Compact graphical representation of phylogenetic data and metadata with GraPhlAn

Francesco Asnicar, George Weingart, Timothy L Tickle, Curtis Huttenhower, Nicola Segata

The increased availability of genomic and metagenomic data poses challenges at multiple analysis levels, including visualization of very large-scale microbial and microbial community data paired with rich metadata. We developed GraPhlAn (Graphical Phylogenetic Analysis), a computational tool that produces high-quality, compact visualizations of microbial genomes and metagenomes. This includes phylogenies spanning up to thousands of taxa, annotated with metadata ranging from microbial community abundances to microbial physiology or host and environmental phenotypes. GraPhlAn has been developed as an open-source command-driven tool in order to be easily integrated into complex, publication-quality bioinformatics pipelines. It can be executed either locally or through an online Galaxy web application. We present several examples including taxonomic and phylogenetic visualization of microbial communities, metabolic functions, and biomarker discovery that illustrate GraPhlAn's potential for modern microbial and community genomics.

Compact graphical representation of phylogenetic data and metadata with GraPhlAn

Francesco Asnicar [1], George Weingart [2], Timothy L Tickle [3], Curtis Huttenhower [2,3], Nicola Segata [1]

1. Centre for Integrative Biology (CIBIO), University of Trento, Italy
2. Biostatistics Department, Harvard School of Public Health, USA
3. Broad Institute of MIT and Harvard, USA

Abstract

The increased availability of genomic and metagenomic data poses challenges at multiple analysis levels, including visualization of very large-scale microbial and microbial community data paired with rich metadata. We developed GraPhlAn (Graphical Phylogenetic Analysis), a computational tool that produces high-quality, compact visualizations of microbial genomes and metagenomes. This includes phylogenies spanning up to thousands of taxa, annotated with metadata ranging from microbial community abundances to microbial physiology or host and environmental phenotypes. GraPhlAn has been developed as an open-source command-driven tool in order to be easily integrated into complex, publication-quality bioinformatics pipelines. It can be executed either locally or through an online Galaxy web application. We present several examples including taxonomic and phylogenetic visualization of microbial communities, metabolic functions, and biomarker discovery that illustrate GraPhlAn's potential for modern microbial and community genomics.

Introduction

Modern high-throughput sequencing technologies provide comprehensive, large-scale datasets that have enabled a variety of novel genomic and metagenomic studies. A large number of statistical and computational tools have been developed specifically to tackle the complexity and high-dimensionality of such datasets and to provide robust and interpretable results. Visualizing data including thousands of microbial genomes or metagenomes, however, remains a challenging task that is often crucial to driving exploratory data mining and to compactly summarizing quantitative conclusions.

In the specific context of microbial genomics and metagenomics, next-generation sequencing in particular produces datasets of unprecedented size, including thousands of newly sequenced microbial genomes per month and a tremendous increase in genetic diversity sampled by isolates or culture-free assays. Displaying phylogenies with thousands of microbial taxa in hundreds of samples is infeasible with most available tools. This is especially true when sequencing profiles need to be placed in the context of sample metadata (e.g. clinical information). Among recently developed tools, iTOL (Letunic & Bork 2007;

Letunic & Bork 2011) targets interactive analyses of large-scale phylogenies with a moderate amount of overlaid metadata, whereas Krona (Ondov et al. 2011) instead emphasizes hierarchical quantitative information typically derived from metagenomic taxonomic profiles. Neither of these tools provides an automatable environment in which very large phylogenies can be combined with high-dimensional metadata such as microbial community abundances, host or environmental phenotypes, or microbial physiological properties.

In particular, a successful high-throughput genomic visualization environment for modern microbial informatics must satisfy two criteria. First, software releases must be free and open-source to allow other researchers to verify and to adapt the software to their specific needs and to cope with the quick evolution of data types and datasets size. Second, visualization tools must be command-driven in order to be embedded in computational pipelines. This allows for a higher degree of analysis reproducibility, but the software must correspondingly be available for local installation and callable through a convenient interface (e.g. API or general scripting language). Local installations have also the advantage of avoiding the transfer of large or sensitive data to remote servers, preventing potential issues with the confidentiality of unpublished biological data. Neither of these criteria, of course, prevent tools from also being embeddable in web-based interfaces in order to facilitate use by users with limited computational expertise (Blankenberg et al. 2010; Giardine et al. 2005; Goecks et al. 2010; Oinn et al. 2004), and all such tools must regardless produce informative, clear, detailed, and publication-ready visualizations.

