HUMAN MICROBIOME VISUALIZATION USING 3D TECHNOLOGY*

JASON H. MOORE
Institute for Quantitative Biomedical Sciences, Departments of Genetics and Community and Family Medicine, Dartmouth Medical School, Lebanon, NH 03756
Email: jason.h.moore@dartmouth.edu

RICHARD COWPER SAL.LARI
Institute for Quantitative Biomedical Sciences, Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756
Email: richard.cowper.sal.lari@Dartmouth.edu

DOUGLAS HILL
Institute for Quantitative Biomedical Sciences, Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756
Email: douglas.hill@Dartmouth.edu

PATRICIA L. HIBBERD
Department of Pediatrics, Division of Global Health, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114
Email: patricia.hibberd@gmail.com

JULIETTE C. MADAN
Department of Pediatrics, Division of Neonatology, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756
Email: juliette.c.madan@hitchcock.org

High-throughput sequencing technology has opened the door to the study of the human microbiome and its relationship with health and disease. This is both an opportunity and a significant biocomputing challenge. We present here a 3D visualization methodology and freely-available software package for facilitating the exploration and analysis of high-dimensional human microbiome data. Our visualization approach harnesses the power of commercial video game development engines to provide an interactive medium in the form of a 3D heat map for exploration of microbial species and their relative abundance in different patients. The advantage of this approach is that the third dimension provides additional layers of information that cannot be visualized using a traditional 2D heat map. We demonstrate the usefulness of this visualization approach using microbiome data collected from a sample of premature babies with and without sepsis.

1. Introduction

1.1. *The Human Microbiome*

The primary goal of the human microbiome project is to understand the role that symbiotic microorganisms play in determining health and disease [1,2]. This is a staged effort that includes

* This work is supported by the Hitchcock Foundation Joshua B. Burnett M.D. Research Career Development Award, the Hearst Foundation, the Dartmouth Center for Clinical and Translational Science, the Norris-Cotton Cancer Center, the Institute for Quantitative Biomedical Sciences and NIH grants LM009012 and RR018787.

1) construction of draft assemblies of reference genomes, 2) creation of reference microbiome data sets, 3) determination of the full human microbiome and 4) determination of the global diversity of the human microbiome. Significant progress toward these goals has been made. For example, Wu et al. [3] carried out a phylogenetic analysis of 56 microbes demonstrating the need for a comprehensive encyclopedia of microbial genomes. A recent study reports the results of the initial sequencing of 178 microbial genomes that will help provide a reference for human microbiome studies [4]. Costello et al. [5] assayed the spatial and temporal variation of the human microbiome from up to 27 different sites in seven to nine subjects. This study demonstrated that bacterial flora varied significantly across body sites and time. The rapid advances in the development of high-throughput sequencing technologies will make it feasible to accomplish many of these goals over the next few years.

1.2. *The Fecal Microbiome of Preterm Infants*

The ultimate goal of these baseline genomic studies is to provide the framework for relating microbial diversity and composition to clinical endpoints. One important application of this technology is to determine whether the human microbiome will be useful for predicting outcomes in infants born prematurely. Colonization of the neonatal intestine happens rapidly after birth, is dependent on delivery method [6], but may be delayed in infants born prematurely. Further, premature infants are more likely to be colonized by pathogenic bacteria [7-9]. Our working hypothesis is that the fecal microbiome of preterm infants will be useful for predicting their clinical course and might provide potential time points for intervention to ameliorate disease risk. Previous work in this area has focused, for example, on neonatal necrotizing enterocolitis (NEC) in preterm infants. This is an inflammatory disorder that may lead to death and has an incidence of one to three per 1000 live births. A study by Wang et al. [10] sequenced 16S rRNA from the fecal samples of 20 preterm infants and found that those with NEC had less diversity and a higher abundance of *Gammaproteobacteria*. Although not conclusive, this study provides a baseline for beginning to think about how the microbiome might influence susceptibility to NEC and other clinical endpoints in preterm infants such as sepsis.

1.3. *A Role for Visualization in Human Microbiome Studies*

The biocomputing challenges of microbiome analysis are both diverse and numerous. This is partly due to the volume of sequence data that is generated and the hierarchical complexity of the microbial data itself. Examples of prior biocomputing work in this area include the development of algorithms for identifying human gut-specific protein families [11], reference genome databases [4], computational inference of function using genomic context [12] and efficient taxonomic profiling [13]. These studies and others are providing the computational methodologies that will be necessary to accurately and efficiently analyze human microbiome data.

