edgeRv4 with expanded functionality and improved support for small counts and larger datasets

Lizhong Chen

Smyth Lab, Bioinformatics Division

WEHI

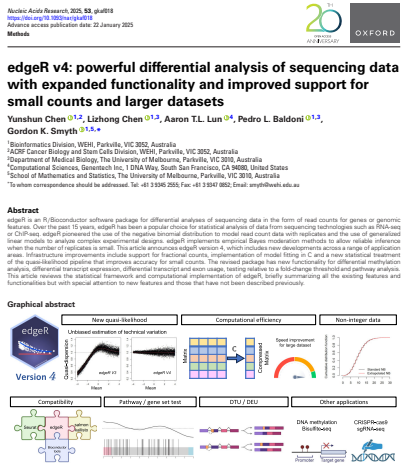
September 17, 2025

Quasi-likelihood pipeline

- Mean-variance relationship

$$\mathrm{var}[y_{gi}] = \sigma_g^2 \mu_{gi} + \psi_g \mu_{gi}$$

- Quasi-dispersion σ_g^2, accounting for technical overdispersion
- Negative binomial dispersion ψ_g, accounting for biological overdispersion
- Estimation of ψ_g is global, and we estimate $\hat{\psi}$ for all genes using highly expressed genes



Nucleic Acids Research, 2025, **53**, gkaf018
https://doi.org/10.1093/nar/gkaf018
Advance access publication date: 22 January 2025
Methods

edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets

Yunshun Chen [1,2], Lizhong Chen [1,3], Aaron T.L. Lun [4], Pedro L. Baldoni [1,3], Gordon K. Smyth [1,5,*]

[1]Bioinformatics Division, WEHI, Parkville, VIC 3052, Australia
[2]ACRF Cancer Biology and Stem Cells Division, WEHI, Parkville, VIC 3052, Australia
[3]Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia
[4]Computational Sciences, Genentech Inc, 1 DNA Way, South San Francisco, CA 94080, United States
[5]School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia
[*]To whom correspondence should be addressed. Tel: +61 3 9345 2555; Fax: +61 3 9347 0852; Email: smyth@wehi.edu.au

Abstract

edgeR is an R/Bioconductor software package for differential analyses of sequencing data in the form of read counts for genes or genomic features. Over the past 15 years, edgeR has been a popular choice for statistical analysis of data from sequencing technologies such as RNA-seq or ChIP-seq. edgeR pioneered the use of the negative binomial distribution to model read count data with replicates and the use of generalized linear models to analyze complex experimental designs. edgeR implements empirical Bayes moderation methods to allow reliable inference when the number of replicates is small. This article announces edgeR version 4, which includes new developments across a range of application areas. Infrastructure improvements include support for fractional counts, implementation of model fitting in C and a new statistical treatment of the quasi-likelihood pipeline that improves accuracy for small counts. The revised package has new functionality for differential methylation analysis, differential transcript expression, differential transcript and exon usage, testing relative to a fold-change threshold and pathway analysis. This article reviews the statistical framework and computational implementation of edgeR, briefly summarizing all the existing features and functionalities but with special attention to new features and those that have not been described previously.

Graphical abstract

Introduction

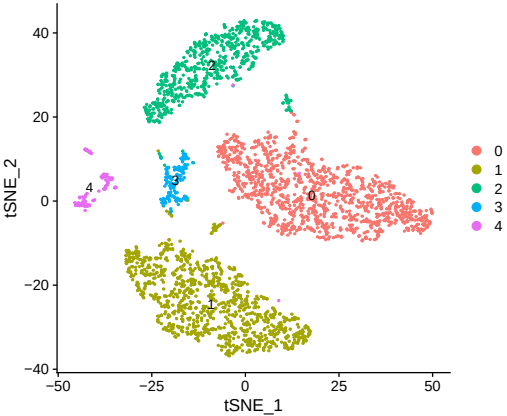
Next generation sequencing (NGS) has revolutionized biomedical research over the past 15–20 years. RNA-seq has become the standard technology for profiling gene and transcript expression [1, 2], while other technologies such as ChIP-seq, ATAC-seq, CUT&Tag, bisulfite sequencing (BS-seq) and Hi-C allow high-resolution exploration of the molecular mechanisms by which expression is regulated [3].