Results and Discussion

GraPhlAn is a new tool for compact and publication-quality representation of circular taxonomic and phylogenetic trees with potentially rich sets of associated metadata. It was developed primarily for microbial genomic and microbiome-related studies in which the complex phylogenetic/taxonomic structure of microbial communities needs to be complemented with quantitative and qualitative sample-associated metadata.

Plotting taxonomic trees with clade annotations

The simplest structures visualizable by GraPhlAn include taxonomic trees (i.e. those without variable branch lengths) with simple clade or taxon nomenclature labels. These can be combined with quantitative information such as taxon abundances, phenotypes, or genomic properties. GraPhlAn provides separate visualization options for trees (thus potentially unannotated) and their annotations, the latter of which (the annotation module) attaches metadata properties using the PhyloXML format (Han & Zmasek 2009). This annotation and subsequent metadata visualization process (**Fig. 1**) can be repeatedly applied to the same tree.

The GraPhlAn tree visualization (plotting module) takes as input a tree represented in any one of the most common data formats: Newick, Nexus (Maddison et al. 1997), PhyloXML (Han & Zmasek 2009), or plain text. Without annotations, the plotting module generates a simple version of the tree (**Fig. 1A**),

but the process can then continue by adding a diverse set of visualization annotations. Annotations can affect the appearance of the tree at different levels, including its global appearance ("global options" e.g. the size of the image, **Fig. 1B**), the properties of subsets of nodes and branches ("node options" e.g. the color of a taxon, **Fig. 1C**), and the background features used to highlight sub-trees ("label options" e.g. the name of a species containing multiple taxa, **Fig. 1D**). A subset of the available configurable options includes the thickness of tree branches, their colors, highlighting background colors and labels of specific sub-trees, and the sizes and shapes of individual nodes. Wild cards are supported to share graphical and annotation details among sub-trees by affecting all the descendants of a clade or its terminal nodes only. These features in combination aim to conveniently highlight specific sub-trees and metadata patterns of interest.

Additional taxon-specific features can be plotted as so-called external rings when not directly embedded into the tree. External rings are drawn just outside the area of the tree and can be used to display specific information about leaf taxa, such as abundances of each species in different conditions/environments or their genome sizes. The shapes and forms of these rings are also configurable; for example, in **Fig. 1E** ("set external ring options"), the elements of the innermost external ring are triangular, indicating the directional sign of a genomic property. The second, third, and fourth external rings show leaf-specific features, using a heatmap gradient from blank to full color. Finally, the last external ring is a bar-plot representing a continuous property of leaf nodes of the tree.