Despite these advances, there are still many biocomputing needs. For example, a typical data set might include a list of hundreds of bacterial species that are hierarchically organized into different groups, including genus, families, orders classes and phyla. One goal is to relate abundance of bacteria at these different taxonomic levels to clinical endpoints. This is further

complicated by information about genes and pathways that are present in each of the bacterial species and how these relate to various clinical endpoints. The genomic information of the host can also be added to the analysis. The ultimate challenge is to put these many different layers of information together in a statistical or machine learning analysis to identify the clinically useful patterns. Although not yet routine, this type of biocomputing analysis will be in high demand in the near future.

The working hypothesis of the present study is that the inherent hierarchical complexity of human microbiome data, and the need to relate these many layers of information to clinical endpoints, will necessitate the development of intuitive user interfaces for visual exploration and analysis. In other words, the Excel-based spreadsheet paradigm will not provide the level of human-computer interaction that is necessary to both understand a complex data set and inform data mining and machine learning analyses. The goals of the present study were to develop a methodology for visualizing multiple dimensions of human microbiome information using 3D technology that is both intuitive and interactive. We present here a three-dimensional (3D) heat map methodology and software that builds on the familiarity and success of the conventional two-dimensional (2D) heat map and the power of commercial video game development engines and 3D technology. The ultimate goal of these studies is to provide comprehensive visual analytics methodology and software for facilitating human microbiome analysis.

2. Methods

2.1. *A 3D Heat Map*

Heat maps have become a popular and useful method for visualizing high-dimensional data (http://en.wikipedia.org/wiki/Heat_map) and were introduced more than fifty years ago by Sneath [14] for biological problems. Eisen et al. [15] popularized the heat map for visualizing the results of clustering genomics data. A heat map consists of a 2D grid or matrix of colored squares where each square represents an observation of a random variable and the color of the square is proportional to the value of that observation. It is common to order the squares by additional categorical data such as tissue of origin and gene on the two axes. Our working hypothesis is that adding an additional dimension (z-axis) to the traditional 2D heat map will provide the opportunity to visualize additional layers of information that will enhance the visual discovery process. To test this hypothesis we developed a 3D heat map methodology and software package using a commercial video game engine. We apply it here to human microbiome visualization.

There are many reasonable platforms for developing 3D visualization software. OpenGL (http://www.opengl.org) with a C++, Java or scripting front end and a user-interface toolkit has all the necessary elements. A virtual reality modeling language, such as X3D (http://www.web3d.org/x3d), with scripting capabilities and free viewers, could also be used. The "Processing" visual programming environment (http://processing.org) provides a rapid prototyping environment with the ability to use Java libraries. Each choice has advantages and disadvantages. We chose here a video game development environment because game engines are explicitly designed around interactivity and immersion in a 3D environment. The ability to interactively explore a heat map visualization as you would a video game environment was an

important feature. We chose the Unity3D (http://unity3d.com) development tool because it uses Mono, the open-source, cross-platform .NET implementation, so we would not be limited to code libraries supplied by the vendor. For a reasonable licensing fee we could distribute royalty-free tools that run on Windows and Macintosh machines. Unity makes GUI code easy to write, enabling rapid prototyping, and the work-flow for incorporating assets from other tools such as Maya and Photoshop is straightforward. An additional advantage is that Unity can use Direct3D on Windows machines, which allows users to employ off-the-shelf drivers to see 3D heat maps in stereo on suitable equipment. Using OpenGL we would have to explicitly code the view from each eye to produce stereo. The ability to easily see 3D heat maps in stereo is important given the widespread availability of 3D televisions and computer monitors.

