

Slides are available online: <https://ecotoxxplorer.github.io/>

RNAseq Data Analysis

Jianguo (Jeff) Xia, Assistant Professor

jeff.xia@mcgill.ca | www.xialab.ca

McGill University, QC

Omics Data Analysis Overview

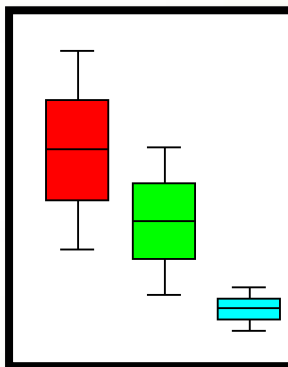
Data processing & quality check

Platform-specific

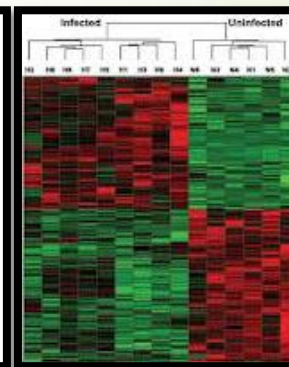
- Microarray
- NGS
- Mass spec

Statistical analysis & visualization

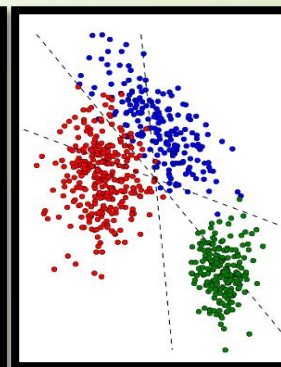
Comparison



Clustering



Classification



Functional interpretation

Omics-specific

- Pathway analysis
- Enrichment analysis
- Ontology analysis

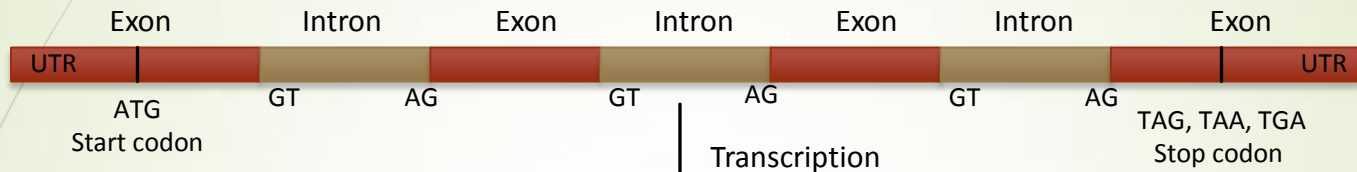


RNaseq Big Picture

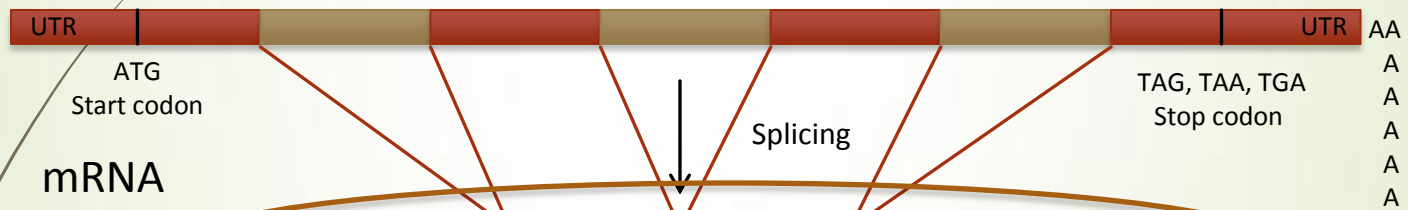
1. Scientific question (biological)
2. Study design (biological/statistical)
3. Conducting experiment (biological)
4. Sequencing
5. Read mapping and transcript identification
6. Pre-processing & normalization
7. Finding differentially expressed genes
8. Interpreting the results

