Identify long non-coding (lnc) using RNA-seq

July 10 2023

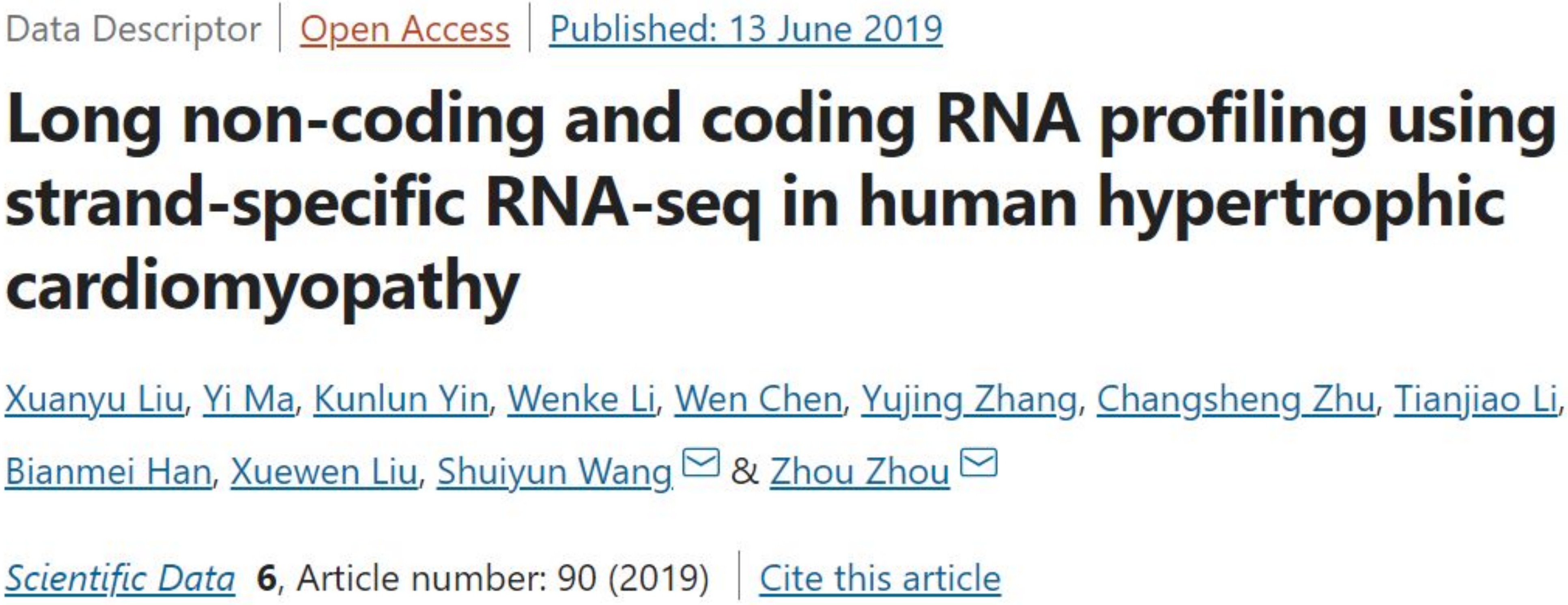
Phuc Loi Luu, PhD

Loi.lp@pacificinformatics.com.vn

Luu.p.loi@googlemail.com

Content

- **“Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy”**
- **Homework**

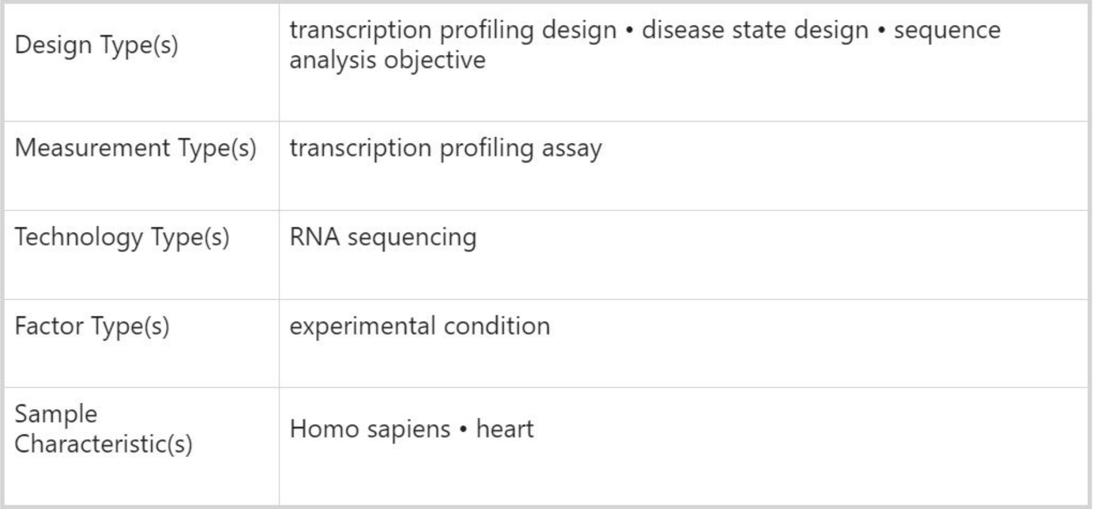


Data Descriptor | Open Access | Published: 13 June 2019

Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy

Xuanyu Liu, Yi Ma, Kunlun Yin, Wenke Li, Wen Chen, Yujing Zhang, Changsheng Zhu, Tianjiao Li, Bianmei Han, Xuewen Liu, Shuiyun Wang ✉ & Zhou Zhou ✉

Scientific Data **6**, Article number: 90 (2019) | Cite this article



Design Type(s)	transcription profiling design • disease state design • sequence analysis objective
Measurement Type(s)	transcription profiling assay
Technology Type(s)	RNA sequencing
Factor Type(s)	experimental condition
Sample Characteristic(s)	Homo sapiens • heart