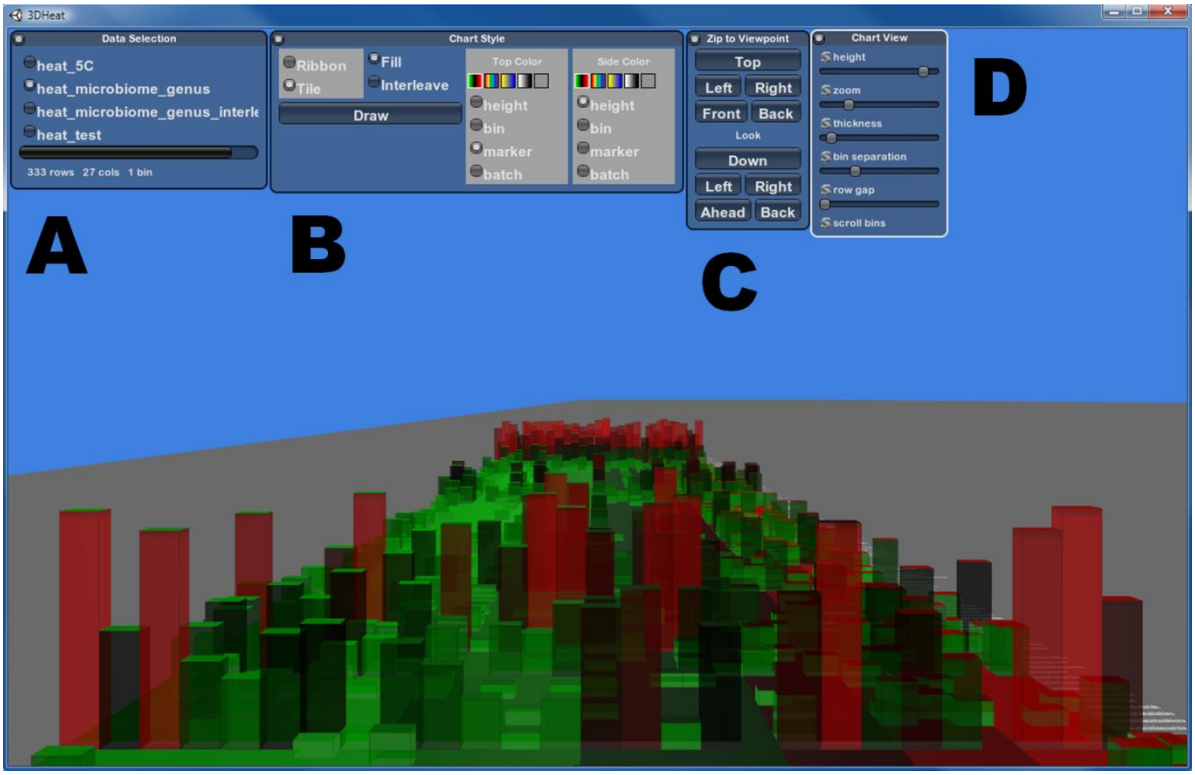

Fig. 1. Screenshot of the 3D heat map application showing menus for data selection (A), chart style (B) viewpoint (C) and chart sizing options (D). Each menu can be minimized or hidden.

A potential disadvantage of the Unity framework is that it makes low-level control of the Graphics Processing Unit (GPU) more difficult. If our primary objective had been to render massive amounts of data, we would have chosen a toolkit that allowed finer control over what is stored on the GPU, to minimize transfers between the CPU and GPU. Our principle goal though is insight through exploration and interaction. Unity allows us to render 120,000 data points (1,440,000 triangles) at 47 frames per second on a Mac Pro with two 2.8 GHz Quad-Core Intel Xeon processors and an ATI Radeon HD 2600 graphics card holding 256 MB of video memory. This scales reasonably well to about 500,000 data points corresponding to a dataset with 10,000

rows and 50 columns. This allows smooth motion and very good responsiveness in exploring datasets of moderate size.

Figure 1 provides a screenshot of the 3D heat map application with the various menus that control what data is being viewed, the style of the heat map, the viewpoint and features of the heat map itself. The menus can be minimized or hidden to make full use of the screen. The chart style menu provides options to view the heat map as ribbons where each data point is connected as in a time series or in the traditional tile or square view. There is an additional option (shown) to fill the tiles or ribbons to make solid objects. This menu also allows the user to map different color schemes for different data to the tops and sides of the 3D objects. The 3D heat map application is freely available by request from the authors or for download from http://Sourceforge.net/3dheatmap.

Figure 2 provides an example of our 3D heat map with some hypothetical data. The leftmost panel shows the data in 2D with a pattern visible as defined by the sorting of the columns (x-axis) and the rows (y-axis). The middle and rightmost panels show the same in 3D with bars on a z-axis that are proportional to intensity. Note that the sides of the bars are colored on a yellow to blue scale. This is an example of how the extra dimension can be used to visualize additional layers of information in parallel without needing to switch between perspectives. Further, the tops of the bars could be colored to represent yet another layer of information.

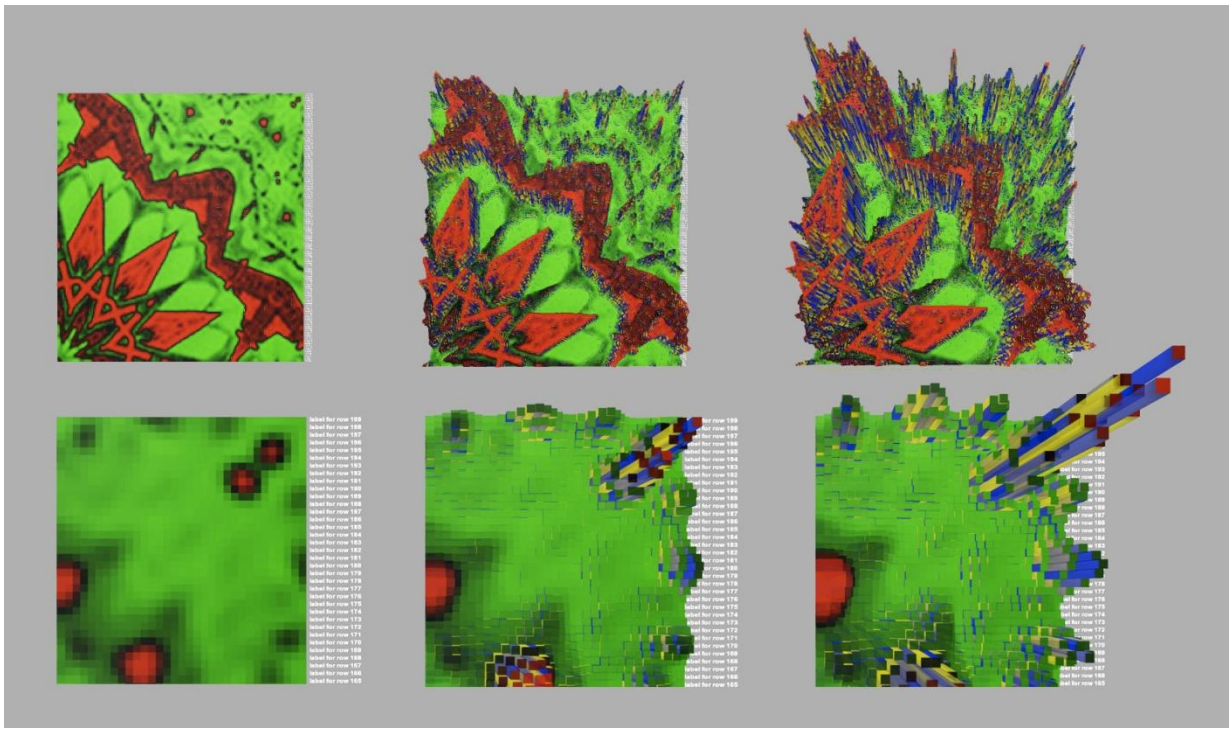

Fig. 2. 3D heat map visualization of hypothetical data. Note the additional layers of information provided by the sides of the bars when illustrated in 3D. Note that he bottom panels are a subset of the top panels corresponding to the upper right corner.

2.2. *Application of the 3D Heat Map to Human Microbiome Visualization*

We applied our 3D heat map method and software to the visualization and interactive exploration of a fecal microbiome data set from infants born prematurely. IRB approval was obtained from the Dartmouth Center for the Protection of Human Subjects in April 2009. Subjects' parents provided informed consent. Six very low birth-weight infants were enrolled within two days of birth for the study and inclusion criteria included birth weight of 501-1500 grams without major congenital or genetic anomalies. Serial stool samples were collected weekly, beginning with the first stool or meconium passed. Stool samples were aliquotted and stored at -80C and bacterial DNA was extracted using the MoBio Powersoil bacterial DNA isolation kit. DNA was quantified and then 454 pyrosequencing was performed at the Josephine Bay Paul Marine Biological Laboratories in Woods Hole, Massachusetts. High throughput sequencing was performed at the Josephine Bay Paul Center and overseen by Dr. Mitch Sogin. Pyrosequencing was targeted at the bacterial 16S the Titanium Roche α-R&D 454 amlicon informatics pipeline to analyze the bacterial community composition of samples.