RNA-seq

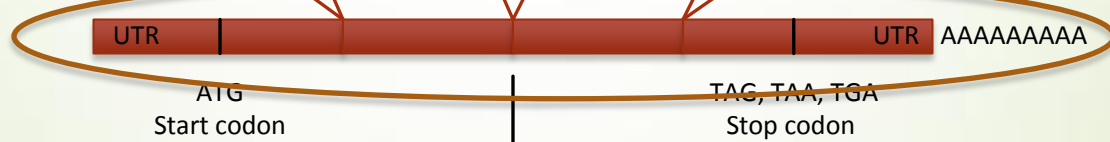
DNA



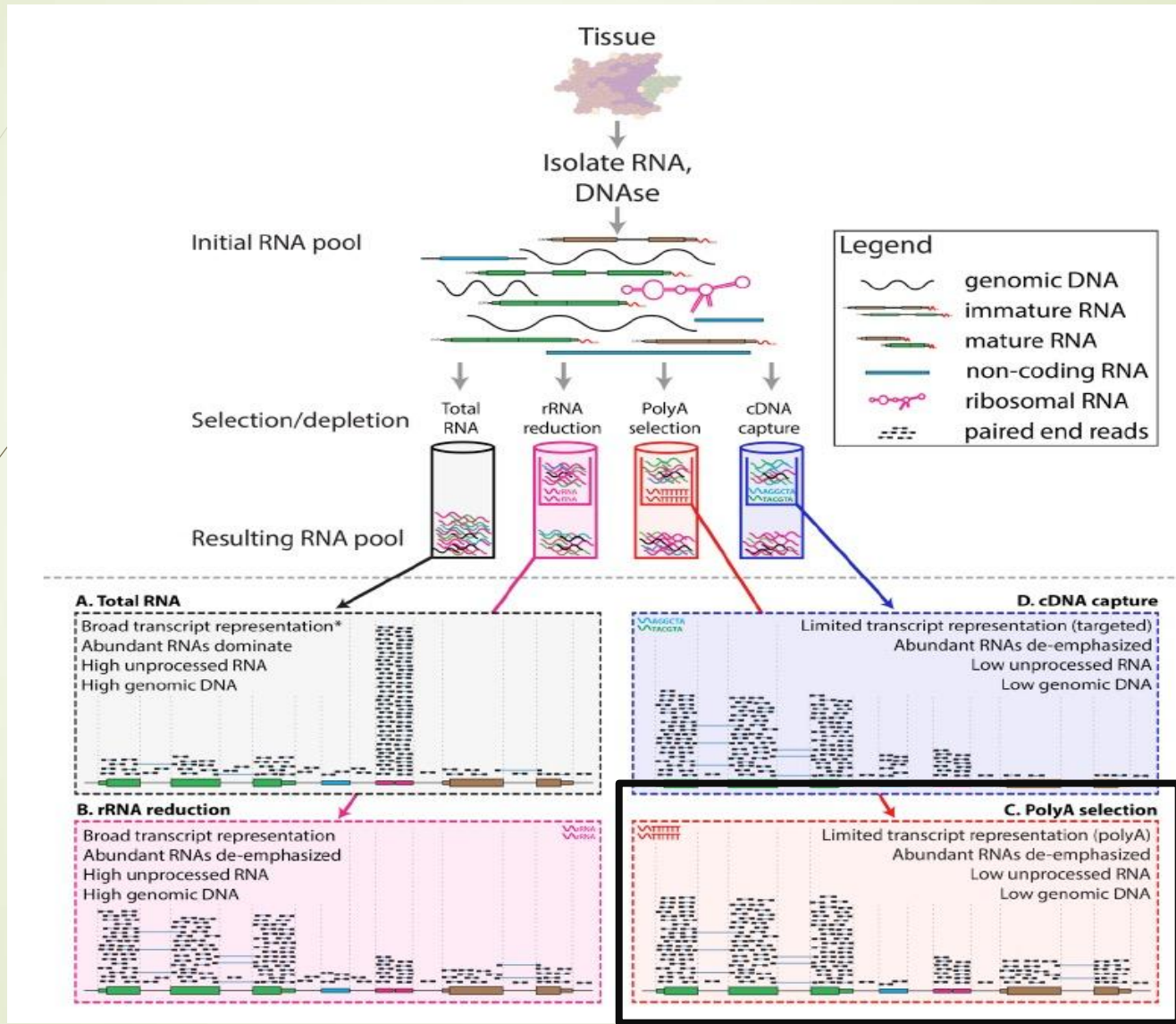
Pre-mRNA



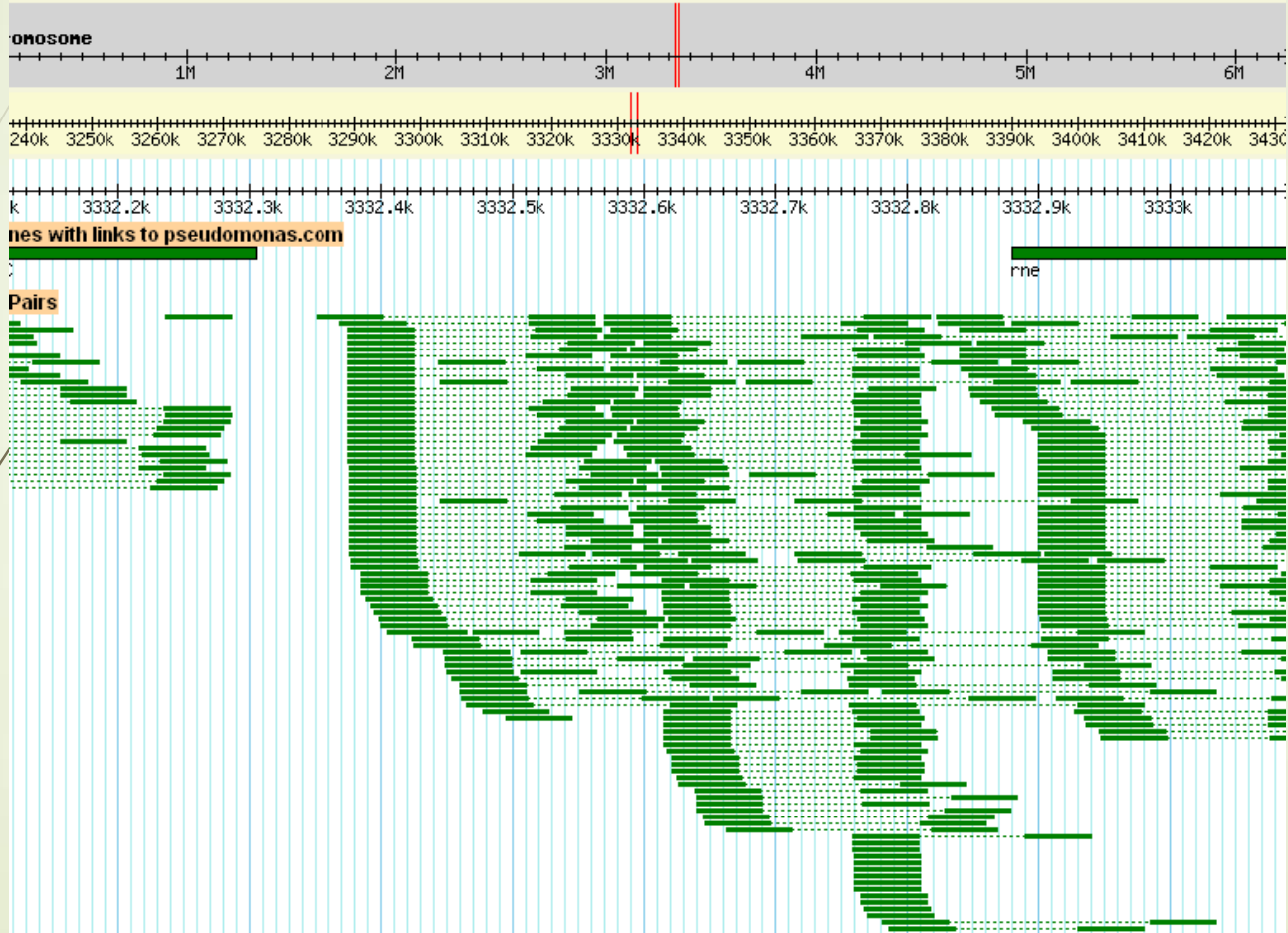
mRNA



sequencing



Mapped Reads




Read mapping & transcript identification

FASTQ Files

SAM/BAM

Need reference
genome and gene
annotation files