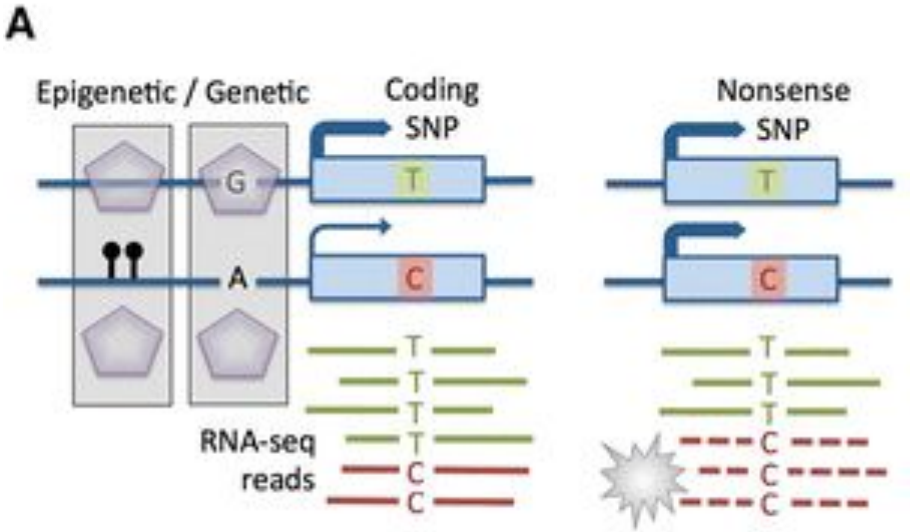
https://www.nature.com/articles/s41597-019-0094-6

A strand-specific RNA-seq

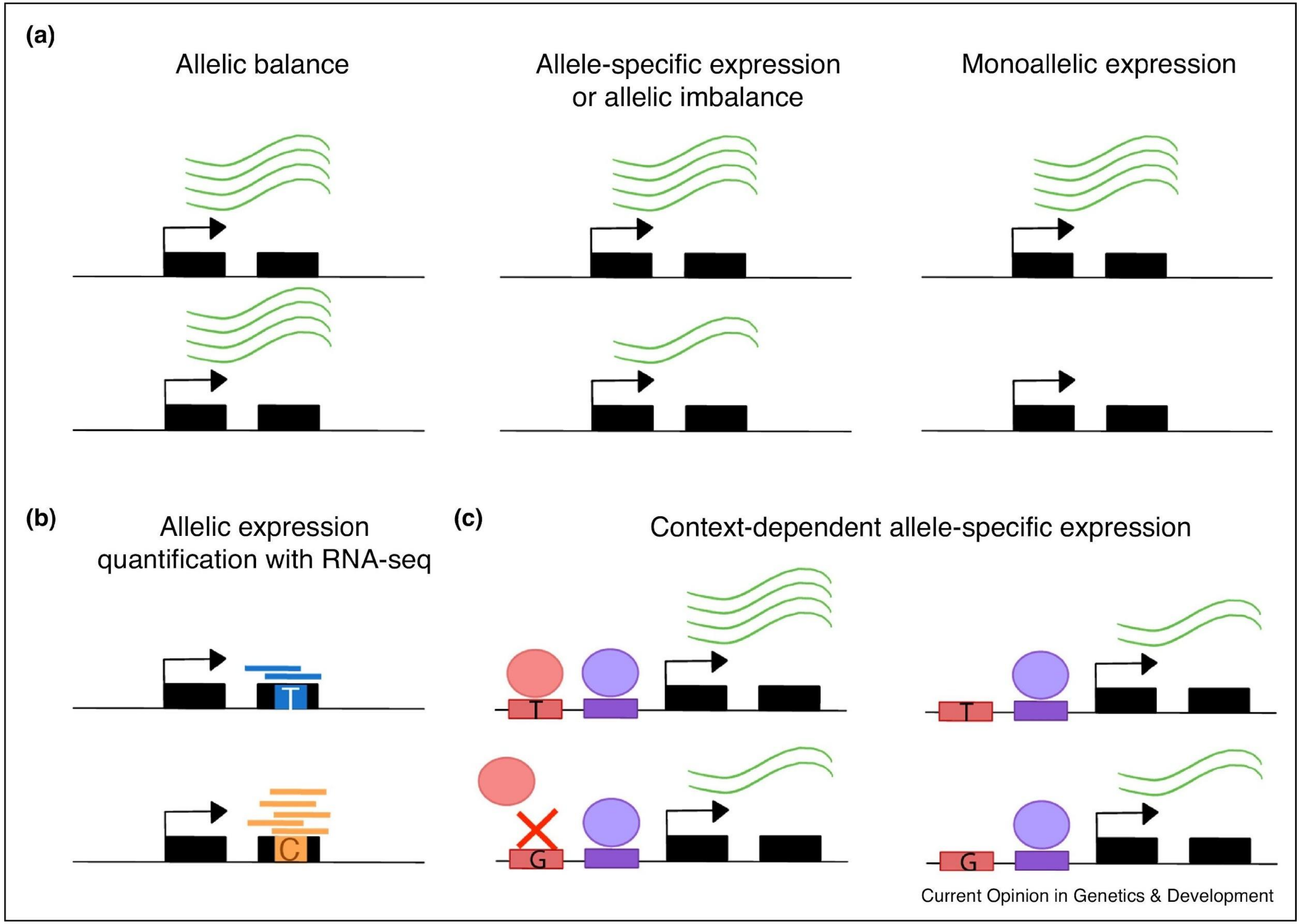
- RNA-seq > microarray
  - novel transcript identification through de novo assembly
  - splice junction identification
  - allele-specific expression analysis
- Standard RNA-seq protocol vs strand-specific RNA-seq retains strand of origin information

□ a greater resolution for sense/antisense profiling, which is essential for antisense lncRNA identification

