Supplementary Information

Table 1. Number of cells belonging to each re-analysis unbiased cluster

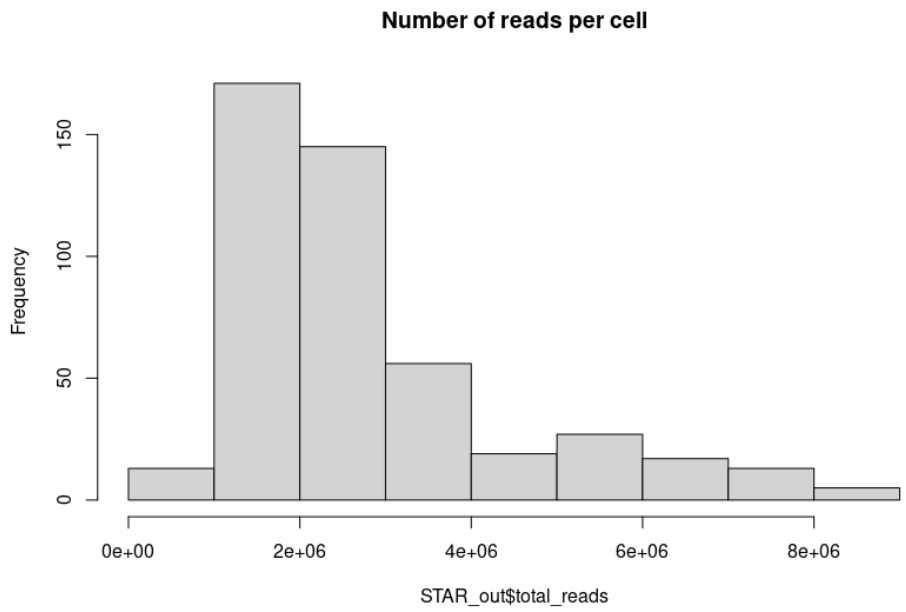
Cluster Number	Number of Cells
1	31
2	169
3	22
4	52
5	65
6	38
7	15
8	56
9	18

The cluster number and number of respective cells was taken directly from the Mclust output summary of the fit.

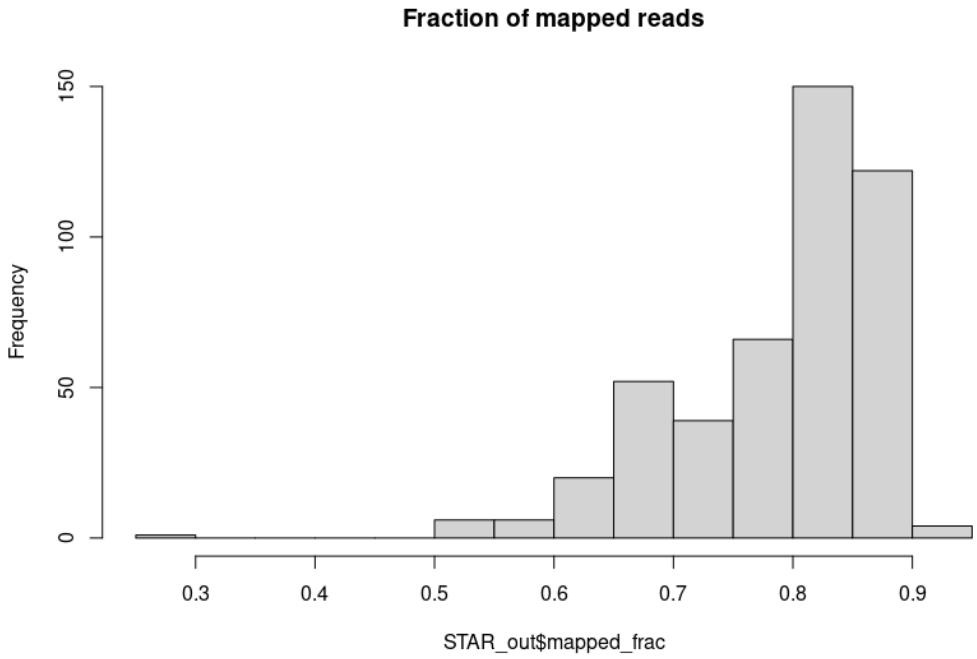
Table 2. Top 20 enriched genes for each of the re-analysis unbiased groups

Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
DCX	OLR1	SYT1	STMN2	PLP1	FGF14	SLC1A2	PLP1	GRIK2
BCL11A	TMSB15A	DCLK1	DCX	VIP	OPCML	NRCAM	CLU	PLP1
SOX11	CENPF	STMN2	SOX11	ATP1A2	CNTN5	PTPRZ1	NEAT1	KCNIP4
DAPK1	CLEC7A	KCNQ5	BCL11A	RNF219-AS1	KCNQ5	TMEM108	SYNPR	SPARCL1
KIAA0319	PLEK	NEDD4L	DLX6-AS1	FOXO1	ASIC2	CADPS	B2M	NTRK2
EPHA4	TMEM98	LINGO2	MCTP1	AXL	GALNTL6	STMN2	VIP	CNTN1
TTTY14	AC010729.1	CHRM3	SLC38A1	PMP22	LINGO2	SLC8A1	RP11-138A9.2	RIMS2
ATP1A2	IL1B	DLX6-AS1	GNAI1	F3	GRIN2B	NELL2	ATP1A2	KCNQ5
GNG2	IFI44L	CKB	GABRA1	GJA1	SCN2A	DCX	NR3C2	CADPS
PLXNA2	TM4SF1	FRMPD4	ADD2	MBP	CNTN4	ATRNL1	ZBTB16	GALNTL6
EZR	ZNF300P1	RYR3	GNG2	ARHGAP42	CHRM3	SLC44A5	CRYAB	MYT1L
LRCH1	C3AR1	NMNAT2	CACNA1E	S100B	FRMPD4	KCNMA1	TMEM144	ASIC2
SEPT3	CLCA4	MGAT4C	GRIA3	SLCO1C1	DLX6-AS1	KCND2	MAP3K5	LINGO2
NETO2	DHRS9	KIRREL3	LINC01122	RP11-138A9.1	STXBP5L	GDAP1	ARAP2	GRIN2B
SMS	C21orf91	GABRB2	AGPAT4	TF	PAK3	GALNT13	SPP1	CHRM3
AKAP12	TOP2A	CACNA1B	REEP1	LINC00499	MEG3	MEIS2	HTRA1	SCN2A
ARHGAP44	C3	KCNJ3	SYT4	AC016831.7	KCNJ3	PCDH7	COBL	ALCAM
DPYSL3	PTPRC	KCNC2	SEZ6L	ETNPPL	GABRB2	SOX11	PPAP2B	MEG3
C3orf70	TFAP2C	RYR2	PLCL2	GLDN	XKR4	MEG3	AXL	FRMPD4
ZNF804A	SAMSN1	SATB2	BCAT1	SPATA6	PCDH11X	CHL1	DOCK5	CNTNAP5

Differential expression for each of the unbiased groups was measuread against the remaining cell population. The results were filtered for adjusted P-values <0.05 and the top 20 logfc genes for each cluster were recorded.

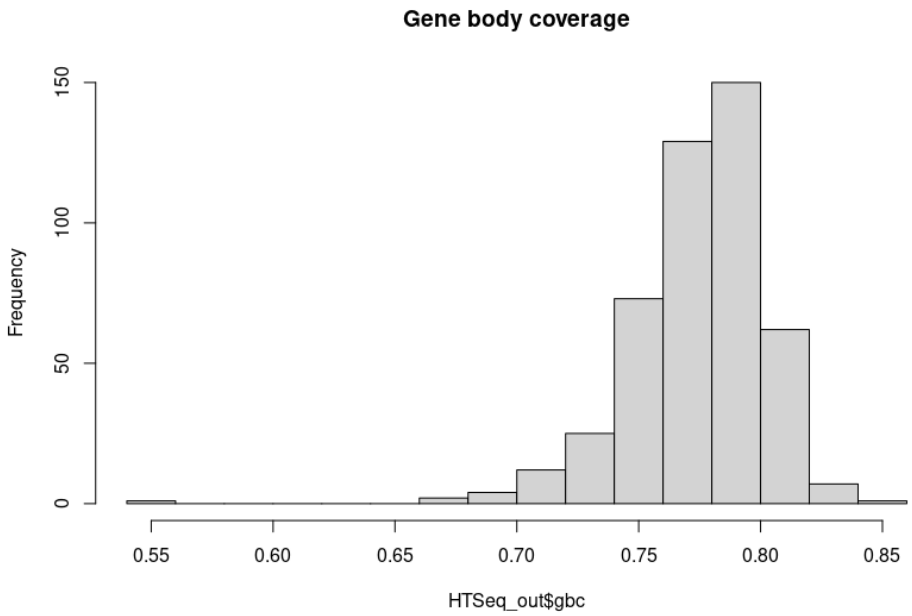


SI Figure 1. Re-analysis number of reads per cell. MultiQC was used to access the STAR output logs to compile a summary containing the reads per each cell.