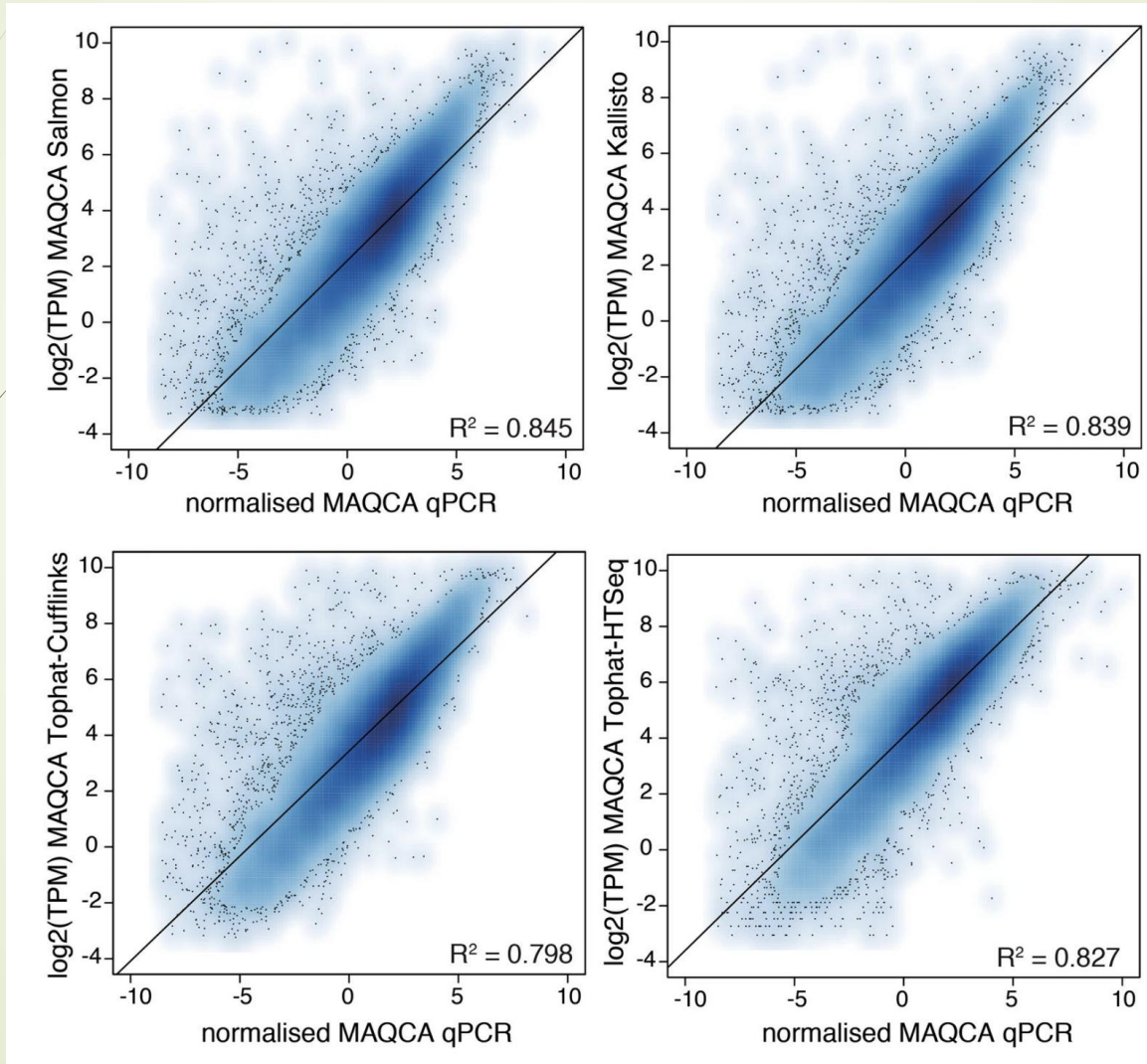
Raw
Counts



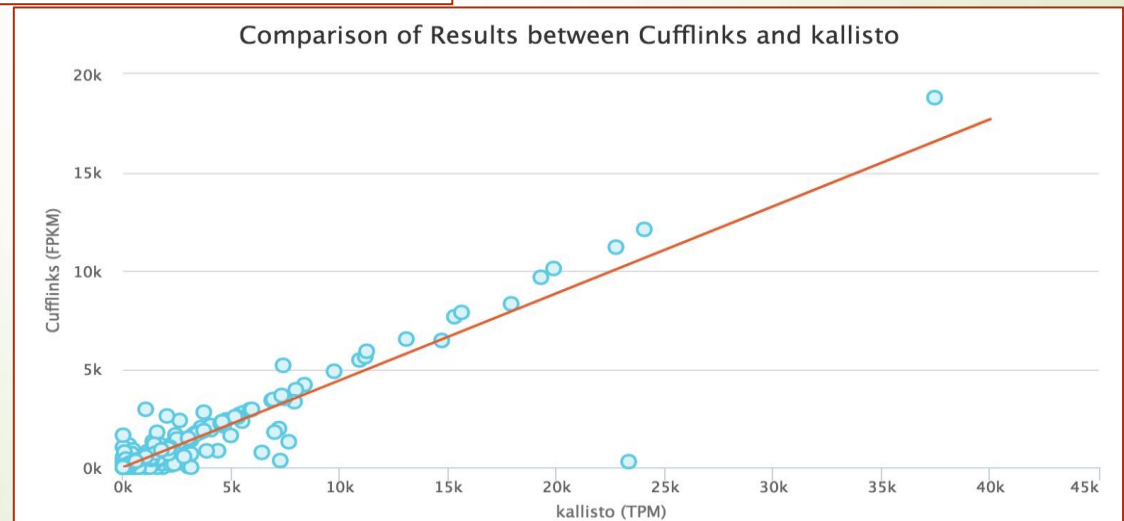
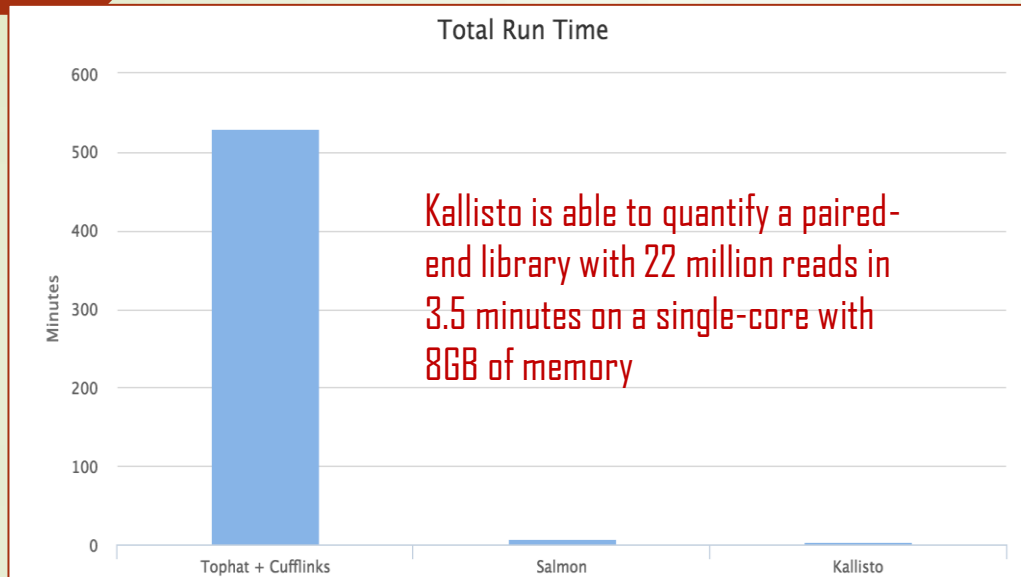
Major Pipelines - from reads to transcript counts

- 1st generation
 - Tophat-Cufflinks
- 2nd generation
 - Tophat/STAR-HTSeq
- 3rd generation
 - Kallisto / Salmon