Received: January 17, 2024. **Revised:** November 22, 2024. **Editorial Decision:** January 6, 2025. **Accepted:** January 8, 2025

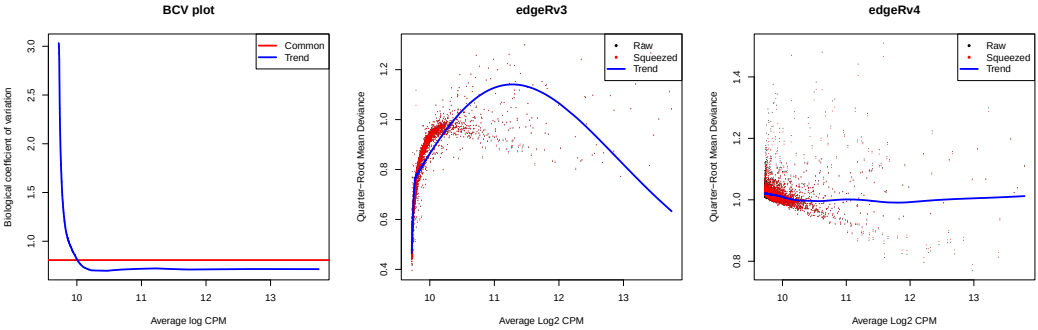
Application to single-cell RNA-seq data

```
# Size
> dim(y)
[1] 9996 3302
>
> # Seurat clusters
> cls <- so@meta.data$seurat_clusters
> des <- model.matrix(~ 0 + cls)
>
> # edgeRv3 pipeline
> system.time(y1 <- estimateDisp(y, des, tagwise = FALSE))
   user  system elapsed
 302.15    3.16  305.91
> system.time(fit1 <- glmQLFit(y1, des, legacy = TRUE))
   user  system elapsed
  19.36    0.32   19.74

> # edgeRv4 pipeline
> system.time(fit0 <- glmQLFit(y, des, legacy = FALSE))
   user  system elapsed
  76.78    0.60   77.56
```



Adjusted deviance statistics



Highly variable gene (HVG) selection

- Null hypothesis: a single population assumption
- Under null hypothesis, the variance is

$$\text{var}[y_{gi}] = \sigma_g^2 \mu_{gi} + \psi_g \mu_{gi}^2$$

- Assume $\sigma_g^2 = \sigma^2$ are the same for all genes
- Biological variation is measured by $\hat{\psi}_g$
- HVGs are those genes with large $\hat{\psi}_g$

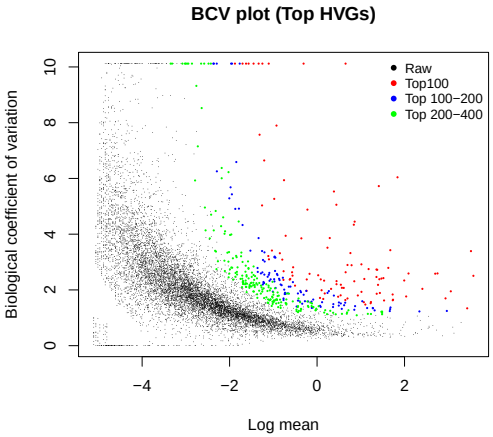
$$\hat{\psi}_g > \hat{\psi}$$

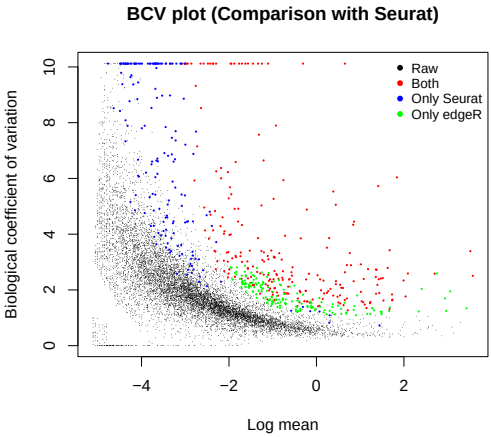
- The HVGs can be classified into two categories
- Null hypothesis is accepted but $\hat{\psi}_g$ is large
- Null hypothesis is rejected that μ_{gi} varies and results in large $\hat{\psi}_g$
- edgeRv4 pipeline does not estimate $\hat{\psi}_g$
- edgeRv4 pipeline does estimate $\hat{\sigma}_g^2$ and $\hat{\sigma}^2$
- edgeRv4 performs goodness of fit test

