LIONS Manual

v.0.1

Artem Babaian, Richard Thompson

2018-11-29

Contents

1 LIONS Pipeline Architecture

LIONS is a bioinformatic analysis pipeline which brings together a few pieces of software and some home-brewed scripts to annotate a paired-end RNAseq library against a reference TE annotation set (such as Repeat Masker).

East Lion processes a bam file input, re-aligns it to a genome, builds an ab initio assembly using Tophat2. This assembly is then proccessed and local read searches are done at the 5' ends to find additional transcript start sites and quality control the 5' ends of the assembly. The output is a file-type <library> .lions which annotates the intersection between the assembly, a reference gene set and repeat set.

West Lion compiles different .lions files, groups them into biological catagories (i.e. Cancer vs. Normal or Treatment vs. Control) and compares and analyzes the data to create graphs and meaningful interpretation of the data.

2 LIONS Installation

Download/clone the LIONS repo: https://github.com/ababaian/LIONS/archive/master.zip

2.1 LIONS Container (Docker)

LIONS can be installed as a Docker container. Navigate to the $LIONS folder containing the 'Dockerfile'. Build container with:

```
cd \$LIONS/
Docker build -t lions .
```

You still will need to download the LIONS resource files. (See below)

2.2 LIONS Command Line

Users with experience of the linux commandline may wish to download the package from github and run it directly without using Docker. This is especially useful for cluster-computing.In this case ensure the following software is installed on your system:

- Python3 and pysam
- Bowtie2
- Tophat2
- Java v8 or higher
- Samtools v0.1.18
- R v3.5.0 or higher
- Bedtools v2.25.0
- Cufflinks v2.2.1

2.3 Initializing LIONS Resources

1. If you don't have LIONS downloaded already; clone it

   ```
   git clone git@github.com:ababaian/LIONS.git
   ```

2. Initialize genomic resources for LIONS in./LIONS/resources copy 'example' folder to '<genomeName>' folder

3. Populate the resource files:
   a) In ./LIONS/resources/<genomeName>/genome/ add a <genomeName>.fa genome sequence file
   b) In ./LIONS/resources/<genomeName>/annotation/ add the ucsc annotation file
   c) In ./LIONS/resources/<genomeName>/repeat/ add the ucsc repeatMasker file

   ```
   # genomeName = hg38 ========================
   # Genome
   http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz

   # Gene Annotation (RefSeq)
   https://s3-us-west-2.amazonaws.com/lionproject/resources/hg38/refseq_hg38.ucsc.g

   # Repeat Masker
   https://s3-us-west-2.amazonaws.com/lionproject/resources/hg38/rm_hg38.ucsc.gz
   ```

```
# genomeName = hg19 ========================
# Genome
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit

# Gene Annotation (RefSeq)
https://s3-us-west-2.amazonaws.com/lionproject/resources/hg19/refSeq_hg19.ucsc.z

# Repeat Masker
https://s3-us-west-2.amazonaws.com/lionproject/resources/hg19/rm_hg19.ucsc.zip
```

(The UCSC files are downloaded from the UCSC Table Browser as the 'all fields from selected table' output format)

4. Initialize the project parameters in parameter.ctrl:
   a) Give your project a name and fill out all the parameters software designated in the software list.

5. This file contains all the *system-specific* parameters for LIONS to run on your computer. Go through this file and ensure the parameters make sense and the software pointers are compatible.
   a) Give your project a name and fill out all the parameters as you see fit.
   b) In the software list, refer LIONS to the paths for each software on your system. If it's in $BIN already then you can simply type the command; LIONS will create links in it's own folder structure to the software designated in the software list.

3 Running LIONS

Once the container has been built, lions can be run inside an interactive docker container which can be started using the following command:

```
Docker run -ti lions
```

It's a good idea to create a local LIONS directory ($LIONS) outside of the container with all the resources and parameter files set-up. You can then mount the $LIONS directory into the container using *-v*.

This command will open a (bash) terminal to the container, in the LIONS base directory.

```
cd \$LIONS
Docker run -v \$LIONS:/LIONS -ti lions
```

where '$LIONS' is the directory on the host machine containg the data for analysis and '/LIONS' is the directory inside the container where the data will be accessed.

e.g. `Docker run -v /home/artem/LIONS-master:/LIONS -ti lions` will allow the user to access to the contents of /home/artem/LIONS-master within the container.

