RNASeq dataset QC Wiki

Reference genome: hg38

Fasta file <ftp://ftp.ensembl.org/pub/release-85/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz>

GTF file: <ftp://ftp.ensembl.org/pub/release-85/gtf/homo_sapiens/Homo_sapiens.GRCh38.85.chr.gtf.gz>

All original fastq files are stored at Orchestra:

/groups/sorger/Marc/RNAseq\_data/RNAseq\_BrCaPorfiling\_201604/run1\_20160419

/groups/sorger/Marc/RNAseq\_data/RNAseq\_BrCaPorfiling\_201604/run2\_20160420

Report of FastQC 0.11.5 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

FastQC report is generated to have an initial view of the data, and generally the quality(Per base/tile sequence quality) of each data is good for further analysis in the report. The full QC report is stored at Orchestra:

/groups/sorger/cchris/qc\_out

QC with Trimmomatic 0.36 (<http://www.usadellab.org/cms/?page=trimmomatic>)

The trimming process includes:

Paired End Mode

Remove leading low quality or N bases (below quality 3) (LEADING:3)

Remove trailing low quality or N bases (below quality 3) (TRAILING:3)

Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)

Drop reads below the 36 bases long (MINLEN:36)

The trimmed fastq files are stored at Orchestra:

/groups/sorger/cchris/run\_qc

Aligner: STAR 2.5.2a

(<http://bioinformatics.oxfordjournals.org/content/early/2012/10/25/bioinformatics.bts635>)

Genome indices are built by STAR in mode “--runMode genomeGenerate” with the fasta file and GTF file mentioned above.

Alignment is performed with mode: “--outSAMtype BAM Unsorted --quantMode GeneCounts”.

Aligned BAM files are stored at Orchestra:

/groups/sorger/cchris/align\_tophat

BAM files to RPKM: R package Rsubread and edgeR

(

[https://www.bioconductor.org/packages/devel/bioc/manuals/Rsubread/man/Rsubread.pdf,](https://www.bioconductor.org/packages/devel/bioc/manuals/Rsubread/man/Rsubread.pdf,https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf))

[https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf)](https://www.bioconductor.org/packages/devel/bioc/manuals/Rsubread/man/Rsubread.pdf,https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf))

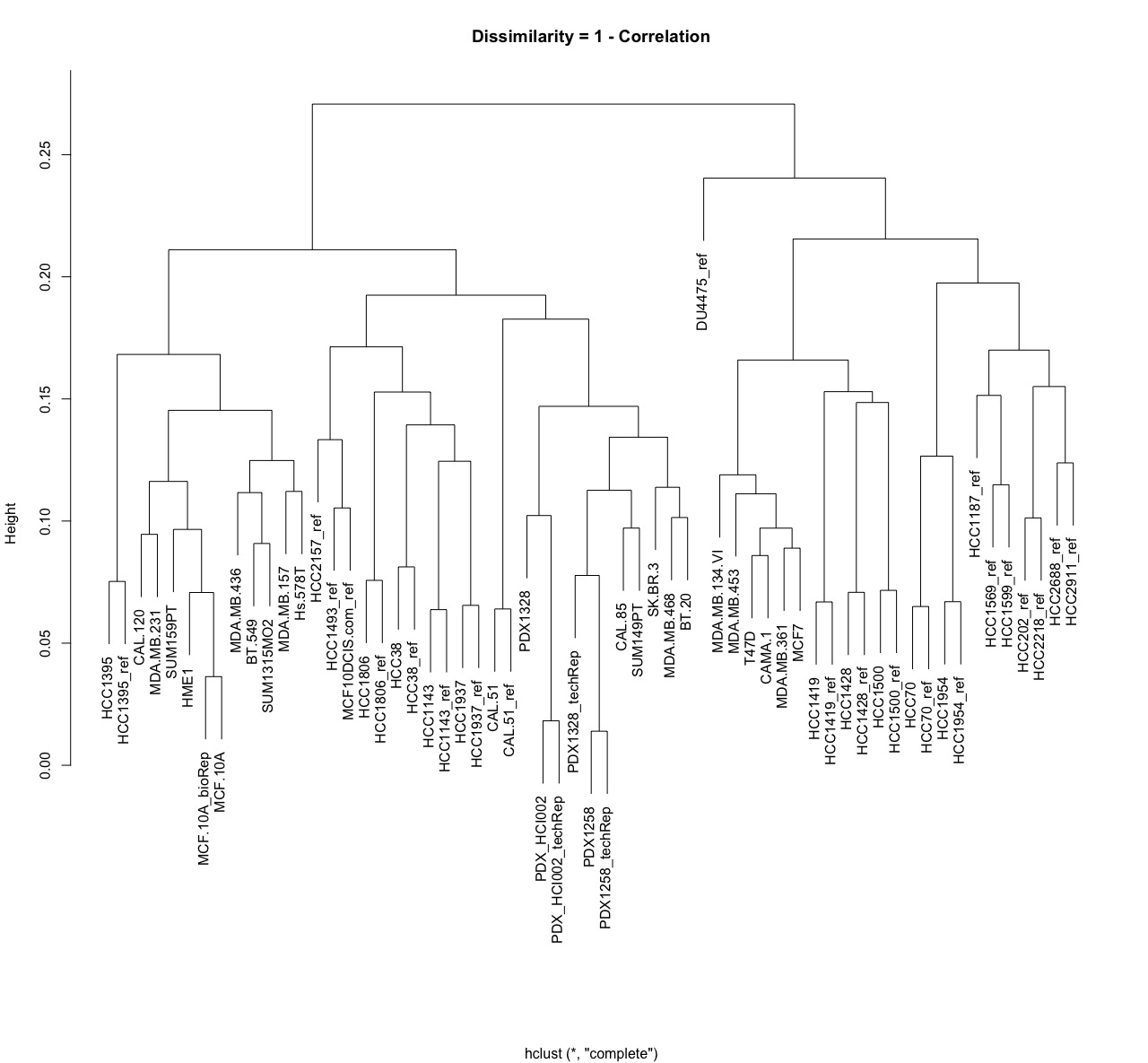
Feature counts are garneted from BAM files by the “featureCounts” function of Rsubread, and RPKM is calculated by the “rpkm” function of edgeR. The results are uploaded on synapse: <https://www.synapse.org/#!Synapse:syn7072782>

Row names are HGNC gene symbol, and column names are cell lines based on the first column of the reference: <https://www.synapse.org/#!Synapse:syn7068205>

Hierarchy clustering is performed in a combination of our dataset and RNAseq from Genentech:

<http://www.nature.com/nbt/journal/v33/n3/abs/nbt.3080.html>

First we removed columns that have 0 variance and use (1 - spearman correlation among samples) as distance, cell lines appeared in reference are those end with “\_ref”.



For another approach, log-transformed RPKM values from 2 datasets are first normalized by sva (<https://bioconductor.org/packages/release/bioc/html/sva.html)> to remove batch effects, then Euclidean distance and complete agglomeration method are used for the clustering.