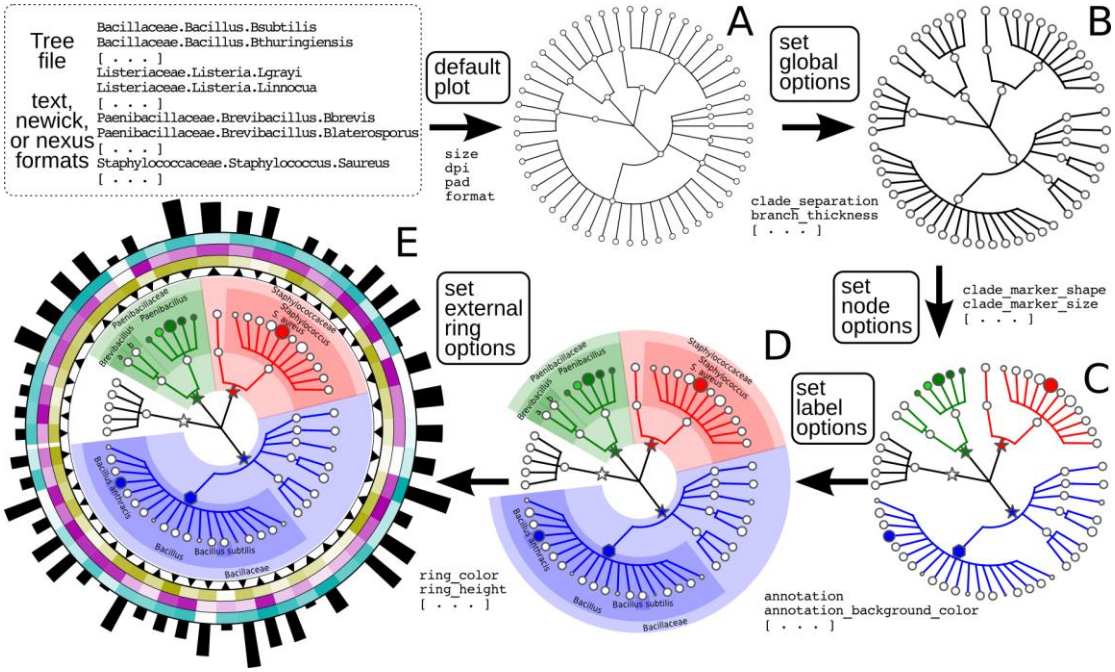

Figure 1. Schematic and simplified example of GraPhlAn visualization of annotated phylogenies and taxonomies. *The software can start from a tree in Newick, Nexus, PhyloXML, or plain text formats. The "default plot" (**A**) produces a basic visualization of the tree's hierarchical structure. Through an annotation file,*

*it is possible to configure a number of options that affect the appearance of the tree. For instance, some global parameters will affect the whole tree structure, such as the color and thickness of branches ("set global options", **B**). The same annotation file can act on specific nodes, customizing their shape, size, and color ("set node options", **C**). Labels and background colors for specific branches in the tree can also be configured ("set label options", **D**). External to the circular area of the tree, the annotation file can include directives for plotting different shapes, heatmap colors, or bar-plots representing quantitative taxon traits ("set external ring options", **E**).*

Compact representations of phylogenetic trees with associated metadata

Visualizing phylogenetic structures and their relation to external metadata is particularly challenging when the dimension of the internal structure is large. Mainly as a consequence of the low cost of sequencing, current research in microbial genomics and metagenomics needs indeed to visualize a considerable amount of phylogenetic data. GraPhlAn can easily handle such cases, as illustrated here in an example of a large phylogenetic tree (3,737 taxa, provided as a PhyloXML file in the software repository, see Availability section) with multiple types of associated metadata (**Fig. 2**).

Specifically, we used GraPhlAn to display the microbial tree of life as inferred by PhyloPhlAn (Segata et al. 2013), annotating this evolutionary information with genome-specific metadata (**Fig. 2**). In particular, we annotated the genome contents related to seven functional modules from the KEGG database (Kanehisa et al. 2011), specifically two different ATP synthesis machineries (M00157: F-type ATPase and M00159: V/A-type ATPase) and five modules for bacterial fatty acid metabolism (M00082: Fatty acid biosynthesis, initiation, M00083: Fatty acid biosynthesis elongation, M00086: acyl-CoA synthesis, M00087: beta-Oxidation, and M00088: Ketone body biosynthesis). We then also annotated genome size as an external circular bar plot.

As expected, it is immediately visually apparent that the two types of ATPase are almost mutually exclusive within available genome annotations, with the V/A-type ATPase (module M00159) present mainly in *Archaea* and the F-type ATPase (module M000157) mostly characterizing *Bacteria*. Some exceptions are easily identifiable: *Thermi* and *Clamydophilia*, for instance, completely lack the F-type ATPase, presenting only the typically archaea-specific V/A-type ATPase. As previously discussed in the literature (Cross & Müller 2004; Mulkidjanian et al. 2007), this may due to the acquisition of V/A-type ATPase by horizontal gene transfer and the subsequent loss of the F-type ATPase capability. Interestingly, some species such as those in the *Streptococcus* genus and some *Clostridia* still show both ATPase systems in their genomes.

With respect to fatty acid metabolism, some clades - including organisms such as *Mycoplasmas* - completely lack any of the targeted pathways. Indeed, *Mycoplasmas* are the smallest living cells yet discovered, lacking a cell wall (Razin 1992) and demonstrating an obligate parasitic lifestyle. Since they primarily exploit host molecular capabilities, *Mycoplasmas* do not need to be able to fulfill all typical cell functions, and this is also indicated by the plotted very

short genome sizes. *Escherichia*, on the other hand, has a much longer genome, and all the considered fatty acid metabolism capabilities are present. These evolutionary aspects are well known in the literature, GraPhlAn permits them and other phylogeny-wide genomic patterns to be easily visualized for further hypothesis generation.

