User Manual

VM version

1. For quick testing, you can download our VM virtualBox UClncR.ova, which packages all dependencies. However, its functions are limited due to insufficient memory and lack of parallel computing.

2. Install the VM on Windows

    a. Download and Install Oracle VirtualBox (https://www.virtualbox.org/wiki/Downloads)
    b. Open the OVA image you downloaded in step 1 from VirtualBox.
    c. Ubuntu is installed in the VM virtualBox with UClncR (the sudo password is 'mayo' in case for additional package installation or screen unlock)
    d. The UClncR is under: /home/chen/UClncR/src
    e. To test the pipeline: change directory to "test" and the configuration files are pre-made according to the VM. You can start the pipeline by this command: /home/chen/UClncR/src/UClncR /home/chen/UClncR/test/run.info
    f. Note that due to the huge aligned bam file we only included the data for chr1 for testing purpose. To get the complete and correct result, you can download the whole bam file from our website or use your own aligned bam file from HISAT2 or Tophat.

Source Version

System requirements (parentheses are versions we used):
Operating System:

Linux OS (CentOS 6) 64bit with Minimum 10 GB and 100 GB storage (storage is project dependent)

Other dependencies/external modules:

Java (1.6), Perl (5), Python (2.7), R (3.02), R libraries (MASS, mixtools), StringTie (1.2.0), samtools (0.1.19), BEDTools (2.17.0), cufflinks (2.1.10).

Installation:

From source code:

1. Download UClncR.v1.0.0.tar.gz and unpack.

2. Open info/tool.info and set paths to "Other dependencies/external modules" listed as "System requirements".

3. Download dataset ENCFF782IVX_std.bam. The sample is from Encode project (GM12878), we aligned with hisat2 on hg19 (please open the BAM file header for details).

Files and directories in the package:

1. UClncR: the main script file
2. info: directory that stores sample configuration files (change to fit your environment and project)
    1. run.info : input parameters to the pipeline (open the config/run.info for details)
    2. tool.info : contains paths to the needed tools and packages for the pipeline. the path to this file is set in run.info
    3. sample.info : contains paths to individual samples, path to this file is set in run.info
3. modules : external modules and packages (pre-packaged)
4. README : README file
5. sebnif-1.3rc: modified scripts from Sebnif used to process individual samples.

Usage:

/path/to/UClncR /path/to/run.info

Parameters to set run.info (open the info/run.info to edit):

- SPECIES: currently, only 'human' is supported.
- REF_ANNO: annotation reference, currently, only 'Gencode' is supported
- MULTI_LEN: minimum length for multi-exon novel lncRNAs.
- SINGLE_LEN: minimum and maximum lengths for single exon novel lncRNAs
- FRFE: cut-off for multi-exon expression (refer to Sebnif for details)
- MODEL: log scale of single exon expression follows either Guassian, Gamma, or GMM distribution (refer to Sebnif for details). If set to 'auto', the distribution model will be automatically determined by the program.
- REPEAT: single exon transcript located in repeat region will be filtered out, a value between 0 and 1 (default=0.05 which means if the repeat regions share more than 5% of the transcript length, the transcript will be discarded, set '0' to disable this option, refer to Sebnif for details)

- sebnifNONCODING: noncoding potential cut-off for Sebnif.
- cpatNONCODING: noncoding potential cut-off for CPAT.
- NONCODING_RETAIN_MODE: UNION – lncRNA passes non-coding potential cut-off of either CPAT or iSeeRNA; INTERSECTION – lncRNA passes non-coding potential cut-off of both CPAT and iSeeRNA
- MIN_MAPPING_QUALITY: score cut-off for aligned reads mapping quality.
- SARURATION: cut-off for maximum read depth (refer to Stringtie for details)

Parameters to set tool.info (open the info/tool.info to edit):

- UCLINCR_SOURCECODE_DIR: your source code directory that specify the internal modules.

Output Files:

- Main_Document.html: final summary of analysis in HTML format.
- lncRNA_raw_counts.xls: merged lncRNAs (from all samples) with raw read counts.
- lncRNA_rpkm.xls: merged lncRNAs (from all samples) with normalized read counts/RPKM.
- samples/sample_id/LincRNAFinder.final.gtf: predicted novel lncRNAs in GTF format for individual samples.

Notes:

1. If you install on Ubuntu, all python libraries included in this package have to be re-installed, and C complied executable files have to be re-built from source code.
2. To deal with large datasets, we support parallel-computing. However, the VM version only supports single-thread.
3. The released package only includes needed reference files for hg19. To set up HG38, please contact us for instructions.

Questions:

Please contact Chen.Xianfeng@mayo.edu, Nair.Asha@mayo.edu, Sun.Zhifu@mayo.edu