Allele-specific expression

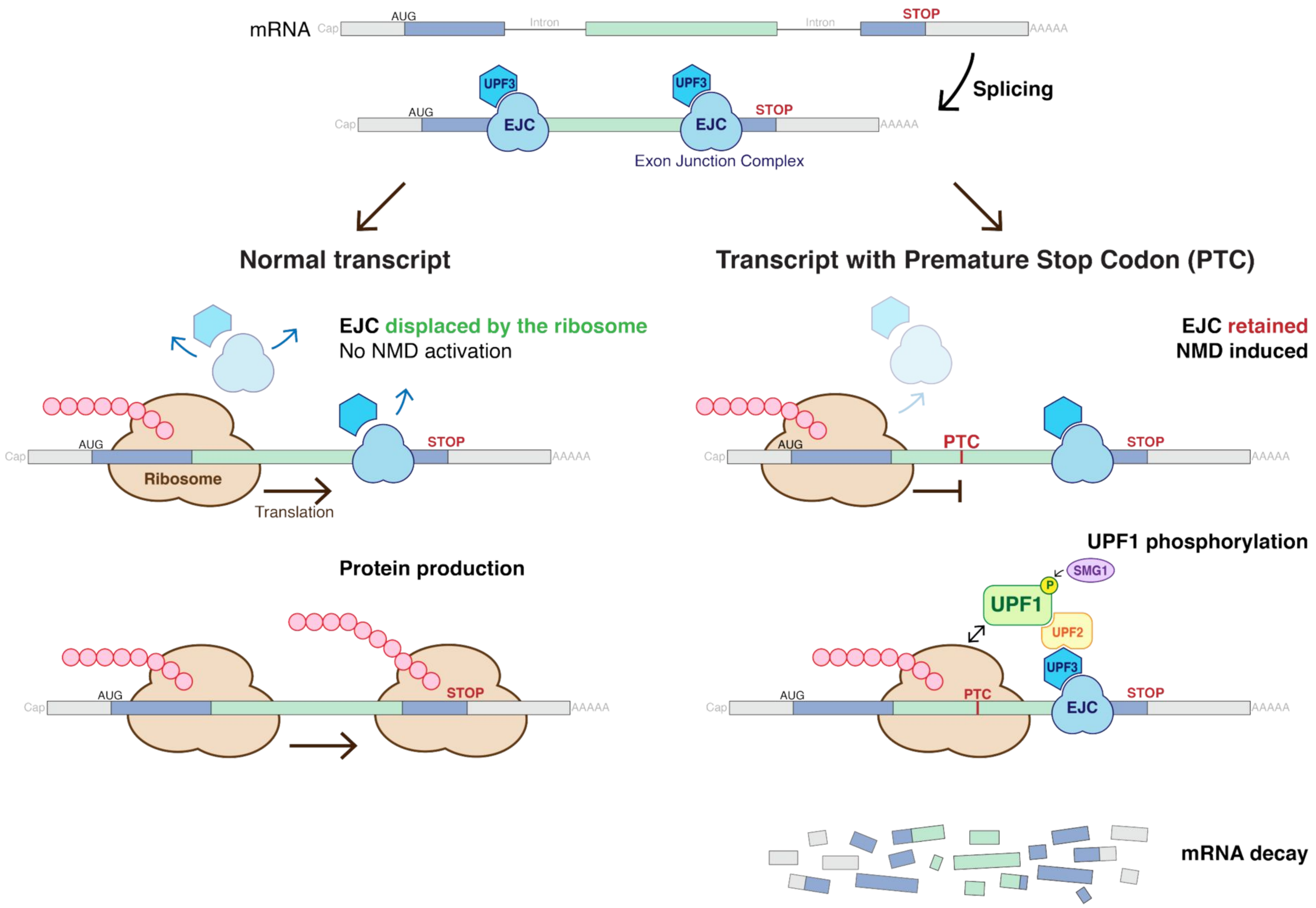


https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0762-6


(a)
Allelic balance
Allele-specific expression or allelic imbalance
Monoallelic expression
(b)
Allelic expression quantification with RNA-seq
T
C
(c)
Context-dependent allele-specific expression
T
G
T
G
Current Opinion in Genetics & Development

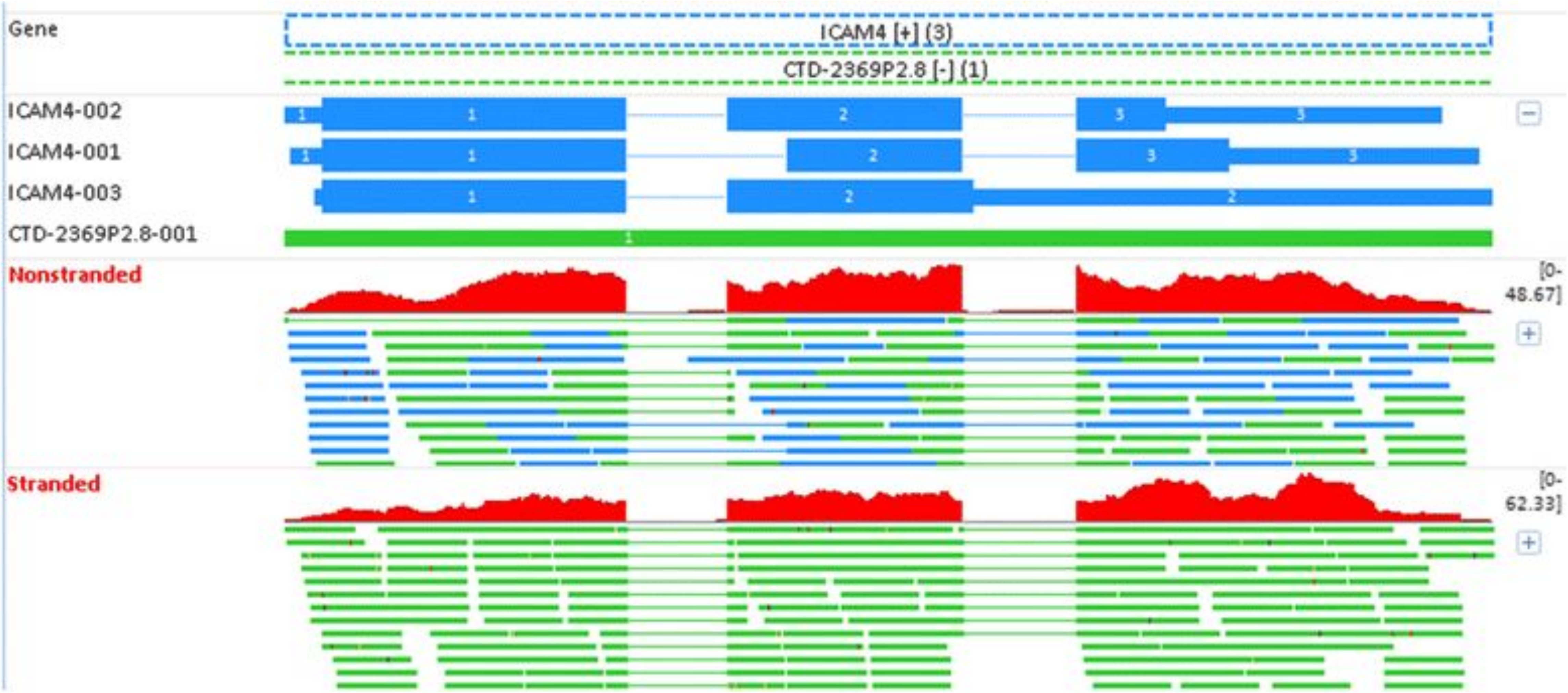
https://www.sciencedirect.com/science/article/pii/S0959437X20301465

