ANOVA-Like Differential Expression tool for high throughput sequencing data

Greg Gloor[*1]

[1]Dep't of Biochemistry, University of Western Ontario

[*]ggloor@uwo.ca

24 July 2018

Package

ALDEx2 1.10.3

Contents

1 Why the ALDEx2 package?

Fundamentally, many high throughput sequencing approaches generate similar data: reads are mapped to features in each sample, these features are normalized, then statistical difference between the features composing each group or condition is calculated[1]. The standard statistical tools used to analyze RNA-seq, ChIP-seq, 16S rRNA gene sequencing, metagenomics, etc. are fundamentally different for each approach despite the underlying similarity in the data structures. In most cases the values expected by, and modelled by, these tools is counts.[2]

ALDEx2 breaks with this approach. Fundamentally, ALDEx2 models the data as the *probability* of observing the count[3]. In general, the observed data are single technical replicates, and the single observed count of a feature is one example from a distribution of examples that could have been observed under a repeated sampling model. The total read depth for a sample contains only information on the precision, and nothing else. ALDEx2 provides a consistent framework that encompasses essentially all high throughput sequencing data types by modelling the data as a log-ratio transformed probability distribution rather than counts (Fernandes et al. 2014).

2 Introduction to ALDEx2

This guide provides an overview of the R package ALDEx version 2 (ALDEx2) for differential (relative) abundance analysis of proportional data[4]. The package was developed and used initially for multiple-organism RNA-Seq data generated by high-throughput sequencing platforms (meta-RNA-Seq)[5], but testing showed that it performed very well with traditional RNA-Seq datasets, 16S rRNA gene variable region sequencing[6]] and selective growth-type (SELEX) experiments[7]. In principle, the analysis method should be applicable to nearly any type of data that is generated by high-throughput sequencing that generates tables of per-feature counts for each sample(Fernandes et al. 2014): in addition to the examples outlined above this would include ChIP-Seq or metagenome sequencing. We will be including examples and citations for application on these types of problems as we move forward.

The ALDEx2 package in Bioconductor is modular and is suitable for the comparison of many different experimental designs. This is achieved by exposing the underlying centre log-ratio transformed Dirichlet Monte-Carlo replicate values to make it possible for anyone to add the specific R code for their experimental design — a guide to these values is outlined below.

ALDEx2 estimates per-feature technical variation within each sample using Monte-Carlo instances drawn from the Dirichlet distribution. This distribution maintains the proportional nature of the data and returns a multivariate probability distribution. ALDEx2 uses the centred log-ratio (clr) transformation that ensures the data are scale invariant and sub-compositionally coherent[5]. The scale invariance property removes the need for a between sample data normalization step since the data are all placed on a consistent numerical co-ordinate. The sub-compositional coherence property ensures that the answers obtained are consistent when parts of the dataset are removed (e.g., removal of rRNA reads from RNA-seq studies or rare OTU species from 16S rRNA gene amplicon studies). All feature abundance values are expressed relative to the geometric mean abundance of all features in a sample. This is conceptually similar to a quantitative PCR where abundances are expressed relative to a standard: in the case of the clr transformation, the standard is the per-sample geometric mean abundance. See Aitchison (1986) for a complete description.

[1] Fernandes et al. (2014)

[2] Gierliński et al. (2015)

[3] Fernandes et al. (2013)

[4] all high throughput sequencing data are compositional (Gloor et al. 2017) because of constraints imposed by the instruments

[5] Macklaim et al. (2013)

[6] Bian et al. (n.d.)

[7] McMurrough et al. (2014);Wolfs et al. (2016)

[5] Aitchison (1986) The statistical analysis of compositional data ISBN:978-930665-78-1

In extreme cases we observe that the centre log-ratio can be asymmetric in the data. This occurs when the data are extremely asymmetric, such as when one group is largely composed of features that are absent in the other group. In this case the geometric mean will not accurately represent the appropriate basis of comparison for each group. ALDEx2 incorporates four methods to deal with this footnote[6]{Wu et al (in prep)}. The first is to include as the denominator for the geometric mean those features that are relatively invariant across all samples. This is termed the 'iqlr' method, and takes as the denominator the geometric mean of those features with variance calculated from the clr that are between the first and third quartile. This approach can be used until the asymmetry becomes so severe that more than 25% of the features are asymmetric between the groups. The iqlr approach has the advantage that it gives essentially the same answer as using the entire set of features in symmetric datasets. The second approach, termed the 'zero' approach uses a different denominator for each group. The per-group denominator is the set of non-zero features in the group. This method introduces much stronger assumptions, but is helpful when the two groups under comparison are very different. The third approach, termed the 'lvha' approach identifies those features that are in the bottom quartile in variance across samples and are in the top quartile in relative abundance in every sample. This approach attempts to find 'housekeeping' features akin to those used as internal standards for qPCR. The final approach allows the user to specify a vector of row indices for the input data that are to be used as the denominator. In this case, the user is making an assumption regarding which features are invariant. In the case of RNA-seq, this might include a set of 'housekeeping' genes that are though to be invariant under the perturbation being tested. IMPORTANT: all rows must contain one or more counts when the user defines the row indices to ensure the appropriate rows are chosen.