Benchmarking against qPCR

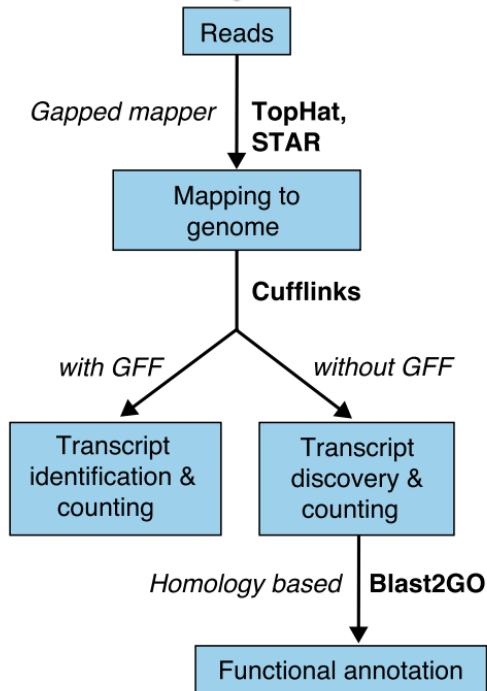


Pseudo-alignment

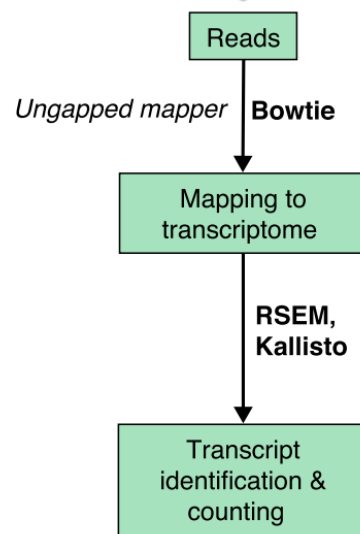


From reads to counts

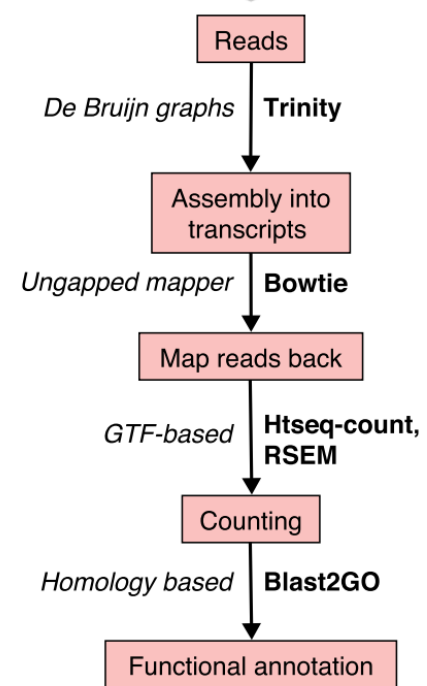
(a) Genome mapping



(b) Transcriptome mapping



(c) Reference-free assembly



RNAseq differential expression analysis

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

[C Trapnell](#), [A Roberts](#), [L Goff](#), [G Pertea](#), [D Kim](#)... - Nature protocols, 2012 - nature.com

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify ...

☆ [🔗](#) Cited by 4975 [Related articles](#) [All 40 versions](#)

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

[MD Robinson](#), [DJ McCarthy](#), [GK Smyth](#) - Bioinformatics, 2010 - academic.oup.com

... edgeR. is designed for the analysis of replicated **count-based expression** data and is an implementation of methodology developed by Robinson and Smyth (2007, 2008) ... Finally, **differential expression** is assessed for each **gene** using an exact test analogous to Fisher's exact ...

☆ [🔗](#) Cited by 7875 [Related articles](#) [All 20 versions](#)

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

[S Anders](#), [DJ McCarthy](#), [Y Chen](#), [M Okoniewski](#)... - Nature protocols, 2013 - nature.com

... In addition, **count-based** methods that operate at the exon level, which share the NB framework ... These methods give a direct readout of **differential** exons, **genes** whose exons are used unequally ... in the hope that these factors (or surrogates of them) can be **differentiated** from the ...

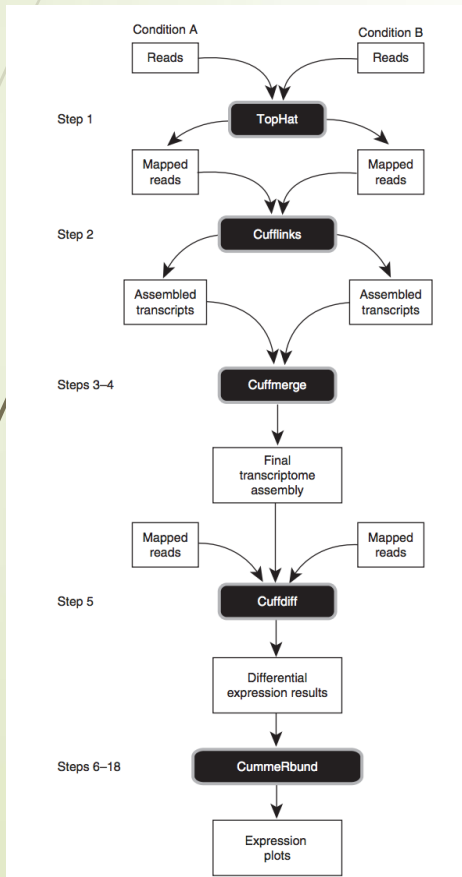
☆ [🔗](#) Cited by 554 [Related articles](#) [All 21 versions](#)

Two widely used step-by-step tutorials (I)

PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

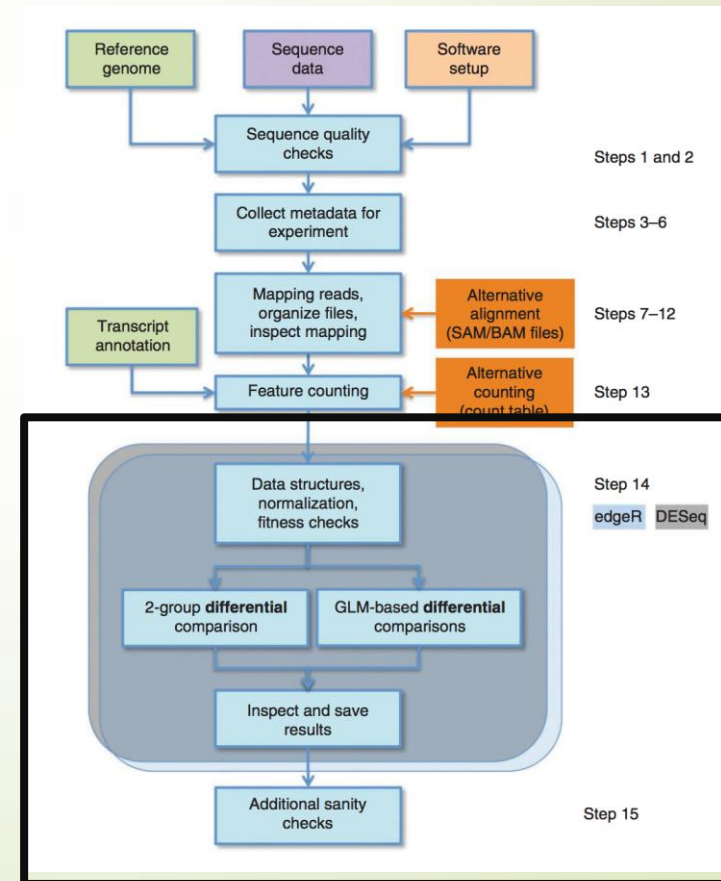
Cole Trapnell^{1,2}, Adam Roberts³, Loyal Goff^{1,2,4}, Geo Pertea^{5,6}, Daehwan Kim^{5,7}, David R Kelley^{1,2}, Harold Pimentel³, Steven L Salzberg^{5,6}, John L Rinn^{1,2} & Lior Pachter^{3,8,9}



PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders¹, Davis J McCarthy^{2,3}, Yunshun Chen^{4,5}, Michal Okoniewski⁶, Gordon K Smyth^{4,7}, Wolfgang Huber¹ & Mark D Robinson^{8,9}





Task overview

1. QC (sanity check)
 - Sequence depth
 - PCA/MDS plot
 - MSD plot
2. Data Filtering
 - Filtering genes
 - Filtering samples
3. Differential Expression Analysis
4. Visual exploration & functional analysis

Data Input

- A raw count table (not RPKM/FPKM)
- Sample names in columns
- Transcript names in rows
- Group labels states with '#CLASS:' (can be more than 1)

