Author's response to reviews

Title: The Biological Observation Matrix (BIOM) Format or: How I Learned To Stop Worrying and Love the Ome-ome

Authors:

Daniel McDonald (daniel.mcdonald@colorado.edu)
Jose C Clemente (jose.clemente@gmail.com)
Justin Kuczynski (justinak@gmail.com)
Jai Ram Rideout (jr378@nau.edu)
Jesse Stombaugh (jesse.stombaugh@colorado.edu)
Doug Wendel (douglas.wendel@colorado.edu)
Andreas Wilke (wilke@mcs.anl.gov)
Susan Huse (shuse@mbl.edu)
John Hufnagle (jhufnagle@mbl.edu)
Folker Meyer (folker@anl.gov)
Rob Knight (rob.knight@colorado.edu)
J Gregory Caporaso (gregcaporaso@gmail.com)

Version: 2 **Date:** 19 April 2012

Author's response to reviews: see over

Dear Dr. Edmunds,

Attached please find our revised manuscript and response to the reviewer's comments. As we discussed, I approached Jonathan Eisen about writing a commentary on our paper, and specifically on the "ome-ome" idea and the proliferation of "omics terms" in the biomedical literature. He mentioned that he is interested in principle, so we have left our discussion of this in the text. If it's OK with you, I'd like to send him this draft of the manuscript. At that point I'll confirm his interest in writing a commentary, and he can use this draft to base his commentary on.

We have pasted the full text of the reviewers' comment below, and provided our responses in red. Thank you again for considering our manuscript for publication in GigaScience.

Sincerely,

J. Gregory Caporaso

Reviewer: Frederick Matsen

Minor essential revision:

I have one real suggestion, which is to separate the description of the format from its implementation in Python. The history of formats that are defined wholly or partially by a reference implementation is not pretty. The description on

http://www.biom-format.org/documentation/biom_format.html

seems complete to me, and I think that link should be given in this manuscript as the specification.

Similarly, I'm a bit surprised by the decision to tie the BIOM format version to the version number of the biom-format software. If the goal is to parsers in lots of languages, won't they at some point be at different versions? I think that versioning the spec is the appropriate thing to do here.

We see the reviewer's point about decoupling the version of the biom-format package and the version of the specification, and will act on that suggestion. We have updated the text describing this point in the paper. If it acceptable to the reviewer and the editor, we would like to synchronize the change to the software and specification with the publication of the paper. At that time, we will create a 1.0.0 version of the BIOM specification, which will be under independent versioning from the `biom-format` project. The reason we would like to wait until the publication of the paper is that we are already planning a 1.0.0 release of the `biom-`

`format` project at that time, and, given that several software development groups either have or are in the process of integrating the BIOM format into their tools, we feel that doing a release now and an additional release in the near future will be disruptive of that process (and could thus hinder the adoption of BIOM).

We thank the reviewer for this very valuable suggestion.

Discretionary revisions:

Abstract:(1) "JSON-derived": I would prefer "JSON-based" rather than "JSON-derived" as the latter makes it sound like you are extending JSON in some way.

Modified as requested.

(2) "bioinformatics bottleneck": I would prefer something more precise like "the file format incompatibility problem".

We have opted to keep this terminology for consistency with the cited references in this section.

Background:(3) 3rd paragraph: It seems to me that you could put in a more compelling example than rarefaction, which is pretty trivial to code up. What about regression?

We like this example because we think the applicability is immediately clear to researchers coming from diverse areas, including areas where rarefaction is less common but nonetheless useful. It is unclear to us that regression is less trivial than rarefaction. We thank the reviewer for the suggestion, but have chosen not to replace the example at this point.

(4) 3rd paragraph, sentence starting "For example, rarefaction...": I suggest a semicolon or a period rather than a colon.

We have reworded this sentence for clarity.

(5) 5th paragraph, "The sparse representation...": is "1% non-zero values" a typo? It sounds like you are storing lots of zeroes.

We have reworded this sentence for clarity.

(6) 5th paragraph: Please specify which version of the GPL is used.

Fixed.

Analyses:(7) 2nd paragraph: "omics data" rather than "omic data"

Fixed.

Discussion:(8) 2nd paragraph: "yet unknown OTUs" rather than "yet unknown observations".

Fixed.

Availability of software:(9) give SF link and please specify which version of the GPL is used.

The link provided here redirects to the main GitHub page (we moved the repository from SF to GitHub to better support a collaborative development environment). The GPL version is now noted.

Box 2:(10) This does seem like a lot of deprecated code to put in a paper, but I suppose that's the point.

