Bioinformatics support for the Tübiom community gut microbiome project

Sina Beier[1], Anna Górska[1, 2], Patrick Grupp[1], Theresa Anisja Harbig[1], Isabell Flade[3], and Daniel H. Huson[1]

[1]**ZBIT Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany**
[2]**International Max Planck Research School From Molecules to Organisms, Max Planck Institute for Developmental Biology and Eberhard Karls University Tuebingen, Spemannstr. 35 - 39, 72076 Tübingen, Germany**
[3]**CeMeT Center for Metagenomics GmbH, Paul-Ehrlich-Straße 23, 72076 Tübingen, Germany**

ABSTRACT

The Tübiom project is a community-based project aimed at constructing a large, representative reference database of human gut microbiome profiles. The goal is to collect 10 000 profiles, along with detailed metadata on each participants health and lifestyle. All samples will be processed using identical sequencing and analysis protocols to ensure comparability. The project has four technical components: sequencing, sequence analysis, data storage and visualization.
The project website http://www.tuebiom.de allows interested people to learn about the project, order a sampling kit and fill-in the metadata questionnaire. Once a sample as been submitted and processed, a participant can explore the taxonomic profile of their gut microbiome and compare it to the typical profile of different comparison groups.This community-based project also hopes to engage participants in science-propagation, spread knowledge about the gut microbiome and the importance of this area of research.
Here we provide a brief introduction to the Tübiom project and describe the bioinformatics framework that we have developed for it.

Keywords: 16S amplicon sequencing, human gut microbiome, sequence analysis, software, MEGAN

INTRODUCTION

The gut microbiome composition is important for human health (Sekirov et al. (2010)), including immune response (Willing et al., 2011), reaction to bacterial infection (Caballero and Pamer, 2015; Ferreira et al., 2011; Buffie and Pamer, 2013) and response to treatment with specific drugs (Viaud et al., 2014; Iida et al., 2013; Willmann et al., 2015). The human gut microbiome varies signficantly between different individuals, and is heavily influenced both by diet and weight (Ravussin et al., 2013) and also general lifestyle choices and medication (Hernández et al., 2013).

There are numerous studies that investigate the human gut microbiome composition and function. The results of these studies are usually not comparable with each other. When analyzed together the samples tend to cluster by the study rather then the biological characteristics of the samples (Lozupone et al., 2013). Publications often do not contain enough details on the sequencing and bioinformatical protocols to allow one to reproduce the analyses.

Thus, comparing new data with published data or even comparing multiple published datasets might fail to identify biological variance in the samples. Scientists are testing various DNA extraction kits (Brooks et al., 2015) and sequencing platforms. There remain many unidentified biases that cannot be corrected for before comparative analysis.

Tübiom is a joint project of CeMeT GmbH, Tübingen University Hospital (Institute for Medical Microbiology and Hygiene) and University of Tübingen (Department of Algorithms in Bioinformatics). The aim of the project is to build a large database of human gut microbiome profiles, together with detailed metadata. Full comparability of the data is ensured by identical sample preparation,

sequencing protocol and a standardized bioinformatics pipeline with a fixed database for comparison.

We hope to use this data to establish significant correlations between gut microbial composition and human health. This project is similar to the American Gut Project (McDonald et al., 2015), which is focused on recruiting paying participants in the USA.

We have constructed a web platform to advertise and inform about the Tübiom project, recruit participants, collect their metadata and present user-specific results. Participants can obtain insights into their gut microbiome and compare their taxonomic profile to the average profiles of different groups that vary in terms of health or lifestyle. Thus, the project fits into the popular *quantified self* philosophy that propagates collecting various data about oneself in order to be more aware about ones body and to manage ones life and health more effectively. The project has been approved by the Ethics Board of Tübingen University Hospital and participation is free.

Scientists constantly need to explain and justify their work and its implications for society. As the Tübiom project is directed at the general public, it does not only pursue scientific goals but also aims at popularizing knowledge and awareness about microbiology, metagenomics and personal health. The Tübiom project seeks to provide background knowledge and to present the collected data in a way that is understandable for participants with varying levels of background knowledge.