Nonsense-mediated decay



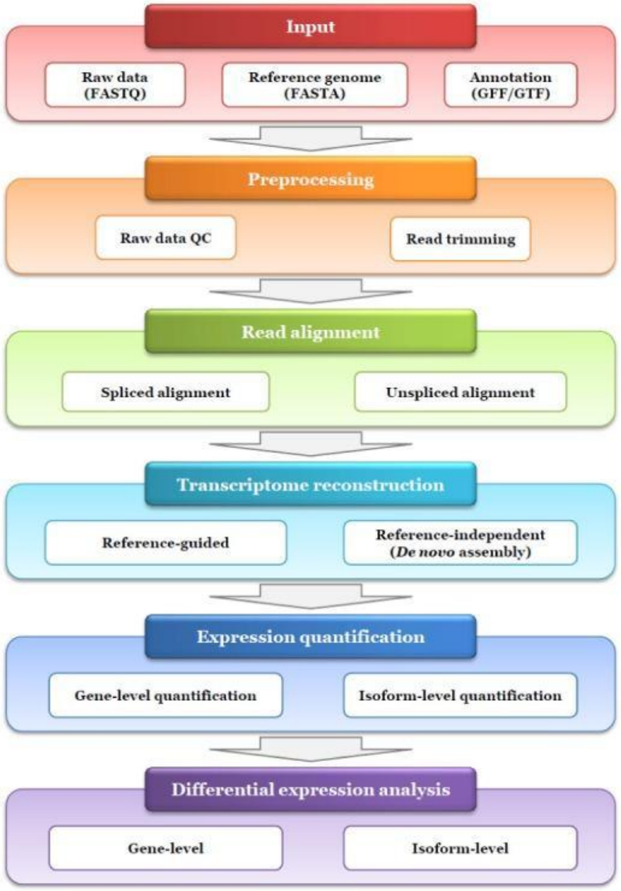
https://en.wikipedia.org/wiki/Nonsense-mediated_decay

Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap

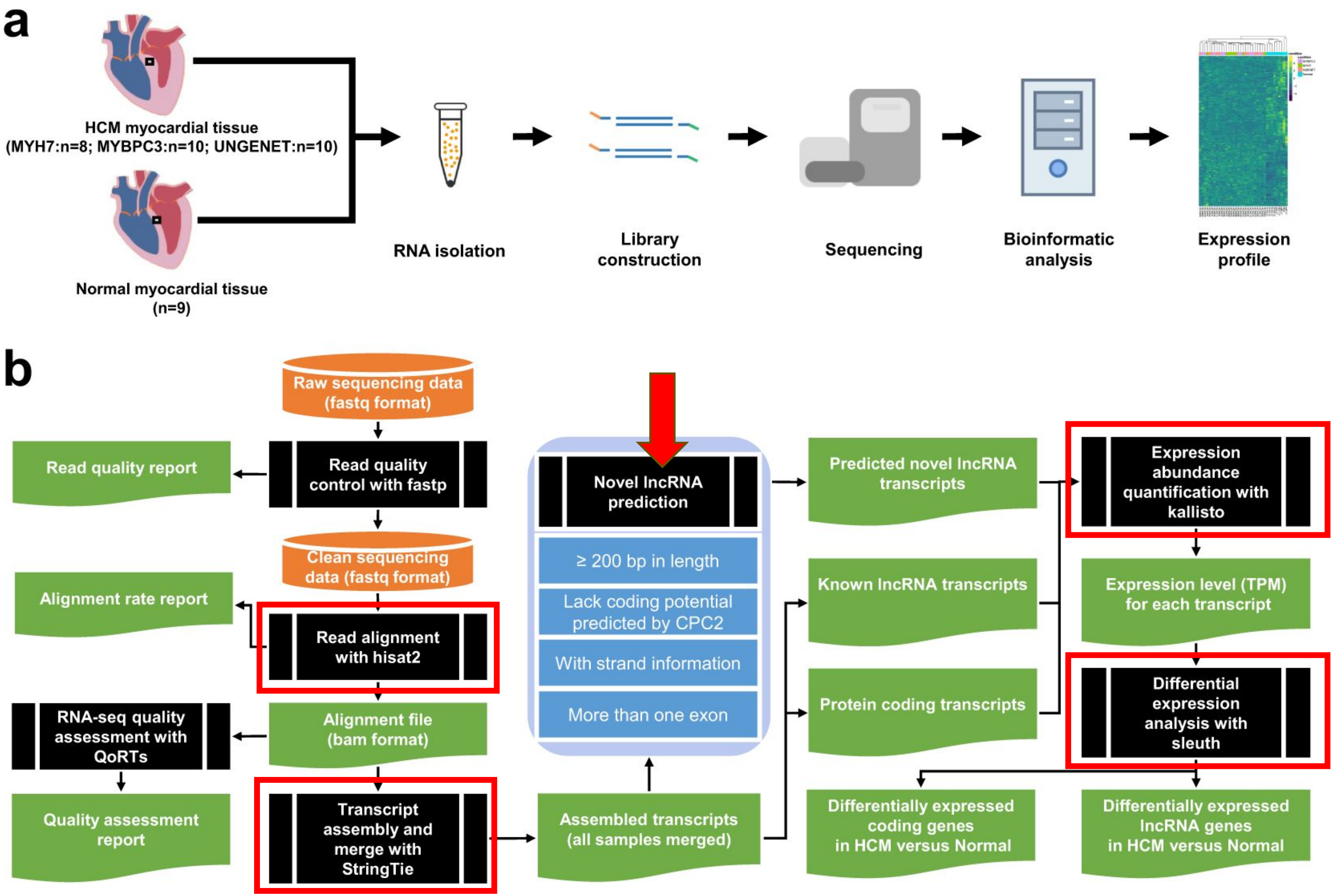


The mapping profiles for *ICAM4* (intercellular adhesion molecule 4) in Replicate PFE1. The gene *ICAM4* is on the "+" strand, and 100 % contained within *CTD-2369P2.8* in the "−" strand. In non-stranded RNA-seq, the ambiguous reads in overlapping regions are excluded from counting, which explains why there is no expression for *ICAM4*. However, the ambiguous reads can be perfectly resolved in stranded RNA-seq. By considering the read direction, all reads can be counted to *ICAM4* because they are reverse complementary to *ICAM4*, but not *CTD-2369P2.8.* All genes, transcripts, and sequence reads are colored in blue if they are in the "+" strand and colored in green if in the "−" strand.

