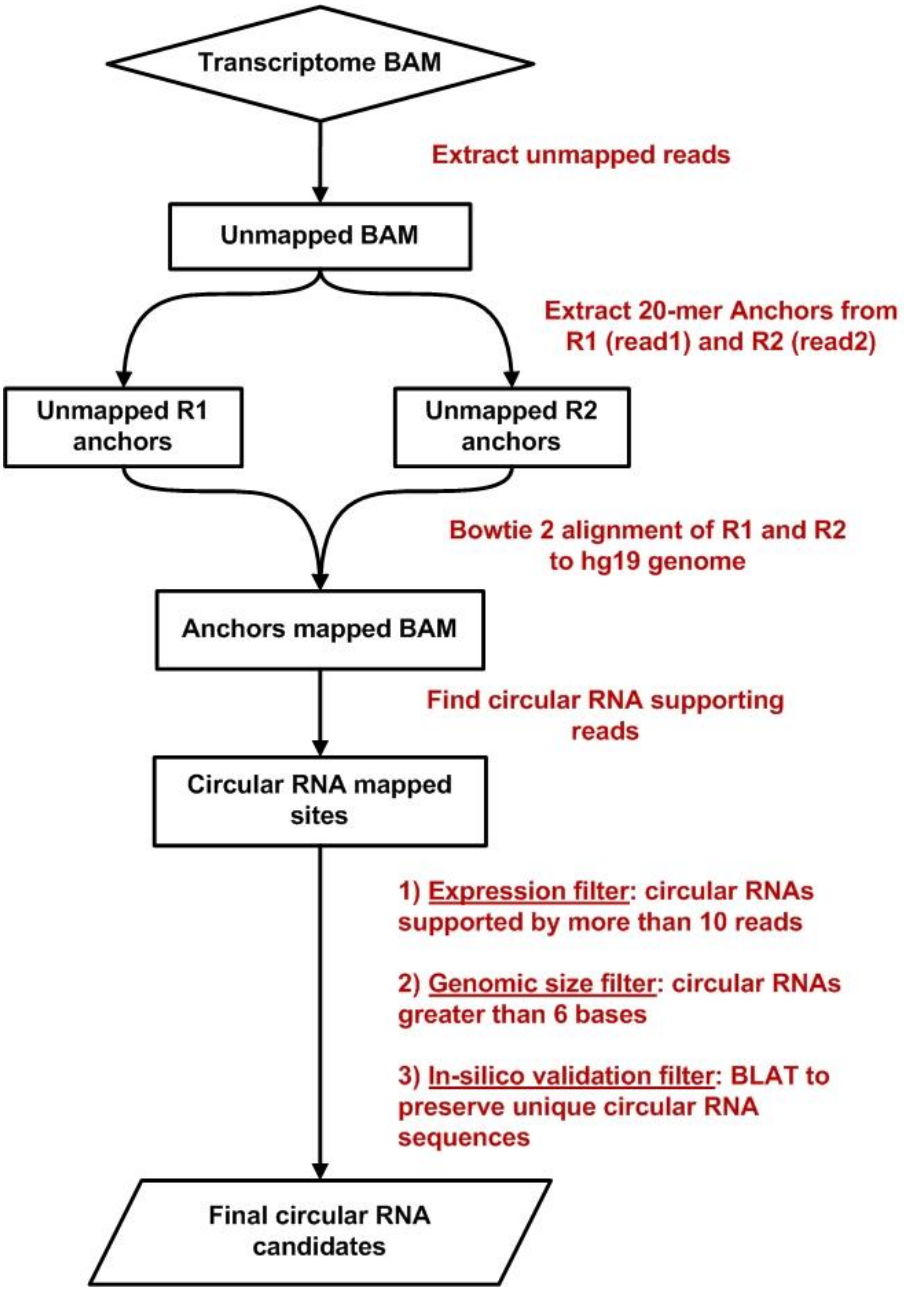
them promising candidates to explore novel diagnostic and therapeutic targets in diseases.