Here we provide a brief introduction to the Tübiom project and then describe the bioinformatics pipeline that we have developed for it.

BACKGROUND

DNA sequences from environmental or biological samples provide unique insights in to the composition, structure and function of microbial communities, allowing one to study unculturable organisms. There are two main sequencing strategies used for studying microbial communities: amplicon sequencing, which is usually applied to the 16S rRNA gene (Peace et al., 1986), and shotgun sequencing (Handelsman, 2004).

Shotgun sequencing allows one to study both the taxonomic and functional content of microbiome samples, whereas 16S rRNA sequencing only allows taxonomic profiling. The sequence of the 16S rRNA gene is overall well conserved, with nine variable regions (Van de Peer et al., 1996). Primers are designed to target the intermediate conserved regions and it is possible to sequence only the end or beginning of a conserved region and a long stretch of the following variable region. Taxonomic profiling using 16S rRNA amplicon sequences is cheaper and faster to perform than taxonomic profiling using shotgun sequencing, therefore in the Tübiom Project we use 16S rRNA sequencing for taxonomic profiling.

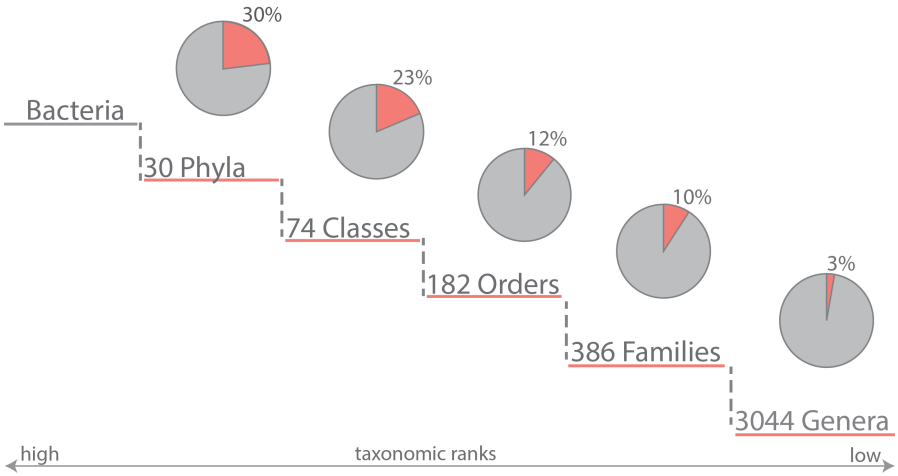

Figure 1. We conduct analysis on five taxonomic ranks within the Bacterial domain. For each rank, we report the relative number of members of the taxa in this rank found so far in at least one sample of the Tübiom project.

Regardless of whether one uses amplicon sequencing or shotgun sequencing, the resulting dataset will consist of millions of short reads. These reads are compared against a database of reference sequences. Based on the result of such a comparison, each read is then assigned to some taxon, at

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2382v1 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

some rank of the taxonomy (see Fig. 1). A taxonomic profile is then obtained by summarizing the number of the reads assigned to each taxon at a given taxonomic rank.

METHODS

The entire processing pipeline, including the samples handling, sequencing, computing and presentation is implemented using infrastructure provided by CeMeT. Data-processing is fully automated. Fig. 2 depicts the general setup of the project.