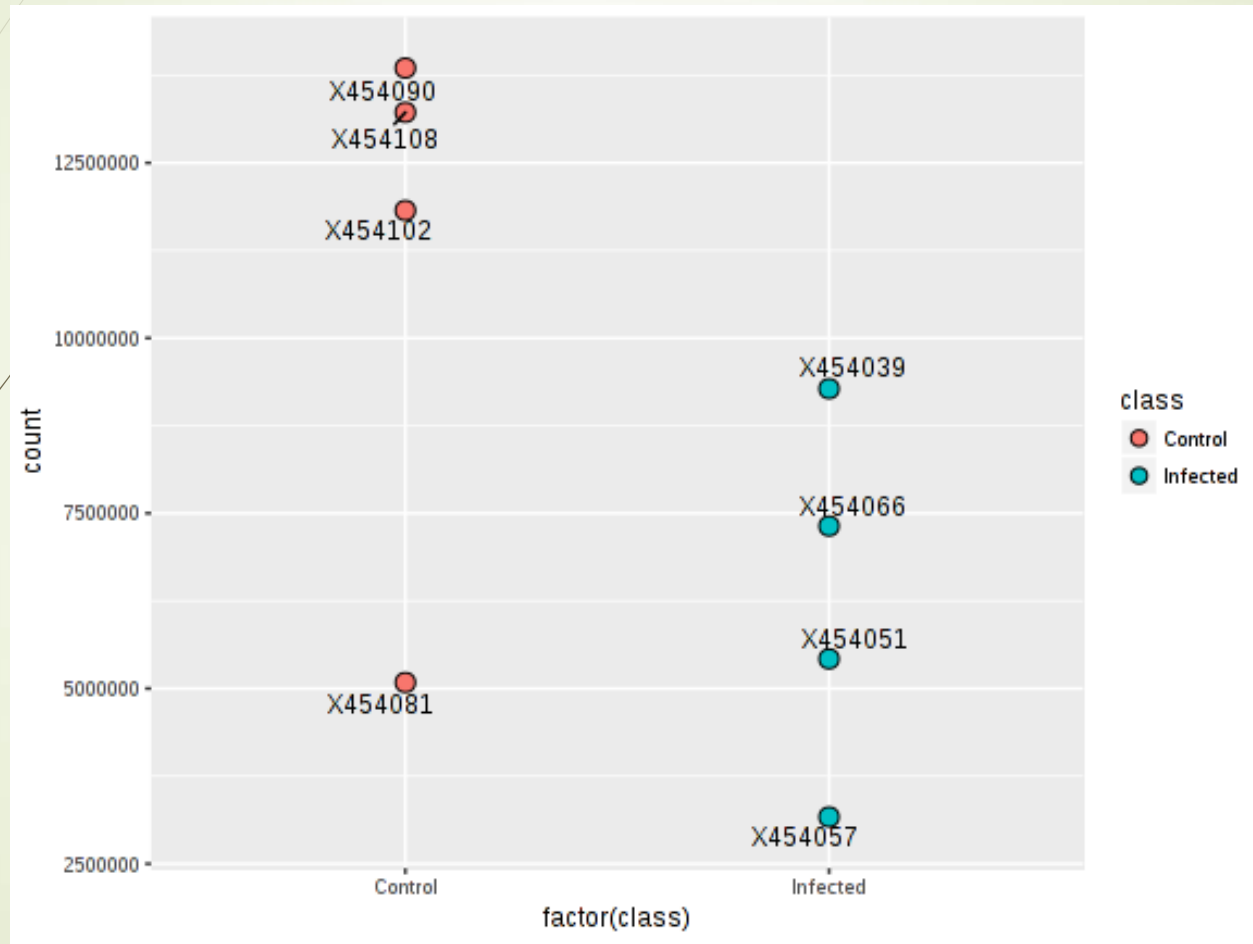
#NAME	ME1-MEP1	ME1-HAP1	ME1-MAP1	ME1-MDP1	ME1-SDP1	ME1-MBP1	ME1-SEP1	ME1-HBP1	ME1-HDP1	ME1-HEP1	ME1-SBP1	ME1-SCP1	ME1-MCP1	ME1-HCP1	ME1-SAP1
#CLASS:DOSE	Medium	High	Medium	Medium	Control	Medium	Control	High	High	High	Control	Control	Medium	High	Control
ENSDARP00000109818.1	1020	557	483	950	645	625	779	574	693	589	607	699	705	800	812
ENSDARP00000140768.1	229	70	134	279	188	149	164	162	67	74	171	226	210	165	71
ENSDARP00000113099.1	660	384	364	814	545	395	610	332	428	366	310	476	462	434	526
ENSDARP00000112079.2	1700	875	776	1452	914	1076	1109	1019	995	846	826	1112	1237	1363	1028
ENSDARP00000062256.3	1788	918	845	1571	1005	1166	1174	1069	1034	887	898	1204	1316	1431	1077
ENSDARP00000130517.1	2945	1595	1352	2440	1660	1881	1887	1728	1835	1641	1443	1850	2041	2384	2123
ENSDARP00000016784.5	5	4	4	9	8	4	3	3	7	8	5	5	4	2	6
ENSDARP00000117321.1	827	402	488	861	584	541	622	541	437	400	589	633	637	657	451
ENSDARP00000060918.4	1221	546	679	1300	922	742	913	777	602	563	841	1016	940	905	567
ENSDARP00000139249.1	746	362	452	769	520	488	565	502	391	374	532	572	575	578	397
ENSDARP00000139249.1	748	362	451	769	522	487	565	500	391	373	534	572	577	581	398
ENSDARP00000031533.6	239	90	153	275	229	144	250	199	135	122	145	238	182	161	105
ENSDARP00000076098.3	227	92	149	272	225	132	248	192	134	117	135	232	176	159	102
ENSDARP00000122329.2	1067	401	533	1099	841	734	846	693	527	502	637	956	781	715	427
ENSDARP00000122329.2	1041	415	526	1082	813	728	855	692	530	500	619	957	780	701	446
ENSDARP00000120705.2	1278	477	639	1523	1117	875	1030	840	588	550	839	1371	1026	836	484
ENSDARP00000092857.3	194	88	121	251	178	137	159	131	119	116	118	160	155	166	120
ENSDARP00000127116.1	172	83	139	260	203	131	148	131	106	115	120	189	158	151	95
ENSDARP00000044330.6	4980	2765	2960	3877	3006	4124	3871	3469	2777	3526	3668	4065	4274	4377	4088
ENSDARP00000056576.5	3584	2130	1999	2308	1803	3001	2763	2400	2157	2859	2515	2548	2992	3322	3293
ENSDARP00000105912.2	4333	2419	2448	2854	2278	3583	3454	2879	2460	3210	3189	3262	3675	3890	3774
ENSDARP00000105912.2	3786	2219	2105	2491	1965	3188	2948	2521	2244	2911	2730	2834	3198	3440	3412
ENSDARP00000134273.1	3567	2116	1968	2298	1800	2999	2742	2374	2121	2830	2527	2574	3008	3274	3265
ENSDARP00000110143.2	4637	2885	2570	5389	3701	2949	3557	2591	3269	2657	2689	3555	3201	3610	3890
ENSDARP00000072374.4	6634	3785	3744	7841	5459	4165	4842	3799	4281	3443	3816	5245	4661	4902	5176
ENSDARP00000091473.4	4647	2927	2535	5239	3540	2976	3633	2729	3303	2709	2665	3435	3158	3688	3896
ENSDARP00000128480.1	25	10	14	30	25	18	31	17	16	18	22	18	16	25	12
ENSDARP00000053920.4	1336	568	632	1196	888	888	1244	742	614	622	951	1263	1060	881	688
ENSDARP00000096667.3	1097	471	447	907	669	708	952	579	480	502	718	934	822	731	586