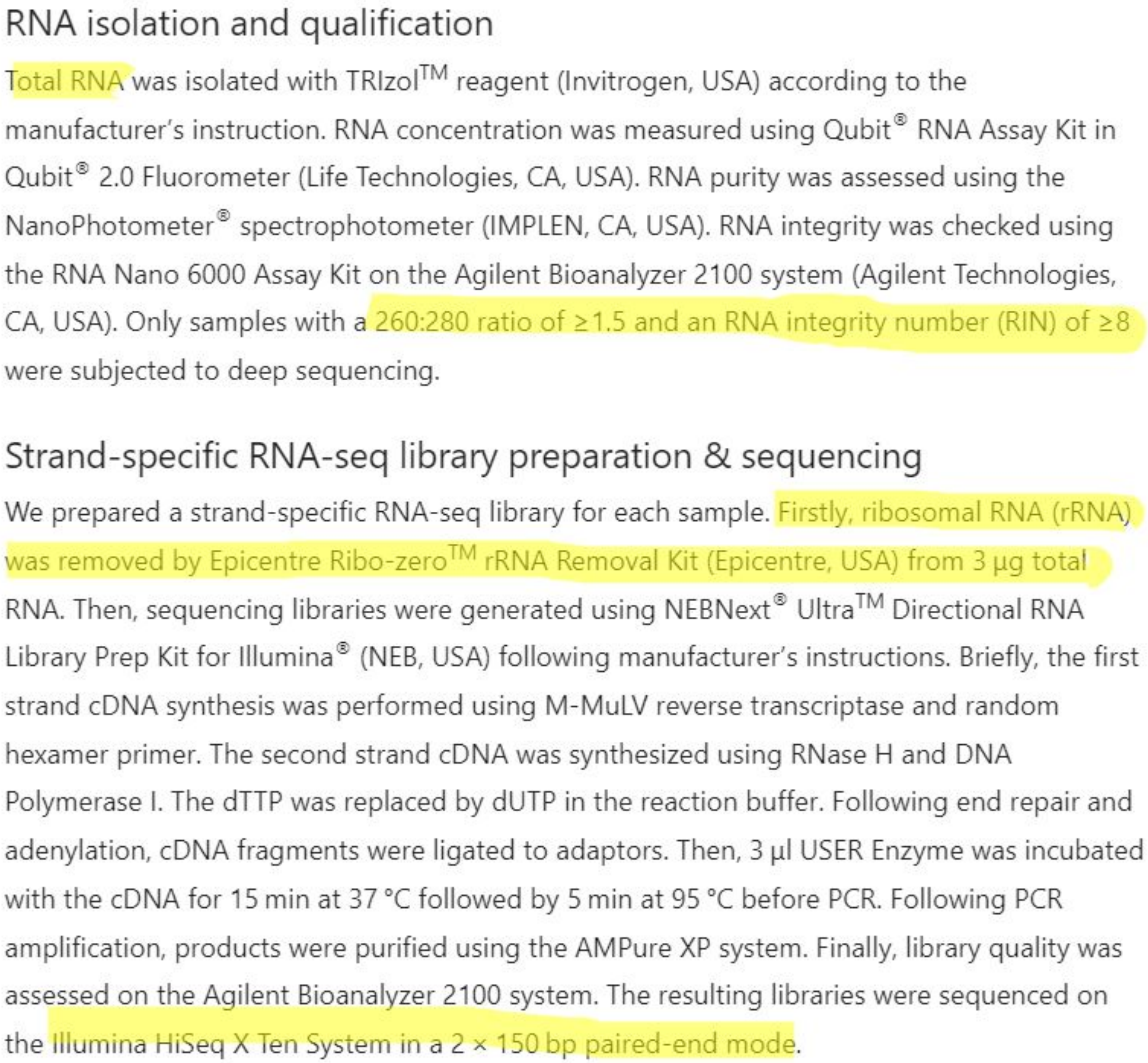
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1876-7



https://www.rna-seqblog.com/review-of-rna-seq-data-analysis-tools/


a
HCM myocardial tissue
(MYH7:n=8; MYBPC3:n=10; UNGENET:n=10)
Normal myocardial tissue
(n=9)
RNA isolation
Library construction
Sequencing
Bioinformatic analysis
Expression profile
b
Raw sequencing data (fastq format)
Read quality report
Read quality control with fastp
Clean sequencing data (fastq format)
Alignment rate report
Read alignment with hisat2
RNA-seq quality assessment with QoRTs
Alignment file (bam format)
Quality assessment report
Transcript assembly and merge with StringTie
Novel lncRNA prediction
≥ 200 bp in length
Lack coding potential predicted by CPC2
With strand information
More than one exon
Assembled transcripts (all samples merged)
Predicted novel lncRNA transcripts
Known lncRNA transcripts
Protein coding transcripts
Expression abundance quantification with kallisto
Expression level (TPM) for each transcript
Differential expression analysis with sleuth
Differentially expressed coding genes in HCM versus Normal
Differentially expressed lncRNA genes in HCM versus Normal

Overview of the experimental procedure. (**a**) Schematic representation of the experimental workflow. The sampling position is indicated by a black rectangular. RNA isolation and library preparation for all samples were performed in the same batch. HCM: hypertrophic cardiomyopathy; GENETUN: Genetically undiagnosed HCM; MYBPC3: HCM patient with mutation in *MYBPC3*; MYH7: HCM patient with mutation in *MYH7*; NORMAL: Normal heart. (**b**) Bioinformatic analysis workflow.