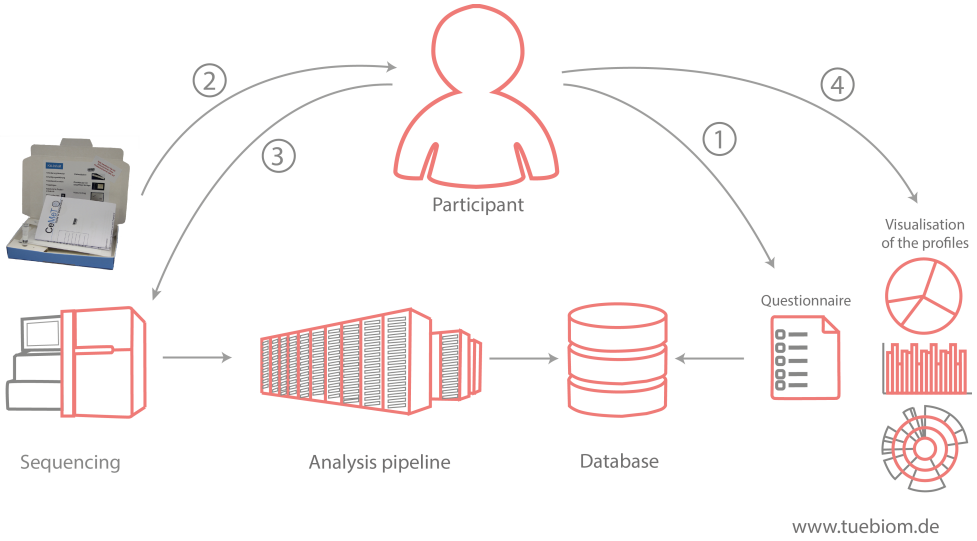

Figure 2. (1) A participant registers with the website and requests a kit. (2) CeMeT sends a kit to the participant. (3) The participant uses the kit to collect a sample and sends it to CeMeT. The DNA is extracted and sequenced, and then the bioinformatics pipeline computes a taxonomic profile and stores it in a database. (4) The participant can interactively explore the result on the Tübiom website.

Kit

The sampling kit (also shown in Fig. 2) contains a swab and tube for sample collection. In addition, it contains a plastic bag with absorbent material and a preaddressed return envelope to send back the sample. The kit also includes all forms necessary for participation and participant information, including a step-by-step description of the sampling process.

Questionnaire

We designed a questioniare to collect the metadata associated with each sample. This is based on the literature factors that are believed to be correleated with the composition of the human gut microbiome. Most of the questions are multiple choice questions in order to simplify automatic parsing and database entry generation. The participants can enter additional information, ensuring that the questionnaire is flexible enough to cope with special situations of individual participants. The questionaire has four sections, general information, lifestyle, health and medication, and captures metadata on ≈ 20 topics.

Groups of participants, average profiles

We intend to use the metadata as phenotype markers in future correlation analyses. Currently they are used to construct average profiles for specific phenotypical groups. The average profiles can be viewed by the participants and compared with their personal samples, allowing a participant to compare themselve with the average vegan, say.

By design, our database allows the construction of very specific comparison groups using logical expressions, such as “vegan and age>60 and runs at least 3 times a week”. However, using such specific comparison groups will only make sense once enough samples have been processed so as to populate such a group with a sufficient number of samples.

We have predefined 31 different groups and allow users to compare their sample against any group for which at least 5 samples are available. At present, the participants can compare against the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2382v1 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

following groups: all samples in the database, vegan, normal weight and colitis. Both back-end and visualization support modification and extension of the groups.

Sequencing parameters

16S rDNA amplicon sequencing is conducted by CeGaT GmbH on an Illumina MiSeq using Nextera XT v2 chemistry and producing 2x250 bp paired reads. In-house primers and a specifically developed single cycle PCR routine are used to generate amplicons for the variable regions V3 and V4.

