# Processing pipeline

### **Datasets processed using this version of the pipeline:**

* GSE62074; We considered two samples of HCC1419 cells that were untreated and another two treated with Lapatinib for only 9 days before harvesting (drug tolerant persisters, DTPs).

### **Processing script:**

* <https://github.com/dcic/RNA-seq-pipelines/blob/master/DCIC/RNA-Seq_annotated_pipeline.R>

### **Reference genome:**

* hg19; FASTA file is downloaded from UCSC as a part of illumina distribution: http://support.illumina.com/sequencing/sequencing\_software/igenome.html

### **Feature definitions:**

* A gene is considered as a feature. We downloaded the gene annotation gtf file from UCSC with the genome version “hg19” (hg19, GRCh37 Genome Reference Consortium Human Reference 37 (GCA\_000001405.1)). For our example, we found 26832 mapped genes.

### **Pre-alignment (fastq) QC:**

* FastQC (version: v0.9.4)

### **Samples removed based on pre-alignment QC:**

* None

### **Criteria for removing samples based on pre-alignment QC:**

* None

### **Aligner information and parameter specifications:**

* software: TopHat (version: v2.0.3), Bowtie2 (version: 2.2.9), Samtools (version: 1.3.1)
* "-p" indicate number of threads to run which is 1; -G: gtf file location; --library-type: fr-unstranded.

### **Post-alignment (bam) QC:**

* RNA-SeQC (version: 1.1.7)

### **Samples removed based on post-alignment QC:**

* None

### **Criteria for removing samples based on post-alignment QC:**

* Pearson’s correlation coefficient between different samples
* Remove a sample if it’s correlation is low with the other samples.

### **Read counts software and parameters:**

* R packages:GenomicAlignments (version: 1.6.3), GenomicFeatures (version: 1.22.13)
* Download refGene as a TranscriptDb(TxDb) object for “hg19” and download exons by gene using the refGene table.
* Counting reads using summarizeOverlaps with mode='Union' and inter.feature=TRUE.

### **Read counts QC:**

* Counts per million (CPM)

### **Samples removed based on read counts QC:**

* None

### **Criteria for removing samples based on read counts QC:**

* Genes with a count per million (CPM) value greater than 1 in more samples than the smaller sample size between two groups are retained for the subsequent analyses and other genes are filtered out.

### **Batch information and batch effect adjustments:**

* None

### **Statistical analysis software, model, parameters:**

* software: R package:edgeR (version:3.12.1)
* A simple two group comparison using Fisher’s exact test.
* Normalization: trimmed mean of M-values(TMM); dispersion="auto"