3 Ways to install

There are multiple ways to download and install the most current of ALDEx2. ALDEx2 will run with only the base R packages and is capable of running several functions with the 'parallel' package if installed. It has been tested with version R version 3, but should run on version 2.12 onward. ALDEx2 will make use of the BiocParallel package if possible, otherwise, ALDEx2 will run in serial mode.

4 Quick example with 'selex' example data and 2 groups:

Case study a growth selection type (SELEX) experiment[8]. This section contains an analysis of a dataset collected where a single gene library was made that contained 1600 sequence variants at 4 codons in the sequence. These variants were cloned into an expression vector at equimolar amounts. The wild-type version of the gene conferred resistance to a topoisomerase toxin. Seven independent growths of the gene library were conducted under selective and non-selective conditions and the resulting abundances of each variant was read out by sequencing a pooled, barcoded library on an Illumina MiSeq. The data table is included as selex_table.txt in the package. In this data table, there are 1600 features and 14 samples. The analysis takes approximately 2 minutes and memory usage tops out at less than 1Gb of RAM on a mobile i7 class processor. For speed concerns we use only the first 400 features. The command used for ALDEx2 is presented below:

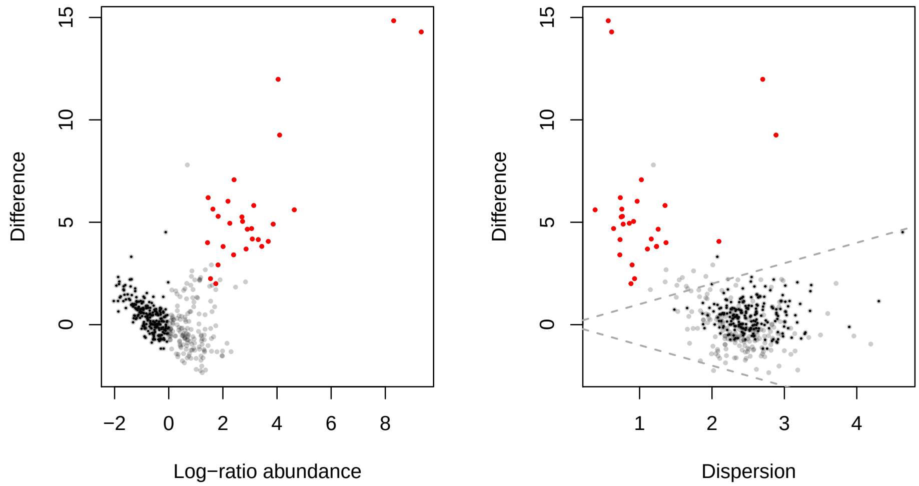
[8] McMurrough et al. (2014)

First we load the library and the included selex dataset. Then we set the comparison groups. This must be a vector of conditions in the same order as the samples in the input counts table.

```
library(ALDEx2)
data(selex)
#subset only the last 400 features for efficiency
selex <- selex[1:400,]

conds <- c(rep("NS", 7), rep("S", 7))
x.all <- aldex(selex, conds, mc.samples=16, test="t", effect=TRUE,
     include.sample.summary=FALSE, denom="all", verbose=FALSE)
## Warning in aldex.clr.function(reads, conds, mc.samples, denom, verbose, :
## values are unreliable when estimated with so few MC smps
## [1] "computing center with all features"

par(mfrow=c(1,2))
aldex.plot(x.all, type="MA", test="welch", xlab="Log-ratio abundance",
    ylab="Difference")
aldex.plot(x.all, type="MW", test="welch", xlab="Dispersion",
    ylab="Difference")
```



ALDEx2 is now modular, offering the user the ability to build a data analysis pipeline for their experimental design. However, for two sample tests and one-way ANOVA design, the user can run the aldex wrapper. This wrapper will link the modular elements together to emulate ALDEx2 prior to the modular approach. Note that if the test is kw', theneffectshould be FALSE. If the test ist', then effect should be set to TRUE. The t' option evaluates the data as a two-factor experiment using both the Welch's t and the Wilcoxon rank tests. Thekw' option evaluates the data as a one-way ANOVA using the glm and Kruskal-Wallace tests. All tests include a Benjamini-Hochberg correction of the raw P values. The data can be plotted onto Bland-Altmann (MA) or effect (MW) plots[7] for for two-way tests using the 'aldex.plot' function. See the end of the modular section for examples of the plots.

```
sessionInfo()
## R version 3.5.0 (2018-04-23)
```