Analysis pipeline

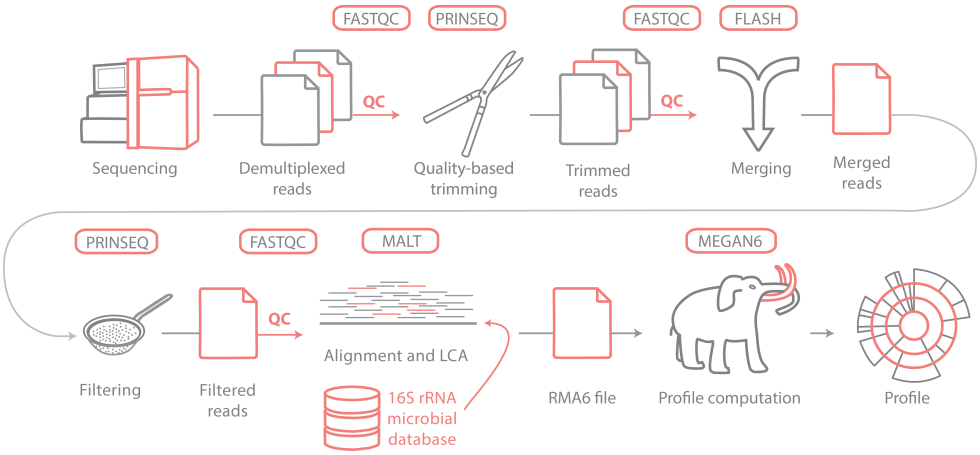

Figure 3. The Tübiom analysis pipeline, starting from raw reads and providing taxonomic profiles for all samples which are the input for the project database.

The input to our analysis pipeline are sequenced and demultiplexed datasets in FastQ format. The output consists of taxonomic profiles that provide bacterial count data in the MEGAN6-specific RMA6 format. The pipeline (Fig. 3) is based on established tools for sequence analysis and in-house developed Python scripts.

Parameters for the pipeline have been tested on standardized datasets generated using the same sequencing protocol and same sequencer as is to be used throughout the project, thus providing a gold standard for analysis of the data. This is necessary, because changes in protocols or instrument is known influence the results Kim and Yu (2014).

After each step in the pipeline, we perform a quality control step and some sanity checks. In more details, quality control is handled by FastQC[1] and is applied to the raw reads, to the reads after trimming, and to the reads after merging and length filtering.

Read trimming before merging and length filtering is performed using PRINSEQ (Schmieder and Edwards, 2011). Parameters like minimal read length, minimal average quality in a sliding window and window size are set in a configuration file. The parameters have been optimized for a gold standard analysis for average output of the used sequencing pipeline, but can be changed to adapt to different sequencing strategies or salvage bad quality sequencing runs if so desired.

After trimming, the paired reads are merged using FLASH (Magoc and Salzberg, 2011) and finally aligned to NCBI's 16SMicrobial database (downloaded September 2015) using the Alignment tool MALT (Herbig et al., 2016). MALT uses the LCA algorithm to perform taxonomic classification of all reads based on their alignments. Alignment parameters are optimized to meet our gold standard and can be adapted in the configuration.

Profile computation

In the last step of the the analysis pipeline, MEGAN (Huson et al., 2016) is used to construct taxonomic sample profiles for all participants samples, which are then normalized with respect to the abundances of reads and saved dynamically to a PostgreSQL database. Guided by the sample metadata, we compute group profiles from the sample profiles and save them to the database.

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2382v1 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

Platform

The Tübiom project uses various software tools and frameworks, in particular, all data are contained in a PostgreSQL[2] 9.4.5 database. The heart of the database are the samples contributed by project participants and the group profiles derived from those.

The database is filled both from inside CeMeT, when taxonomic profiles are computed and entered, and from outside, when participants answer the questionnaire. Participants answer one questionaire for each sample collected, as some parts of the metadata might change between sampling events.

Django[3] 1.9.1 is used as a web-framework, allowing security and dynamic website generation. Django requires Python 2.7 and a connector to the PostgreSQL database `psycopg2`. For serving the website on the Internet we use Apache HTTPD[4], with the `mod_wsgi` module, in order to allow theconnection to the Django framework.

Tübiom website is available only in German language, since the majority of participants are German-speaking. We plan to add English translation in the near future.

Data visualization

The results of the taxonomic analysis are dynamically visualized using JavaScipt and the D3 package, a powerful library for creating interactive visualizations.

