(GIGA)n SCIENCE

TECHNICAL NOTE Open Access

EMPeror: a tool for visualizing high-throughput microbial community data

Yoshiki Vázquez-Baeza[1], Meg Pirrung[2], Antonio Gonzalez[3] and Rob Knight[3,4,5*]

Abstract

Background: As microbial ecologists take advantage of high-throughput sequencing technologies to describe microbial communities across ever-increasing numbers of samples, new analysis tools are required to relate the distribution of microbes among larger numbers of communities, and to use increasingly rich and standards-compliant metadata to understand the biological factors driving these relationships. In particular, the Earth Microbiome Project drives these needs by profiling the genomic content of tens of thousands of samples across multiple environment types.

Findings: Features of EMPeror include: ability to visualize gradients and categorical data, visualize different principal coordinates axes, present the data in the form of parallel coordinates, show taxa as well as environmental samples, dynamically adjust the size and transparency of the spheres representing the communities on a per-category basis, dynamically scale the axes according to the fraction of variance each explains, show, hide or recolor points according to arbitrary metadata including that compliant with the MIxS family of standards developed by the Genomic Standards Consortium, display jackknifed-resampled data to assess statistical confidence in clustering, perform coordinate comparisons (useful for procrustes analysis plots), and greatly reduce loading times and overall memory footprint compared with existing approaches. Additionally, ease of sharing, given EMPeror's small output file size, enables agile collaboration by allowing users to embed these visualizations via emails or web pages without the need for extra plugins.

Conclusions: Here we present EMPeror, an open source and web browser enabled tool with a versatile command line interface that allows researchers to perform rapid exploratory investigations of 3D visualizations of microbial community data, such as the widely used principal coordinates plots. EMPeror includes a rich set of controllers to modify features as a function of the metadata. By being specifically tailored to the requirements of microbial ecologists, EMPeror thus increases the speed with which insight can be gained from large microbiome datasets.

Keywords: Microbial ecology, QIIME, Data visualization

Findings

Background

Rapid increases in sequencing capacity are greatly expanding our ability to understand the microbial world: scaling from a handful of samples to hundreds, or thousands, allows a rich picture of trends over temporal and spatial scales that were previously unattainable. Human microbiome studies are not the only beneficiaries of this ability to perform increased sampling: large-scale patterns are now being discovered in communities ranging from soils [1] to oceans [2] including the efforts from the International Census of Marine Microbes (ICoMM). We can now process thousands of samples in a single sequencing run [3], and in turn computational tools must also scale to fulfill these needs [4].

Although data visualization is an empowering tool that allows an efficient understanding of information [5], it remains a major challenge in this area of study, specifically because with more samples comes richer information relating the samples to one another (this contextual information is often referred to as "sequence metadata") and to the study design itself. When analyzing large numbers of samples, researchers need to know the patterns that link specific samples or microbes to overall patterns of diversity, and to different metadata variables: this is typically critical for usable visualizations. A well

* Correspondence: rob@spot.colorado.edu
[3]BioFrontiers Institute, University of Colorado at Boulder, Boulder, CO 80309, USA
[4]Department of Chemistry & Biochemistry, University of Colorado at Boulder, 80309 Boulder, CO, USA
Full list of author information is available at the end of the article