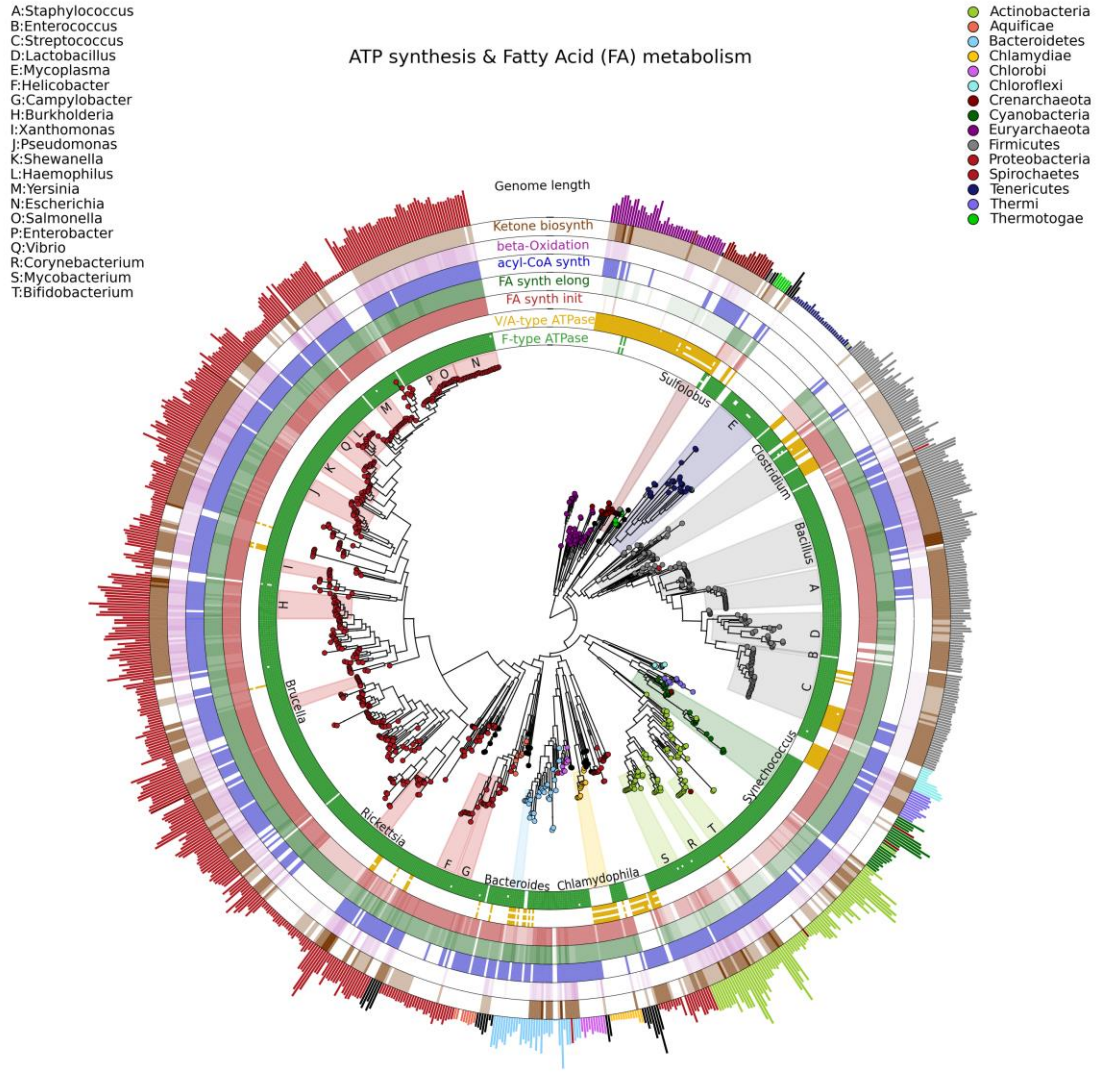

Figure 2. A large, 3,737 genome phylogeny annotated with functional genomic properties. *We used the phylogenetic tree built using PhyloPhlAn (Segata et al. 2013) on all available microbial genomes as of 2013 and annotated the presence of ATP synthesis and Fatty Acid metabolism functional modules (as annotated in KEGG) and the genome length for all genomes. Colors and background annotation highlight bacterial phyla, and the functional information is reported in external rings. ATP synthesis rings visualize the presence (or absence) of each module, while Fatty Acid metabolism capability is represented with a gradient color. Data used in this image are available as indicated in the "Datasets used" paragraph, under "Materials and Methods" section.*