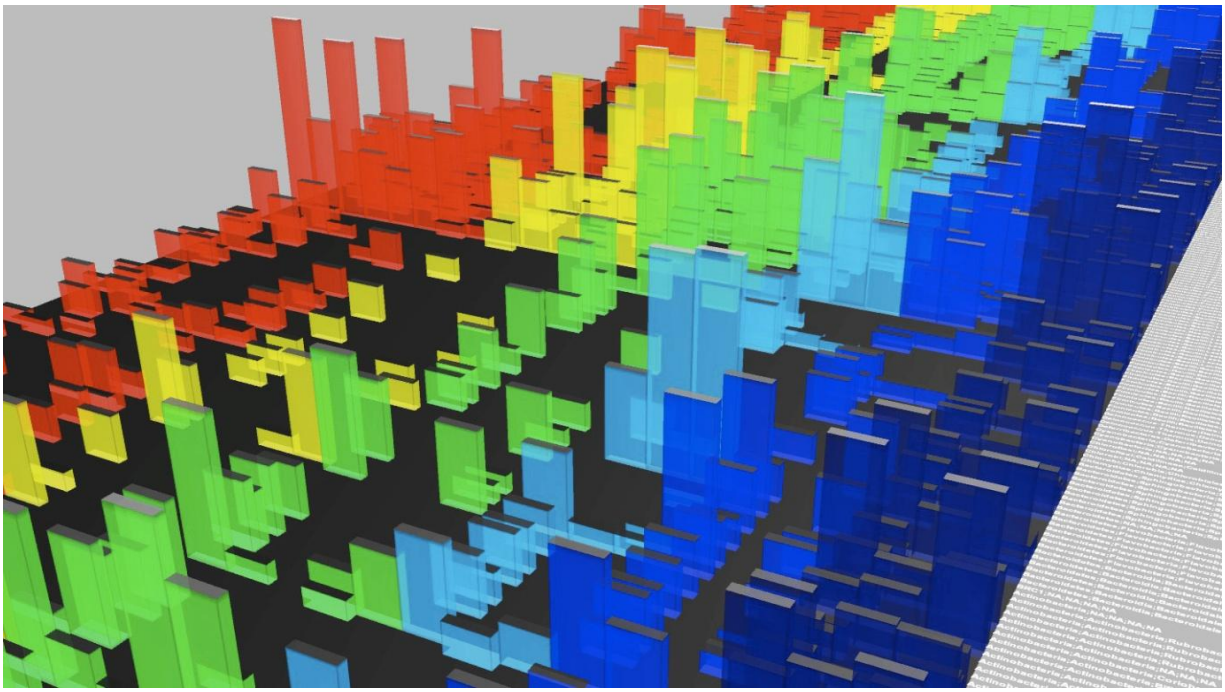

Fig. 3. 3D heat map visualization of fecal microbiome data from six premature infants. The patients and their different time points are ordered on the x-axis while bacterial species are ordered on the y-axis with the species name in white text (side). The bars in the z-axis represent relative abundance of each bacterial species for each specific patient and time point. The tops of the bars are colored in grayscale to also reflect relative abundance with lighter colors indicating higher abundance. This corresponds to the colors used in the 2D heat map. The sides of bars are color-coded by patient.

Our goal was to compare 2D and 3D heat map representations of the microbiome data from serial samples from these six patients. The 2D heat map was used to visualize relative abundance

of bacteria (colored squares) with patient and time on the x-axis and bacterial species on the y-axis. For the 3D heat map we also visualized abundance of bacteria as colored squares organized by patient and time on the x-axis and bacterial specie on the y-axis. In addition, we extended bars for each colored square into the z-axis according to relative abundance with higher bars being more abundant. We added an additional layer of information about samples by coloring the four sides of the 3D bars extending into the z-axis. This demonstrates the ability to include additional layers of information in 3D space to facilitate exploration and interpretation. While it is possible to add additional symbols to a 2D heat map, it is much easier to see and explore in 3D. Symbols and other shapes would also significantly enhance the 3D visualization and would be easy to implement within the video game development framework.

3. Results

Figure 3 illustrates the 3D heat map of bacterial abundance for the six patients (x-axis and side color on each bar) over different time points for each bacterial species measured (y-axis). Note that the 3D perspective allows at least five layers of information to be visualized simultaneously. The five layers include the three axes, the top color of the bars and the side color of the bars. For example, the top color could be used to indicate the presence of sepsis in an infant at a particular time point. The ability to include clinical data with microbiome data will facilitate the visual discovery of patterns that otherwise would not be visible in a 2D heat map representation.

Not only does the 3D heat map allow multiple dimensions of information to be displayed, the video game technology allows the user to interactively explore the 3D space using the keyboard or a 3D mouse that facilitates movement in all three dimensions. The ability to 'fly' through a 3D visualization allows all of the information to be easily explored from multiple different angles. This sort of exploration and interactive visualization is not possible with a typical 3D bar plot as implemented in Microsoft Excel or other similar software packages.

Figure 4 specifically compares a 2D heat map (right panel) with the 3D heat map representation. In both panels the bacterial abundance is colored in grayscale. We kept the grayscale color-coding of abundance on the tops of the bars in the 3D heat map in addition to representing abundance on the z-axis to facilitate direct comparison to the 2D heat map. However, as mentioned above, the tops of the bars could be used to color-code additional information such as a clinical covariate. Resetting graphing parameters and assigning data to each of the dimensions can be done literally "on the fly" as the speed and direction of flight are unchanged by the update. This allows the user to explore multiple projections of the data without losing their current point of view.