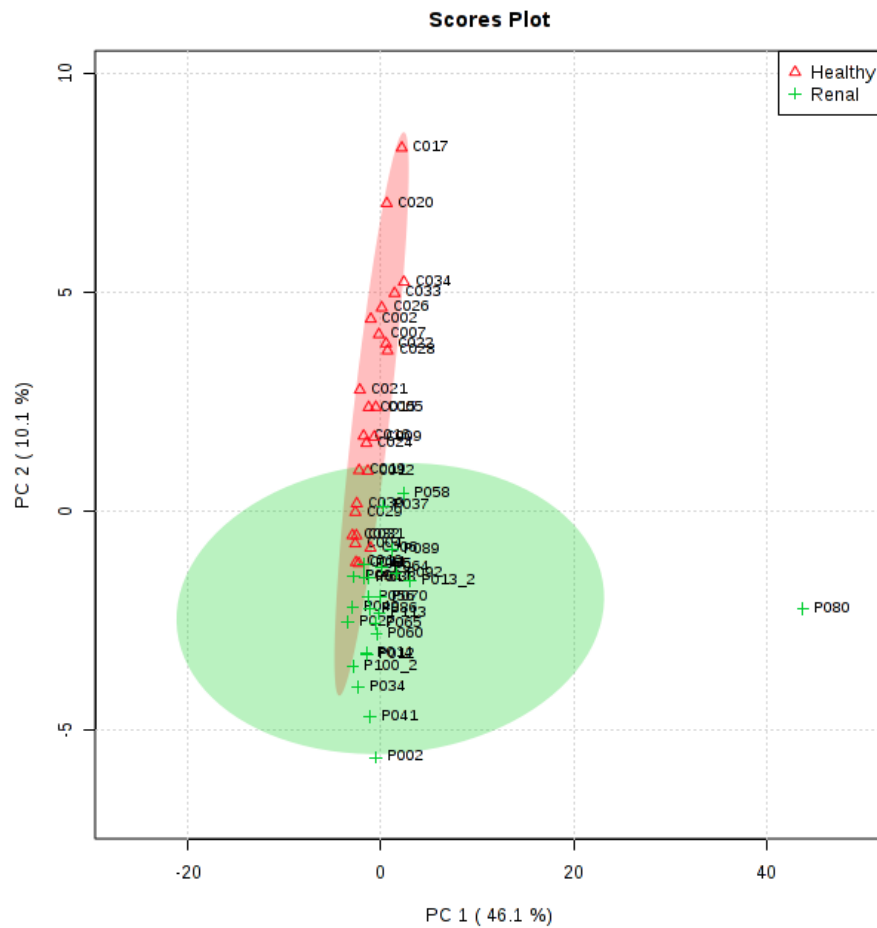
Data filtering

- **Very small values**
 - Close to baseline or detection limit
- **Near-constant values**
 - Throughout the experiment conditions
- **Low repeatability**
 - QC/Spike-ins
- **Unannotated transcripts?**
 - Will not contribute to functional analysis

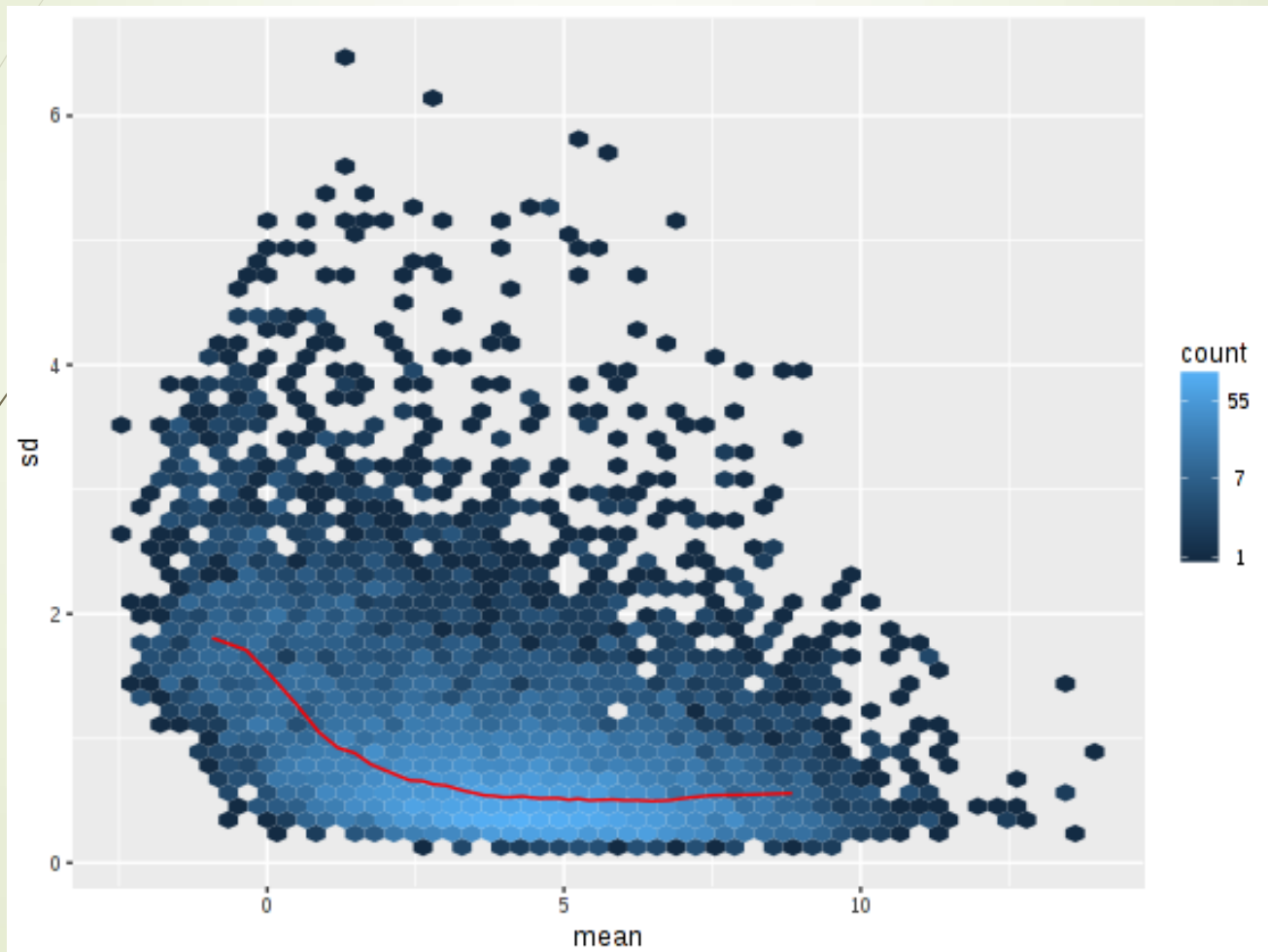
Uneven sequence depth?