$$\hat{\sigma}_g^2 > \hat{\sigma}^2 \approx \hat{\psi}_g > \hat{\psi}$$

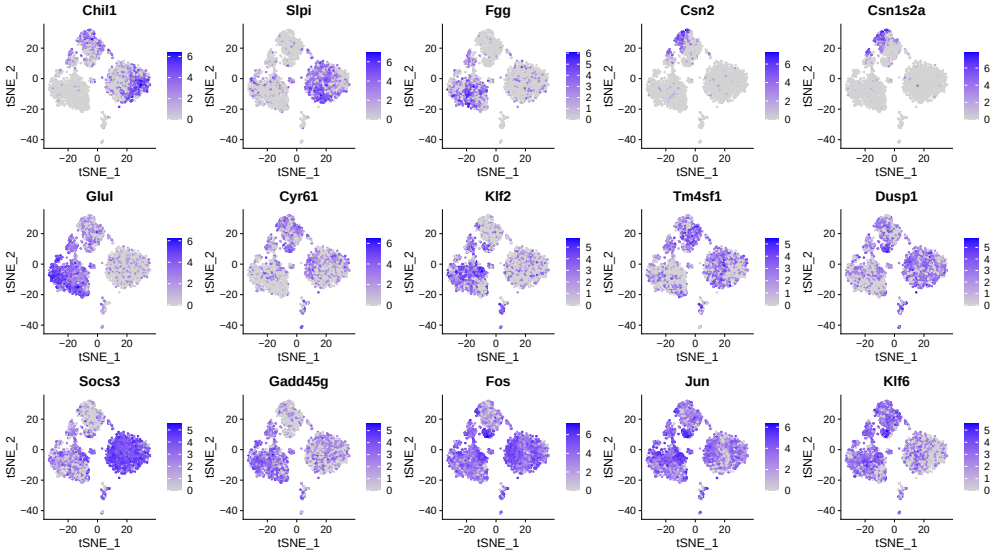
by adjusted deviance statistics

Results of selected HVGs





Examples of selected HVGs

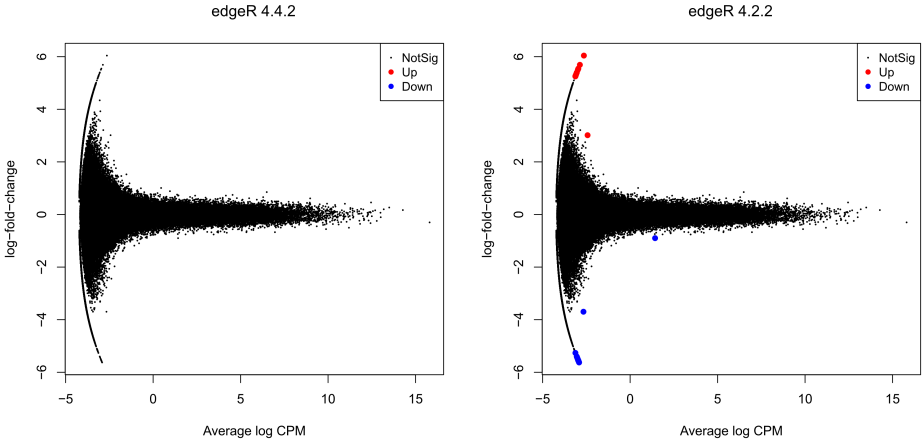


Prior estimation of empirical Bayes process

- Suppose we have $\hat{\sigma}_g^2 \sim s_0^2 \times F_{d_0,d_g}$ The problem is to estimate s_0^2, d_0.
- (Smyth 2004) Moment estimators on s_0^2 and d_0.
- (Sartor et al., 2006) Prior trend on s_0^2 using splines.
- (Phipson et al., 2016) Robust estimators on s_0^2 and d_0.