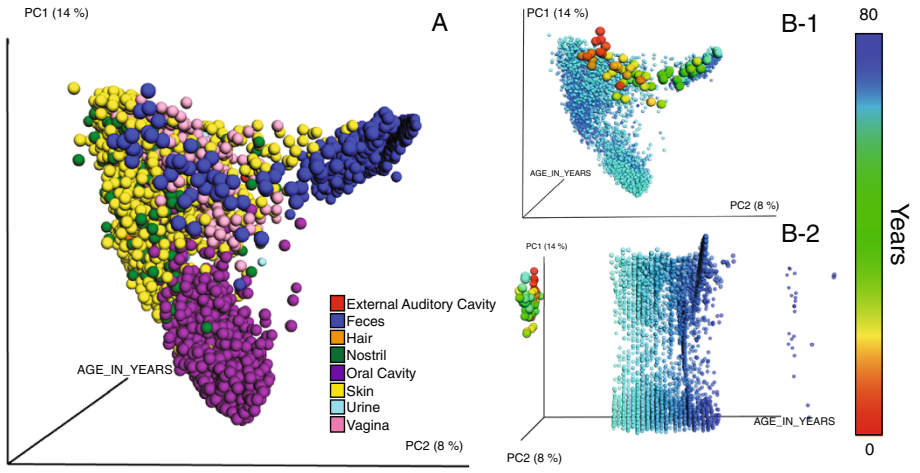

Figure 1 EMPeror display showing the combination of the datasets described in [3,20-22], consisting of 5740 samples representing human auditory canal, skin, nostril, feces, vagina, urine, hair and oral body habitats. (A) Data colored by body habitat; **(B-1)** Principal coordinate 1 (PC1) vs. PC2 with the same the data colored according to the age of the subjects (a continuous variable). **(B-2)** PC1 vs. an explicit time axis. The results allow us to see by immediate visual inspection that the body habitats are remarkably different between each other and that this is consistent through time as a human reaches adulthood.

know ecological metric to quickly compare the microbial composition of the samples is beta diversity, which collates them by creating a distance matrix of these differences. Ordination methods, such as Principal Coordinates Analysis (PCoA) [6] are useful for dimensionality reduction and widely used in different fields to conceptualize distance matrices, however determining how to visualize the samples to reveal clear patterns often remains a challenge. Figure 1A shows the samples colored by the body site each belong to, a common approach that will make evident the main differences explained in the first two axes of variation; however, when integrating metadata in the coloring patterns (Figure 1B-1, B-2), the plot clearly shows the age differences between the samples of an infant, compared to the samples belonging to healthy human adults.

There are several existing methods for displaying PCoA results, but none to date are specifically designed to account for the common use cases in this research field; furthermore, each of the most representative solutions allots different limitations. For example, QIIME [7], an open source framework for upstream and downstream analysis of microbial community samples generated via high-throughput sequencing instruments, typically generates 3D plots using KiNG [8] originally designed as a molecular graphics viewer, which requires static files containing each metadata field to be produced in advance, replicating the coordinates for each of these categories and resulting in long load times and large file sizes when the metadata are rich. SpotFire [9] is a very expensive commercial solution, beyond the budget of many research laboratories. Generic packages that provide 3D plotting functionalities such as MATLAB [10], Mathematica [11], R [12], Excel [13] or Matplotlib [14] can always be used, but custom code or manual approaches are typically required to relate each point to a specific visual feature intended to highlight a given variable. Consequently, this could become a time-consuming process, which as a side effect compromises its reliability, reusability and reproducibility. Moreover, none of the previously mentioned applications are specifically modeled to support the workflows of the modern microbial ecologist. Allowing the user to choose among metadata coloring dynamically, and separating coloring from visibility, has a surprisingly large

Table 1 Studies used to create Figure 1

Title	General description	Collected samples	Reference
Moving pictures of the human microbiome	Samples from two subjects are collected for up to 15 months in three body sites (oral, skin and gut)	1964	[3]
Bacterial community variation in human body habitats across space and time	Samples from healthy adult human samples from eight subjects of up to 27 body sites	585	[20]
Structure, function and diversity of the healthy human microbiome	Samples from 242 healthy adult human samples from up to eighteen different body sites	3131	[21]
Succession of microbial consortia in the developing infant gut microbiome	Gut samples collected biweekly from an infant through the first 2.5 years of life	60	[22]

effect in encouraging interactive exploration, understanding and analysis, and often allows insights into the main factors, as well as more subtle ones, structuring the data to be obtained much more rapidly.

