The ethanomics NGS data analysis suite
Version August 12, 2012

To install the scripts and get them running you must do the following:

Install dependencies:
1. Install samtools. http://samtools.sourceforge.net/
2. Install bedtools. http://code.google.com/p/bedtools/
3. Install the latest version of R (it is preinstalled on Macs).
4. Install the R/Bioconductor packages:
    a) DESeq
    http://www.bioconductor.org/packages/2.6/bioc/html/DESeq.html
    b) goseq
    http://www.bioconductor.org/packages/2.6/bioc/html/goseq.html
    c) GO.db
    www.bioconductor.org/packages/2.10/data/annotation/html/GO.db.html
    d) org.Mm.eg.db
    http://www.bioconductor.org/packages/2.10/data/annotation/html/org.Mm.eg.db.html
    e) org.Hs.eg.db
    http://www.bioconductor.org/packages/2.10/data/annotation/html/org.Hs.eg.db.html
    f) If you are not working on human or mouse the genome wide annotation for your genome.
    http://www.bioconductor.org/packages/2.10/data/annotation/
    If your genome is not available at the above link than the goseq gene ontology portion of ezDESeq will not work but the rest of the script will.

5. Install HTseq. http://www-huber.embl.de/users/anders/HTSeq/doc/index.html
6. Install MACS. http://liulab.dfci.harvard.edu/MACS/
7. Install CEAS. http://liulab.dfci.harvard.edu/CEAS/

IMPORTANT. All the above installation must be in your PATH variable!!!!!

Unpack the ethanomics NGS data analysis suite with the following command:
tar -zxvf ethanomicsNGS.tgz
(if you are reading this you probably have already done this)

Prepare reference genomes
1. Place the 'genomes' directory found in the ethanomicsNGS directory in your home directory. The path to the 'genomes' directory must be:
~/genomes
2. Download iGenomes for hg19, mm9 or whatever species you want.
http://tophat.cbcb.umd.edu/igenomes.html

3. The following instructions are for hg19. If you are using mm9 the process is exactly the same.

a) Place the directories found in the hg19 directory from iGenomes into the hg19 directory found in the 'genomes' directory created in step 1 above.

b) You must have the following paths to these reference files:

Bowtie Index: ~/genomes/hg19/Sequence/BowtieIndex/genome
GTF file: ~/genomes/hg19/Annotation/Archives/archive-current/Genes/genes.gtf
Genome index file: ~/genomes/hg19/Sequence/WholeGenomeFasta/genome.fa.fai

c) O.k. that was the hardest part

4. If you are using a genome other than hg19 or mm9 but there is an available iGenome, pretty much all the scripts will still run with just the iGenome files.

5. If you are using a genome that does not have a iGenome. Just create the file structure in step 3b for your Bowtie index, GTF file and Genome index file and pretty much all the scripts will run.

Install the ethanomics NGS scripts

1. Place the NGS_scripts directory found in the ethanomicsNGS directory and place it in your home directory. The path should be ~/NGS_scripts

That's it. You're ready to go!!!

ezDESeq.sh

Version August 12, 2012

This script runs DESeq in a nice easy automated way. It does the following:

1) Maps reads to the genome/transcriptome with Tophat.
2) Removes reads mapped to chrM.
3) Assigns mapped reads genes using HTseq-count.
4) Merges HTseq-count output into a counts table.
5) Runs DESeq using the script DESeqWrapper.R.
6) Runs goseq using the script GOseqWrapper.R.
7) Makes bedGraph files for visualization on a genome browser.

To run ezDESeq.sh

1. Create a directory with the name of your experiment, e.g. 'malakia' or whatever you like.

2. In the experiment directory (e.g malakia) create two more directories. One for each condition, e.g. 'control' and 'treated'. You can use whatever two names you like.