Yes, this is the point we are trying to make. We are happy to move this to supplementary material, but leave this decision to the editors.

Reviewer: Josh D Neufeld

Major compulsory revisions
From the title onward, the manuscript is distracted by the concept of the "ome-ome", which is an observation that the (over)use of "omics" terms (e.g. genomics, metagenomics, etc) has increased in the literature since ~1990. There is no need to have this coined terminology be a focus of the manuscript. The authors could simply comment in the text that, as everyone knows, omics-based approaches have become commonplace in biological research. Figure 1 is completely unnecessary, the title (although catchy) is silly and distracting, and too much of the methods/results is wasted on this ome-ome notion. The title should be modified to something like "The Biological Observation Matrix (BIOM) format for unifying omics analysis in the biological sciences". Figure 1 should be deleted along with the associated results, methods, discussion and acknowledgements.

We thank the reviewer for this comment. After discussion of this point with the editor we have decided to reduce the discussion of this point, in favor of highlighting some of the discussion in a commentary on the topic.

Instead of the "ome-ome" discussion, Figure 1 would be more valuable as a visual representation of the biom file format so the reader has a sense of what the file would look like. Showing a file format example or abstraction modified from the following website would be ideal: http://www.biom-format.org/documentation/biom_format.html#example-biom-files

We have added a new Box (Box 3) that contains an example BIOM file derived from the page the reviewer references. We additionally point the reader to more examples in the caption for this box.

Throughout the manuscript, 16S rRNA gene analysis in QIIME received sufficient attention but a concern was that MG-RAST and VAMPS (and alternative omics platforms) are not given comparable attention throughout. For example, benefits to transcriptomics, genomics or proteomics research contexts could be specifically mentioned and the cross-platform applicability could be better explained. The suggested new Figure 1 legend could be clear about general format differences between these major omic approaches.

We have expanded our discussion of the benefits of this format in the third paragraph of the *Background* section to mention examples related to proteomics and transcriptomics.

We have included a note on format differences between omics techniques in the new Box 3 legend. The key point here is that there are no differences: there is a required `type` value associated with BIOM tables, but this is included so that tools that use the BIOM format can restrict which data types they accept, which is sometimes necessary in order to provide valid results.

As indicated by the authors, a major computational benefit offered by the BIOM format is the "compression" of sparse tables. This argument should be contrasted with the computational improvements gained by processing compressed files (.bzip/.gzip) directly because this is a common work-around in processing large text-based data files.

We have added a paragraph to the *Analyses* section to compare the BIOM format versus to compressed (e.g., gzipped) tab-separated text. We thank the reviewer for highlighting this point, which we inadvertently omitted from the discussion.

Finally, the authors' terminology is imprecise when referring to aspects of the biom format. For example, the terms 'software package', 'project' and 'standard' are used interchangeably. Instances where the authors are referring specifically to the implementation requirements should use "biom-format standard" or "biom standard". Broader references, including software utilities, web page and other infrastructure, should use "biom-format project" or an analogous term.

We have checked all mentions of these terms in the text and clarified our terminology.

Minor essential revisions (use "find" for location in manuscript due to lacking page/line #s):
The use of emphasis (italics, single ['] and double ["]) is inconsistent throughout. Suggest using ["].
Fixed.

Should read "as done in [16]; a researcher...".

Fixed.

Should read “In many existing software packages [e.g. 14, 15], contingency tables...”
Fixed.

Replace “won’t” with “will not”.
Fixed.

Replace semicolons in the first sentence of the Discussion with commas.
Fixed.

Replace “...causing an explosion in...” with anything else.
Fixed.

Delete “etc.”
Fixed.

None of the references are correctly formatted (e.g. journal abbreviations, issue numbers).
We have reformatted the references with the EndNote style provided by the journal.

Figure 2: x-axis should end at 100.1 MB and y-axis should be correspondingly adjusted.
Fixed. This was adjusted and now has a maximum value of 200 MB on both axes (the reviewer’s suggestion of 100.1 MB cut off several points on the y-axis).

Supplementary figure of compression ratio: italicize “R” and remove shadow from symbols.
Fixed.

Reviewer: Sarah Hunter
1) Major Compulsory Revisions
1.1) A brief explanation of the decision to base BIOM on JSON should be included, detailing the benefits/disadvantages that this brings vs other file formats.
Added. We have expanded our discussion of this in *BIOM file format* section.