> Alternatively you can load the resource files and data files into container at start-up.
>
> ```
> -v <resource_directory>:/LIONS-docker/resources/<genomeName> \
> -v <data_directory>:/LIONS-data/
> ```
>
> for example:
>
> ```
> -v ~/hg19:/LIONS-docker/resources/hg19 \
> -v ~/ENCODE_bams/:/LIONS-data
> ```
>
> this will avoid adding resources to the image eachtime the container is run, or swelling the container size.

3.1 Running LIONS using a parameter file

The default parameter file for LIONS is `/LIONS/controls/parameter.ctrl`.

However, the user can run LIONS with with a specified parameter file as follows:

```
bash lions.sh <path/to/parameter/ctrl>
```

N.B. As the docker container will not carry across changes to the file structure, such as saved output files, it is advised that the mount point identified with the `-v` option is used to save output files and modified `parameter.ctrl` and `input.list` files, in addition to supplying input files.

3.2 Running LIONS using commandline options

Lions can be run andline th commandline options using the `lions_cli.sh` script. Any options not provided on the commandline will be taken from the default parameter file.

`./lions_opt.sh [options]`

Where options include:

-h — --help Show commandlines usage

-P — --parameter Define user parameter.ctrl file, see section **??**; ***N.B.*** This option must be defined first in order that the other options are used correctly.

-b — --base <path> Sets the LIONS base directory to be ¡path¿, defaults to $PWD.

-p — --project <project-name> Set name of the project, and therefore the output directory name in the projects folder, to ¡project_name¿

-i — --inputlist <path/to/input.list> Identifies the input file containing the sample names and input files for the analysis, see section **??**

--callsettings <pre-set-name> Tells LIONS which settings to use for the TE-initiation. Expected values include:

'oncoexapatation' Detect high abundance TE isoforms

'transcriptomeANN' Artifical Neural Network based classifier

'screenTE' High sensitivity, low specificity detection

'driverTE' TE-initiations as main drivers of gene-expression

'custom' Use custom settings defined below

default setting is **'oncoexaptation'**.

-I — --index <INDEX> Sets the LIONS Resource index, see section **??**.

--geneset <geneset.file> Options identifies a ucsc formatted file containing a gene set to be used in the analysis, see section **??**.

--repeatmasker <repeat.file> Identifies a ucsc formatted file containing the repeatMasker annotation for the genome, see section **??**.

--systemctrl <system.ctrl> Identifies the control file used for passing system settings to LIONS, see section **??**

--bowtie <?> ?Additional Bowtie options? .. or command?

--alignbypass # binary switch to tell LIONS whether to run a new alignment on input files, or to use that provided in the input files. Default is '1'.

0 Calculate a new alignment for the input

1 Do not re-calculate the alignment, simply create symbolic link to the bam file in <input.list> within the LIONS folder architecture

--**inread #** The Inner Read distance to be used by Bowtie during alignment (Bowtie's `-r` option).

--**threads #** The number of parallel threads to be used by Tophat/Bowtie and Cufflinks during the analysis. ***N.B.*** presently it is necessary to ensure that the `--threads` option is passed after the `--systemctrl`, if both are used, in order to ensure the threads option is used correctly throughout the analysis. This will be corrected in future versions.

--**denovo #** Tell Cufflinks whether to carry out a *de novo* or *ab initio* assembly, LIONS defaults to *de novo*.

0 *Ab Initio*

1 *De Novo*

--**minfrags #** Set Cufflink's 'minimum fragments per transfrag', default is 10.

--**multiflag #** Set Cufflink's 'multi-mapping transfrag fragments' option, default is 0.75.

--**mintrim #** Set Cufflink's 'minimum coverage to attempt 3' trimming' option, default is 10.

--**trimdrop #** Set Cufflink's 'minimum Trim dropoff in fraction' option.

--**merger #** Set Cufflink's 'merge radius in bp', default is 50 bp.

--**quality** <**qual**> RNAseq Analysis Quality, defined by the string 'q#.F#' where q = quality cutoff; F bam flags to discard, *e.g.* 'q10.F772' and 'q1.F1796'.

4 LIONS File Architecture

4.1 ./

base directory: LIONS is a self-contained pipeline and references needed by LIONS from the system are linked within this directory.