- Two major challenges for edgeRv4 QL method using adjusted deviance statistics
- $d_{g,adj}$ is not a constant, and may vary a lot.
- Many $d_{g,adj}$ can be small.
- A two-steps method is proposed to improve empirical Bayes hyperparameter estimation
- It is implemented in fitDistUnequalDF1() in limma

Prior estimation of empirical Bayes process



* Null simulation for DTE (only filter zeros), edgeR 4.2 fails to control FDR (df.prior = Inf)

Marker gene selection

- Marker genes are used to identify clusters with a positive logFC
- Cluster specific is preferred, one gene one cluster
- It can be a marker gene set, and the combination specifies cluster
- Assume clusters are well defined, edgeR performs one vs the average of others test
- For one sample, edgeR can perform on the single-cell level
- For multiple samples, pseudo-bulk approach is recommended

Marker gene selection

```
> # contrast matrix
> contr.matrix <- matrix(-1/4,5,5)
> diag(contr.matrix) <- 1
> contr.matrix
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,]  1.00 -0.25 -0.25 -0.25 -0.25
[2,] -0.25  1.00 -0.25 -0.25 -0.25
[3,] -0.25 -0.25  1.00 -0.25 -0.25
[4,] -0.25 -0.25 -0.25  1.00 -0.25
[5,] -0.25 -0.25 -0.25 -0.25  1.00
>
> # Test for cluster 2 (LP cells)
> qlf <- glmQLFTest(fit0, contrast = contr.matrix[,3])
> topTags(qlf)[,-(1:4)]
Coefficient:  -0.25*cls0 -0.25*cls1 1*cls2 -0.25*cls3 -0.25*cls4
         logFC   logCPM        F        PValue           FDR
Spp1  5.995375 13.90512 3207.176  0.000000e+00  0.000000e+00
Trf   5.297083 14.00792 3675.661  0.000000e+00  0.000000e+00
Csn3  5.194536 11.68842 2828.526  0.000000e+00  0.000000e+00
Plet1 4.371423 12.03224 3521.563  0.000000e+00  0.000000e+00
Cd14  4.014644 10.90419 2389.327  0.000000e+00  0.000000e+00
Lcn2  3.531909 11.76725 2493.848  0.000000e+00  0.000000e+00
Mfge8 3.517283 11.75730 2775.851  0.000000e+00  0.000000e+00
Cst3  3.089129 11.46859 2020.506  0.000000e+00  0.000000e+00
Mgst1 2.565137 11.32096 1751.716  0.000000e+00  0.000000e+00
Clu   2.926397 11.26094 1741.443 7.674257e-318 7.671188e-315
```

- p-values are not reliable because of the inter-correlation among cells
- Rank of genes is reasonable so we can choose top DE genes as potential marker genes
- Top DE genes may not be cluster specific
- For logFC cutoff, a treat-style method is recommended

differential splicing (differential transcript usage)

Dividing out quantification uncertainty enables assessment of differential transcript usage with limma and edgeR

Pedro L. Baldoni[1,2,†], Lizhong Chen[1,2,†], Mengbo Li[1,2], Yunshun Chen[1,3,3], and Gordon K. Smyth[1,2,*]

[1]Bioinformatics and Computational Biology Division, WEHI, Parkville, VIC 3052, Australia, [2]Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia, [3]ACRF Cancer Biology and Stem Cells Division, WEHI, Parkville, VIC 3052, Australia,

[†]These authors contributed equally to this work.

