Gene Quantification Pipeline from Guigo Lab CRG

Input: pairs of long RNA-seq FASTQs. The library accessions are attached ("library.accessions.list"). [Note: Alternatively we can provide FASTQ file accessions].

Each FASTQ pair is fed to STAR 2.3.1z12 (https://github.com/alexdobin/STAR/releases). The corresponding command is specified in "STAR_command" section together with the accompanying parameter section "STAR_Parameters". $genomeDir is the directory where the genome index (GI) to map against is. There is one GI per sex, and each library is mapped against the GI of its biosample's sex, if known. If unknown, the mapping is done against the male genome [Note: We can provide both GIs, which are ~30GB each]. The GIs also index gencode 19 transcript annotations, in order to improve read mappings across splice junctions. The output of STAR is one BAM file per library (i.e. FASTQ pair).

We then proceed with the transcript quantifications, which are performed with Flux Capacitor 1.6.1 (FC, http://sammeth.net/confluence/display/FLUX/Home). FC takes as input (1) a BAM file, and (2) a GTF file. The GTF file consists in the official Gencode 19 GTF, from which short RNA annotations, i.e. transcripts with short RNA biotype (see "small_tx_types" section) have been subtracted. The FC command is specified in the "FLUX_command" section. The output of FC is a GTF file containing transcript records. Each record contains its inferred expression values (in both raw read count and RPKM units) as GTF attributes (9th field). Gene quantifications (also in both "reads" and "RPKM" units) are then obtained by summing up the expression values of all transcripts that compose it, according to the input annotation.

Each quantified gene is then scored by npIDR (non-parametric IDR, https://github.com/pervouchine/npIDR, see "npIDR_command" section on its "reads" value (raw read count) across pairs of replicates of the same experiment.

A matrix of gene RPKM and npIDR across all pairs of replicates is then built from the values obtained in each experiment. More precisely the matrix contains genes as rows and pairs of replicates of the same experiment as columns (identified in the header using the following format: labExpId1,labExpId2:rnaExtract.cell.localization). Values of the matrix are formatted in the following way: RPKM1:RPKM2:npIDR where RPKM1 is the gene RPKM in labExpId1, RPKM2 is the gene RPKM in labExpId2 and npIDR is the npIDR value. We usually consider as reliable, gene quantifications for which npIDR<=0.1 in a given experiment.

STAR COMMAND

STAR --parametersFiles STAR_Parameters --readFilesCommand zcat --readFilesIn $fastq1 $fastq2 --genomeDir $genomeDir --outFileNamePrefix $workDir/ --outStd SAM | samtools view -b -u -S - | samtools sort -m 20000000000 - $workDir/$outputName

Variables:
$fastq1 = Fastq file read 1
$fastq2 = Fastq file read 2
$genomeDir = STAR genome directory path
$workDir = Working directory. This may be different from
the final output directory (eg: when running on a cluster)
$outputName = The name of the bam file (without the .bam extention).
This should match the LID

NOTES:
1 - The STAR parameters "genomeLoad" and "runThreadN" can be set however you like,
depending on the environment where you are running STAR
2 - We use samtools 0.1.18
3 - iYou can of course change the value of "samtools sort -m xxx" to a smaller value,
but be aware of it's disk usage (especially on a cluster node)

IF RUNNING ON A CLUSTER
1 - I reserve about 55GB of memory for each job. STAR needs to load the genome into memory,
plus 20GB for samtools sort, plus extra wiggle room. You can of course reduce this if
you reduce the samtools sort memory reservation.
2 - Decompressing the fastq files in STAR uses threads. So rememeber to reserve 2 extra slots.
EG: If running STAR with 10 threads, reserve 12.
3 - If using a cluster node's local storage temporarily, be sure to reserve it. Eg: ~50GB or more

STAR PARAMETERS

genomeLoad LoadAndRemove
runThreadN 6
outSAMunmapped Within
outFilterType BySJout
outFilterMultimapNmax 20
alignSJoverhangMin 8
alignSJDBoverhangMin 1
outFilterMismatchNmax 999
outFilterMismatchNoverLmax 0.04
alignIntronMin 20
alignIntronMax 1000000
alignMatesGapMax 1000000

FLUX COMMAND

flux-capacitor --sort-in-ram --output $fluxOut.gtf --read-strand MATE2_SENSE --input $bamFile --annotation $annotationGtf --annotation-mapping PAIRED_STRANDED

small_tx_types

Mt_rRNA
Mt_tRNA
Mt_tRNA_pseudogene
miRNA
miRNA_pseudogene
misc_RNA
misc_RNA_pseudogene
rRNA
rRNA_pseudogene
scRNA_pseudogene
snRNA
snRNA_pseudogene
snoRNA
snoRNA_pseudogene
tRNA
tRNA_pseudogene
tRNAscan

npIDR COMMAND

R --slave --args $replicate1$replicate2.matchedReadCounts.txt $replicate1$replicate2.iIDR.out 1 2 < npIDR.r