PeerJ Preprints NOT PEER-REVIEWED

Normalization of metatranscriptomic and metaproteomic data for differential gene expression analyses: The importance of accounting for organism abundance

Author: Manuel Kleiner

Affiliations:

Energy Bioengineering and Geomicrobiology Group, Department of Geoscience, University of Calgary, Calgary, Canada

Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, USA

Correspondence: manuel.kleiner@ucalgary.ca

Article Format: Perspective Piece

Keywords: Metaproteomics, metatranscriptomics, transcriptomics, proteomics, differential gene expression analyses

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2846v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

PeerJ Preprints NOT PEER-REVIEWED

Abstract

Metatranscriptomics and metaproteomics make it possible to measure gene expression in microbial communities. So far these approaches were mostly used to get a general overview of the dominant metabolism and physiologies of community members. Recently, environmental microbiologists have started using metatranscriptomics and metaproteomics to look at gene expression differences between different environments or conditions. This has been mostly done by using makeshift adaptations of pure culture focused differential transcriptomics and proteomics approaches. However, since meta-omics data has many more variables attached to it as compared to pure culture derived data, such makeshift adaptations are problematic at best. One particular challenge is posed by the data normalization strategies used to account for technical and biological variables in meta-omic data. Here I discuss the most common normalization strategy for transcriptomic and proteomic data and why it is not valid by itself for meta-omic data. I provide logical proof that variation in species abundances between samples is an additional variable that must be accounted for during normalization of meta-omic data. Finally, I show how the existing normalization methods for transcriptomic and proteomic data can be augmented to be applicable to meta-omic data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2846v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

Main Text

In the last decade technological advances in sequencing technology and mass spectrometry have made it possible to measure gene expression in microbial communities on a large scale. The respective approaches have been termed metatranscriptomics and metaproteomics [1, 2]. Metatranscriptomics is an umbrella term for methods that measure the transcription levels in microbial communities and metaproteomics is the corresponding term for methods that measure the protein abundances in microbial communities. The outputs of both methods are tables which list gene expression values for individual genes (rows) across multiple samples (columns). See the first worksheet in the supplemental table for a simulated example (Supplementary Table S1). For metatranscriptomics, the expression values are usually based on the counting of reads mapped to a set of reference genes/genomes. For metaproteomics, the expression values are based on the number of spectra matching to reference protein sequences (spectral counting based methods) or on the chromatographic peak intensities of peptides that match to reference protein sequences [3]. These raw counts or intensities are usually converted into proportional (relative) data that gives individual gene expression values as a fraction of 1. This conversion process is part of the data normalization discussed below.

Initially metatranscriptomics and metaproteomics were mostly used for discovery based studies that addressed the question which genes are expressed in the community and which proteins are the most important players [4, 5]. In more recent years researchers have started to use these methods for a more in depth investigation of how gene expression differs between different environmental sites, seasons or real or artificially induced changes (e.g. [6-8]). So we are now entering an era in which we start applying differential metatranscriptomics and metaproteomics. So far most differential meta-omics studies have used makeshift adaptations of well-established differential transcriptomics and proteomics methods that were developed for single-organism applications.

Metatranscriptomics and metaproteomics come with their own specific set of methodological challenges including, for example, sample extraction biases, contaminants, the construction of suitable reference databases and problems with database redundancies. These challenges are or will be discussed elsewhere [9-11].

Here I will discuss data normalization for differential gene expression analyses of metatranscriptomes and metaproteomes, which differs in part from the normalization steps required for differential transcriptomics and proteomics. To make samples comparable on a gene expression level for transcriptomics and proteomics the necessity for two normalization steps is widely accepted [12-15]: (i) In the first normalization step, the expression values are adjusted for the gene/protein sequence length, which can for example be done by simple division of the expression values by gene length. This normalization step is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2846v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

justified by the fact that both metatranscriptomics and the metaproteomics will yield higher raw expression values (read counts, spectral counts or summed peptide intensities) for larger transcripts/proteins. (ii) In the second normalization step, the expression values are adjusted for variations in the sum of expression values for each sample (column). After this normalization step the sum of expression values for each sample should be identical across all samples (e.g. if you normalize to %, the sum of each column should be 100). This normalization step is justified and needed because of technical variations between sample runs. In nextSeq based metatranscriptomics each sample will for example yield a different number of total reads, while in metaproteomics variation between runs can lead to difference in total spectral counts or peptide intensities. These normalization steps have been implemented in many different forms for both transcriptomics and proteomics and are reviewed elsewhere [12-14]. Suitable implementations of this normalization scheme for transcriptomics are the transcripts per million (TPMs) metric [12] and for proteomics either normalized spectral abundance factors (NSAFs) [14, 16] or for peptide intensities MaxLFQ [17].

For metatranscriptomes and metaproteomes an additional level of variation needs to be considered when comparing expression differences between genes of individual organisms. This additional level is variation of organism abundances between samples. Here an important differentiation has to be made, as the kind of normalization required in meta-omes very much depends on the exact question asked:

(a) If your question is of the type: "Does the expression of *gene*A contribute a higher number of transcripts/protein mass to COMMUNITY1 as compared to COMMUNITY2?" OR "Which genes differ in contribution to total community transcript number or protein mass between COMMUNITY1 and COMMUNITY2?", then the above described two-step normalization scheme for transcriptomics and proteomics is perfectly adequate. To give two concrete example for such questions (1) "Is the human structural protein collagen enriched in the intestinal microbiome samples of PERSON1 versus PERSON2?" and (2) "Are Carbohydrate-Active Enzymes (*CAZymes*) overall more abundantly expressed in COMMUNITY1 as compared to COMMUNITY2

(b) If your question is of the type: "Is the expression of *gene*A from SPECIESX higher in SPECIESX in COMMUNITY1 as compared to SPECIESX in COMMUNITY2?" OR "Which genes differ in expression between COMMUNITY1 and COMMUNITY2 on the species level?", then the above described two-step normalization scheme for transcriptomics and proteomics by itself is not valid. As I will prove here, an additional normalization step is needed after the two-step normalization to account for variation in species/strain abundances between samples.

Generally, there are at least two ways to provide evidence or proof for this. First, one could generate empirical data using two or more mock communities made with the same species, but different species abundances. This data could then be used to validate normalizations methods. Such mock community studies have helped to validate other omics methods for environmental microbiology in the past e.g. methods for quantitative metagenomic sequencing [18, 19]. The second approach that one can use in this case is to do a thought experiment to show that the comparison of expression values is invalid if the data is not corrected for variation in species abundance in each sample (and valid if the correction is done). I will use simulated datasets that represent two extreme cases for this thought experiment.

To re-iterate the assumptions:

(1) Gene expression is measured for a microbial community with >1 species.
(2) Gene expression values have been normalized to gene length and the sum of expression values in each sample (column).
(3) We ask a question of the type (b) above.

Proof:

In the first worksheet of the supplemental table the simplest case of a microbial community is shown: one with only two community member species. To keep it simple, I assume that for each of the two species gene expression was detected for 50 genes and that the expression of all genes is identical. To emphasize the importance of replication for differential omics [20], I show 6 replicate columns; although for the purpose of this proof replication is not really relevant.

To show the effect of relative species abundance in the community on gene expression data I have simulated the gene expression data for two distinct species abundance profiles. Samples 1 through 6 come from a community in which both species have the same abundance (1:1 abundance ratio). In samples 7 to 12 the same exact gene expression patterns are shown, but expression values have been adjusted to be coming from a very different species abundance profile (species ratio is 20:1). Without the need for statistical tests, it becomes immediately clear that the expression of individual genes would be considered to be different between the two community types. This proves that for type (b) questions two-step normalized data is not sufficient.

How to normalize expression data for species abundances?

Now the question is of course how to actually normalize the data to species abundance. The simplest way is to normalize the expression values for each sample and species to a constant value (i.e. the sum of expression values for each species in each sample should be the same after normalization), which make expression values comparable across samples as the effect of different species abundance profiles is removed. A simple implementation of this is shown in the second worksheet of the supplemental table. An implementation of this procedure for spectral counting based metaproteomics was published by Mueller et al. [21] and has been used in many other metaproteomics studies [7, 22, 23]. One important thing to check before normalizing to species/strains is that there are enough measurements (e.g. read counts, spectral counts) for the species/strain to be normalized to. This is crucial to avoid skewing the data simply because there are only very few transcripts/proteins to be considered for the respective species/strain.