Visualizing microbiome biomarkers

GraPhlAn provides a means for displaying either phylogenetic (trees with branch lengths) or taxonomic (trees without branch length) data generated by other metagenomic analysis tools. For instance, we show here examples of GraPhlAn

plots for taxonomic profiles (**Fig. 3**), functional profiles (**Fig. 4**), and specific features identified as biomarkers (**Fig. 3** and **4**). In these plots, GraPhlAn highlights microbial sub-trees that are found to be significantly differentially abundant by LEfSe (Segata et al. 2011), along with their effect sizes as estimated by linear discriminant analysis (LDA). To enhance biomarker visualization, we annotated them in the tree with a shaded background color and with clade names as labels, with decreasing font sizes for internal levels. To represent the effect size, we scaled the node color from black (low LDA score) to full color (high LDA score).

Fig. 3 shows the taxonomic tree of biomarkers (significantly differential clades) resulting from a contrast gut metagenome profiles from the Human Microbiome Project (HMP) (Huttenhower et al. 2012) and MetaHIT samples (Qin et al. 2010). Only samples from healthy individuals in the latter cohort were included. The filtered dataset was analyzed using LEfSe (Segata et al. 2011) and the cladogram obtained using the *export2graphlan* script provided with GraPhlAn and discussed in the following section. As expected, the image highlights that *Firmicutes* and *Bacteroides* are the two most abundant taxa in the healthy gut microbiome (David et al. 2014; Wu et al. 2011). The *Bacteroidetes* phylum contains many clades enriched in the HMP dataset, while *Firmicutes* show higher abundances for MetaHIT samples. GraPhlAn can thus serve as a visual tool for inspecting specific significant differences between conditions or cohorts.