Circ-Seq is an integrated bioinformatics workflow for identifying and characterizing circRNAs using high-throughput transcriptome sequencing data. Briefly, it improves the circRNA identification methodology developed by Memczak et.al 2013 by applying filters to exclude false positives from the final catalog of candidate circRNAs. Circ-Seq also helps users prioritize the final list of circRNAs by annotating them with exon information on the location (exon boundary or within exons) of their 3' (head) and 5' (tail) ends. Circ-Seq processes unmapped reads of the transcriptome, obtained either from the MAP-RSeq workflow or any other RNA-Seq alignment software. These reads are checked for evidence of alignment in a circular RNA specific (3′ to 5′) fashion. Filters on expression, genomic size and validation are applied to report legitimate circRNA candidates. Circ-Seq provides a circRNA quantification report and a FASTA file that contains 50-base nucleotide sequences containing the 3´–5´ fused junction of circRNAs in the final report.

The package for the Circ-Seq workflow can be downloaded from the website:

http://bioinformaticstools.mayo.edu/research/circ-seq/

The Circ-Seq bioinformatics workflow flowchart is represented below:



Local Installation to your system

Please see the “README.CircSeq” file included with the Circ-Seq source distribution (circRNA_OpenGridEngine.tgz) for a step by step guide to installing all the software required to run the workflow. This installation

allows you to run the workflow more efficiently and take advantage of parallel computing 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.

Reference Files

In order to run Circ-Seq, a local copy of the human genome reference is required. The reference genome assembly should contain sequences in FASTA format (.fai, .dict, and bowtie2 index files are also needed). The hg19 reference can be downloaded from the following link: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz

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

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

The BLAT 2bit file for the hg19 reference genome is also required for Circ-Seq. The file can also be downloaded from the UCSC genome bioinformatics downloads page at the following link:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/

Running Circ-Seq

To run Circ-Seq, a configuration file is required. More details on this file is provided in following sections.

Once the configuration file is created, Circ-Seq can be invoked using the following commands, either in standalone or sun grid engine (SGE) / open

grid engine (OGE) mode. Please provide the complete path to the configuration file while running the command.

Single machine

If the workflow is being run on a single machine then circseq_wrapper.sh will continue running until the entire workflow is finished.

```
./circseq_wrapper.sh `pwd`/config_standalone.txt
```

OGE cluster environment

If it is being run on an OGE cluster then circseq_wrapper.sh will quickly run and submit all of the OGE jobs.

```
./circseq_wrapper.sh `pwd`/config.txt
```

In either case an email will be sent to the user upon completion.

Configuration File

The configuration file (config.txt) contains various information and parameters specific to the current run of Circ-Seq. Here is an example of a config.txt file:

```
#### Input parameters
WORKFLOW=Circ-Seq
RUN_NAME=Example run
EMAIL=user.name@institution.edu
BAM_DIR=/path/to/unmapped/BAMfile/directory/
SAMPLENAMES=test_sample
CIRC_DIR=/path/to/output/circRNA/directory/
ANCHOR_SIZE=20

#### Reference files
```

```
REF_BOWTIE=/path/to/hg19
REF_GENOME=/path/to/hg19.fa
REF_GENOME_DIR=/path/to/directory/containing/hg19.fa/
BLAT_REF=/path/to/hg19.2bit
EXON_START_BOUNDARY=/path/to/exon_start_boundary.bed
EXON_END_BOUNDARY=/path/to/exon_end_boundary.bed
INTRON_START_BOUNDARY=/path/to/intron_start_boundary.bed
INTRON_END_BOUNDARY=/path/to/intron_end_boundary.bed
EXONS=/path/to/exons.bed

#### Tools
BOWTIE=/path/to/bowtie 2.1.0/
PYTHON=/path/to/python/2.7/bin
SAMTOOLS=/path/to/samtools/0.1.18
SGE=/path/to/oge/ge2011.11/bin/linux-x64/
BEDTOOLS=/path/to/BEDTools/2.16.2/bin/
BLAT=/path/to/blat/34_x64
WORKFLOW_PATH=/path/to/where/Circ-Seq/scripts/are/located/

#### env variables
PYTHONPATH=/path/to/Linux-x86_64/lib/python2.7

#### queue
QUEUE=4-days

#### Queue parameters
NOTIFICATION_OPTION=ae
THREADS=8
```

Details of the configuration parameters are provided below:

