.. run the RNASeq pipeline

Monti lab has generated semi-automated pipeline that runs a suite of tools required to perform analysis using high-throughput RNA Sequencing data. The pipeline takes list of user-specified parameters in the form of a parameter file (which also contains pointers to the pipeline scripts folder and raw files to be run using the pipeline) and does a complete run, with added flexibility in how programs are utilized (skipping certain steps, etc). An html report is generated, summarizing the output/results from each step, with links to the associated data (log files, QC reports etc.)

**Prerequisites**

* An up to date version is located in the montilab project directory

/restricted/projectnb/montilab-p/projects/superfund/project4/scripts/RNASeq\_pipeline

* The pipeline was extensively tested on SCC4 and uses the currently installed modules
* To run the pipeline we assume that the python2.7 and R-3.0.0 modules are loaded:
  + module load R/R-3.0.0\_gnu-4.4.6
  + module load python2.7/Python-2.7.3\_gnu446

**Parameters and file specifications**

**Phenotype file (example in the tar ball)**

The phenotype file is a separate tab delimited file with the following specifications:

* First row contains the header information
* The first column contains the location of the raw file (for paired end sequencing data the first 2 columns)
* The second column (third column for paired end) contains the sample names, which are used for output files
* The sample names should be called sample\_names in the header, otherwise the pipeline won't automatically create an R/Bioconductor ExpressionSet
* In addition to raw file location and sample name the file can also contain any number of additional phenotype columns
* The last column is used for a PCA that is automatically created using the FPKM normalized data

Example:

FastQ\_files sample\_name genotype pregnenolone\_treated Group

/restricted/projectnb/montilab-p/projects/superfund/project5/RNASeq2/raw/Sample\_Ctr-MO\_D1.R1.fastq.gz Sample\_Ctr-MO\_D1.R1 control no Ctrl\_DMSO

/restricted/projectnb/montilab-p/projects/superfund/project5/RNASeq2/raw/Sample\_Ctr-MO\_D2.R1.fastq.gz Sample\_Ctr-MO\_D2.R1 control no Ctrl\_DMSO

/restricted/projectnb/montilab-p/projects/superfund/project5/RNASeq2/raw/Sample\_Ctr-MO\_D3.R1.fastq.gz Sample\_Ctr-MO\_D3.R1 control no Ctrl\_DMSO

/restricted/projectnb/montilab-p/projects/superfund/project5/RNASeq2/raw/Sample\_Ctr-MO\_D4.R1.fastq.gz Sample\_Ctr-MO\_D4.R1 control no Ctrl\_DMSO

/restricted/projectnb/montilab-p/projects/superfund/project5/RNASeq2/raw/Sample\_Ctr-MO\_PN1.R1.fastq.gz Sample\_Ctr-MO\_PN1.R1 control yes Ctrl\_PN

**Parameter file**

All other parameters are specified in the parameter file. This file follows the format <KEY> = <VALUE>, with all test after a '#' removed. It does not allow spaces within keys or values. An example is available in the tar ball.

The example file has a description for each parameter but some of the more important ones are explained here:

* working\_dir : specifies the location where, the intermediate results, deliverables and final report will be output.
* raw\_filenames: file with location of raw files and phenotype information as described above
* paired: Specifies if the data are paired end
* clean\_run: Indicates if the pipeline should be run in resume mode or from scratch. If it runs in resume mode it will only run the steps of the pipeline that haven't finished
* verbose: output the main log also onto the console
* aligner: tophat (standard) or skip if one wants to start with bam files, which are then specified in the raw file directory. In the future there might be multiple aligners available.
* QC\_and\_trim\_only: runs only the initial preprocessing step to make sure the adapter clipping was successful before continuing. One you confirm the QC looks alright you can run the pipleline in resume mode, which the skips the fastqc runs and adapter clipping.
* There are also parameters for every single module (most important are the genome and annotation files) that have to be specified correctly. Common annotations are stored under: /restricted/projectnb/montilab-p/CBMrepositoryData/annot

**Running the pipeline**

Once the parameter file has be specified correctly the pipeline can be simply run:

python2.7 <script\_directory>/RNASeq\_pipeline\_prototype.py -p my\_parameter\_file.txt

**Output**

* The pipeline outputs a report similar to this one:
  + <http://smonti.bumc.bu.edu/~montilab/zoho/superfund/project5/report/index.html>
* A tab delimited text file that contains the FPKM data from cufflinks
* And the raw counts from HTSeq both as tab delimited text file and as R/Bioconductor ExpressionSet: