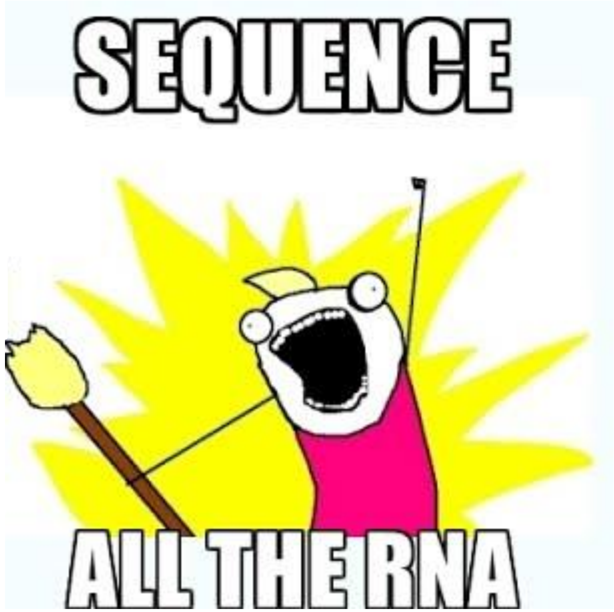
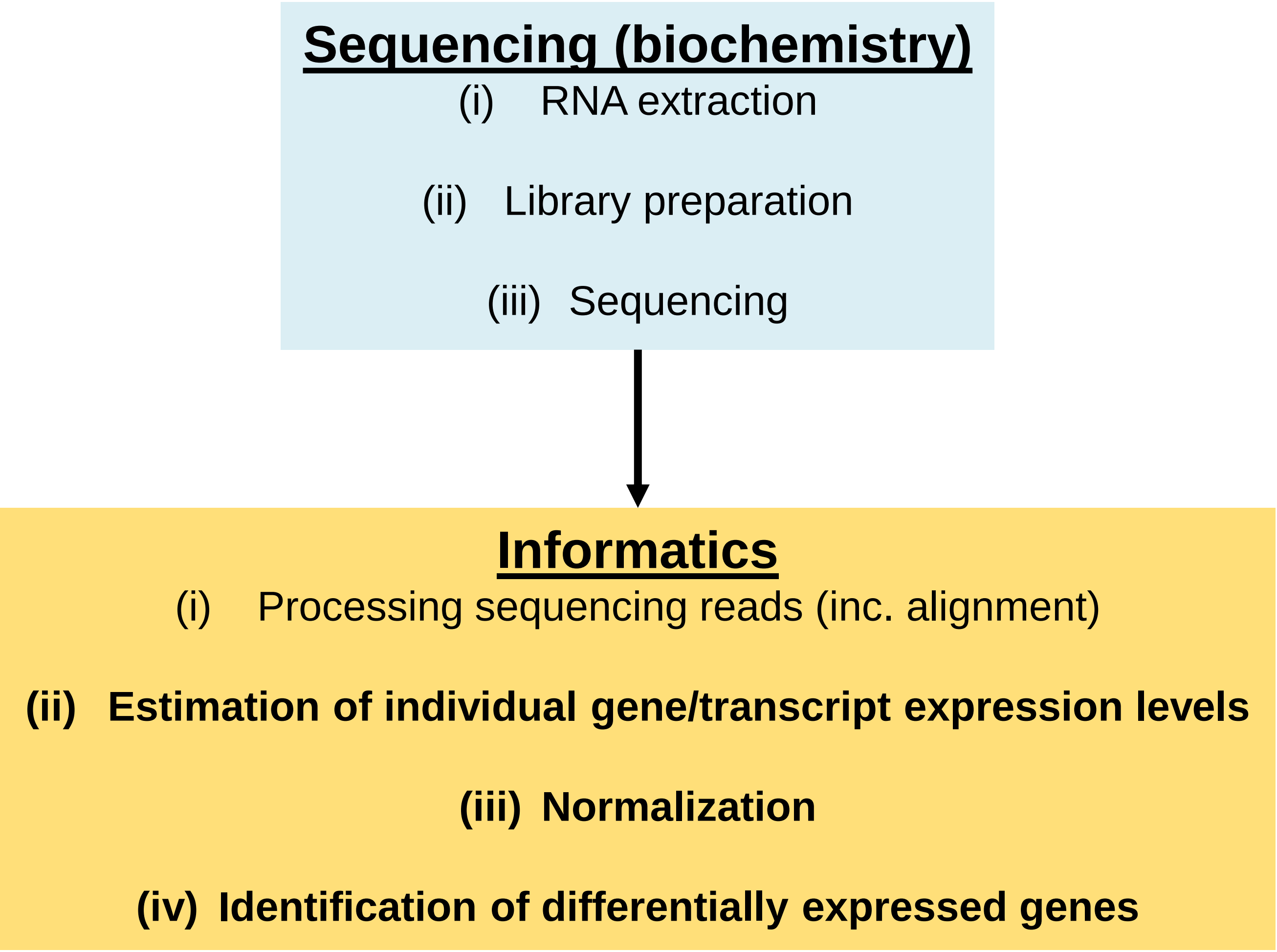
Differential Gene Expression Analysis

Daniel P. Depledge, Ph.D



So you've got some RNA-Seq data… Now what?



Sequencing (biochemistry)

(i) RNA extraction

(ii) Library preparation

(iii) Sequencing

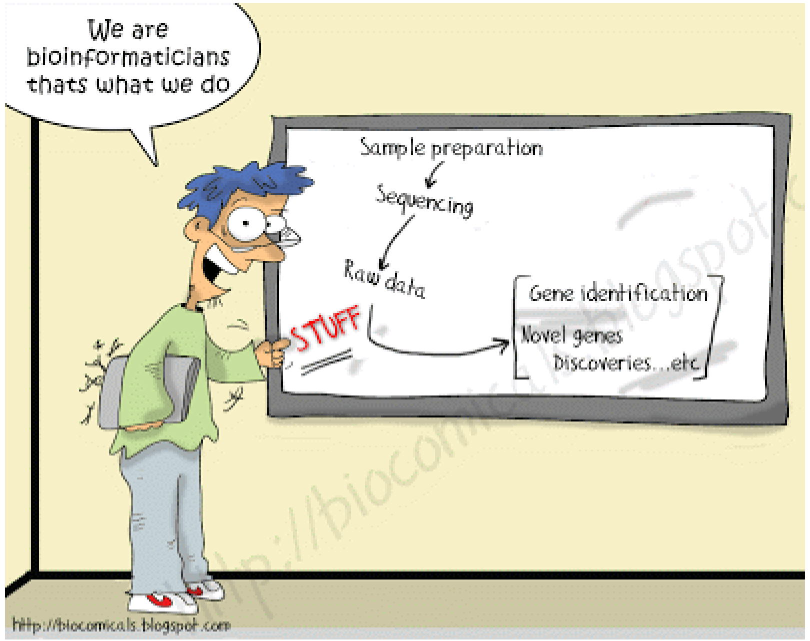
Informatics

(i) Processing sequencing reads (inc. alignment)

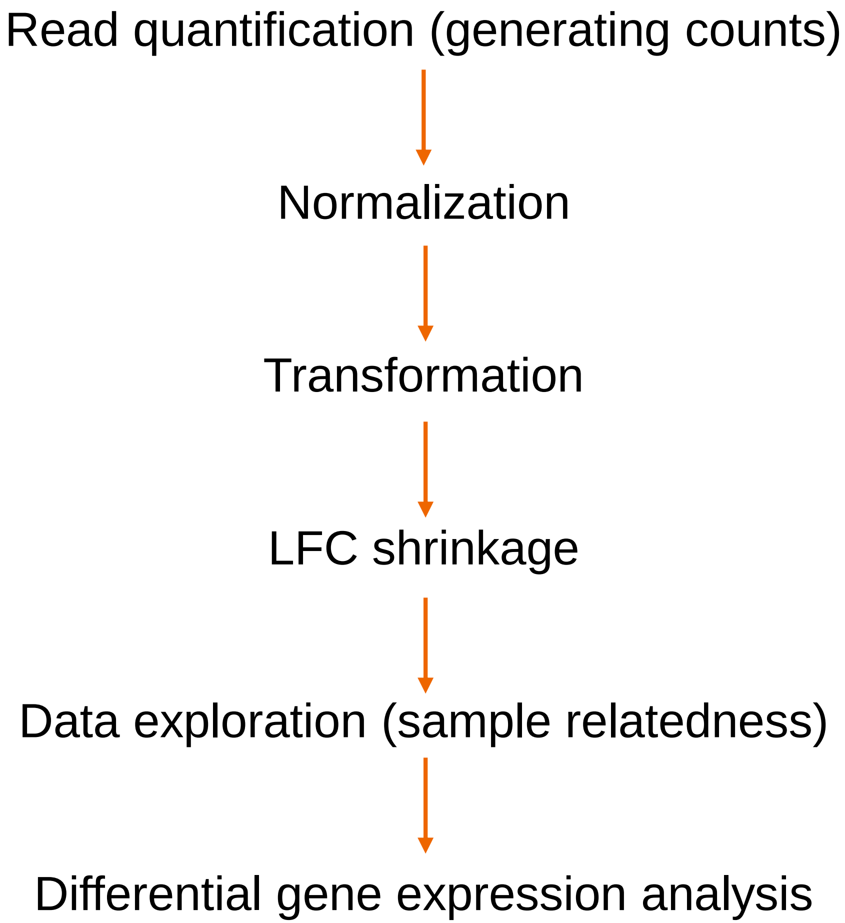
(ii) Estimation of individual gene/transcript expression levels

(iii) Normalization

(iv) Identification of differentially expressed genes



Pathways to analysis



Read quantification (generating counts)

Normalization

Transformation

LFC shrinkage

Data exploration (sample relatedness)

Differential gene expression analysis

Normalizing and quantifying gene counts

- While the number of sequenced reads is known, the total RNA library and its complexity is unknown and variation between samples may be due to contamination as well as biological reasons

- The purpose of normalization is to eliminate systematic effects that are not associated with the biological differences of interest

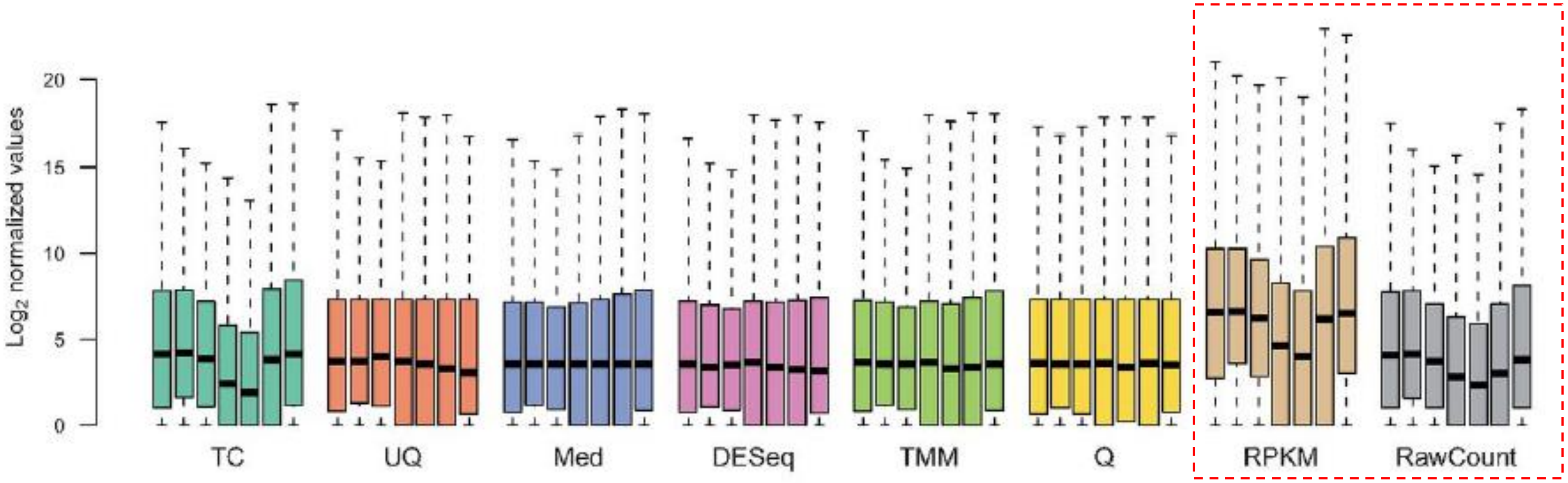
- Given a uniform sampling of a diverse transcript pool, the number of sequenced reads mapped to a gene depends on:
  - its own expression level
  - its length
  - the sequencing depth
  - the expression of all other genes within the sample

- To compare gene expression values between two conditions, we calculate the fraction of reads assigned to each gene, relative to the total number of reads and with respect to the entire RNA repertoire which may vary drastically from sample to sample

Normalization methods – final word

TL;DR

RLE and TMM are widely considered the best methods and if properly implemented give near identical results…



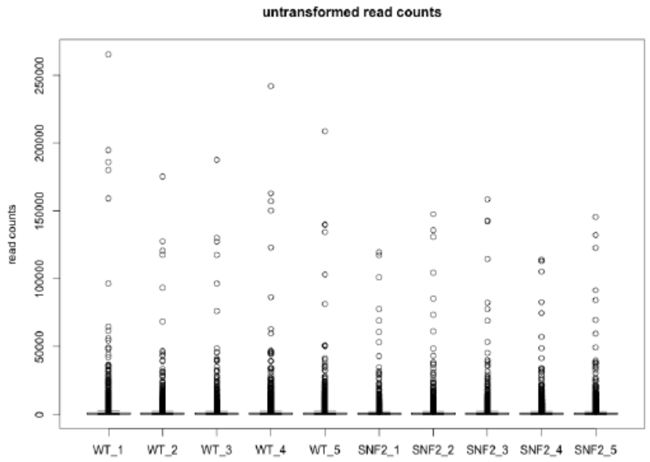
total count

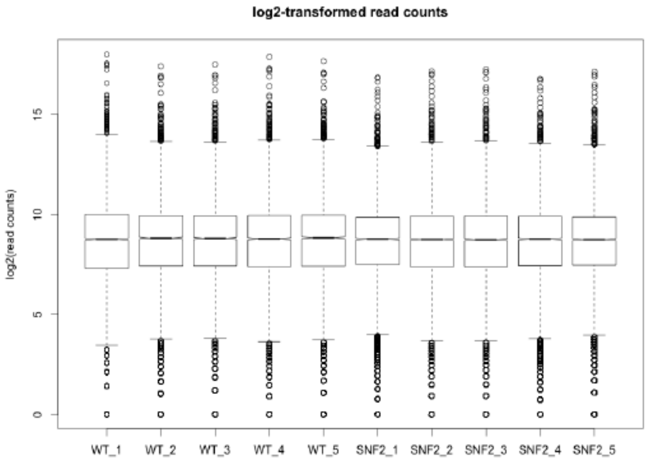
median

quantile

Transformation

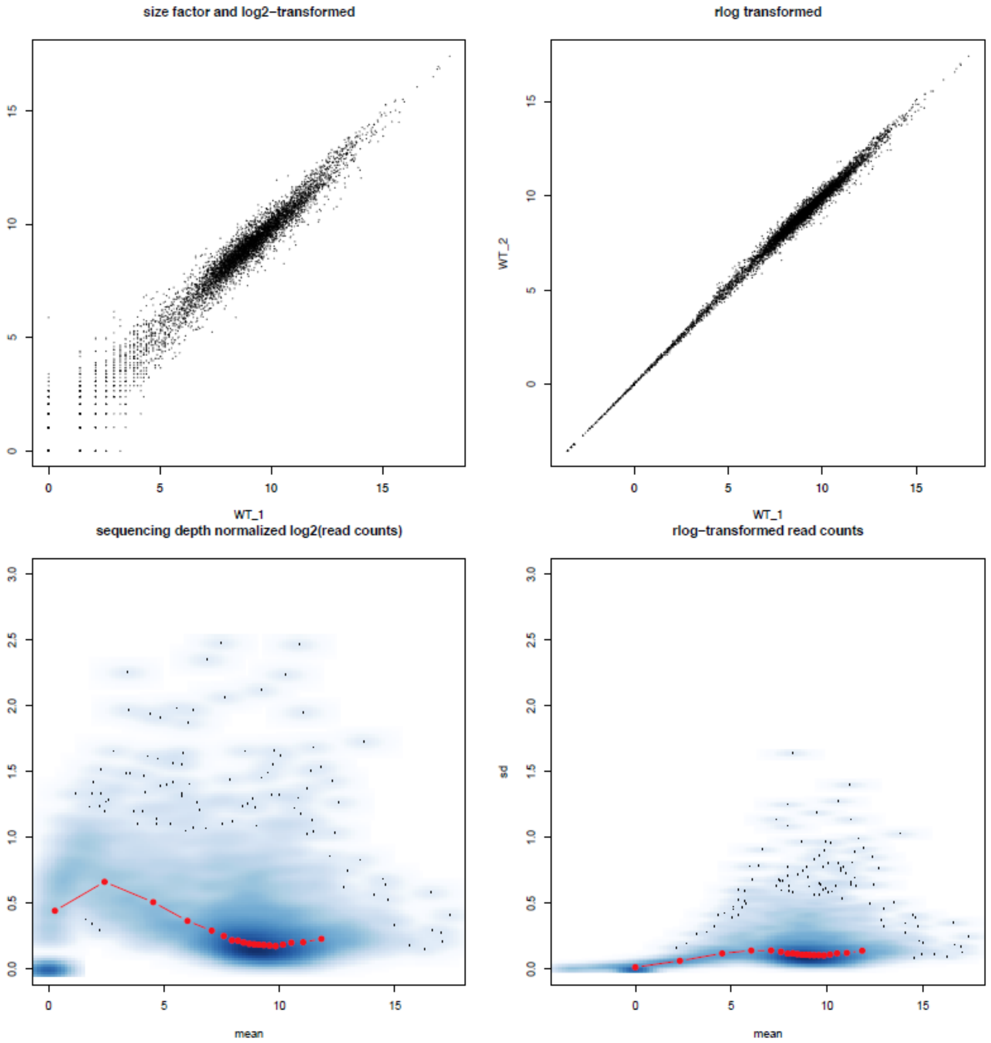
- Most models used for differential gene expression testing operate on the raw count values
- ***Conversely***, many downstream analyses (including clustering) work better if the read counts are transformed to the log scale
- Log2 transformation is generally performed in addition to sequencing depth normalization





Heteroskedacity

- Many statistical tests and analyses assume that data is homoskedastic, i.e. that all variables have similar variance
- However, data with large differences among the sizes of the individual observations often shows heteroskedastic behavior
- To reduce the amount of heteroskedasticity, DESeq2 and edgeR offer several means to shrink the variance of low read counts
- They do this by using the dispersion-mean trend that can be observed for the entire data set as a reference
- Consequently, genes with low and highly variable read counts will be assigned more homogeneous read count estimates so that their variance resembles the variance observed for the majority of the genes



Violation of heteroskedacity

Heteroskedacity restored by shrinking variance

Data exploration *(aka sanity checks…)*

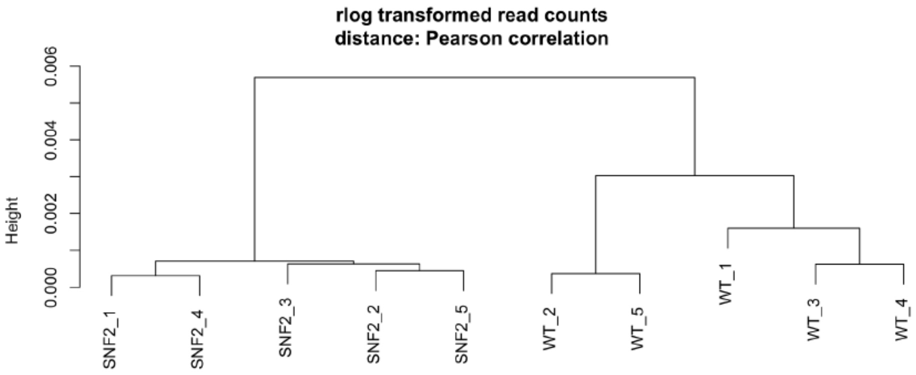
A crucial step before diving into the identification of differentially expressed genes is to check whether expectations about basic global patterns are met:

- technical and biological replicates should show similar expression patterns
- expression patterns of differing experimental conditions should be more dissimilar

Three general approaches

1. **Pairwise correlation** – use Pearson correlation coefficient, r, to evaluate similarity between bio/tech replicates
   - correlation score > 0.9 (ENCODE recommendations
   - easily achieved in R using cor() function

2. **Hierarchical clustering**
   - Determines whether different samples can be clustered in an unsupervised fashion
   - Hierarchical clustering requires two decisions:
     - How should the (dis)similarity between pairs be calculated?
     - How should the (dis)similarity be used for the clustering?
   - A common way to assess the (dis)similarity is the Pearson correlation coefficient
   - Alternatively, the Euclidean distance is often used as a measure of distance between two vectors of read counts
   - Euclidean distance is strongly influenced by differences of the scale: if two samples show large d distance more than the distance based on the Pearson correlation coefficient.



Data exploration *(aka sanity checks…)*

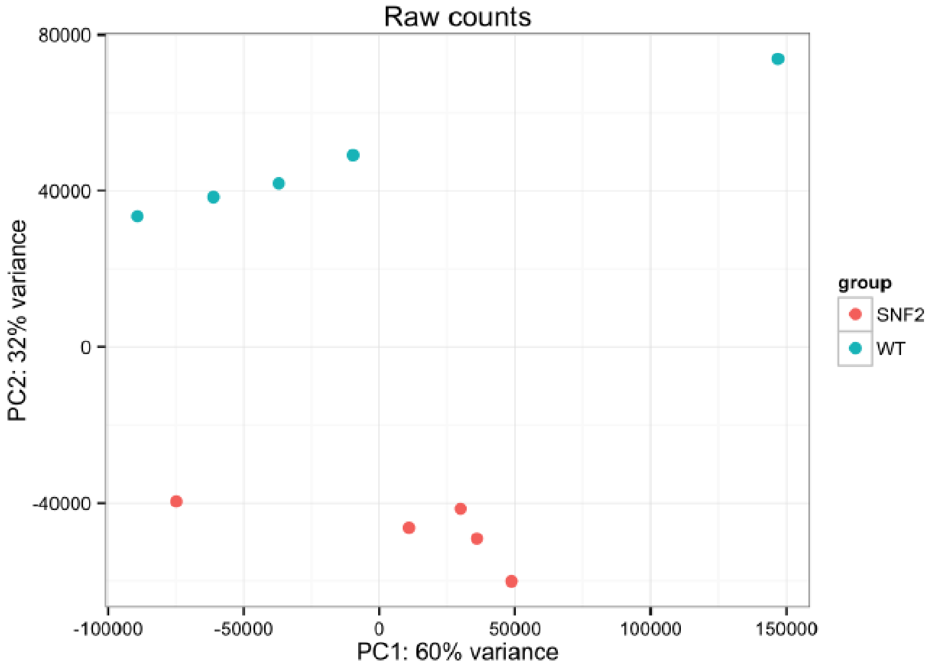
3. Principal component analyses (PCA)

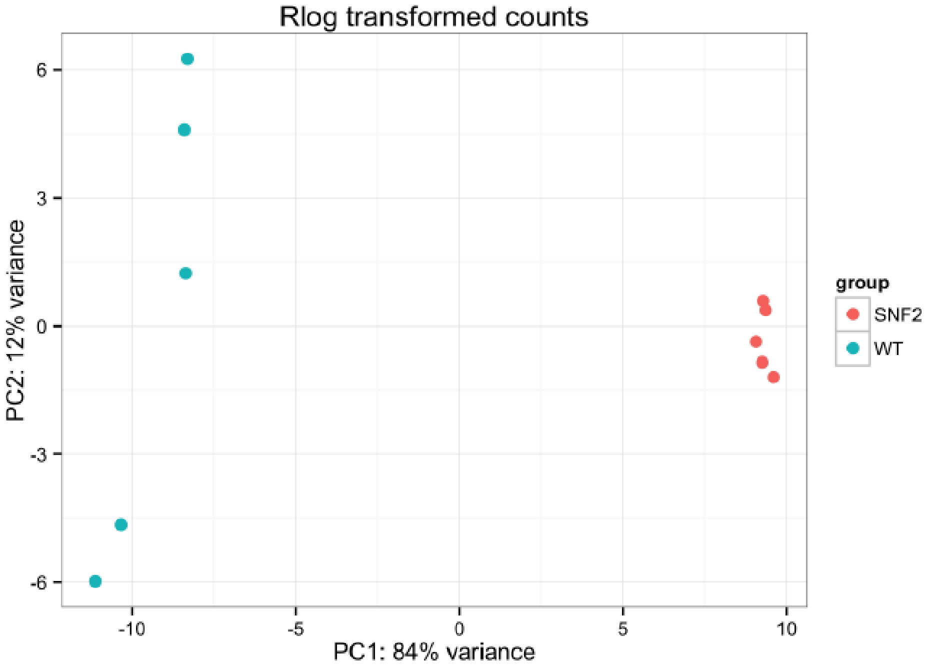
- A complementary approach to determine whether samples display greater variability between experimental conditions than between replicates of the same treatment is principal components analysis

- It is a typical example of dimensionality reduction approaches that have become very popular in the field of machine learning

- The goal is to find groups of features (e.g., genes) that have something in common (e.g., certain patterns of expression across different samples), so that the information from thousands of features is captured and represented by a reduced number of groups

- Most commonly, the two principal components explaining the majority of the variability are displayed





PCA and clustering should be done on normalized and transformed read counts, so that the high variability of low read counts does not occlude potentially informative trends.

Differential gene expression (DGE) analysis – an overview

DGE tools perform two basic tasks:

1. Estimate the magnitude of differential expression between two or more conditions based on read counts from replicated samples, i.e., calculate the fold change of read counts, taking into account the differences in sequencing depth and variability

2. Estimate the significance of the difference and correct for multiple testing

The best performing tools tend to be edgeR, DESeq/DESeq2, and limma-voom

- DESeq and limma-voom tend to be more conservative than edgeR (better control of false positives)

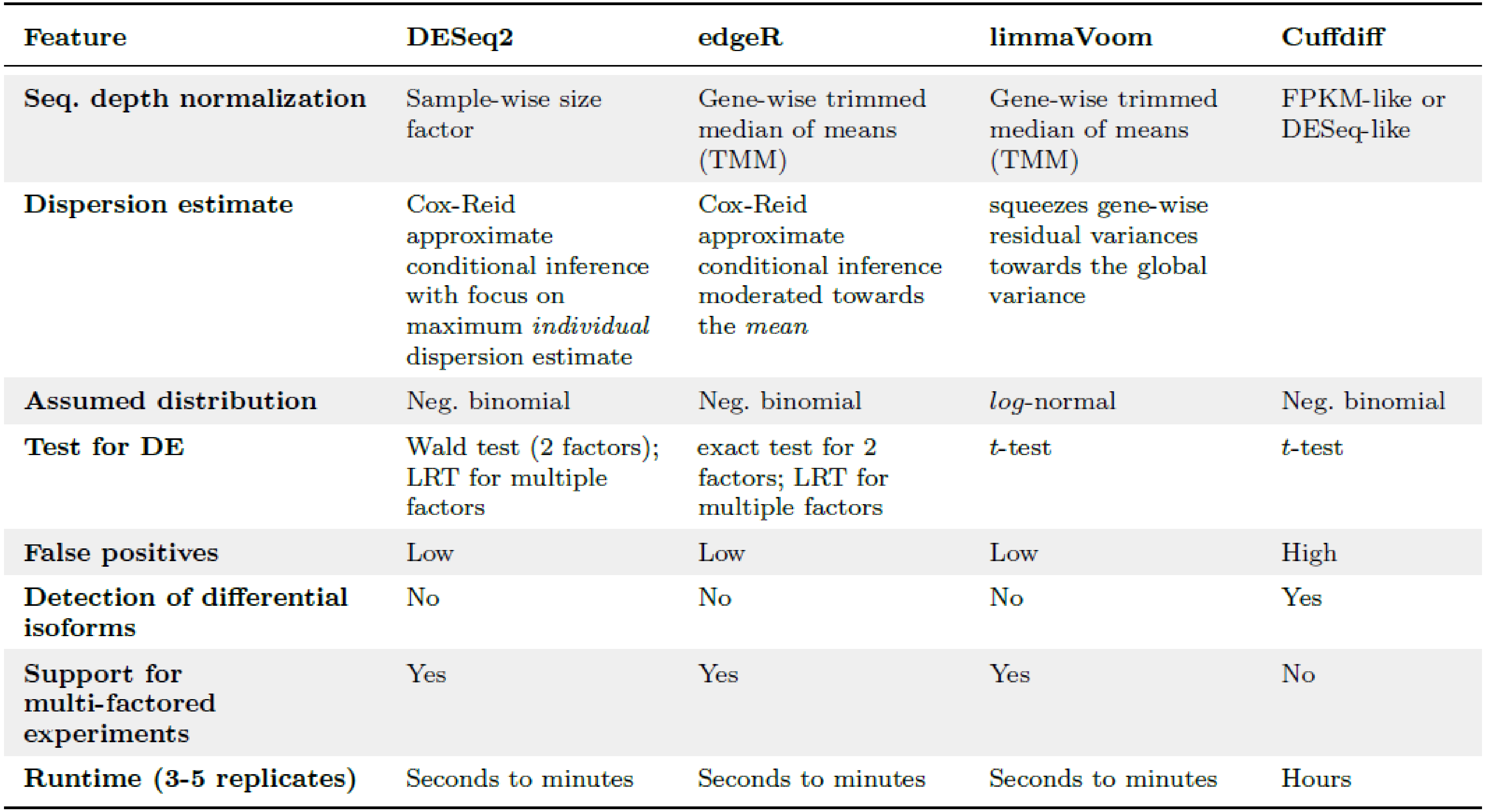
- edgeR is recommended for experiments with fewer than 12 biological replicates

- All statistical methods developed for read counts rely on approximations of various kinds
  - e.g. assumptions must be made about the data properties.

- edgeR and DESeq, for example, assume that the majority of the transcriptome is unchanged between the two conditions.

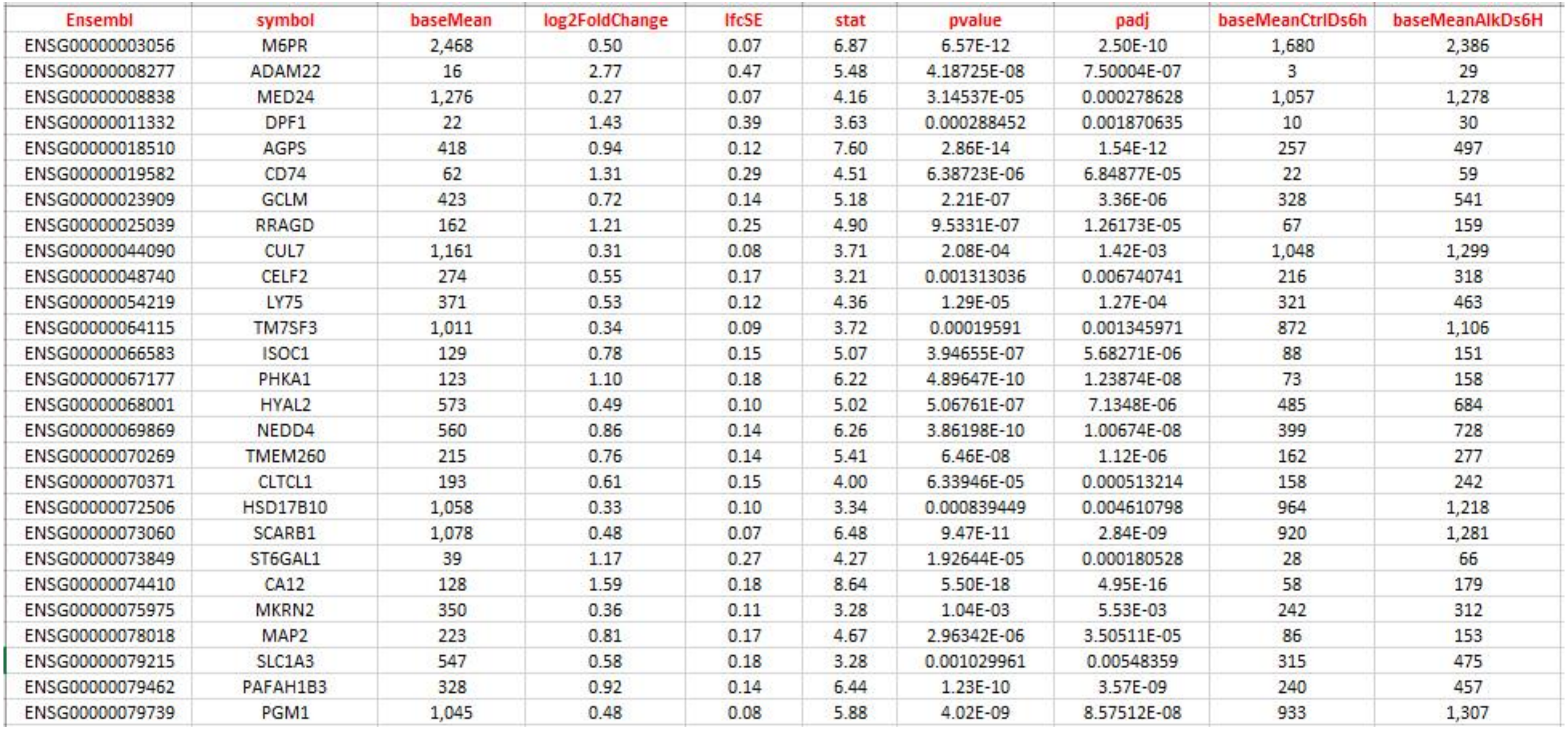
- If this assumption is not met by the data, both log2 fold change and the significance indicators are most likely incorrect!

Differential gene expression (DGE) analysis – an overview



Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Dispersion estimate	Cox-Reid approximate conditional inference with focus on maximum *individual* dispersion estimate	Cox-Reid approximate conditional inference moderated towards the *mean*	squeezes gene-wise residual variances towards the global variance	
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Wald test (2 factors); LRT for multiple factors	exact test for 2 factors; LRT for multiple factors	*t*-test	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

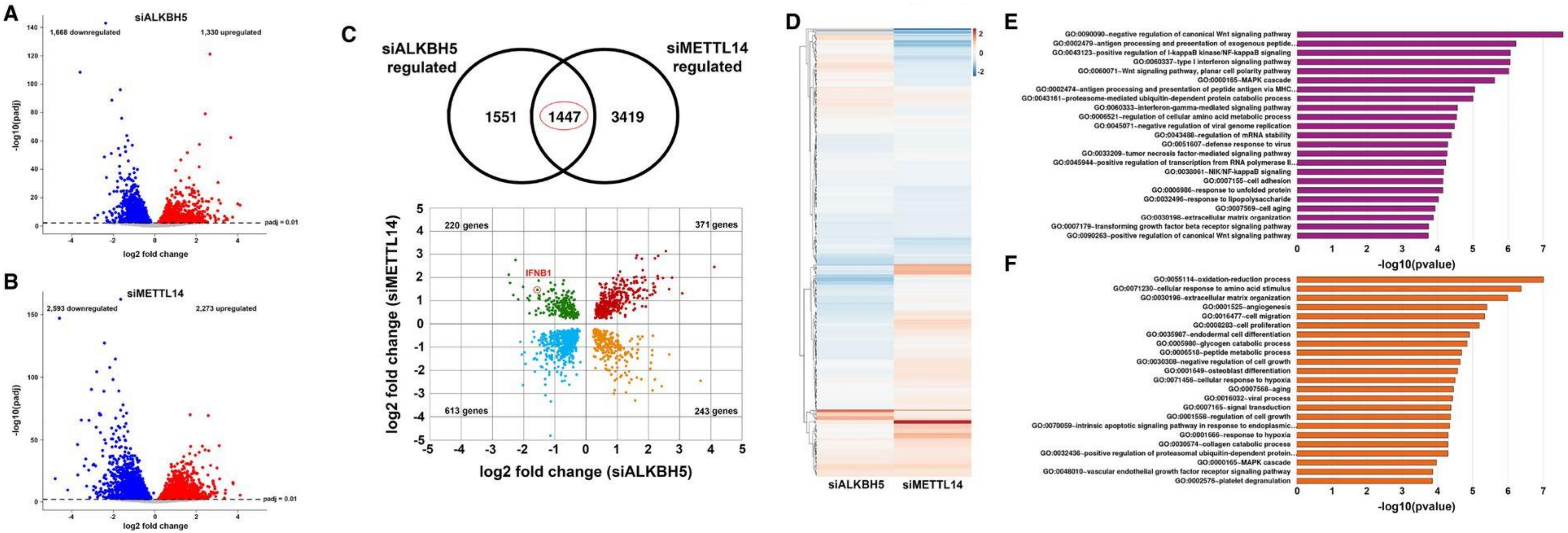
DGE analysis outputs



Ensembl	symbol	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	baseMeanCtrlDs6h	baseMeanAlkDs6H
ENSG00000003056	M6PR	2,468	0.50	0.07	6.87	6.57E-12	2.50E-10	1,680	2,386
ENSG00000008277	ADAM22	16	2.77	0.47	5.48	4.18725E-08	7.50004E-07	3	29
ENSG00000008838	MED24	1,276	0.27	0.07	4.16	3.14537E-05	0.000278628	1,057	1,278
ENSG00000011332	DPF1	22	1.43	0.39	3.63	0.000288452	0.001870635	10	30
ENSG00000018510	AGPS	418	0.94	0.12	7.60	2.86E-14	1.54E-12	257	497
ENSG00000019582	CD74	62	1.31	0.29	4.51	6.38723E-06	6.84877E-05	22	59
ENSG00000023909	GCLM	423	0.72	0.14	5.18	2.21E-07	3.36E-06	328	541
ENSG00000025039	RRAGD	162	1.21	0.25	4.90	9.5331E-07	1.26173E-05	67	159
ENSG00000044090	CUL7	1,161	0.31	0.08	3.71	2.08E-04	1.42E-03	1,048	1,299
ENSG00000048740	CELF2	274	0.55	0.17	3.21	0.001313036	0.006740741	216	318
ENSG00000054219	LY75	371	0.53	0.12	4.36	1.29E-05	1.27E-04	321	463
ENSG00000064115	TM7SF3	1,011	0.34	0.09	3.72	0.00019591	0.001345971	872	1,106
ENSG00000066583	ISOC1	129	0.78	0.15	5.07	3.94655E-07	5.68271E-06	88	151
ENSG00000067177	PHKA1	123	1.10	0.18	6.22	4.89647E-10	1.23874E-08	73	158
ENSG00000068001	HYAL2	573	0.49	0.10	5.02	5.06761E-07	7.1348E-06	485	684
ENSG00000069869	NEDD4	560	0.86	0.14	6.26	3.86198E-10	1.00674E-08	399	728
ENSG00000070269	TMEM260	215	0.76	0.14	5.41	6.46E-08	1.12E-06	162	277
ENSG00000070371	CLTCL1	193	0.61	0.15	4.00	6.33946E-05	0.000513214	158	242
ENSG00000072506	HSD17B10	1,058	0.33	0.10	3.34	0.000839449	0.004610798	964	1,218
ENSG00000073060	SCARB1	1,078	0.48	0.07	6.48	9.47E-11	2.84E-09	920	1,281
ENSG00000073849	ST6GAL1	39	1.17	0.27	4.27	1.92644E-05	0.000180528	28	66
ENSG00000074410	CA12	128	1.59	0.18	8.64	5.50E-18	4.95E-16	58	179
ENSG00000075975	MKRN2	350	0.36	0.11	3.28	1.04E-03	5.53E-03	242	312
ENSG00000078018	MAP2	223	0.81	0.17	4.67	2.96342E-06	3.50511E-05	86	153
ENSG00000079215	SLC1A3	547	0.58	0.18	3.28	0.001029961	0.00548359	315	475
ENSG00000079462	PAFAH1B3	328	0.92	0.14	6.44	1.23E-10	3.57E-09	240	457
ENSG00000079739	PGM1	1,045	0.48	0.08	5.88	4.02E-09	8.57512E-08	933	1,307

DGE output (text file)

DGE analysis outputs


A
siALKBH5
1,668 downregulated
1,330 upregulated
-log10(padj)
padj = 0.01
log2 fold change
B
siMETTL14
2,593 downregulated
2,273 upregulated
-log10(padj)
padj = 0.01
log2 fold change
C
siALKBH5 regulated
siMETTL14 regulated
220 genes
371 genes
613 genes
243 genes
IFNB1
log2 fold change (siMETTL14)
log2 fold change (siALKBH5)
D
siALKBH5
siMETTL14
E
GO:0090090~negative regulation of canonical Wnt signaling pathway
GO:0002479~antigen processing and presentation of exogenous peptide...
GO:0043123~positive regulation of I-kappaB kinase/NF-kappaB signaling
GO:0060337~type I interferon signaling pathway
GO:0060071~Wnt signaling pathway, planar cell polarity pathway
GO:0000165~MAPK cascade
GO:0002474~antigen processing and presentation of peptide antigen via MHC...
GO:0043161~proteasome-mediated ubiquitin-dependent protein catabolic process
GO:0060333~interferon-gamma-mediated signaling pathway
GO:0006521~regulation of cellular amino acid metabolic process
GO:0045071~negative regulation of viral genome replication
GO:0043488~regulation of mRNA stability
GO:0051607~defense response to virus
GO:0033209~tumor necrosis factor-mediated signaling pathway
GO:0045944~positive regulation of transcription from RNA polymerase II...
GO:0038061~NIK/NF-kappaB signaling
GO:0007155~cell adhesion
GO:0006986~response to unfolded protein
GO:0032496~response to lipopolysaccharide
GO:0007569~cell aging
GO:0030198~extracellular matrix organization
GO:0007179~transforming growth factor beta receptor signaling pathway
GO:0090263~positive regulation of canonical Wnt signaling pathway
-log10(pvalue)
F
GO:0055114~oxidation-reduction process
GO:0071230~cellular response to amino acid stimulus
GO:0030198~extracellular matrix organization
GO:0001525~angiogenesis
GO:0016477~cell migration
GO:0008283~cell proliferation
GO:0035987~endodermal cell differentiation
GO:0005980~glycogen catabolic process
GO:0006518~peptide metabolic process
GO:0030308~negative regulation of cell growth
GO:0001649~osteoblast differentiation
GO:0071456~cellular response to hypoxia
GO:0007568~aging
GO:0016032~viral process
GO:0007165~signal transduction
GO:0001558~regulation of cell growth
GO:0070059~intrinsic apoptotic signaling pathway in response to endoplasmic...
GO:0001666~response to hypoxia
GO:0030574~collagen catabolic process
GO:0032436~positive regulation of proteasomal ubiquitin-dependent protein...
GO:0000165~MAPK cascade
GO:0048010~vascular endothelial growth factor receptor signaling pathway
GO:0002576~platelet degranulation
-log10(pvalue)