./lions.sh <parameter.ctrl> The master script from which the entire pipeline is ran The script reads and processes all files in input.list and all parameters can be controlled from parameter.ctrl

4.2 ./controls

Control Files: This folder contains project and system-specific parameters for running LIONS. There are three main files which need to be set-up, LIONS run parameters, system parameters and input RNA-seq libraries.

./controls/parameter.ctrl A bash script which defines global project-specific variables such as Project Name, library input list etc... The .sysctrl and .list file are defined here as well

./controls/system.sysctrl A bash script which defines global variables for all LIONS scripts. Also defines system-specific variables such as System Name, number of CPU cores etc...

./controls/input.list A three column tab-delimited file defining:

<Library_Name> <LibPath> <Grouping>

where `Grouping` is one of:

1 control

2 experimental

3 other

.. and `<libPath>` is either a paired end bam file, *e.g.* '/home/libPath.bam', or a pair of sorted fastQ files (as a comma seperated list), *e.g.* '/home/lib.fq1,/home/lib.fq2'.

4.3 ./bin

LIONS internal folder for symbolic links to binaries needed by the pipeline and script to initialize the folder. Make sure to set the correct commands for the software list in parameters.ctrl for your system

4.4 ./projects

For each <Project_Name> a single folder will be initilized in which the data will be organized.

./projects/<Project_Name> The main directory for this project. Each individual library in the input will have a folder generated here called <Library>.

./projects/<Project_Name>/logs Folder contains run-specific information such as input file at time of run and a copy of the input parameter file at time of run.

./projects/<Project_Name>/Analysis(_RUNID) Not implemented yet. This contains all data analysis for a run of LIONS. All graphs and project-wide .lions files are stored here

./projects/<Project_Name>/<Library> Library-specific data and primary analysis files.

./projects/<Project_Name>/<Library>/<Library>.lcsv Raw output file from LIONS containing all possible TE-Exon interaction data. This will include initiations, exonizations and terminations along with many calculated values about these loci from which LIONS will sort initiation events from the others. <Library>.pc.lcsv is the same file with additional information about overlapping protein coding genes.

./projects/<Project_Name>/<Library>/<Library>.lions Initiation only TE-exon data from post-sorting. This file is the complete list of transcripts initiated by TEs in this library. This data is passed on to the West Lion protocol to compare TE usage between libraries.

./projects/<Project_Name>/<Library>/alignment tophat2-generated alignment and the re-aligned .bam file which will be used for analysis. Also contains flagstats and log files. Note: Once the alignment is generated it will not be re-generated even if you change the alignment parameters in parameter.ctrl. To re-make alignments simply start the project with a new name or delete these files.

./projects/<Project_Name>/<Library>/assembly cufflinks-generated assembly in 'transcripts.gtf'

./projects/<Project_Name>/<Library>/expression The output from a series of custom scripts 'RNAseqPipeline' which will generate wig files and perform RPKM calculations on a series

./projects/<Project_Name>/<Library>/resources Charlie-foxtrot of library-specific files used to calculate a score of parameters in the pipeline.

4.5 ./scripts

Scripts: all scripts to run lions are held here except for the controlling lions.sh script which is in the base folder. Check initializeScripts.sh for complete list of scripts The main scripts are

./scripts/eastLion.sh Alignment, Assembly, Chimeric Detection pipeline

./scripts/westLion.sh LIONS analysis pipeline

4.6 ./resources

Input files containg resource information needed for LIONS to run the analysis. The geneset and RepeatMasker files should be set in the `parameter.ctrl` file.

./resources/<INDEX>/annotation/<GENESET> A ucsc formatted file containing a gene set to be used in the analysis (i.e. look for overlapping genes to transcripts) Download from: https://genome.ucsc.edu/cgi-bin/hgTables The standard annotation set used was RefSeq 'RefGene' table.

./resources/<INDEX>/genome/<INDEX>.fa The only requisite file here for running LIONS is a fasta formatted genome. This could be a symbolic link. LIONS will generate the other files necessary from INDEX.fa. If you have the .bt2 index files already generated you can symbolically link them in this folder to skip regenerating them.