[*]To whom correspondence should be addressed. Tel: +61 3 9345 2555; Fax: +61 3 9347 0852; Email: smyth@wehi.edu.au

Abstract

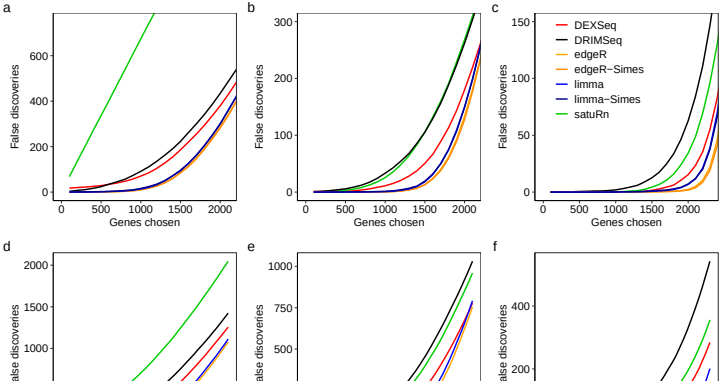
Differential transcript usage (DTU) refers to changes in the relative abundance of transcript isoforms of the same gene between experimental conditions, even when the total expression of the gene doesn't change. DTU analysis requires the quantification of individual isoforms from RNA-seq data, which has a high level of uncertainty due to transcript overlap and read-to-transcript ambiguity (RTA). Popular DTU analysis methods do not directly account for the RTA overdispersion within their statistical frameworks, leading to reduced statistical power or poor error rate control, particularly in scenarios with small sample sizes. This article presents limma and edgeR analysis pipelines that account for RTA during DTU assessment. Leveraging recent advancements in the limma and edgeR Bioconductor packages, we propose DTU analysis pipelines optimized for small and large datasets with a unified interface via the diffSplice function. The pipelines make use of divided counts to remove RTA-induced dispersion from transcript isoform counts and account for the sparsity in transcript-level counts. Simulations and analysis of real data from mouse mammary epithelial cells demonstrate that the diffSplice pipelines provide greater power, improved efficiency, and improved FDR control compared to existing specialized DTU methods.

Introduction

RNA sequencing (RNA-seq) has revolutionized biomedical research by enabling comprehensive profiling of the transcriptome, providing insights into gene expression regulation across diverse biological contexts, including cancer, immunology, and developmental biology. A common task in RNA-seq data analysis is to identify genomic features that have altered expression levels across conditions, such as treatments, disease status, or genotypes. Differential expression (DE) analysis has traditionally focused on genes as the primary units of expression [1]. However, genes often express multiple transcript isoforms (transcripts) via alternative splicing, a process in which gene exons are joined in different combinations, resulting in distinct messenger RNA products [2, 3, 4]. Recent computational and statistical developments now allow fast and accurate detection of differential transcript expression (DTE) [5, 6]. Yet, transcriptional changes resulting from alternative splicing rarely occur in isolation, as biological processes often affect multiple expressed transcripts of a gene simultaneously. Examples of such processes include alternative splicing via transcription start site variation and isoform switching via exon skipping [7]. These phenomena often occur in the context of cancer, where an oncogene transcript replaces a major transcript due to DNA damage or epigenomic modifications [8, 9]. It is therefore of key interest for biomedical researchers to identify those genes for which any differential splicing event has occurred, resulting in changes in the relative abundance of expressed transcripts for that gene between conditions.

Differential splicing can be assessed either at the level of exons via differential exon usage (DEU) or at the level of transcripts (RNA isoforms) via differential transcript usage (DTU). In DEU analyses, RNA-seq reads are aligned to a reference genome with a splice-aware aligner, reads are counted for exons, and

1



Future work

- Treat analysis - testing logFC relative to a threshold
- Sample weights - accounting for the variations in sample quality
- New quasi-likelihood pipeline for Methylation analysis

Acknowledgement

Smyth Lab

Gordon Smyth

Pedro Baldoni

Mengbo Li

Hannah Coughlan

Waruni Abeysekera

Davidson Lab

Alex Yan

Chen Lab

Andy Chen

Visvader Lab

Jane Visvader

WEHI brighter together

Australian Government
National Health and Medical Research Council

NHMRC

WEHI
brighter together

Thank you

WEHI_research WEHIresearch WEHImovies WEHI_research Walter and Eliza Hall Institute