Normalization to species could actually be abolished if only the reference genome/protein sequences of the organism of interest were used for generating the expression profile data by read mapping or spectral counting. However, using only a subset of reference sequences for the generation of expression data carries the danger of reads or spectra falsely mapping to this reference due to the absence of the potentially better matching reference sequences of the other community members. For metatranscriptomics this can be alleviated by using very strict read mapping criteria, i.e. only use counts from reads mapped with very high identity. For metaproteomics, the strategy of only using the target organism reference genome cannot be recommended, because spectra that would match non-uniquely to multiple sequences if the complete database were used, may match uniquely to a single protein sequence if a limited set of sequences is used (for more details on the so called protein inference problem see [24]).

There are several alternative approaches to data intrinsic normalization that could be used. First, abundance profile data obtained with other methods, e.g. 16S rRNA amplicon sequencing or metagenomic sequencing, could be used to correct expression values for each sample. However, this kind of data might bring its own skews and biases into the normalization procedure. Second, spiking in of known amounts of mRNA or protein into samples prior to extraction allows estimating transcript or protein abundances in relation to the standard. This spike in strategy can provide absolute per cell quantification if cell numbers are determined prior to extraction [10, 25].

A normalization of expression values to housekeeping genes, which is sometimes used for transcriptomic and proteomic data [26] can currently not be used for metatranscriptomics and metaproteomic data. A housekeeping gene based normalization requires that the housekeeping gene in question is quantified as a function of cell number or cell mass for all conditions that will be considered in a differential expression experiment. In theory, such a correlation of cell number with housekeeping gene expression could be measured for members of a microbial community e.g. by using a combination of mRNA FISH with 16S rRNA FISH, however, the effort required for this seems prohibitive, particularly since much simpler methods are already available.

What comes after normalization?

PeerJ Preprints NOT PEER-REVIEWED

Of course, the normalization steps are only a small part of the workflow for looking at gene expression differences. After normalization of the data, simple checks should be done to test the overall validity of the data and to discover potential sample mixups and alike. This can, for example, be done by hierarchical clustering or principal component analysis of samples based on expression values. Here you should see a separation of samples based on the sampling sites or conditions used. If all seems in order, one can proceed with statistical testing for differential gene expression. Statistical methods for differential gene expression analyses, which have to account for the multiple testing problem, missing values and the fact that the normalized gene expression data represents compositional data, are discussed elsewhere [15, 27-29].

Acknowledgments

I am grateful to The Symbiosis Department at the Max Planck Institute for Marine Microbiology for extensive discussions on normalization that inspired the conception of this manuscript, Maxim Rubin-Blum, Juliane Wippler and Tjorven Hinzke for feedback on the manuscript. MK was supported by an NSERC Banting Postdoctoral Fellowship.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2846v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017

PeerJ Preprints NOT PEER-REVIEWED

References

1. Warnecke F, Hess M: **A perspective: Metatranscriptomics as a tool for the discovery of novel biocatalysts**. *J Biotechnol* 2009, **142**:91-95.
2. Hettich RL, Sharma R, Chourey K, Giannone RJ: **Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities**. *Curr Opin Microbiol* 2012, **15**:373-380.
3. Nahnsen S, Bielow C, Reinert K, Kohlbacher O: **Tools for label-free peptide quantification**. *Mol Cell Proteomics* 2013, **12**:549-556.
4. Williams TJ, Cavicchioli R: **Marine metaproteomics: deciphering the microbial metabolic food web**. *Trends Microbiol* 2014, **22**:248-260.
5. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF: **Systems Biology: Functional analysis of natural microbial consortia using community proteomics**. *Nat Rev Microbiol* 2009, **7**:196-205.
6. Hawley AK, Brewer HM, Norbeck AD, Paša-Tolić L, Hallam SJ: **Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes**. *Proc Natl Acad Sci USA* 2014, **111**:11395-11400.
7. Ponnudurai R, Kleiner M, Sayavedra L, Petersen JM, Moche M, Otto A, Becher D, Takeuchi T, Satoh N, Dubilier N, Schweder T, Markert S: **Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis**. *ISME J* 2017, **11**:463-477.
8. Shi W, Moon CD, Leahy SC, Kang D, Froula J, Kittelmann S, Fan C, Deutsch S, Gagic D, Seedorf H, Kelly WJ, Atua R, Sang C, Soni P, Li D, Pinares-Patiño CS, McEwan JC, Janssen PH, Chen F, Visel A, Wang Z, Attwood GT, Rubin EM: **Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome**. *Genome Res* 2014, **24**:1517-1525.
9. Wilmes P, Heintz-Buschart A, Bond PL: **A decade of metaproteomics: Where we stand and what the future holds**. *Proteomics* 2015, **15**:3409-3417.
10. Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan L-K, Meng J, Durham BP, Shen C, Varaljay VA, Smith CB, Yager PL, Hopkinson BM: **Sizing up metatranscriptomics**. *ISME J* 2013, **7**:237-243.
11. Kuske CR, Hesse CN, Challacombe JF, Cullen D, Herr JR, Mueller RC, Tsang A, Vilgalys R: **Prospects and challenges for fungal metatranscriptomics of complex communities**. *Fungal Ecology* 2015, **14**:133-137.
12. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A: **A survey of best practices for RNA-seq data analysis**. *Genome Biol* 2016, **17**:13.
13. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Brief Bioinform* 2012, **14**:671-683.
14. McIlwain S, Mathews M, Bereman MS, Rubel EW, MacCoss MJ, Noble WS: **Estimating relative abundances of proteins from shotgun proteomics data**. *BMC Bioinf* 2012, **13**:308.
15. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D: **Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data**. *Genome Biol* 2013, **14**:3158.
16. Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL, Washburn MP: **Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors**. *Methods* 2006, **40**:303-311.
17. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M: **Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ**. *Mol Cell Proteomics* 2014, **13**.

18. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, Ngan CY, Cheng J-F, Tringe SG, Woyke T: **Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community**. *BMC Genomics* 2015, **16**:856.
19. Kleiner M, Hooper LV, Duerkop BA: **Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes**. *BMC Genomics* 2015, **16**:7.
20. Prosser JI: **Replicate or lie**. *Environ Microbiol* 2010, **12**:1806-1810.
21. Mueller RS, Denef VJ, Kalnejais LH, Suttle KB, Thomas BC, Wilmes P, Smith RL, Nordstrom DK, McCleskey RB, Shah MB, VerBerkmoes NC, Hettich RL, Banfield JF: **Ecological distribution and population physiology defined by proteomics in a natural microbial community**. *Mol Syst Biol* 2010, **6**:374.
22. Hamann E, Gruber-Vodicka H, Kleiner M, Tegetmeyer HE, Riedel D, Littmann S, Chen J, Milucka J, Viehweger B, Becker KW, Dong X, Stairs CW, Hinrichs K-U, Brown MW, Roger AJ, Strous M: **Environmental Breviatea harbour mutualistic *Arcobacter* epibionts**. *Nature* 2016, **534**:254-258.
23. Justice NB, Pan C, Mueller R, Spaulding SE, Shah V, Sun CL, Yelton AP, Miller CS, Thomas BC, Shah M, VerBerkmoes N, Hettich R, Banfield JF: **Heterotrophic archaea contribute to carbon cycling in low-pH, suboxic biofilm communities**. *Appl Environ Microbiol* 2012, **78**:8321-8330.
24. Nesvizhskii AI, Aebersold R: **Interpretation of shotgun proteomic data: The protein inference problem**. *Mol Cell Proteomics* 2005, **4**:1419-1440.
25. Zauber H, Schüler V, Schulze W: **Systematic evaluation of reference protein normalization in proteomic experiments**. *Frontiers in Plant Science* 2013, **4**:25.
26. Ferguson RE, Carroll HP, Harris A, Maher ER, Selby PJ, Banks RE: **Housekeeping proteins: A preliminary study illustrating some limitations as useful references in protein expression studies**. *Proteomics* 2005, **5**:566-571.
27. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J: **The Perseus computational platform for comprehensive analysis of (prote)omics data**. *Nat Meth* 2016, **13**:731-740.
28. Fernandes AD, Reid JNS, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB: **Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis**. *Microbiome* 2014, **2**:15-15.
29. Weiss SJ, Xu Z, Amir A, Peddada S, Bittinger K, Gonzalez A, Lozupone C, Jesse R Zaneveld, Vazquez-Baeza Y, Birmingham A, Knight R: **Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data**. *PeerJ PrePrints* 2015, **3**:e1157v1151

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2846v1 | CC BY 4.0 Open Access | rec: 2 Mar 2017, publ: 2 Mar 2017