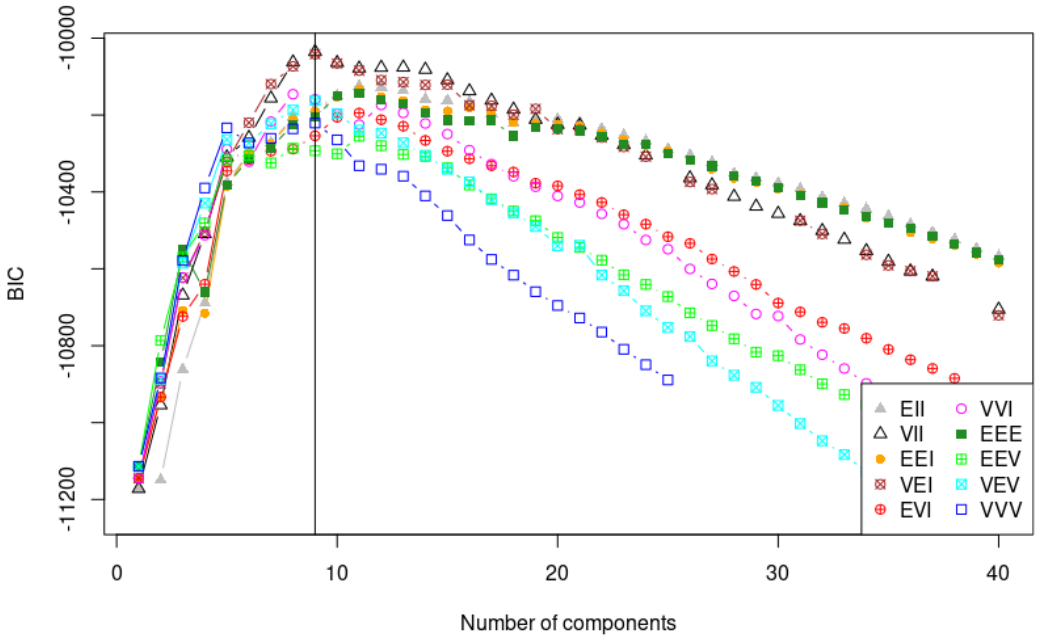


SI Figure 2. Re-analysis fraction of mapped reads. MultiQC was used to access the STAR output logs to compile a summary containing the percentage of mapped reads per each cell, which was converted to fraction.

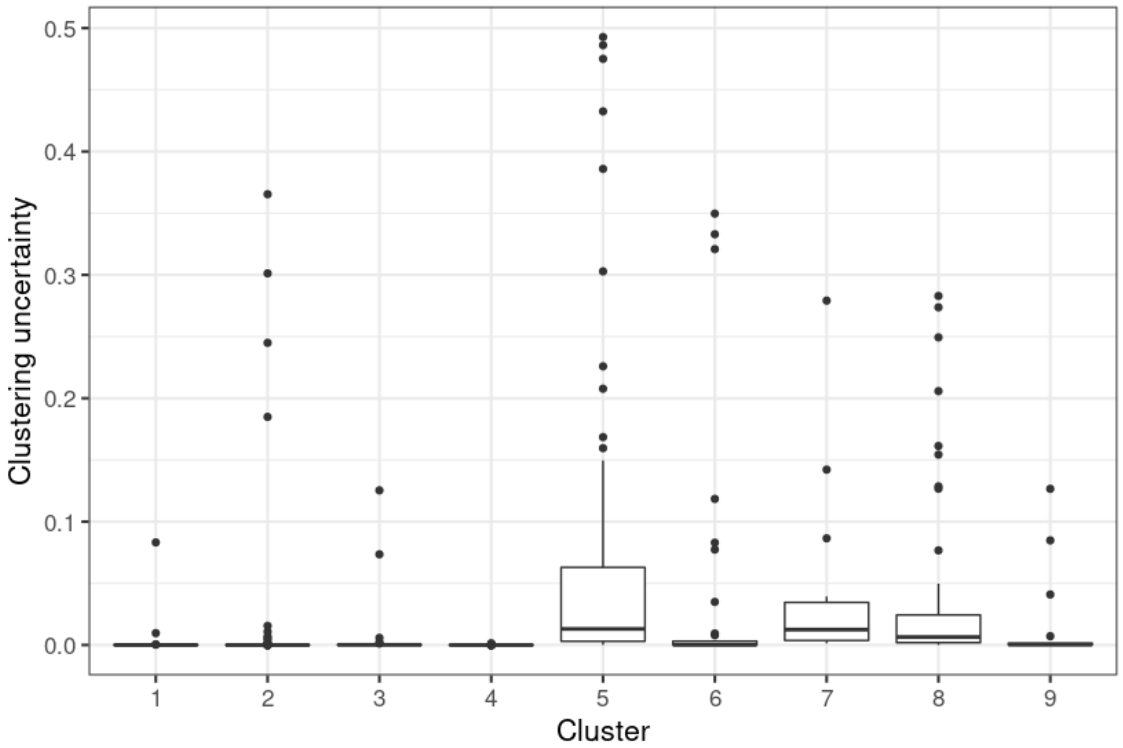
SI Figure 3. Re-analysis gene body coverage



SI Figure 3. Re-analysis gene body coverage. MultiQC was used to access the HTSeq output tsv files to compile a summary containing the gene body coverage per cell.



SI Figure 4. Bayesian information criterion (BIC) plot generated in R package Mclust. A spherical, varying volume model (VII) was used to a fit 9 components.



SI Figure 5. Uncertainty plot generated in R package Mclust depicts the uncertainty of data points for each cluster generated.

Pipeline1 Supplementary Methods

1. Data extraction and reformatting;

1.1 Script '1_fastq_extract.sh' was run to extract SRA files from NCBI short read archive (project SRP057196) using SRA Toolkit 3.0.0.

1.2 Replacement of the .1 .2 extensions with /1 /2 was required so that the reads are recognised as pairs in Prinseq. This was achieved by running script '2_file_reformat.sh' which called python script 'idadaptor.py' to reformat the fastq files.

2. Trimming

2.1 Script '3_prinseq_1.sh'; run to remove reads (prinseq-lite-0.20.4) with a length less than 30 bp, removethe first 10 bp from 5' end, trim reads with quality below phred score 25 on 3' end, filter for low complexity reads.

Prinseq parameters:
-min_len 30 -trim_left 10 -trim_qual_right 25 -lc_method entropy -lc_threshold 65

2.2 Script '4_fastqc_extract.sh' used to QC trimmed fastq files then the FastQC (v0.11.2) out put generated was parsed for over-represented sequences via running script '5_cutadapt.sh' that called python script or_1.py/or_2.py to extract over represented sequences from each set of paired files and apply further trimming with these sequences using Cutadapt (version 2.0). This version of cutadapt is later than the publication date but is the first version following a major bug fix; used to overcome running issues with installation of earlier versions.
https://cutadapt.readthedocs.io/en/stable/changes.html#v2-0-2019-03-06

Cutadapt parameters;
-e 0.15 -m 30

2.3 Prinseq was again utilised to remove lone pairs of reads under 30 bp in length (script '6_prinseq_2.sh', moving forward with only paired read outputs for the next stage of trimming.

Prinseq parameters:
-min_len 30 -out_bad null

2.4 Script '7_trimgalore.sh' enabled TrimGalore (version 0.4.1) to remove nextera adapters from paired end fastq files.
TrimGalore parameters:
--nextera --stringency 1

3. Genome indexing

Script '8_STAR_genomegenerate.sh' was run to generate a genome index using STAR (version 2.4.0h) with GRCh37 release-81 ensembl publications (for both the primary assembly fasta files and GTF annotation) as inputs.

Star parameters:

--runMode genomeGenerate --runThreadN 16 --genomeDir

/scratch/spectre/s/sesm2/SRP/genome_directory --genomeFastaFiles /scratch/spectre/s/sesm2/SRP/Homo_sapiens.GRCh37.75.dna.primary_assembly.fa --sjdbGTFfile /scratch/spectre/s/sesm2/SRP/Homo_sapiens.GRCh37.75.gtf --sjdbOverhang 74 --genomeChrBinNbits 12

4. Alignment and counts conversion

4.1 Script '9_STAR_map.sh' utilised STAR (version 2.4.0h) was used to align each paired set of fastq files to produce a SAM output.

STAR parameters:

--runThreadN 16 \-outFilterType BySJout \--outFilterMultimapNmax 20 \-- alignSJoverhangMin 8 \--alignSJDBoverhangMin 1 \--outFilterMismatchNmax 999 \--outFilterMismatchNoverLmax 0.04 \--alignIntronMin 20 \--alignIntronMax 1000000 \--alignMatesGapMax 1000000 \--outSAMstrandField intronMotif

4.2 Script '10_HTSeq.sh' was run to produce counts from each SAM file using HTSeq (version 0.6.1), incorparating the GRCh37 release-81 ensembl GTF as reference input. The output mapped counts to ensembl gene names.

HTSeq parameters:
-m intersection-nonempty \-s no

4.3 Script '11_CPM_nolog10.sh' was then run to call python script 'cpm_mod_nolog10.py' to convert all HTSeq count outputs to counts per million (CPM) and remove the last five rows containing '__no_feature', '__ambiguous', '__too_low_aQual', '__not_aligned' and '__alignment_not_unique'.

5. Data analysis

Generation of a counts matrix and all subsequent analysis was conducted in R (version 4.1.3) using script 'P1_analysis.R'.See script annotation for more detail on program version, release dates, installation, methodology and parameters.