1.2) An overview of the format, as specified at http://biom-format.org/documentation/biom_format.html would assist the reader in understanding the text better (e.g. the compression efficiencies of the sparse vs dense formats). At the very least, this format description should be linked from the text.
Added. We have added a link to the file format in the *BIOM file format* section and added *Box 3* to illustrate the file format in the text.

1.3) Analyses, paragraph 1 - re-write for clarity required

Authors state that the discrepancy in file sizes arise from "the matrix positions that must be stored with all counts in the sparse representation". It's not currently possible to understand why this is the case with the information currently provided in the text - see my previous point.
Done. This description has been expanded, and references the new *Box 3*.

1.4) Discussion, paragraphs 2 and 3
Highly repetitive when compared to the background/introduction section. These two paragraphs could also be merged together, with much more specific discussion of why BIOM in particular is a necessary step. For example, there is no mention of the challenges faced by any format if it is to be adopted by the wider research community (and how the authors propose to meet these challenges).
We thank the reviewer for this suggestion. We have expanded our discussion to cover challenges that will be associated with the adoption of the BIOM standard, as well as maintaining a community software development environment. At this point we have kept the last two paragraphs separate as we think this reads better then several variants that we experimented with.

2) Minor Essential Revisions
2.1) Background, paragraph 2
I question the usefulness of Fig 1 and the Medline mining to illustrate the increasing numbers of categories of omics data. I'm not convinced that this is truly representative (isn't it reflecting, instead, scientists' penchant for leaping on a bandwagon regarding names? e.g. I wouldn't call the "kinome" or the "O-GlcNAc-ome" entirely new datatypes - they're a subset of the proteome, surely?). I would be satisfied with the authors simply asserting instead that there are increasing numbers of omic approaches to analysis, illustrated with some examples of newer omics data types. The authors could then remove reference to the MEDLINE mining.
We thank the reviewer for this comment. After discussion of this point with the editor we have decided to reduce the discussion of this point, in favor of highlighting some of the discussion in a commentary on the topic.

2.2) Background, paragraph 3 - requires edit for clarity
A brief description of what a contingency table is would be beneficial, prior to the different omics examples. A potentially more readable sentence would start "Despite the different types of data involved in the various comparative omics techniques (e.g. metabolomics, proteomics or microarray-based transcriptome analyses), they all share an underlying, core data type: the sample by observation contingency table. A contingency table is... [brief explanation followed by omics examples, as in the text].
Changed.

2.3) Background, paragraph 2 - edit for clarity
Suggestion: "A common data format will facilitate the sharing and publication of comparative omics data and associated metadata, as well as improving the

interoperability of comparative omics software. It will enable rapid advances in omics fields by allowing researchers to focus on data analysis instead of formatting data for transfer between different software packages or reimplementing existing analysis workflows to support their specific data types."
Changed.

2.4) Background, paragraph 3 - edit for clarity and brevity
Suggestion: "However, many techniques are applicable aross data types, for example rarefaction analyses (i.e. collector curves). These are frequently applied in microbiome studies to compare how the rate of incorporation of additional sequence observations affects the rate at which new OTUs are observed. This is done to determine whether an environment is approaching the point of being fully sampled (e.g. [14]). Similarly, they can also be applied in comparative genomics [...]" (etc). This whole paragraph could be more concisely written.
The suggested change has been incorporated, and the discussion of metadata has been moved to a new paragraph.

2.5) Background, paragraph 3 - requires clarification
In the sentence "A standard format [...] will support interoperability of these tools and facilitate development and adoption of future analysis pipelines..." it's not immediately clear what tools are being referred to and how exactly it will facilitate pipeline development.
We have clarified this section.

2.6) Background, paragraphs 2 and 3 - re-write required to remove redundancy and improve readability
A few sentences in particular ("A common data format to facilitate sharing and publication of comparative omics data and associated metadata", "The inclusion of high-quality metadata in this format, for example as defined in the MIxS standards [13], is essential for enabling future meta-analyses." and "Additionally, the incorporation of sample and observation metadata allows convenient sharing and archiving of these data within a single file.") are saying related things about metadata - would make sense to try to condense together into a single location in the text.
We have added a new paragraph to Background to discuss metadata issues specifically.

2.7) Background, paragraph 4 - requires clarification
Don't the authors mean "For example, differing representations of samples and observations as either rows or columns, and the mechanism for incorporating sample or observation metadata (if this is possible at all), cause the formats used by different software packages to be incompatible." i.e. the decision itself has nothing to with compatability....
Fixed.