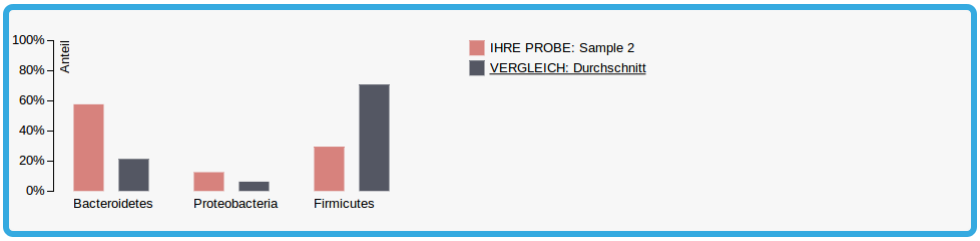

Figure 4. Static header summarizing the profiles for all of the users samples (Ihre probe) compared to the average of all available samples in the database (vergleich: Durschnitt). The header summarizes only the most abundant phyla: *Bacteroidetes, Proteobacteria* and *Firmicutes*.

Participants can view the taxonomic composition of their samples in the results section of the website. At the top of the page a bar chart displays a comparison of the users latest profile to the overall average from all samples currently in the database, on the phylum level (Fig. 4). This chart is intended to catch the users attention and provides a first impression of the results of the analysis. The header is the only static plot.

For a more detailed exploration we designed an interactive visualization where up to five of the profiles can be compared. The chosen profiles can be any subset of a participants own profiles and the comparison groups. The visualization provides two menus (Fig. 5) for the selection of profiles, tabs for selecting the taxonomic rank and five individual plots, each focusing on a different aspect of the analysis results. In order to match every participants interests and background knowledge we provide both simple and more advanced plots.

Three of the five plots focus on the composition of the profiles on the selected taxonomic rank. The first and simplest plot (Fig. 5) is a bar chart similar to the header where four taxa can be chosen and compared. We use a stacked bar chart (Fig. 6) and a heat map (Fig. 7) for the comparison of all taxa on the selected taxonomic rank where each stack of the bar chart and respectively each row of the heatmap correspond to one profile.

Since we also want to provide information about the diversity and hierarchy of the profiles we included “sunburst” plots (Fig. 8). In sunbursts, the composition of higher taxonomic ranks (inner rings) is shown in increasing detail (lower ranks) using concentric rings.

The final plot shows the similarities between profiles for the selected taxonomic rank. The user can chose one of the already preselected profiles and the cosine similarities between this profile and the other selected profiles are visualized using a bar chart.

All of the charts are equipped with tool tips and descriptions. The average analysis result view is visible to prospective participants, before signing up and ordering their own analysis. Although we do not support downloading of the sequencing data we are developing a pdf printable version of the report.

[2]http://www.postgresql.org
[3]http://www.djangoproject.com
[4]http://httpd.apache.org

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2382v1 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