Figure 5 compares a specific portion of the 2D and 3D heat maps from Figure 4. Here, the rows highlighted with the red asterisk are for a bacterial species from the Veillonellaceae family. This family belongs to the order Clostridiales and are characterized by gram-negative obligate anaerobes. The top panel of Figure 5 shows the traditional 2D heat map of the data with lighter squares indicating higher relative abundance. The bottom panel of this figure shows the same data in 3D with the patients color-coded on the sides of the bars. It is clear from the 3D heat map that the yellow and green patients have very similar patterns of bacterial abundance across the different

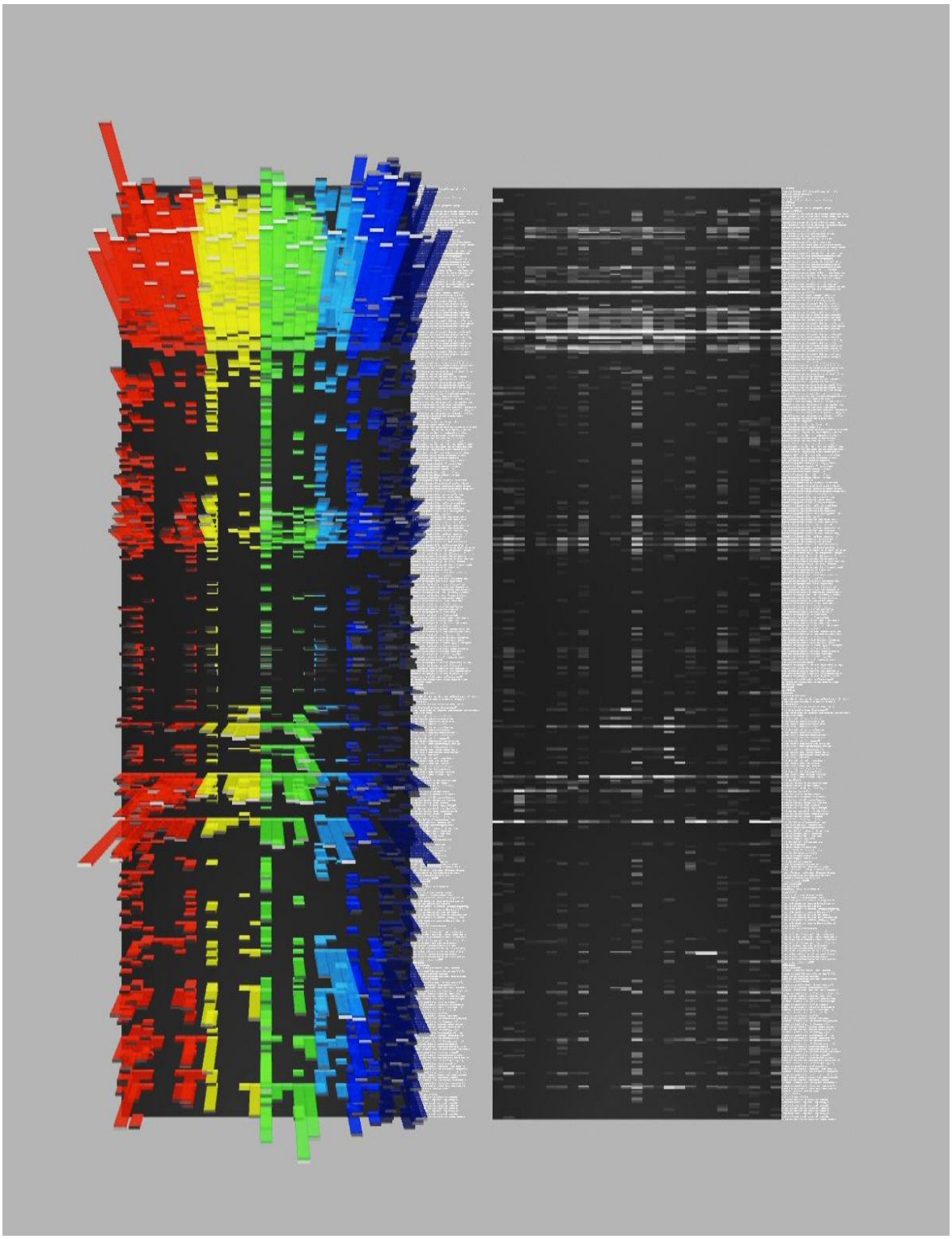

Fig. 4. Comparison of the 2D (right panel) and 3D (left panel) heat maps of bacterial abundance (grayscale in 2D and 3D and z-axis in 3D) in six premature infants. Note the additional layers of information that can be provided the z-axis, the bar top color and the bar side color.

time points. Interestingly, these two patients are twins who received the same diet (maternal breastmilk) and who had similar clinical courses without complications of prematurity. The interactive exploration provided by the 3D video game platform makes these kinds of patterns easy to identify and explore. The colors associated with the additional layers of information make the patient-specific pattern more apparent than in the 2D heat map.