3. Place your fastq files* (gziped is ok) in either the control or treated directories. For example, the path to your fastq files should be like this:
/pathToYourExpDirectory/malakia/control/yourfiles.fastq
/pathToYourExpDirectory/malakia/treated/yourfiles.fastq
4. Open a terminal and change directories using the 'cd' command so that you are in the experiment directory (e.g. malakia).
5. Type sh /NGS_Scripts/ezDESeq.sh in the command line and follow the prompts.
IMPORTANT
*Your fastq files must end in .fastq or .fastq.gz if they are gzipped. If they do not end it .fastq, change their name before running the script!!!

ezMACS.sh
Version August 12th, 2012

This script is a nice easy ChIP-seq data analysis pipeline. It does the following:
1. Maps your reads to a reference genome with Bowtie.
2. Trims your reads with Trimmomatic (optional).
3. Finds enriched 'peaks' with MACS.
4. Makes BED files and trims reads that map off the ends of chrmosomes.
5. Makes bedGraph files normalized by total reads per sample.
4. Runs CEAS (optional).

To run ezMACS.sh
1. Create a directory named after your experiment.
2. Place your two fastq files in the directory (gzipped is ok). One should be the "ChIP sample" and the other the "input" sample.
3. Your fastq files must end in .fastq or .fastq.gz if gzipped. If they do not have either of these ending, change their names so that they end in .fastq (or .fastq.gz if gzipped).
3. Open a terminal and change directories so you are in the directory with you fastq files.
4. Run script by typing sh ~/NGS_scripts/ezMACS.sh
5. Follow prompts

bedEndRepair.pl
Takes the reads that map off the end of chromosomes and repositions them as two base pair reads at the end of the chromosome.

1) make a genome index file with samtools with the following command: 'samtools faidx <ref.fasta>' or make a 2 column tab delimited text file. The first column contains the chromosome name written as it is in column 1 of your bed file (e.g. chr1) and the second column has the length of the corresponding chromosome (an integer). **Note:** If you set up the file structure as described above, iGenomes

contains a genome index file, which should now be at ~/genomes/
/Sequence/REPLACE_WITH_YOUR_GENOME_NAME/WholeGenomeFasta/genome.fa.fai

2) Run with the following command
perl bedEndRepair.pl path/to/your/indexfile.fai path/to/your/bedfile.bed
e.g.
perl bedEndRepair.pl ~/genomes/Sequence/hg19/WholeGenomeFasta/genome.fa.fai ./yourBedFile.bed
3) Output is saved in your current working dirctory with the name yourBedFile.repaired.bed

randomLines.pl
Outputs a random number of lines from a given input file. Often very useful for making control files which a matching number of lines as your experimental file.

IMPORTANT -- This script only works with files that end in .bed. If your file ends in something else, you can modify the script at line 15 to accomodate your file name or change your file name so it ends in .bed.

Usage: perl ~/NGS_Scripts/randomLines.pl <path/to/yourfile> <NumberOfLines>

subsampler.py
Same idea as randomLines.pl but works with fastq files. No my work. It's from here: http://seqanswers.com/forums/showthread.php?t=16505

ezCuffdiff.sh
This script just runs Cuffdiff but save the effort of having to type all the names of your BAM files in command line when you run Cuffdiff.

Usage:
1) Set up file struture as described in ezDESeq readme.txt (This has been done on the Thanos Lab MacPro Cluster)
2) Create a new directory named after you experiment.
3) In that directory create directories for each of your conditions. Up to four conditions are supported at this time.
4) Place your BAM files (must end in .bam) for each condition in the appropriate directory.
5) Open a terminal window and change directories to the directory named after your experiment.
6) Run script by entering "sh ~/NGS_Scripts/ezCuffdiff.sh" in the command line. To run from the Thanos Lab MacPro you only need to type "ezCuffdiff.sh" in the command line.

bedGraphMaker.sh
This script takes a directory of sam or bam files and converts them to bedGraph files scaled by read number.

Usage:
1) Set up file struture as described in ezDESeq readme.txt (This has been done on the Thanos Lab MacPro Cluster)
2) Create a new directory and place all the sam or bam files that you would like to convert into bedGraphs in to it.
3) Open terminal and change directories to the directory with your sam or bam files.
4) Begin script by typing “sh ~/NGS_Scripts/bedGraphMaker.sh” in the command line.
5) Follow prompts.