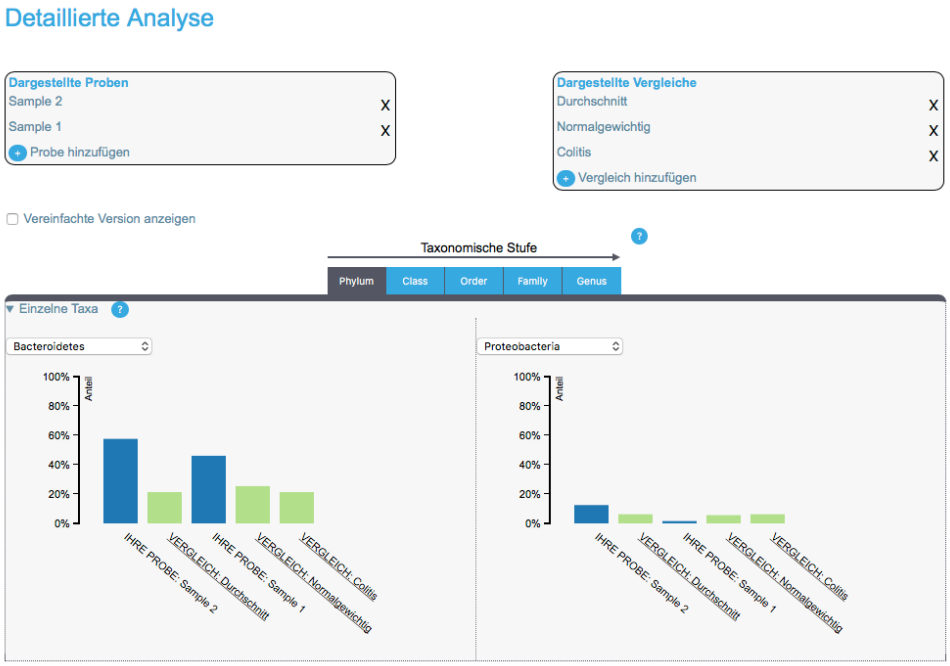

Figure 5. Two menus located just below the static bar chart (Fig. 4). First menu is used to choose profiles from personal samples (Proben) and from group average profiles (Vergleiche). Here, the two personal samples are compared (Sample 1 and Sample 2) to the average profiles computed for all samples in the database (Durchschnitt), samples of people with healthy weight (Normalgewichtig) and samples of people with Colitis. The interactive bar chart allow browsing abundance of the individual taxa (Einzelne Taxa), that can be chosen from the drop-down menu. The taxonomic rank for this bar chart and all following chats is set with a tab. User can view their comparison on five taxonomic levels from Phylum to Genus.

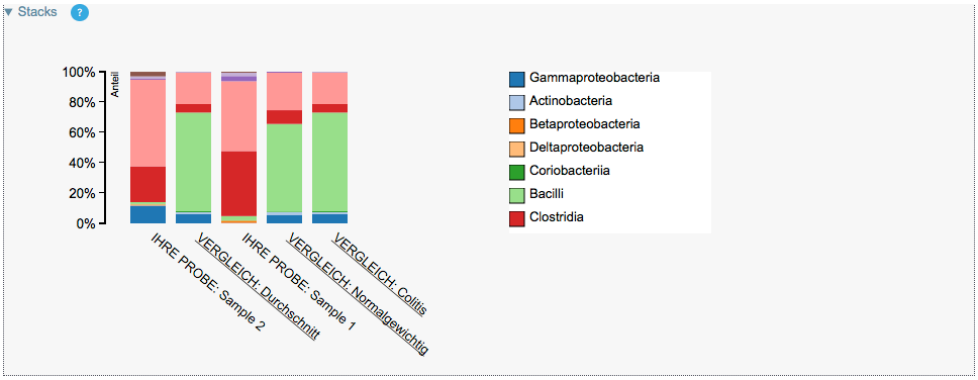

Figure 6. Stack bar chart for five profiles on the class taxonomic level, for the same samples and settings as Fig. 5.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2382v1 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

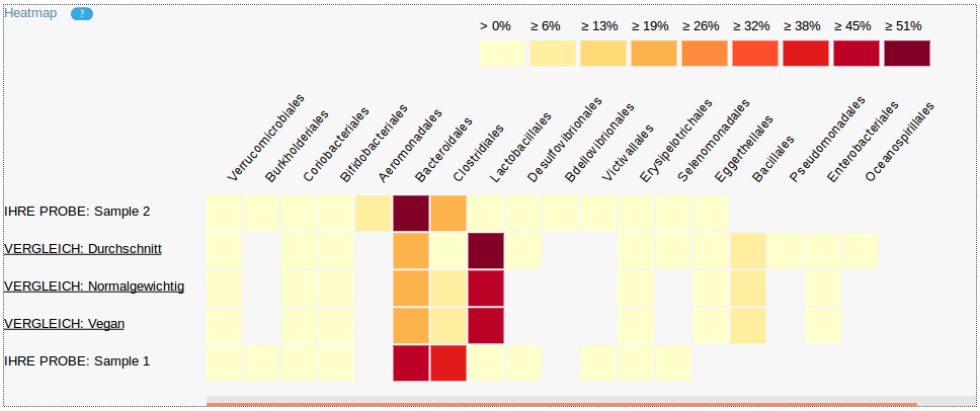

Figure 7. The heatmap showing the relative abundance of taxa in the different samples and groups on a certain taxonomic level. The columns represent all the taxa which are currently displayed. Each rectangle is colored depending on the percentage of the taxa as shown in the legend at the top of the plot.