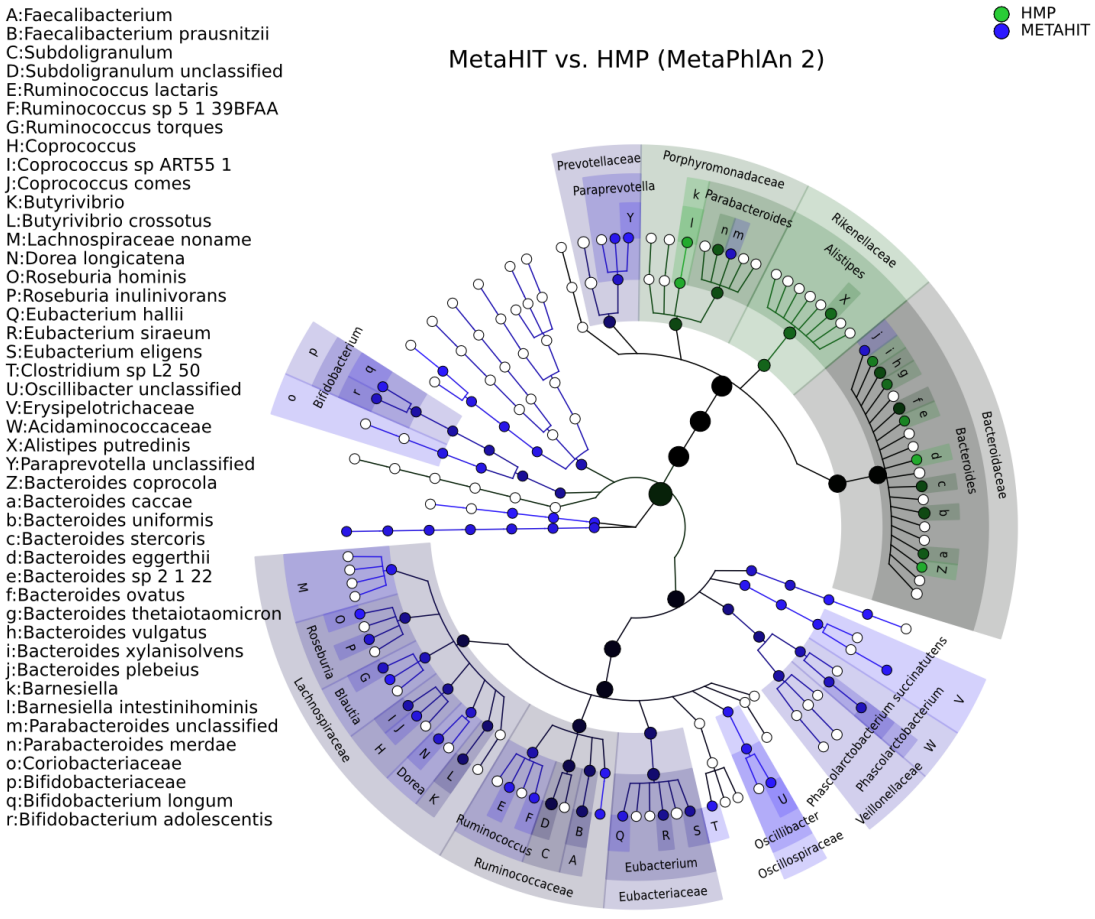

Figure 3. Taxonomic comparison between HMP and MetaHIT stool samples. *The taxonomic cladogram shows a comparison between the MetaHIT and HMP studies limited to samples from the gut (for the latter) and from healthy subjects (for the former). This image has been generated by GraPhlAn using input files from the supporting "export2graphlan" script (see "Materials and Methods") applied on the output of MetaPhlAn2 (Segata et al. 2012b) and LEfSe (Segata et al. 2011). Colors distinguish between HMP (green) and MetaHIT (blue), while the intensity reflects the LDA score, an indicator of the effect sizes of the significant differences. The size of the nodes correlates with their relative and logarithmically scaled abundances. Data used for this image is available as indicated under "Datasets used" paragraph in the "Materials and Methods" section.*

Functional ontologies can be represented by GraPhlAn in a similar way and provide complementary features to the types of taxonomic analyses shown above. Metabolic profiles quantified by HUMAnN (Abubucker et al. 2012) using KEGG (Kanehisa et al. 2014) from the same set of HMP and MetaHIT samples are again contrasted on multiple functional levels in **Fig. 4**. The tree highlights three different broad sets of metabolic pathways: Environmental Information Processing, Genetic Information Processing, and Metabolism, with the last being the largest subtree. More specific metabolic functions are specifically enriched in the HMP cohort, such as Glycolysis and the Citrate cycle, or in the MetaHIT cohort, such as Sulfur Metabolism and Vitamin B6 Metabolism. This illustrates GraPhlAn's use with different types of data, such as functional trees in addition to taxonomies or phylogenies. By properly configuring input parameters of *export2graphlan*, we automatically obtained both **Fig. 3** and **Fig. 4** (bash scripts used for these operations are available in the GraPhlAn software repository).