./resources/<INDEX>/repeat/<REPEATMASKER>.ucsc a ucsc formatted file containing the repeatMasker annotation for the genome. (Download from UCSC genome browswer or format from RepeatMasker) Columns are; bin, swScore, milliDiv, milliDel, milliIns, genoName, genoStart, genoEnd, genoEnd, genoLeft, strand, repName, repClass, repFamily, repStart, repEnd, repLeft, id

<INDEX> : the name of the index set. To be compatible with different genome versions, species and gene sets there can be different sets of data. The <INDEX> global variable is set in parameter.ctrl file.

4.7 ./software

Packaged with LIONS is a few bits of software which will set-up your system to run the pipeline. Namely setuptools and pysam are the most challenging things to install. I found that it's easiest to set-up the pipeline using pip and download the package pysam from there. Pysam is used to read teh bam files in the python scripts.

5 LIONS Error Codes

Error 1 Internal software error - Check last-run software.

5.1 Initialization Codes

Error 2 Initialization file missing or inaccesible - A file is missing or is unreadable. Ensure you have a complete version of LIONS and/or make the missing script readable/exectable

Error 3 A LIONS script is missing - A script is missing from `./LIONS/scripts/`; ensure your copy of LIONS is complete or redownload.

Error 4 Initialization bin missing - A binary is not found on the system. Configure `./LIONS/bin/initializeBin.sh` for your system

Error 5 A resource file is missing or unreadable - Checking/initialization of `./LIONS/scripts`.

Error 6 A Python requisite is missing.

Error 7 The input read file (.bam or .fastq) is non-readable or empty

7A Bam file error

7B FastQ file error. Ensure the two files are comma seperated in the input

5.2 eastLion Error Codes

Error 10 alignment not generated - An attempt was made to generate an alignment but the output file was empty at the end of the script

Error 12 wig not generated - An attempt was made to generate the wig file but the output file wasn't present after the script ran

5.3 westLion Error Codes

Error 15 A lions file wasn't generated - In the run, one of the lions files wasn't generated which means there was an error. Don't run West Lions pipeline.

6 LIONS output definitions

6.1 Output File Types

LIONS produces several outputs from different stages of the analysis in addition to the standard outputs one would expect (.bam / .gtf).

'<library>.lcsv' / '.pc.lcsv' These are LIONS CSV files which contain the raw calculations for all major numeric operations. This includes ALL TE-exon interactions types (Initiation, Exonization and Termination). As such there are usually hundreds of thousands of TEs which have read fragments joining them to some assembled exon.

The '.pc.' pre-suffix means the data has been intersected to the input set of protein coding genes.

Use this file for re-calculating "TE-Initiations" with new parameters.

'<library>.lion' This is the filtered set of TE-exon interactions which have been classified as "TE-Initiations" or TE transcription start sites. This is per-library input.

'<project>.lions' A merged file of several '.lion' files combining biological groups defined in the 'input.list'. A good example of this is merging 10 cancer libraries and 10 normal libraries and outputing only those TE-initiations which are in at least 20% of Cancer and no Normal libraries. These parameters can be changed in the 'paramter.ctrl' input.

'<project>.rslions' The 'rs' is for Recurrent and Specific TE-initiations only. That is if you compare the set of libraries 1 (Normal) vs set 2 (Cancer), this contains only those TE-initiations which occur multiple times in Cancer (recurrant) and do not occur in Normal (specific). As defined by '$cgGroupRecurrence' and '$cgSpecificity' in the 'paramter.ctrl' file.

'<project>.inv.rslions' The '.inv.' pre-suffix is simply the **inverse** of the '.rslion' file. So instead of "Cancer vs. Normal", "Normal vs. Cancer". A necessary control if one makes any conclusions based on enrichment/depletion.

6.2 Output Columns

6.2.1 '<project>.lions'

Most columns should be self-explanatory, some are not.

transcriptID Unique identifier for the transcript (isoform). Usually taken from the assembly/reference transcriptome

exonRankInTranscript For each TE-exon interaction combination (row) which exon in the 'transcriptID' is this row referring to

repeatName The <repeat_name>:<repeat_class>:<repeat_family> taken from input set

coordinates Useful coordinates for visualizing the interaction. It starts/ends in the exon and repeat so when opening in a visualization tool you can see the reads spanning this area.

ER_Interaction: The type of relative intersection in the genome between the exon and the repeat. Definitions are relative to the exon. Can be "Up", "UpEdge", "EInside", "RInside", "Down", "DownEdge".