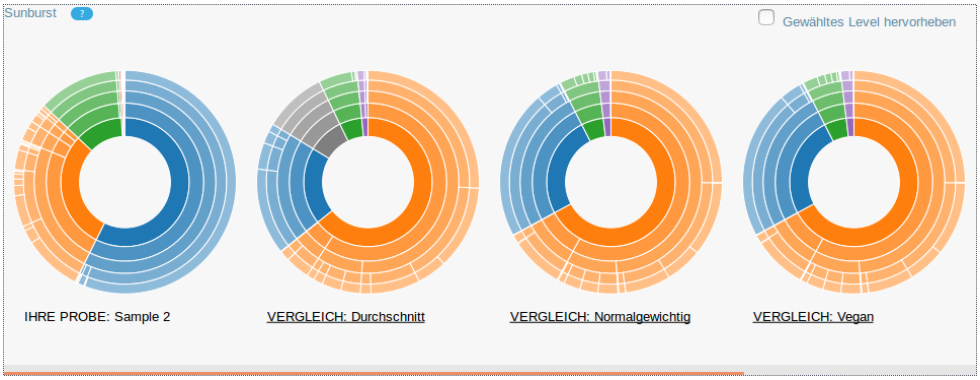

Figure 8. Sunburst showing the hierarchical taxonomic composition of a sample or a group (as labeled below each sunburst). The outer ring represents the genus level, whereas higher taxonomic ranks are more closer to the center. The colors are assigned automatically to distinguish taxa. The description of taxa and percentage appears when user hovers on the ring with mouse.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2382v1 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

Ethics

As every research on human subjects, the Tübiom project was reviewed by the Ethics Committee of the University Hospital Tübingen and approved. The ethics committee consisting of doctors, scientists, lawyers, philosophers and theologians advise scientists in ethical and legal issues of their research. Particular emphasis is placed in the assessment on the data privacy as well as ethical guidelines and regulations.

DISCUSSION

In the first two months after launching the Tübiom project 552 participants signed up and more than 3 000 sampling kits have been send out to interested people.

We have developed a model for generating a database of well defined microbial profiles associated with comprehensive, normalized metadata about the participant. This will benefit research related to the human gut microbiome in health and disease.

It will provide a baseline of the microbiome and its normal variance in healthy people following different lifestyle choices, which is important information for the development of health-relevant biomarkers based on the microbiome. This should also help "treat the healthy" so as to make it easier for them to reach their health goals such as keeping and improving their weight or reducing negative side effects of medication they might have to take.

The Tübiom is not the only community-based gut microbiota quantifying project out there. Unlike the American Gut Project we plan to investigate the change the persons gut microbiome undergoes when the metadata changes, e.g. diet or lifestyle. We store the direct email contact to the participants, so we can reach out to them to ask additional questions when needed.

From the technical aspects, the Tübiom participants represent a different population then British or American gut. Also, there are large differences in the analysis pipeline. We are using the customized pipeline with the in-house performing crucial steps namely MEGAN and MALT.

The bioinformatics framework of the project has been designed so as to allow further development, adjustment and extensions. It is easy to reanalyse all data using updated algorithms or databases. Everything is modular, the four pillars of sequencing, sequence analysis, database storage and visualization have as little interdependence as possible and can thus be changed independently. This allows use to maintain comparability of the data analysis, while keeping the analysis and visualizations state of the art.

The bioinformatics framework can also be used to process shotgun metagenome datasets so as to provide more detailed taxonomic and functional analysis and a subset of the collected samples will be subjected to shotgun sequencing and full metagenome analysis.

ACKNOWLEDGMENTS

We thank Detlef Weigel (MPI Developmental Biology), Ingo Autenrieth and Matthias Willmann (University Hospital Tübingen), and CeGaT GmbH (especially Dirk Biskup and Tim Scheuernbrand) for support and discussions.

REFERENCES

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., Strauss, J. F., Jefferson, K. K., and Buck, G. A. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC microbiology*, 15(1):66.

Buffie, C. G. and Pamer, E. G. (2013). Microbiota-mediated colonization resistance against intestinal pathogens. *Nature reviews. Immunology*, 13(11):790–801.