Data Overview (MDS/PCA)



MSD Plot



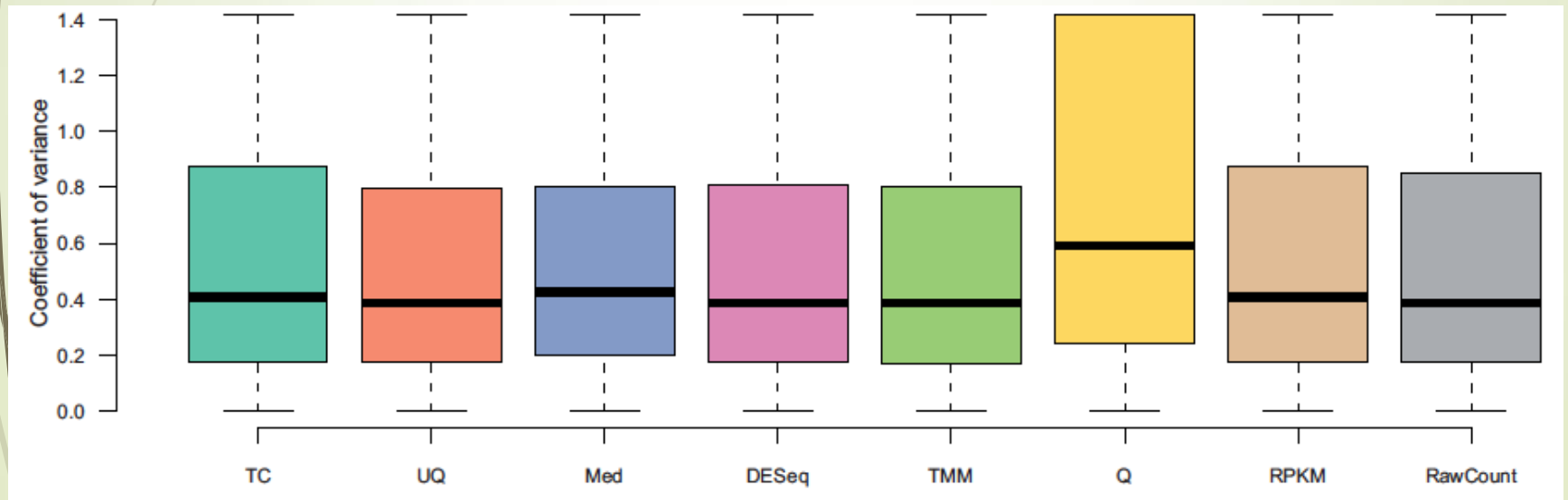


Normalization Methods

- ▶ Many methods have been developed
 - ▶ Total Count (TC)
 - ▶ Upper Quantile (UQ)
 - ▶ Median (Med)
 - ▶ DESeq built-in
 - ▶ Trimmed Mean of M-values (TMM)
 - ▶ Used by edgeR
 - ▶ Quantile (Q)
 - ▶ RPKM