4. Discussion

We have introduced here a 3D heat map method and freely available software package (3dheatmap) for interactive visualization of high-dimensional biomedical data. We have demonstrated the ability of the 3D heat map to visualize at least three more layers of information that the traditional 2D heat map. In addition, the use of a commercial video game engine has made

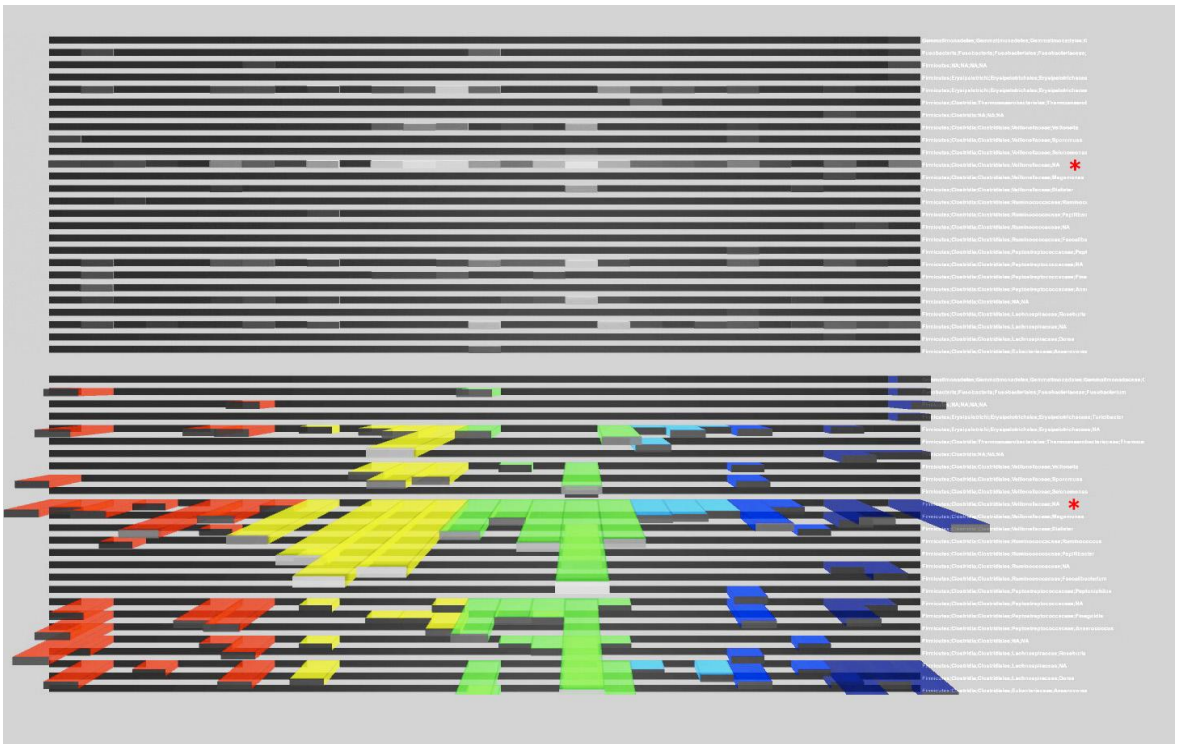

Fig. 5. A zoomed portion of the 2D (top) and 3D (bottom) heat map from Figure 3 highlighting a bacterial species from the Veillonellaceae family that has similar levels of relative abundance across time points within two patients (yellow and green).

it possible to harness the power of video games for interactive exploration of the 3D visualization. We have applied the 3D heat map method to the visualization of human microbiome data and have compared the results with that provided by a 2D heat map.