EMPeror

EMPeror is a thoroughly tested and open-source software package with an interactive user interface and hardware-accelerated graphics, implemented with HTML5, WebGL, Javascript and Python, and tightly integrated with QIIME [7] and PyCogent [15]. EMPeror's command line interface accepts QIIME principal coordinates files and metadata mapping files, and produces an interactive 3D visualization that can be delivered in the context of a web page independent of the command line tool. As an example of EMPeror's ability to deal with continuous variables (time, alpha diversity, pH) that are part of the metadata, these factors can be integrated as an explicit axis in the plot, lines connecting subsequent points of single trajectories (treatments, subjects, sites, etc.) or using a colormap to have each sample's color be a function of its position in the gradient. The main features that EMPeror provides are: (1) easily change visibility features of data points in the plots based on metadata; (2) can be easily embedded into other tools, such as Evident [16] as a reusable visualization component; (3) scale to thousands of points with minimal load times (seconds versus many minutes in KiNG); and (4) ability to display auxiliary data to increase the understanding of the intrinsic data patterns; these include: biplots [17], procrustes analysis [18], and jackknifed beta diversity plots [19].

To illustrate the effectiveness of EMPeror, we show the combination of [3,20-22], see Table 1, as generated with the QIIME web application [23]. This combination represents 5,740 samples (spheres), and 120 columns of metadata [24]. In KiNG, the resulting files for both the discrete and gradient coloring result in a size of 1.85 GB, but in EMPeror only 26 MB [25], meaning only 1.3% of the original size, see Figure 1. Additionally, we can easily view the intrinsic age patterns within the data, Figure 1B, both panels.

EMPeror installation instructions can be found in the online documentation (http://qiime.org/emperor/installation_index.html).

Conclusions

EMPeror provides a user-friendly interface and set of tools for visualizing large numbers of microbial community samples associated with increasingly extensive metadata, and interactively manipulating these datasets to add auxiliary data and visualization techniques. Additionally, it contains several user interface features, enabling straightforward modifications and customization of perceptible aspects in the plot plus the incorporation of statistical techniques, which also help increase the ease and speed of exploratory analysis. We believe that EMPeror will have a large impact on the field, especially for large-scale environmental sampling projects, such as the Earth Microbiome Project [26], and large-scale clinical projects, such as the Human Microbiome Project [20].

Availability and requirements

Project name: Emperor
Project home page: http://emperor.colorado.edu
Operating system(s): Platform independent for the graphical user interface; OS X (10.6 and higher) and Linux only for the command line interface.
Programming language: Python and JavaScript.
Other Requirements: Python 2.7, Chrome, QIIME (python libraries only), NumPy, BIOM 1.1.0 and PyCogent.
License: Modified BSD.
Any restrictions to use by non-academics: None.

Availability of supporting data

The example files and additional data sets supporting the results of this article are available in the GigaScience Database [24], as well as from the EMPeror FTP site [25].

Abbreviations

EMP: Earth Microbiome Project; HTML5: HyperText Markup Language, version 5; ICoMM: International Census of Marine Microbes; MIxS: Minimum information about any (x) sequence; NumPy: Numerical Python; PCoA: Principal Coordinates Analysis; PyCogent: Comparative and Genomic Toolkit; QIIME: Quantitative Insights into Microbial Ecology; WebGL: Web Graphics Library.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YVB, MP and AG developed parts of the visualization and backend frameworks for EMPeror. YVB, AG and RK wrote the manuscript. AG, and RK established the initial design and goals of the project. All authors read and approved the final manuscript.

Authors' information

YVB, AG, MP and RK are developers and or leaders of the QIIME project.

Acknowledgements

We thank Jackson Chen, Jai Ram Rideout, Daniel McDonald, William Van Treuren, Jose Antonio Navas-Molina, Nicholas A. Bokulich, Adam Robbins-Pianka and Greg Caporaso for feedback and useful discussion regarding the design and implementation of the software package.
This work was supported in part by the National Institutes of Health, the Crohn's and Colitis Foundation of America, the Alfred P. Sloan Foundation, and the Howard Hughes Medical Institute.

Author details

[1]Department of Computer Science, University of Colorado at Boulder, Boulder, CO 80309, USA. [2]Department of Pharmacology, University of Colorado Denver, Aurora, CO 80045, USA. [3]BioFrontiers Institute, University of Colorado at Boulder, Boulder, CO 80309, USA. [4]Department of Chemistry & Biochemistry, University of Colorado at Boulder, 80309 Boulder, CO, USA. [5]Howard Hughes Medical Institute, 80309 Boulder, CO, USA.

Received: 17 October 2013 Accepted: 20 November 2013
Published: 26 November 2013

References

1. Lauber CL, Hamady M, Knight R, Fierer N: **Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale.** *Appl Environ Microbiol* 2009, **75:**5111–5120.
2. Harris R: **The L4 time-series: the first 20 years.** *J Plankton Res* 2010, **32:**577–583.
3. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, *et al*: **Moving pictures of the human microbiome.** *Genome Biol* 2011, **12:**R50.
4. Gonzalez A, Knight R: **Advancing analytical algorithms and pipelines for billions of microbial sequences.** *Curr Opin Biotechnol* 2012, **23:**64–71.
5. O'Donoghue SI, Gavin AC, Gehlenborg N, Goodsell DS, Heriche JK, Nielsen CB, North C, Olson AJ, Procter JB, Shattuck DW, *et al*: **Visualizing biological data-now and in the future.** *Nat Methods* 2010, **7:**S2–4.
6. Gower JC, Legendre P: **Metric and euclidean properties of dissimilarity coefficients.** *J Classif* 1986, **3:**5–48.
7. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, *et al*: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods* 2010, **7:**335–336.
8. Chen VB, Davis IW, Richardson DC: **KING (kinemage, next generation): a versatile interactive molecular and scientific visualization program.** *Protein Sci* 2009, **18:**2403–2409.
9. TIBCO-Software: **Spotfire.** In *Book Spotfire.* Sommerville, Massachusets: TIBCO Software; 2013.
10. The-MathWorks-Inc: **MATLAB: the Language of Technical Computing.** In *Book MATLAB: The Language of Technical Computing.* Natick, Massachusets: The MathWorks Inc; 2013.
11. Wolfram-Research: **Mathematica, Version 8.0.** In *Book Mathematica, Version 8.0.* Champaign, Illinois: Wolfram Research, Inc; 2010.
12. R-Core-Team: **R: A language and environment for statistical computing.** In *Book R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2013.
13. Microsoft: **Microsoft Excel.** In *Book Microsoft Excel.* Redmond, Washington: Microsoft; 2011.
14. Hunter JD: **Matplotlib: a 2D graphics environment.** *Comput Sci Eng* 2007, **9:**90–95.
15. Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z, *et al*: **PyCogent: a toolkit for making sense from sequence.** *Genome Biol* 2007, **8:**R171.
16. **Evident: elucidating sampling effort for microbial analysis studies.** [http://github.com/qiime/evident]
17. Hewitt KM, Mannino FL, Gonzalez A, Chase JH, Caporaso JG, Knight R, Kelley ST: **Bacterial diversity in two neonatal intensive care units (NICUs).** *PLoS One* 2013, **8:**e54703.
18. Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, Fontana L, Henrissat B, Knight R, Gordon JI: **Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans.** *Science* 2011, **332:**970–974.
19. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R: **Using QIIME to analyze 16S rRNA gene sequences from microbial communities.** *Curr Protoc Microbiol* 2012, **Chapter 1:**Unit 1E 5.
20. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R: **Bacterial community variation in human body habitats across space and time.** *Science* 2009, **326:**1694–1697.
21. HMP-Consortium: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, **486:**207–214.
22. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE: **Succession of microbial consortia in the developing infant gut microbiome.** *Proc Natl Acad Sci USA* 2011, **108**(Suppl 1):4578–4585.
23. **QIIME web application.** [http://www.microbio.me/qiime/]
24. Vázquez-Baeza YP M, Gonzalez A, Knight R: **Example files and supporting material for "EMPeror: an interactive analysis and visualization tool for high throughput microbial ecology datasets".** *GigaScience Database* 2013. http://dx.doi.org/10.5524/100068.
25. **EMPeror ftp page.** [ftp://thebeast.colorado.edu/pub/emperor_files/]
26. Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Desai N, Eisen JA, Evers D, Field D, Feng W, *et al*: **Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project.** *Stand Genomic Sci* 2010, **3:**243–248.

doi:10.1186/2047-217X-2-16
Cite this article as: Vázquez-Baeza *et al.*: EMPeror: a tool for visualizing high-throughput microbial community data. *GigaScience* 2013 **2**:16.