Performance

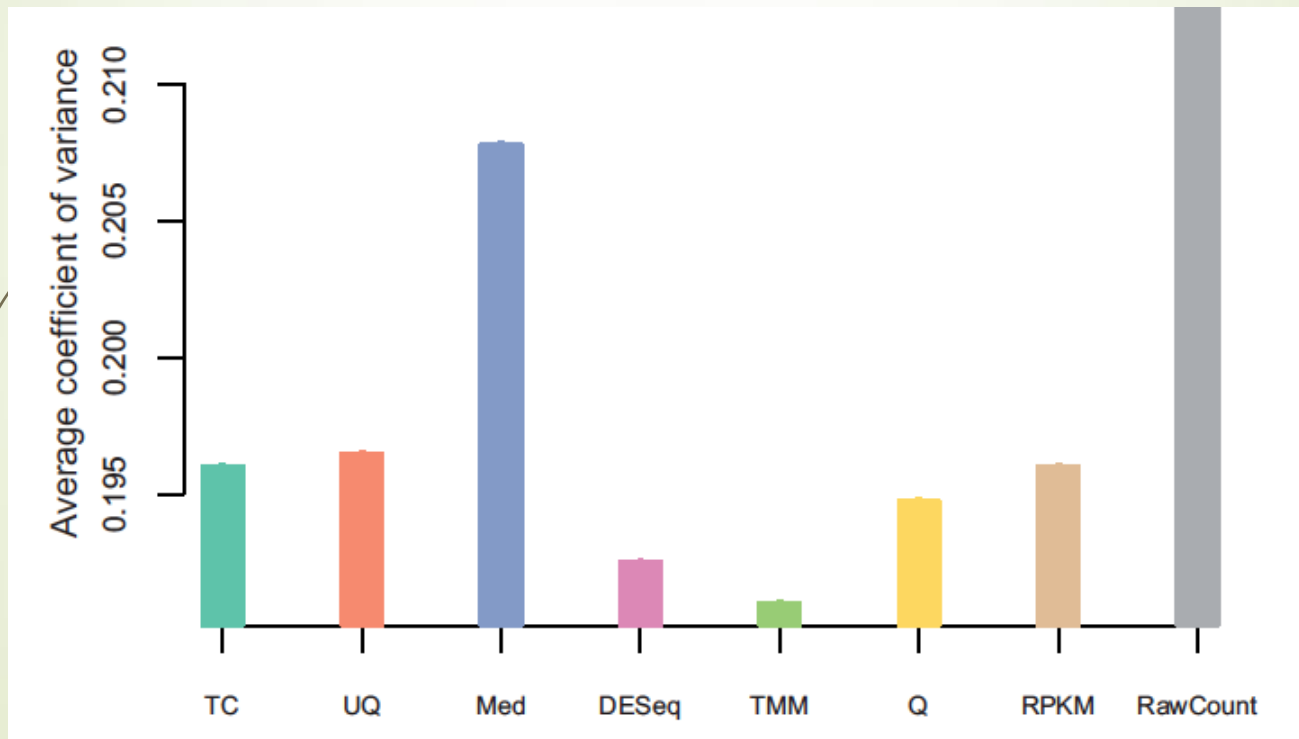
■ Intra-group variance



<http://www.ncbi.nlm.nih.gov/pubmed/22988256>

Performance

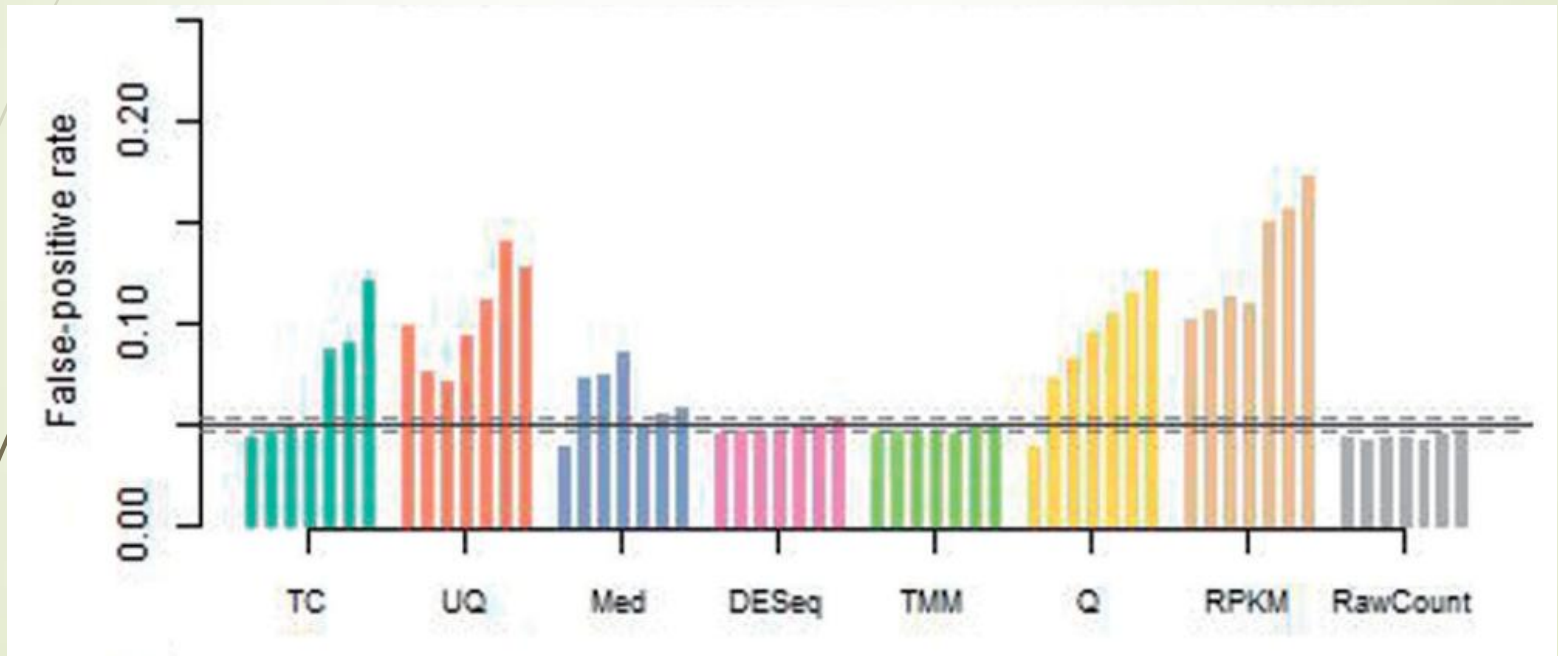
House keeping genes



<http://www.ncbi.nlm.nih.gov/pubmed/22988256>

Performance

False discovery rate





Differential Expression

- **Comparative analysis:**

- Compare gene expression across two or more samples to determine significant differential expressed gene list.

- **Methods in R:**

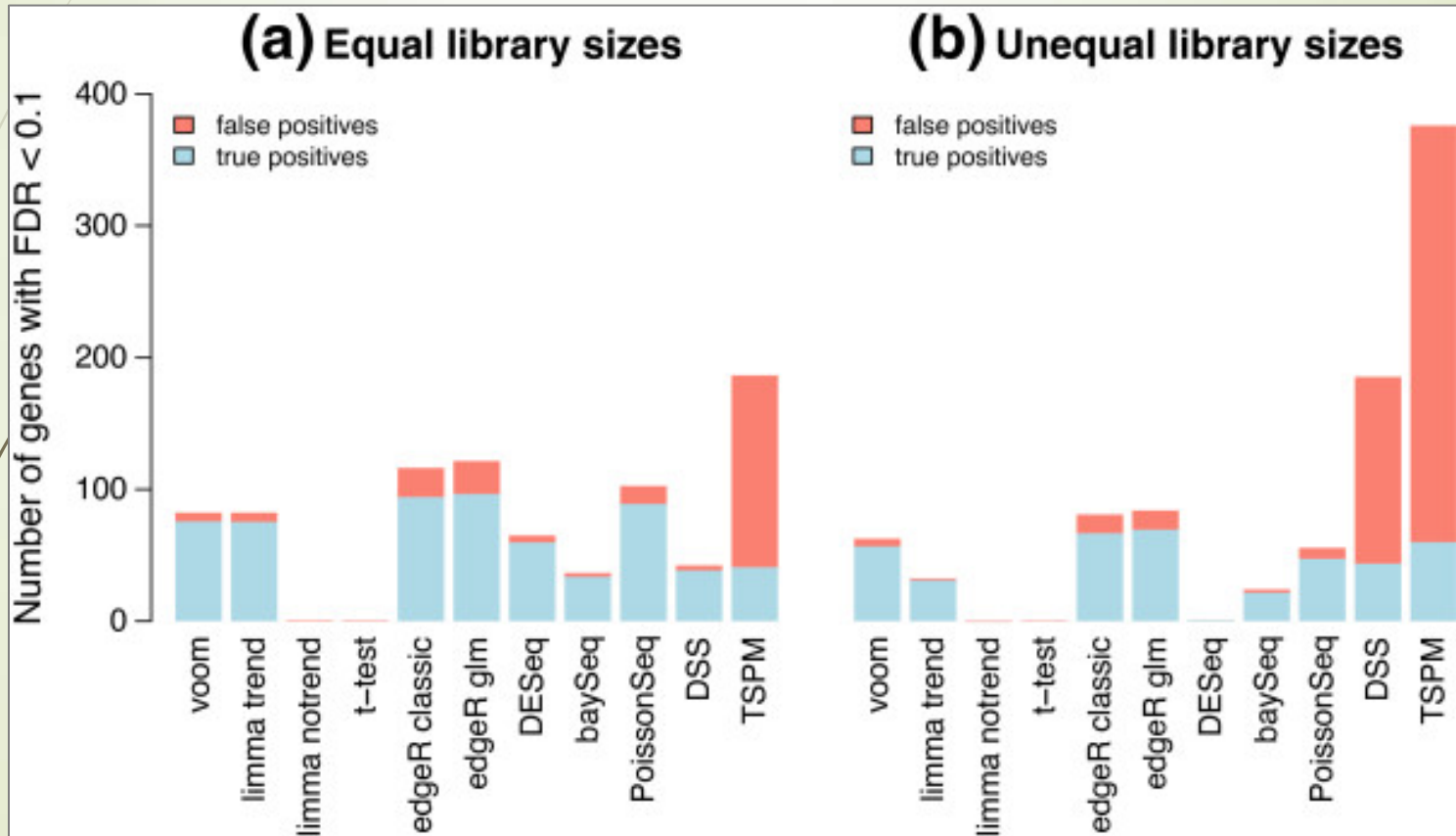
- **Limma**

- Adapted to support RNAseq after voom transformation

- **EdgeR**

- **DESeq2**

EdgeR & Limma (voom)



Code Snippet (edgeR)

```
# count.data is a matrix with genes in rows and samples in column
# cls.lbl is a factor containing class labels for the samples
> require(edgeR);
> design <- model.matrix(~-1 + cls.lbl);
> colnames(design) <- levels(cls.lbl);
> grps.cmp <- paste(levels(cls.lbl)[2], "-", levels(cls.lbl)[1], sep="");
> myargs <- list(grps.cmp, levels = design);
> contrast.matrix <- do.call(makeContrasts, myargs);
> y <- DGEList(counts=count.data, group=cls.lbl)
> y <- calcNormFactors(y)
> y <- estimateGLMCommonDisp(y, design);
> y <- estimateGLMTrendedDisp(y, design);
> y <- estimateGLMTagwiseDisp(y, design);
> fit <- glmFit(y, design);
> lrt <- glmLRT(fit, contrast=contrast.matrix);
> topFeatures <- topTags(lrt, n=Inf)$table;
```

- EdgeR user guide:

<https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf>

Step-by-step RNAseq Analysis

► In R - RNAseq123

- RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR

- <https://www.bioconductor.org/help/workflows/RNAseq123/>

► Using EcoToxXplorer!!

- Free, web/cloud-based tool

- Implemented most steps and algorithms

- Interactive visualization

Demo & hands on



Tutorial @ GitHub:

- <https://ecotoxxplorer.github.io/>



Two websites:

- <http://www.ecotoxxplorer.ca>
- <http://dev.ecotoxxplorer.ca>