Parameters	Default values	Description
WORKFLOW	Circ-Seq	Name of workflow
RUN_NAME	Example run	Name for run being analyzed
EMAIL	user.name@institution.edu	Email address of individual running Circ-Seq
BAM_DIR	/path/to/unmapped/BAM/file	Full path to where the unmapped BAM file is located
SAMPLENAMES	test_sample	Name of sample(s) being analyzed. For analysis with more than one sample, please provide colon separated list, e.g., sample1:sample2

CIRC_DIR	/path/to/output/circRNA/directory/	Name of output directory to which cirRNA results will be written
ANCHOR_SIZE	20	size to which the unmapped reads will be fragmented
REF_BOWTIE	/path/to/hg19	Full path to bowtie2 indexed files for reference genome assembly
REF_GENOME	/path/to/hg19.fa	Full path to hg19 reference genome assembly in FASTA format
REF_GENOME_DIR	/path/to/directory/containing/hg19.fa/	Full path to directory containing reference genome
BLAT_REF	/path/to/hg19.2bit	Full path to the 2bit file for BLAT
EXON_START_BOUNDARY	/path/to/exon_start_boundary.bed	Full path to the .BED file downloaded from the Circ-Seq website
EXON_END_BOUNDARY	/path/to/exon_end_boundary.bed	Full path to the .BED file downloaded from the Circ-Seq website
INTRON_START_BOUNDARY	/path/to/intron_start_boundary.bed	Full path to the .BED file downloaded from the Circ-Seq website
INTRON_END_BOUNDARY	/path/to/intron_end_boundary.bed	Full path to the .BED file downloaded from the Circ-Seq website
EXONS	/path/to/exons.bed	Full path to the .BED file downloaded from the Circ-Seq website
BOWTIE	/path/to/bowtie 2.1.0/	Full path to install directory of bowtie 2.1.0
PYTHON	/path/to/python/2.7/bin	Full path to install directory of python 2.7
SAMTOOLS	/path/to/samtools/0.1.18	Full path to install directory of samtools 0.1.18
SGE	/path/to/oge/ge2011.11/bi	Full path to install directory

	n/linux-x64/	of OGE
BEDTOOLS	/path/to/BEDTools/2.16.2/bin/	Full path to install directory of bedtools 2.16.2
BLAT	/path/to/blat/34_x64	Full path to install directory of BLAT
WORKFLOW_PATH	/path/to/where/Circ-Seq/scripts/are/located/	Full path to install directory of Circ-Seq containing scripts
PYTHONPATH	/path/to/Linux-x86_64/lib/python2.7	Full path to library files for python 2.7
QUEUE		If running Circ-Seq on an OGE cluster, this parameter defines the name of the queue that the jobs should be submitted to.
NOTIFICATION_OPTION	ae	If running Circ-Seq on an OGE cluster, this parameter defines if user would receive (a)bort and (e)rror notifications
THREADS	8	If running Circ-Seq on an OGE cluster, this parameter defines the number of threads that can be used for analysis

Circ-Seq Results

When Circ-Seq completes analysis, an email will be sent to the user indicating that the analysis completion. The email will also include start and end time stamps of the analysis.

The output directory will be organized in the following structure:

```
circRNA_output/
       test_sample/
          test_sample.circRNA.3prime5prime_fused_junction.fasta
```

```
test_sample.circRNA.expressed.txt
test_sample.circRNA.reads
test_sample.sites.log
test_sample.unmapped.anchors.gz
```

The quantification and annotation report for the final set of circRNAs is recorded in `test_sample.circRNA.expressed.txt`

The 50 base fasta sequence of the head-to-tail (3' to 5') fused junction of individual circRNAs in the final quantification report is provided in `test_sample.circRNA.3prime5prime_fused_junction.fasta`

The 20-mer anchors used for bowtie2 alignment, summary statistics on circRNA alignment by Memczak et.al and total number of circRNAs before Circ-Seq filters application are provided in files `test_sample.unmapped.anchors.gz, test_sample.sites.log` and `test_sample.circRNA.reads` respectively.

Contact Information

Please visit the Circ-Seq website for the latest updates:

http://bioinformaticstools.mayo.edu/research/circ-seq/

Please feel free to contact Asha Nair (Nair.Asha@mayo.edu) for technical questions, Matt Bockol (Bockol.Matthew@mayo.edu) for installation, or Dr. Krishna R. Kalari (Kalari.Krishna@mayo.edu) for general questions about the Cir-Seq workflow and circular RNA analysis.