CoDa HTS Workshop Proposal

Greg Gloor

Contents

Analyzing data as compositions

Website: https://github.com/ggloor/CoDa_microbiome_tutorial. This will serve as the central repository for the demonstrated tools and workflows. The repository will be set as release 2.0 at the end of the workshop so that participants will have a permanent public record of what was covered.

We have adapted and developed tools and protocols for the analysis of HTS as compositional count data (Erb and Notredame, 2016; Fernandes et al., 2013, 2014; Quinn et al., 2017a). Analyses conducted under this paradigm are reproducible and robust, and allow conclusions about the relative relationships between features (genes, OTUs, etc) in the underlying environment (Bian et al.; Gloor et al., 2017).

It is possible to replace almost all steps in traditional RNA-seq, metagenomics or 16S rRNA gene sequencing analysis with compositionally appropriate methods (Gloor et al., 2017) that are robust to data manipulations and that provide reproducible insights into the underlying biology and composition of the system.

Objectives and outcomes

The workshop will enable participants to:

1. be able to identify when biological datasets are compositional, and understand the root problems that cause problems when interrogating compositional datasets.
2. understand why HTS data should be analyzed in a compositionally-appropriate framework.
3. know how to install, use and interpret the output from the basic HTS compositional toolkit that consists of compositional biplots, the `propr R` package and the `ALDEx2 R` package.
4. have a frame of reference for more complex compositional tools such as `philr` and concepts such as b-association and balance dendrograms.

Outline

The workshop will be delivered as mixed didactic and participation sessions, with about a 1:4 mixture. Each session will be introduced by a short didactic introduction and demonstration. The remainder of the session will be hands-on learning exercises in the `R` programming environment.

We will demonstrate a test dataset from [@]Schurch:2016aa;@Gierlinski:2015aa] the lab of Dr. Geoffrey Barton that examined the effect of a SNF2 gene knockout *Saccharomyces cervisiae* transcription. This dataset is nearly ideal and simple to understand.

However, participants are invited (expected) to bring their own dataset in the form of a count table with associated metadata for examination.

The outline of this 1-day workshop is:

Start time 9am - Introduction (Gloor, didactic lecture)

- demonstrate and understand the geometry of high throughput sequencing data and how this constrains the analyses
  - demonstrate the pathologies associated with HTS data analyzed using standard methods
  - enable participants to understand why and when the usual methods of analysis are likely to be misleading
  - understand the importance of subcompositional coherence and subcompositional dominance, and how these concepts lead to robust analyses

Start time 9:45 - Probabilities and ratio transformations (Gloor, hands on)

- provide an overview of sequencing as a probabilistic process, and the manipulation of probability vectors using compositional data methods
  - how to generate probability distributions from count data using ALDEx2
  - how to generate and interpret compositionally appropriate data transformations
  - zero replacement strategies for sparse data with the zCompositions R package
  - why count normalization is futile

Break 10:30

Start time 11 - Dimension reduction, outlier identification and clustering (Gloor, hands on)

- demonstrate dimension reduction of compositional data
  - the production and interpretation of a compositional PCA biplot
  - identifying outlier samples
  - learn how to conduct and interpret clustering and discriminate analysis in compositional data
  - fuzzy clustering

Start time 12: Correlation and compositional association (Erb)

- demonstrate compositionally appropriate identification of correlated (compositionally associated) features using the `propr` `R` package (Quinn et al., 2017b)
  - an introduction to compositional association

Start time: 1 lunch break

Start time : 2: - Correlation and compositional association continued (Erb)

Start time 2:30 Differential abundance with ALDEx2 (Gloor)

- demonstrate compositionally appropriate identification of differentially relatively abundant features using the `ALDEx2` `R` package
  - learn how to generate and interpret posterior expected values for differential relative abundance
  - learn how to generate and use standardized effect sizes for differential relative abundance
  - learn how to interpret effect plots as an adjunct to volcano and Bland-Altmann plots

Start time: 3:30 - Working with users' data (Gloor, Erb)

- analyzing users' own data
- troubleshooting users' own datasets
- common problems from the participants will be highlighted and solutions demonstrated

Start time: 4:30- Wrapup (Gloor, Erb)

- review of concepts and strategies
- understand the congruence between the results obtained by the compositional biplot, compositional association and

compositional differential relative abundance
- provide guidance and sources on the proper interpretation of HTS datasets using a compositional paradigm

Finish time 5 pm

Requirements

1. a reasonably up-to-date laptop computer with at leaset 8Gb RAM
2. familiarity with scripting or programming languages, proficency in the `R` programming environment
3. the current version of the `R` programming language installed
4. a number of `R` packages will be used during the workshop. Participants should be familiar with installation of packages from both `Bioconductor` and `CRAN`

Intended Audience and Level

The intended audience for this session is bioinformaticians or computational biologists who use high throughput sequencing with experimental designs that include tag sequencing (eg. 16S rRNA gene sequencing), metagenomics, transcriptomics or meta-transcriptomics.

This is not intended to be an introduction to R for bioinformaticians: attendees should be relatively proficient with R, either using RStudio, or on the command line and should have a plain text editor available. Attendees will use R markdown documents to keep track of their work, and templates will be provided for use. Attendees will be expected to have a laptop with R installed and the following packages and their dependencies: propr (CRAN), ALDEx2 (Bioconductor), omicplotR (Bioconductor), zCompositions (CRAN). Attendees are encouraged to bring their own datasets for analysis, but should be aware that only pairwise (i.e., two condition) experiments will be demonstrated.

Compositional concepts will be at an introductory-intermediate level suitable for participants of any background, but will be more intuitive to those with a grounding in probability and linear algebra.

The practical aspects will be at an intermediate level, suitable for participants with pre-exisiting competency in `R`.

Attendance should be capped at no more than 40 participants.

Organizers and Presenters

Greg Gloor is a Professor of Biochemistry at The University of Western Ontario. He is one of the pioneers in using compositional data analysis to analyze HTS datasets. He is the maintainer of the `ALDEx2 R` package on Bioconductor used for differential relative abundance analysis. He has published original research, methods papers, and reviews that use compositional data analysis methods to interpret HTS datasets using transcriptome, microbiome and meta-transcriptome datasets (Bian et al.; Fernandes et al., 2013, 2014; Gloor et al., 2016a, 2017, 2016b, 2016c; Gloor and Reid, 2016; Goneau et al., 2015; Macklaim et al., 2013; McMillan et al., 2015; Wolfs et al., 2016). He has taught undergraduate and graduate courses in computational biology for almost two decades, and has won awards from both student groups and from faculty-wide competitions. His homepage and CV is at ggloor.github.io

Ionas Erb is a PDF and Bioinformatician at the Centre for Genomic Regulation. He is an active developer of tools to determine compositional association and is a contributor to the `propr R` package on CRAN used to explore correlation in a compositionally appropriate manner. He is an advocate for and active developer of tools that for compositionally-appropriate methods to examine correlation (Erb and Notredame, 2016; Erb et al., 2017; Quinn et al., 2017a)

References

Bian, G., Gloor, G. B., Gong, A., Jia, C., Zhang, W., Hu, J., et al. The gut microbiota of healthy aged chinese is similar to that of the healthy young. *mSphere* 2, e00327–17. doi:10.1128/mSphere.00327-17.

Erb, I., and Notredame, C. (2016). How should we measure proportionality on relative gene expression data? *Theory in Biosciences* 135, 21–36.

Erb, I., Quinn, T., Lovell, D., and Notredame, C. (2017). Differential proportionality - a normalization-free approach to differential gene expression. *bioRxiv*. doi:10.1101/134536.

Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., and Gloor, G. B. (2013). ANOVA-like differential expression (aldex) analysis for mixed population rna-seq. *PLoS One* 8, e67019. doi:10.1371/journal.pone.0067019.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (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. doi:10.1186/2049-2618-2-15.

Gloor, G. B., Macklaim, J. M., and Fernandes, A. D. (2016a). Displaying variation in large datasets: Plotting a visual summary of effect sizes. *Journal of Computational and Graphical Statistics* 25, 971–979. doi:10.1080/10618600.2015.1131161.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology* 8, 2224. doi:10.3389/fmicb.2017.02224.

Gloor, G. B., Macklaim, J. M., Vu, M., and Fernandes, A. D. (2016b). Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian Journal of Statistics* 45, 73–87. doi:doi:10.17713/ajs.v45i4.122.

Gloor, G. B., and Reid, G. (2016). Compositional analysis: A valid approach to analyze microbiome high-throughput sequencing data. *Can J Microbiol* 62, 692–703. doi:10.1139/cjm-2015-0821.

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016c). It's all relative: Analyzing microbiome data as compositions. *Ann Epidemiol* 26, 322–9. doi:10.1016/j.annepidem.2016.03.003.

Goneau, L. W., Hannan, T. J., MacPhee, R. A., Schwartz, D. J., Macklaim, J. M., Gloor, G. B., et al. (2015). Subinhibitory antibiotic therapy alters recurrent urinary tract infection pathogenesis through modulation of bacterial virulence and host immunity. *MBio* 6. doi:10.1128/mBio.00356-15.

Macklaim, M. J., Fernandes, D. A., Di Bella, M. J., Hammond, J.-A., Reid, G., and Gloor, G. B. (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. *Microbiome* 1, 15. doi:doi:10.1186/2049-2618-1-12.

McMillan, A., Rulisa, S., Sumarah, M., Macklaim, J. M., Renaud, J., Bisanz, J. E., et al. (2015). A multi-platform metabolomics approach identifies highly specific biomarkers of bacterial diversity in the vagina of pregnant and non-pregnant women. *Sci Rep* 5, 14174. doi:10.1038/srep14174.

Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2017a). Understanding sequencing data as compositions: An outlook and review. *bioRxiv*. doi:10.1101/206425.

Quinn, T., Richardson, M. F., Lovell, D., and Crowley, T. (2017b). Propr: An R-package for identifying proportionally abundant features using compositional data analysis. *bioRxiv*. doi:10.1101/104935.

Wolfs, J. M., Hamilton, T. A., Lant, J. T., Laforet, M., Zhang, J., Salemi, L. M., et al. (2016). Biasing genome-editing events toward precise length deletions with an rna-guided tevcas9 dual nuclease. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1616343114.