[7] Gloor et al (2016) J. Comp. Graph. Stat. DOI:10.1080/10618600.2015.11311

```
## Platform: x86_64-apple-darwin17.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.6
##
## Matrix products: default
## BLAS: /opt/local/Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.dylib
## LAPACK: /opt/local/Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ALDEx2_1.10.3   BiocStyle_2.8.2
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.17                compiler_3.5.0
##  [3] GenomeInfoDb_1.16.0         XVector_0.20.0
##  [5] bitops_1.0-6                tools_3.5.0
##  [7] zlibbioc_1.26.0             digest_0.6.15
##  [9] evaluate_0.11               lattice_0.20-35
## [11] Matrix_1.2-14               DelayedArray_0.6.1
## [13] yaml_2.1.19                 parallel_3.5.0
## [15] xfun_0.3                    GenomeInfoDbData_1.1.0
## [17] stringr_1.3.1               knitr_1.20
## [19] S4Vectors_0.18.3            IRanges_2.14.10
## [21] stats4_3.5.0                rprojroot_1.3-2
## [23] multtest_2.36.0             grid_3.5.0
## [25] Biobase_2.40.0              survival_2.41-3
## [27] BiocParallel_1.14.2         rmarkdown_1.10
## [29] bookdown_0.7                magrittr_1.5
## [31] MASS_7.3-49                 backports_1.1.2
## [33] htmltools_0.3.6             matrixStats_0.53.1
## [35] BiocGenerics_0.26.0         GenomicRanges_1.32.6
## [37] splines_3.5.0               SummarizedExperiment_1.10.1
## [39] stringi_1.2.4               RCurl_1.95-4.11
```

Bian, Gaorui, Gregory B Gloor, Aihua Gong, Changsheng Jia, Wei Zhang, Jun Hu, Hong Zhang, et al. n.d. "The Gut Microbiota of Healthy Aged Chinese Is Similar to That of the Healthy Young." *mSphere* 2 (5):e00327–17. https://doi.org/10.1128/mSphere.00327-17.

Fernandes, Andrew D, Jean M Macklaim, Thomas G Linn, Gregor Reid, and Gregory B Gloor. 2013. "ANOVA-Like Differential Expression (Aldex) Analysis for Mixed Population Rna-Seq." *PLoS One* 8 (7):e67019. https://doi.org/10.1371/journal.pone.0067019.

Fernandes, Andrew D, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. 2014. "Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-Seq, 16S RRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis." *Microbiome* 2:15.1–15.13. https://doi.org/10.1186/2049-2618-2-15.

Gierliński, Marek, Christian Cole, Pietà Schofield, Nicholas J Schurch, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, et al. 2015. "Statistical Models for Rna-Seq Data Derived from a Two-Condition 48-Replicate Experiment." *Bioinformatics* 31 (22):3625–30. https://doi.org/10.1093/bioinformatics/btv425.

Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. 2017. "Microbiome Datasets Are Compositional: And This Is Not Optional." *Frontiers in Microbiology* 8:2224. https://doi.org/10.3389/fmicb.2017.02224.

Macklaim, M Jean, D Andrew Fernandes, M Julia Di Bella, Jo-Anne Hammond, Gregor Reid, and Gregory B Gloor. 2013. "Comparative Meta-RNA-Seq of the Vaginal Microbiota and Differential Expression by *Lactobacillus Iners* in Health and Dysbiosis." *Microbiome* 1:15. https://doi.org/doi: 10.1186/2049-2618-1-12.

McMurrough, Thomas A, Russell J Dickson, Stephanie M F Thibert, Gregory B Gloor, and David R Edgell. 2014. "Control of Catalytic Efficiency by a Coevolving Network of Catalytic and Noncatalytic Residues." *Proc Natl Acad Sci U S A* 111 (23):E2376–83. https://doi.org/10.1073/pnas.1322352111.

Wolfs, Jason M, Thomas A Hamilton, Jeremy T Lant, Marcon Laforet, Jenny Zhang, Louisa M Salemi, Gregory B Gloor, Caroline Schild-Poulter, and David R Edgell. 2016. "Biasing Genome-Editing Events Toward Precise Length Deletions with an Rna-Guided Tevcas9 Dual Nuclease." *Proc Natl Acad Sci U S A*, December. https://doi.org/10.1073/pnas.1616343114.