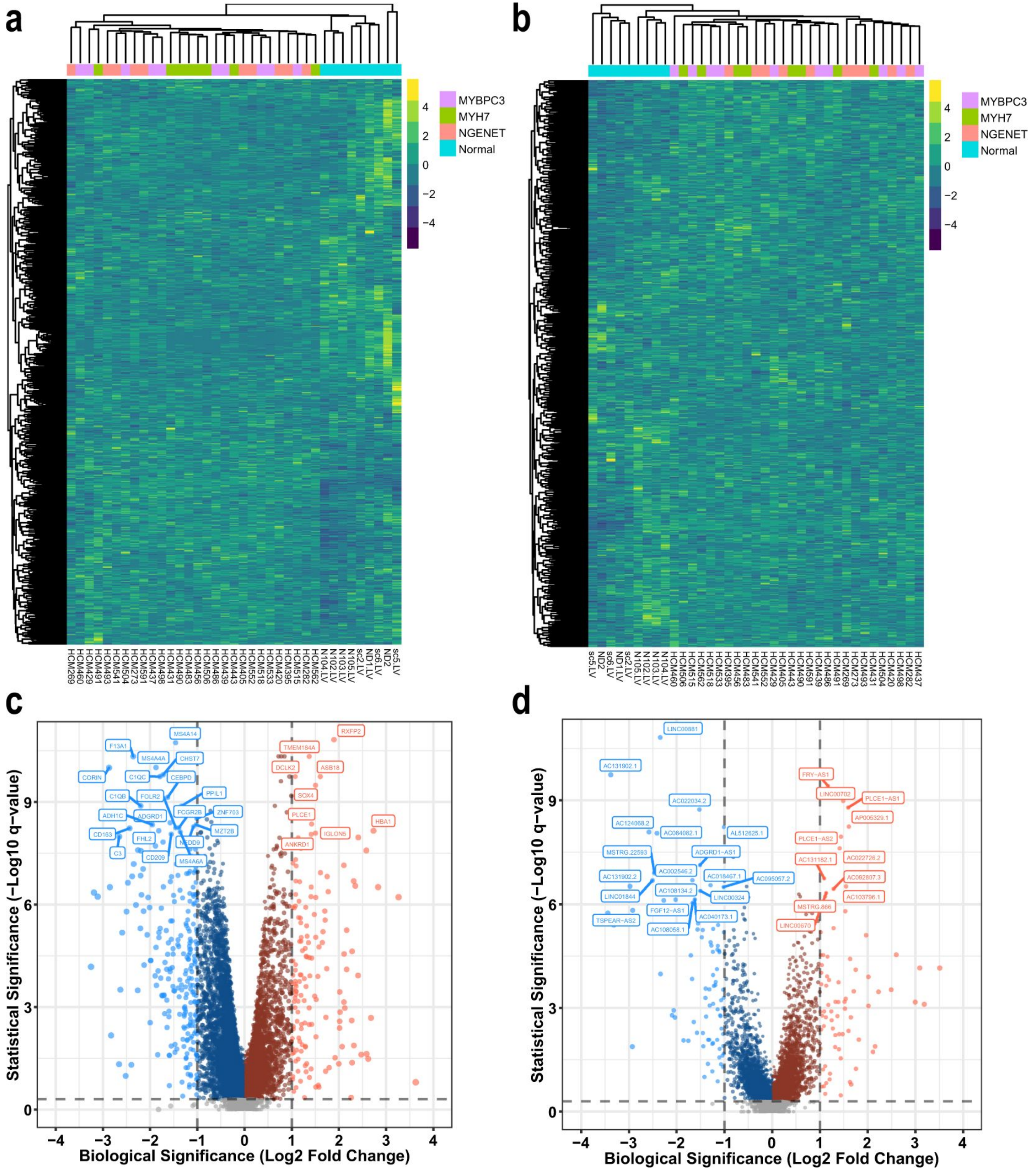


RNA isolation and qualification

Total RNA was isolated with TRIzol$^{\text{TM}}$ reagent (Invitrogen, USA) according to the manufacturer's instruction. RNA concentration was measured using Qubit$^{®}$ RNA Assay Kit in Qubit$^{®}$ 2.0 Fluorometer (Life Technologies, CA, USA). RNA purity was assessed using the NanoPhotometer$^{®}$ spectrophotometer (IMPLEN, CA, USA). RNA integrity was checked using the RNA Nano 6000 Assay Kit on the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Only samples with a 260:280 ratio of ≥1.5 and an RNA integrity number (RIN) of ≥8 were subjected to deep sequencing.

Strand-specific RNA-seq library preparation & sequencing

We prepared a strand-specific RNA-seq library for each sample. Firstly, ribosomal RNA (rRNA) was removed by Epicentre Ribo-zero$^{\text{TM}}$ rRNA Removal Kit (Epicentre, USA) from 3 µg total RNA. Then, sequencing libraries were generated using NEBNext$^{®}$ Ultra$^{\text{TM}}$ Directional RNA Library Prep Kit for Illumina$^{®}$ (NEB, USA) following manufacturer's instructions. Briefly, the first strand cDNA synthesis was performed using M-MuLV reverse transcriptase and random hexamer primer. The second strand cDNA was synthesized using RNase H and DNA Polymerase I. The dTTP was replaced by dUTP in the reaction buffer. Following end repair and adenylation, cDNA fragments were ligated to adaptors. Then, 3 µl USER Enzyme was incubated with the cDNA for 15 min at 37 °C followed by 5 min at 95 °C before PCR. Following PCR amplification, products were purified using the AMPure XP system. Finally, library quality was assessed on the Agilent Bioanalyzer 2100 system. The resulting libraries were sequenced on the Illumina HiSeq X Ten System in a 2 × 150 bp paired-end mode.



Expression profiles of coding and lncRNA genes.

(a) Hierarchical clustering of the samples from the three HCM groups and the normal group based on the expression of coding genes.

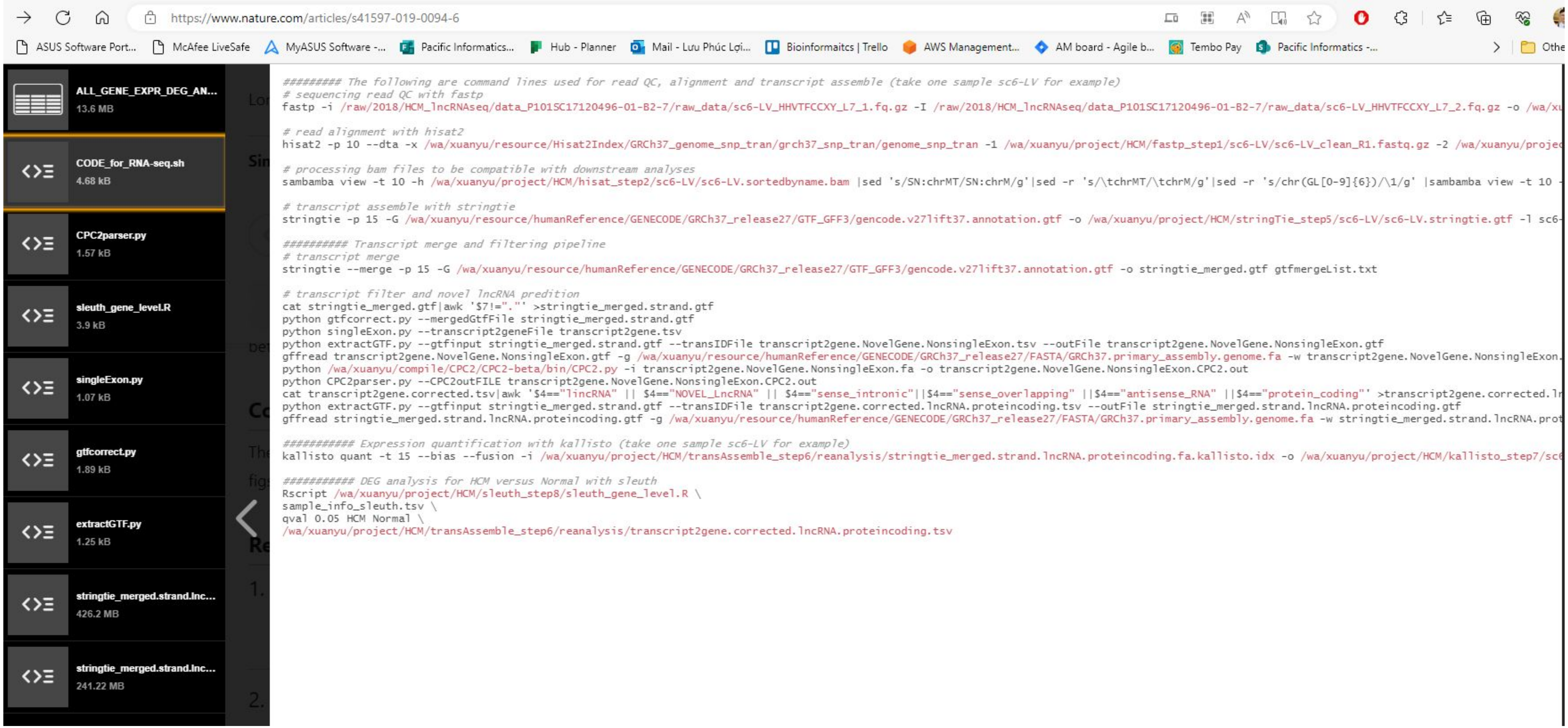
(b) Hierarchical clustering of the samples from the three HCM groups and the normal group based on the expression of lncRNA genes.