Caballero, S. and Pamer, E. G. (2015). Microbiota-Mediated Inflammation and Antimicrobial Defense in the Intestine. *Annu Rev Immunol*, 33:227–256.

Ferreira, R. B. R., Gill, N., Willing, B. P., Antunes, L. C. M., Russell, S. L., Croxen, M. a., and Finlay, B. B. (2011). The intestinal microbiota plays a role in Salmonella-induced colitis independent of pathogen colonization. *PloS one*, 6(5):e20338.

Handelsman, J. (2004). Metagenomics : Application of Genomics to Uncultured Microorganisms Metagenomics : Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4):669–685.

Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., and Huson, D. H. (2016). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv*, page 50559.

Hernández, E., Bargiela, R., Diez, M. S., Friedrichs, A., Pérez-Cobas, A. E., Gosalbes, M. J., Knecht, H., Martínez-Martínez, M., Seifert, J., von Bergen, M., Artacho, A., Ruiz, A., Campoy, C., Latorre, A., Ott, S. J., Moya, A., Suárez, A., Martins dos Santos, V. a. P., and Ferrer, M. (2013). Functional consequences of microbial shifts in the human gastrointestinal tract linked to antibiotic treatment and obesity. *Gut microbes*, 4(4):306–315.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H. J., and Tappu, R. (2016). MEGAN Community Edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *To appear in: PLoS Computational Biology*.

Iida, N., Dzutsev, A., Stewart, C. A., Smith, L., Bouladoux, N., Weingarten, R. a., Molina, D. a., Salcedo, R., Back, T., Cramer, S., Dai, R.-M., Kiu, H., Cardone, M., Naik, S., Patri, A. K., Wang, E., Marincola, F. M., Frank, K. M., Belkaid, Y., Trinchieri, G., and Goldszmid, R. S. (2013). Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science (New York, N.Y.)*, 342(6161):967–970.

Kim, M. and Yu, Z. (2014). Variations in 16S rRNA-based microbiome profiling between pyrosequencing runs and between pyrosequencing facilities. *J. Microbiol.*, 52(5):355–365.

Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., V??zquez-Baeza, Y., Jansson, J. K., Gordon, J. I., and Knight, R. (2013). Meta-analyses of studies of the human microbiota. *Genome Research*, 23(10):1704–1714.

Magoc, T. and Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963.

McDonald, D., Birmingham, A., and Knight, R. (2015). Context and the human microbiome. *Microbiome*, 3(1):52.

Peace, N. R., Stahl, D. A., Lane, D. J., and Olsen, G. J. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. *Advances in Microbial Ecology*, (9):1–55.

Ravussin, Y., Koren, O., Spor, A., LeDuc, C., Gutman, R., Stombaugh, J., Knight, R., Ley, R. E., and Leibel, R. L. (2013). Responses of Gut Microbiota to Diet Composition and Weight Loss in Lean and Obese Mice. *Obesity (Silver Spring)*, 20(4).

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.

Sekirov, I., Russell, S. L., Antunes, L. C. M., and Finlay, B. B. (2010). Gut Microbiota in Health and Disease. *Physiol Rev*, pages 859–904.

Van de Peer, Y., Chapelle, S., and De Wachter, R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic acids research*, 24(17):3381–3391.

Viaud, S., Daillère, R., Boneca, I. G., Lepage, P., Langella, P., Chamaillard, M., Pittet, M. J., Ghiringhelli, F., Trinchieri, G., Goldszmid, R., and Zitvogel, L. (2014). Gut microbiome and anticancer immune response: really hot Sh*t! *Cell death and differentiation*, pages 1–16.

Willing, B. P., Vacharaksa, A., Croxen, M., Thanachayanont, T., and Finlay, B. B. (2011). Altering host resistance to infections through microbial transplantation. *PloS one*, 6(10):e26988.

Willmann, M., El-Hadidi, M., Huson, D. H., Schütz, M., Weidenmaier, C., Autenrieth, I. B., and Peter, S. (2015). Antibiotic selection pressure determination through sequence-based metagenomics. *Antimicrobial Agents and Chemotherapy*, 59(12):7335–7345.