IsExonic ??

ExonsOverlappingWithRepeat A list of <transcriptID:exonRank> which overlap the repeat.

ER / DR / DE / DD / Total A count of the number of TE-Exon sequence fragments which join this rows TE and Exon. ER means that one end overlaps the Exon and one end overlaps the Repeat exclusively, DD means that both ends of the fragment overlap both (dual) exon and repeat ...

Chromosome / EStart / EEnd / EStrand Start, end and strand of the exon

RStart / REnd / RStrand Start, end and strand of the repeat

RepeatRank Relative exon/intron position of the repeat to the contig

UpExonStart / UpExonEnd Coordinates used for calculating expression of genome immediatly adjacent an exon boundary. Useful for quantifying read- through or spurious transcriptional events.

UpThread The number of read 'threads' going upstream of the exon. See Manuscript for a figure explaining this.

DownThread The number of read 'threads' going downstream of the exon. See Manuscript for a figure explaining this.

ExonRPKM RPKM calculation for this exon

ExonMax The maximum coverage count reached within the exon boundaries. Often more reliable measure of expression then RPKM for small exons.

UpExonRPKM / UpExonMax The expression of the exon immediatly upstream of the one this row is referring to. (i.e. Exon 1 expression if the row refers to Exon 2). Useful for quantifying the relative increase in expression when a TE is acting as an alternative promoter into a downstream exon.

RepeatRPKM / RepeatMaxCoverage Expression level within repeat boundaries.

UpstreamRepeatRPKM / UpstreamRepeatMaxCoverage The expression adjacent to the repeat, a test for background expression levels.

RefID When intersecting to a reference gene set, the gene symbol of any genes which intersect the area between the Exon-Repeat coordinates.

RefStrand Strand of the reference genes defined above

assXref The strand-relationship between the reference gene and the contig exon This accounts for anti-sense long non-coding RNA (as), or transcripts which run anti-sense to the reference gene. (s) is sense and (c) means complex, often some combination of multiple genes. (u) means it could not be determined.

Contribution An estimate of the promoter contribution of this Repeat TSS to the expression of gene in total. Calculated with ExonMax and UpExonMax.

UpCov Ratio of the coverage adjacent to an exon and the exon expression

UpExonRatio Ratio of hte expression of the exon and it's upstream exon

ThreadRatio DownThread / UpThread. Set to [10] if dividing by zero.

RepeatID A unique Identifer for each Repeat in the genome (left-most coordinate). Can repeat and thus be used for determining one repeat inititating a trancsript in different assemblies.

LIBRARY Library from which this repeat-exon interaction was calculated from.

6.2.2 '<project>.rslions'

Normal_occ Number of times each TE-initiation was found in the "normal" set of libraries. (Usually set 1)

Cancer_occ Number of times each TE-initiation was found in the "cancer" set of libraries. (Usually set 2)

Library A semi-colon seperated list of the LIBRARY identifiers in which each TE-initiation was found in.

A Appendices

A.1 Running Docker on the command line

Often it is useful to detach the ryunning Docker container from the terminal, This can be acheived in a number of ways. The official Docker method is to detach the terminal using

`Ctrl-p Ctrl-q`

to diconnect, however this is dependent on using the `-t` option when running the container. Once detached the user can reconnect to the container using

```
docker attach <container-id or name >
```

It is possible to name a docker container using the `-n <name>` option to the `docker run` command. However where the container name is not specified or known, the container-id can be found using the `docker ps` command.

The author prefers using GNU Screen as it is capable of presenting a split screen allowing the user to monitor LIONS whilst continuing to work. Screen is invoked using the command `screen` which opens a virtual terminal, the LIONS docker can then be run as described in section **??**.

Screen sessions can be detached using `Ctrl-a d` and reconnected using `screen -r`.

A single screen session allows a user to simultaneously run multiple virtual terminals; known as 'windows'. New windows are created within a session using `Ctrl-a c`, whilst `Ctrl-a n` and `Ctrl-a p` allow the user to switch between windows within a session.

Screen is able to split the terminal view vertically using `Ctrl-a |` or `Ctrl-a V`, depending on the implementation; `Ctrl-a S` can be used to split the window horizontally. Once split the user can switch between view regions using `Ctrl-a Tab`.