In a and b, each row represents a gene, and each column represents a sample. For better visualization, only the expression of 1,000 randomly selected genes are displayed on the heatmap.

(c) Volcano plot showing the differentially expressed coding genes between HCM and normal groups.

(d) Volcano plot showing the differentially expressed lncRNA genes between HCM and normal groups.

In c and d, dots coloured in light red or light blue denote statistically and biologically significant genes being up-regulated or down-regulated, respectively. The dot size reflects the absolute fold change. Only the top 30 DEGs were labelled with gene symbols.

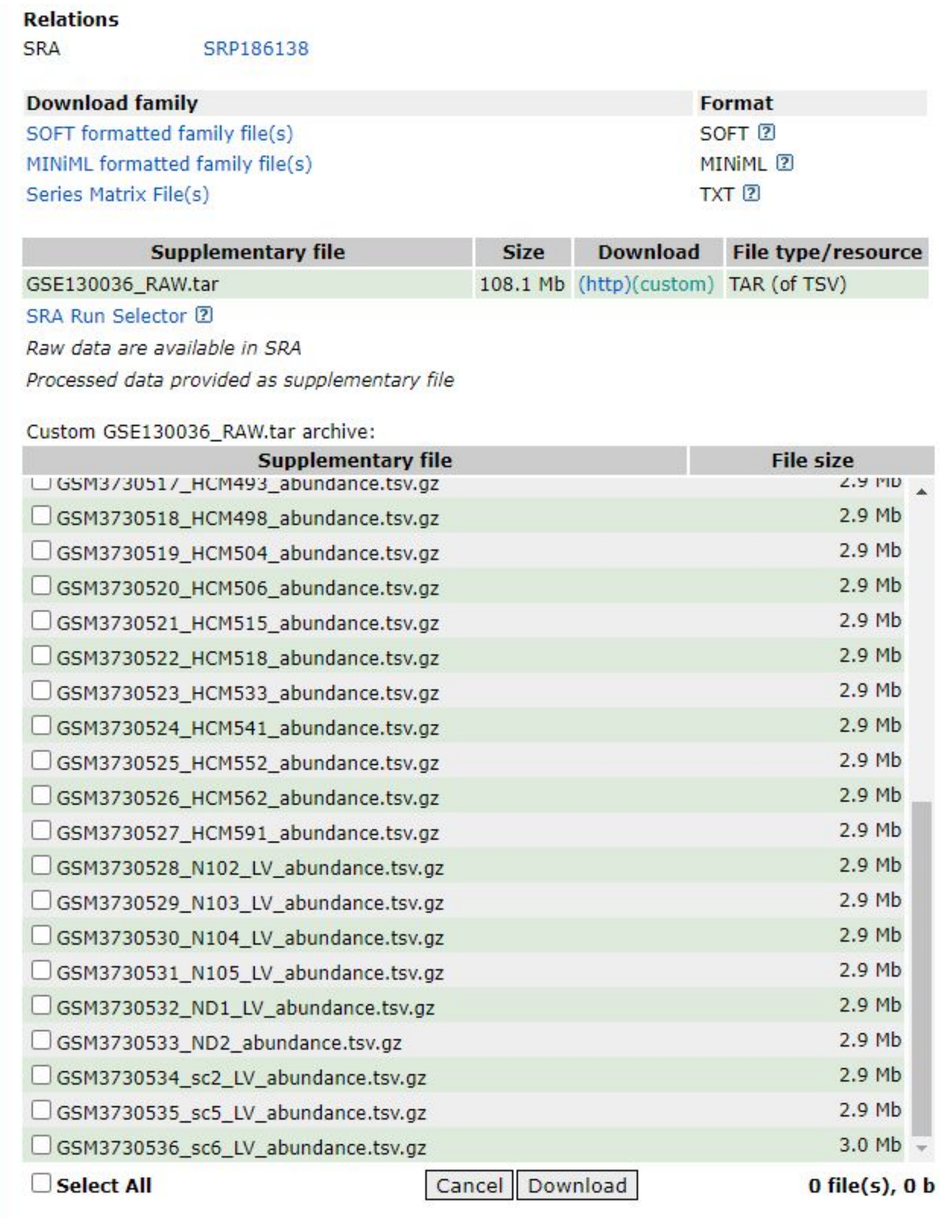


```
######### The following are command lines used for read QC, alignment and transcript assemble (take one sample sc6-LV for example)
# sequencing read QC with fastp
fastp -i /raw/2018/HCM_lncRNAseq/data_P101SC17120496-01-B2-7/raw_data/sc6-LV_HHVTFCCXY_L7_1.fq.gz -I /raw/2018/HCM_lncRNAseq/data_P101SC17120496-01-B2-7/raw_data/sc6-LV_HHVTFCCXY_L7_2.fq.gz -o /wa/xu

# read alignment with hisat2
hisat2 -p 10 --dta -x /wa/xuanyu/resource/Hisat2Index/GRCh37_genome_snp_tran/grch37_snp_tran/genome_snp_tran -1 /wa/xuanyu/project/HCM/fastp_step1/sc6-LV/sc6-LV_clean_R1.fastq.gz -2 /wa/xuanyu/projec

# processing bam files to be compatible with downstream analyses
sambamba view -t 10 -h /wa/xuanyu/project/HCM/hisat_step2/sc6-LV/sc6-LV.sortedbyname.bam |sed 's/SN:chrMT/SN:chrM/g'|sed -r 's/\tchrMT/\tchrM/g'|sed -r 's/chr(GL[0-9]{6})/\1/g' |sambamba view -t 10 -

# transcript assemble with stringtie
stringtie -p 15 -G /wa/xuanyu/resource/humanReference/GENECODE/GRCh37_release27/GTF_GFF3/gencode.v27lift37.annotation.gtf -o /wa/xuanyu/project/HCM/stringTie_step5/sc6-LV/sc6-LV.stringtie.gtf -l sc6-

########## Transcript merge and filtering pipeline
# transcript merge
stringtie --merge -p 15 -G /wa/xuanyu/resource/humanReference/GENECODE/GRCh37_release27/GTF_GFF3/gencode.v27lift37.annotation.gtf -o stringtie_merged.gtf gtfmergeList.txt

# transcript filter and novel lncRNA predition
cat stringtie_merged.gtf|awk '$7!="."' >stringtie_merged.strand.gtf
python gtfcorrect.py --mergedGtfFile stringtie_merged.strand.gtf
python singleExon.py --transcript2geneFile transcript2gene.tsv
python extractGTF.py --gtfinput stringtie_merged.strand.gtf --transIDFile transcript2gene.NovelGene.NonsingleExon.tsv --outFile transcript2gene.NovelGene.NonsingleExon.gtf
gffread transcript2gene.NovelGene.NonsingleExon.gtf -g /wa/xuanyu/resource/humanReference/GENECODE/GRCh37_release27/FASTA/GRCh37.primary_assembly.genome.fa -w transcript2gene.NovelGene.NonsingleExon.
python /wa/xuanyu/compile/CPC2/CPC2-beta/bin/CPC2.py -i transcript2gene.NovelGene.NonsingleExon.fa -o transcript2gene.NovelGene.NonsingleExon.CPC2.out
python CPC2parser.py --CPC2outFILE transcript2gene.NovelGene.NonsingleExon.CPC2.out
cat transcript2gene.corrected.tsv|awk '$4=="lincRNA" || $4=="NOVEL_LncRNA" || $4=="sense_intronic"||$4=="sense_overlapping" ||$4=="antisense_RNA" ||$4=="protein_coding"' >transcript2gene.corrected.ln
python extractGTF.py --gtfinput stringtie_merged.strand.gtf --transIDFile transcript2gene.corrected.lncRNA.proteincoding.tsv --outFile stringtie_merged.strand.lncRNA.proteincoding.gtf
gffread stringtie_merged.strand.lncRNA.proteincoding.gtf -g /wa/xuanyu/resource/humanReference/GENECODE/GRCh37_release27/FASTA/GRCh37.primary_assembly.genome.fa -w stringtie_merged.strand.lncRNA.prot

########### Expression quantification with kallisto (take one sample sc6-LV for example)
kallisto quant -t 15 --bias --fusion -i /wa/xuanyu/project/HCM/transAssemble_step6/reanalysis/stringtie_merged.strand.lncRNA.proteincoding.fa.kallisto.idx -o /wa/xuanyu/project/HCM/kallisto_step7/sc6

########### DEG analysis for HCM versus Normal with sleuth
Rscript /wa/xuanyu/project/HCM/sleuth_step8/sleuth_gene_level.R \
sample_info_sleuth.tsv \
qval 0.05 HCM Normal \
/wa/xuanyu/project/HCM/transAssemble_step6/reanalysis/transcript2gene.corrected.lncRNA.proteincoding.tsv
```