2.8) Background, paragraph 5 - requires clarification

In this sentence: "The sparse representation of the QIIME OTU table with 6164 samples and 7082 OTUs (mentioned in the previous paragraph) contains 1% non-zero values in BIOM format and is over 14x smaller than the same data represented in tab-separated text (Supplementary File 1)." is confusing - surely both files contain 1% non-zero values?
Fixed.

2.9) Background, paragraph 5 - minor correction for readability
Suggestion: "This includes a format validator, a script to easily convert BIOM files to tab-separated text representations (useful when working with spreadsheet programs), and Python objects to support working with this data."
Fixed.

2.10) Box 2 legend - minor edit for clarity
Suggestion: "Comparison of QIIME OTU Table collapsing code with native QIIME OTU table data structures (Panels A-D) and biom-format Table objects with equivalent functionality. [...]"
Fixed.

2.11) Analyses section - re-write required for clarity
I would re-order aspects of paragraphs 1 and 2 to make it more readable. For example, describe the initial data set in the first paragraph (size of OTU tables, density range and median, file compression ratios). In the second paragraph, explain the patterns seen (e.g. explain/describe discrepancies in filesize and when each of the formats is most efficient for compression, incurred overheads with dense vs sparse representations, etc.). At the moment it's a bit of a jumble.
We have reorganized these paragraphs according to the reviewers' suggestion.

2.12) Analyses, paragraph 2 - minor edit for readability
Suggestion: "In the data set we analysed, the density ranges from 1.3% non-zero values to 49.8% non-zero values, with a median of 11.1%. The file compression ratio (tab-separated text file size divided by BIOM file size) increases with decreasing contingency table density for this data set (compression ratio = 0.2 × density-0.8; R2 = 0.9; Supplementary Figure 1)."
We have partially incorporated this change. We left our reference to Figure 2 as it is more specific.

2.13) Discussion, paragraph 1 - minor edit Suggestion: "[...]versions of Linux), and so they should be [...]"
Changed as requested.

2.14) Availability of software - minor edit Suggestion: "It is available under GPL, and is free for all to use"
Changed as requested.

3) Discretionary Revisions

3.1) General: readability - try to keep sentences shorter.

3.2) General: repetitive in parts - "Collectively the ome-ome" crops up multiple times, and "useful for interacting with BIOM data in spreadsheet programs" effectively twice.
Fixed.

3.3) Background, paragraph 4 - edit for clarity
Defining "density" here, rather than later on in the "Analyses" section, makes more sense. Suggested edit: "Additionally, in many of these applications a majority of the values (frequently greater than 90%) in the contingency table are zero. The fraction of the table that have non-zero values is defined as the "density"; thus, a matrix with a low number of non-zero values is said to have a low density."
Done.

3.4) Background, paragraph 4 - minor correction
Suggestion: "[...]marker gene survey OTU tables with many samples (such as the one presented in Supplementary Table 1"
Done.

3.5) Background, paragraph 4 - a semantic query..
In the sentence "[...] meaning that many of the values in the matrix [...] are zero". Is it accurate to refer to these values as "zero" rather than null (i.e. no value/not observed) even if the figure "zero" is used in the file..?
We have clarified that we treat zeros as representing an observation that was not observed in the corresponding sample. We thank the reviewer for pointing out this potential source of confusion.

3.6) Suppl data, Box 1 - minor correction for readbility
Suggestion: "Information on the data type (e.g., OTU Table, Ortholog Table, Metabolite Table) should be included, based on terms from a controlled vocabulary."
Fixed.

3.7) Background, paragraph 5 - minor correction, missing comma
Suggestion: "[...] and metadata in a single, standard file format, BIOM supports [...]"
Fixed.

3.8) Background, paragraph 5 - minor correction
Start a new paragraph just prior to the sentence beginning "To support the use of this file format..."
Fixed.

3.9) Data description, paragraph 1

It might be worth referring directly to an example Qiime formatted file in the supplementary material (see also "Major Compulsory Revisions").
Fixed.

3.10) Analyses, paragraph 2 - minor edit
If density is defined earlier, suggest sentence changed to :"The magnitude of compression [...] is a function of the density of the continency table".
Fixed.

3.11) Discussion, paragraph 1 - minor edit for readability
Suggestion: "[...] and to provide an efficient means for representing biological contingency tables in memory with associated convenient functionality for operating on those tables."
We have opted to leave this sentence as-is.

3.12) Discussion, paragraph 1 - minor edit for clarity
Suggestion: "The core BIOM development group will review these implementations and, if they are fully documented and tested, will add them to the biom-format repository (or grant the developers themselves direct access to the repository)."
Fixed.