While the additional layers of information and the interactive exploration of the visualization move well beyond traditional visualization methods, there are many additional features that need to be added to move this approach from the realm of 'information visualization' to that of 'visual

analytics'. Visual analytics is an emerging discipline that combines visualization methods with data analysis and human-computer interaction [16]. This is distinguished from *scientific visualization* that focuses on the mathematics and physics of visualizing 3D objects and *information visualization* that focuses on methods such as heat maps for showing high-dimensional research results. Heer et al. provide a thorough review of information visualization methods [17]. What makes visual analytics different is the integration of the visualization methods with data analysis. That is, the statistical and machine learning analyses can be launched directly from the visualization and the visualization, in turn, can be changed in a manner that is dependent on the data analysis results. This iterative and synergistic process of visualization and analysis is facilitated by computer hardware technology that makes it easy for the user to interact with the software. For example, new touch-based computer interfaces such as the Microsoft Surface Computer or the Apple iPad could replace the keyboard and mouse as the preferred interface for visual analytics. All of this combined with a 3D visualization screen or wall provides a modern visual analytics discovery environment that immerses the user in their data and research results.

Our future goals are to integrate the R statistical computing platform so that analyses can be launched directly from the 3D heat map application. It might be of interest, for example, to interactively select two different families of bacterial species within the visualization and then launch a statistical analysis comparing the relative abundance of species in the two different groups. The ability to launch analyses directly from the visualization environment opens the door to making discoveries that are inspired by visual cues rather than pre-conceived hypotheses that are dependent on existing knowledge. Of course, rigorously testing this hypothesis is not easy but Perer and Shneiderman have presented design guidelines for evaluating visual analytics software [18]. Their methodology has five phases. First, the domain expert or user is interviewed for one hour to determine their intentions. Second, there is a two hour training phase on use of the software. Third, there is a two to four week early use phase in which the users employ the software and the development team is available for troubleshooting and user support. Fourth, there is another two to four hour period of mature use where the only support that is given is for technical problems with the software. Finally, there is an outcome interview to determine whether the visual analytics software had an impact on the research of the user. Impact can be measured in many different ways but might include the generation of new ideas or hypotheses or new knowledge leading to a scientific publication. Positive impact could also be measured in terms of research efficiency. For example, the visualization approach could allow the researchers to make discoveries faster.

Human microbiome data and related research questions will continue to become more complex. This is especially true once the DNA sequence of the host is added to the mix. Visualization has an important role to play in helping the investigator become familiar with their high-dimensional data in a way that might not be possible with a spreadsheet or database. Visual analytics is an emerging discipline that harnesses the power of visualization technology, data analysis and human-computer interaction. The 3D heat map application we have presented here provides a starting point for developing such a discovery system for microbiome analysis. The use of video game engines and other 3D technology has the potential to make this technology accessible to those not skilled in bioinformatics or biostatistics.

5. Acknowledgments

We would like to thank the patients and their families for participating in the research study described here.

References

1. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Ligget, R. Knight, J. I. Gordon, *Nature*. **449**, 804 (2007).
2. The NIH HMP Working Group, *Genome Res.* **19**, 2317 (2009).
3. D. Wu et al., *Nature*. **462**, 1056 (2009).
4. The Human Microbiome Jumpstart Reference Strains Consortium, *Science*. **328**, 994 (2010).
5. E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, R. Knight, *Science*. **326**, 1694 (2009).
6. M. G. Dominguez-Bello, E. K. Costello, M. Contreras, M. Magris, G. Hidalgo, N. Fierer, R. Knight, *PNAS*. **107**, 11971 (2010).
7. D. A. Goldman, J. Leclair, A Macone, *J Pediatr*. **93**, 288 (1978).
8. I. H. Gewolb, R. S. Schwalbe, V. L. Taciak, T. S. Harrison, P. Panigrahi, *Arch Dis Child Fetal Neonatal Ed.* **80**, F167 (1999).
9. A. Schwiertz, B. Gruhl, M. Lobnitz, P. Michel, M. Radke, M. Blaut, *Pediatr Res.* **54**, 393 (2003).
10. Y. Wang et al., *The ISME J.* **3**, 944 (2009).
11. K. Ellrot, L. Jaroszewski, W. Li, J. C. Wooley, A. Godzik, *PLoS Comp Bio.* **6**, e1000798 (2010).
12. G. Vey, G. Moreno-Hagelsieb, *Mol BioSys.* **6**, 1247 (2010).
13. F. Schreiber, P. Gumrich, R. Daniel, P. Meinicke, *Bioinformatics.* **26**, 960 (2010).
14. P. H. A. Sneath, *J Gen Micro.* **17**, 201 (1957).
15. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *PNAS.* **95**, 14863 (1998).
16. J. Thomas, K. Cook, *Illuminating the Path: Research and Development Agenda for Visual Analytics.* IEEE Press (2005).
17. J. Heer, M. Bostock, V. Ogievetsky, *Comm ACM.* **53**, 59 (2010).
18. A Perer, B. Scneiderman, *IEEE Comp Graph App.* **29**, 39 (2009).