https://www.nature.com/articles/s41597-019-0094-6

Homework

- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130036
- Download the processed data
- Merge these files into a data frame in R
- Call DEGs between HCM vs healthy
- Plot PCA of all data and plot PCA for the DEGs
- Plot volcano plot and heatmap
- Do the gene set enrichment analysis and plot the results



Relations

SRA	SRP186138

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Supplementary file	Size	Download	File type/resource
GSE130036_RAW.tar	108.1 Mb	(http)(custom)	TAR (of TSV)

SRA Run Selector ?

Raw data are available in SRA

Processed data provided as supplementary file

Custom GSE130036_RAW.tar archive:

Supplementary file	File size
☐ GSM3730517_HCM493_abundance.tsv.gz	2.9 Mb
☐ GSM3730518_HCM498_abundance.tsv.gz	2.9 Mb
☐ GSM3730519_HCM504_abundance.tsv.gz	2.9 Mb
☐ GSM3730520_HCM506_abundance.tsv.gz	2.9 Mb
☐ GSM3730521_HCM515_abundance.tsv.gz	2.9 Mb
☐ GSM3730522_HCM518_abundance.tsv.gz	2.9 Mb
☐ GSM3730523_HCM533_abundance.tsv.gz	2.9 Mb
☐ GSM3730524_HCM541_abundance.tsv.gz	2.9 Mb
☐ GSM3730525_HCM552_abundance.tsv.gz	2.9 Mb
☐ GSM3730526_HCM562_abundance.tsv.gz	2.9 Mb
☐ GSM3730527_HCM591_abundance.tsv.gz	2.9 Mb
☐ GSM3730528_N102_LV_abundance.tsv.gz	2.9 Mb
☐ GSM3730529_N103_LV_abundance.tsv.gz	2.9 Mb
☐ GSM3730530_N104_LV_abundance.tsv.gz	2.9 Mb
☐ GSM3730531_N105_LV_abundance.tsv.gz	2.9 Mb
☐ GSM3730532_ND1_LV_abundance.tsv.gz	2.9 Mb
☐ GSM3730533_ND2_abundance.tsv.gz	2.9 Mb
☐ GSM3730534_sc2_LV_abundance.tsv.gz	2.9 Mb
☐ GSM3730535_sc5_LV_abundance.tsv.gz	2.9 Mb
☐ GSM3730536_sc6_LV_abundance.tsv.gz	3.0 Mb

☐ **Select All**

0 file(s), 0 b