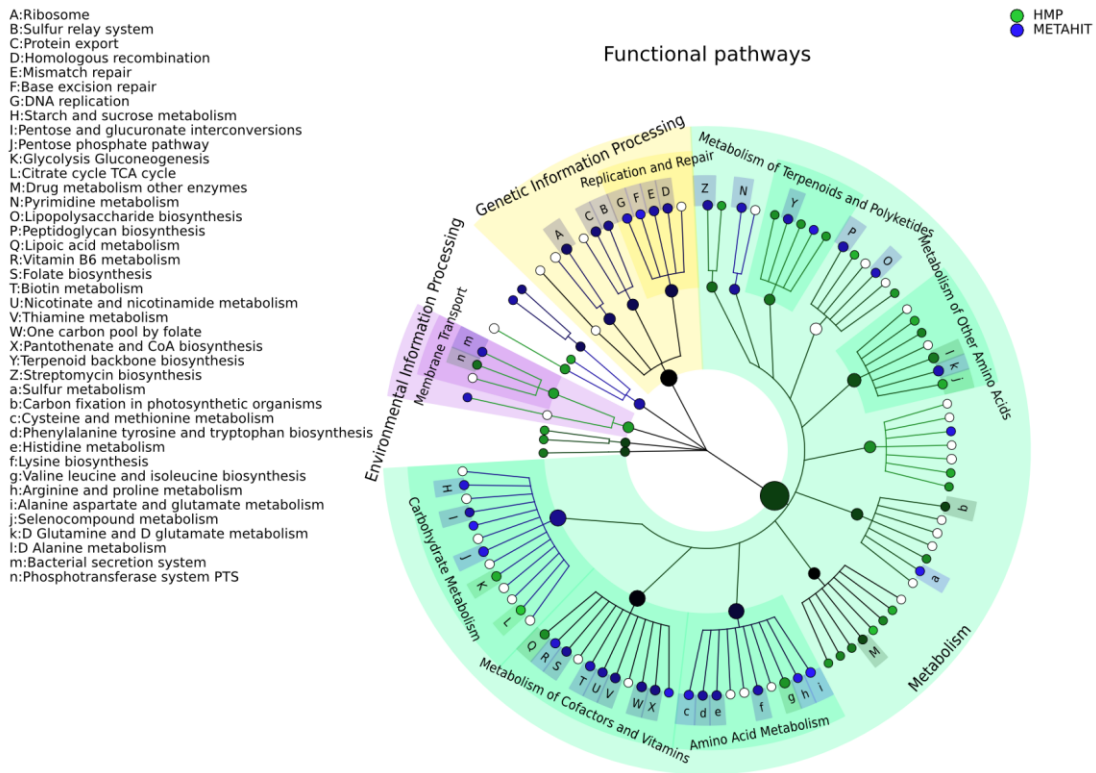

Figure 4. Comparison of microbial community metabolic pathway abundances between HMP and MetaHIT. *Comparison of functional pathway abundances from the HMP (green) and MetaHIT (blue). This is the functional counterpart of the plot in* ***Fig. 3*** *and was obtained applying GraPhlAn on HUMAnN (Abubucker et al. 2012) metabolic profiling. The intensity of the color represents the LDA score, and the sizes of the nodes are proportional to the pathway relative abundance estimated by HUMAnN. Three major groups are automatically highlighted by specifying them to the export2graphlan script: Environmental Information Processing, Genetic Information Processing, and Metabolism. Data used for this image is available as indicated under "Datasets used" paragraph in "Materials and Methods" section.*

Reproducible integration with existing analysis tools and pipelines

Graphical representations are usually a near- final step in the complex computational and metagenomic pipelines, and automating their production is crucial for convenient but reproducible analyses. To this end, GraPhlAn has been developed with command-driven automation in mind, as well as flexibility in the input "annotation file" so as to be easily generated by automated scripts. Depending on the specific analysis, these scripts can focus on a diverse set of commands to highlight the features of interest. Despite this flexibility, we further tried to ease the integration of GraPhlAn by providing automatic offline conversions for some of the available metagenomic pipelines and by embedding it into the well-established Galaxy web framework (Blankenberg et al. 2010; Giardine et al. 2005; Goecks et al. 2010).

In order to automatically generate GraPhlAn plots from a subset of available shotgun metagenomic tools comprising MetaPhlAn (for taxonomic profiling), HUMAnN (for metabolic profiling), and LEfSe (for biomarker discovery), we developed a script named "export2graphlan" able to convert the outputs of these tools into GraPhlAn input files as schematized in **Fig. 5**. This conversion software is also meant to help biologists by providing initial, automated input files for GraPhlAn that can then be manually tweaked for specific needs such as highlighting clades of particular interest. The export2graphlan framework can further accept the widely adopted BIOM format, both versions 1 and 2 (McDonald et al. 2012). This makes it possible to readily produce GraPhlAn outputs from other frameworks such as QIIME (Caporaso et al. 2010) and mothur (Schloss et al. 2009) for 16S rRNA sequencing studies.

A web-based deployment of the GraPhlAn application is available to the public via Galaxy at http://huttenhower.sph.harvard.edu/galaxy/. The Galaxy interface of GraPhlAn consists of four processing modules: (1) *Upload file*, that manages the upload of the input data into Galaxy; (2) *GraPhlAn Annotate Tree*, which allows the user to specify the annotations that will be applied to the final image; (3) *Add Rings to tree*, an optional step to select an already uploaded file in Galaxy that will be used as an annotation file for the external rings; and (4) *Plot tree*, that sets some image parameters such as the size, the resolution, and the output format.

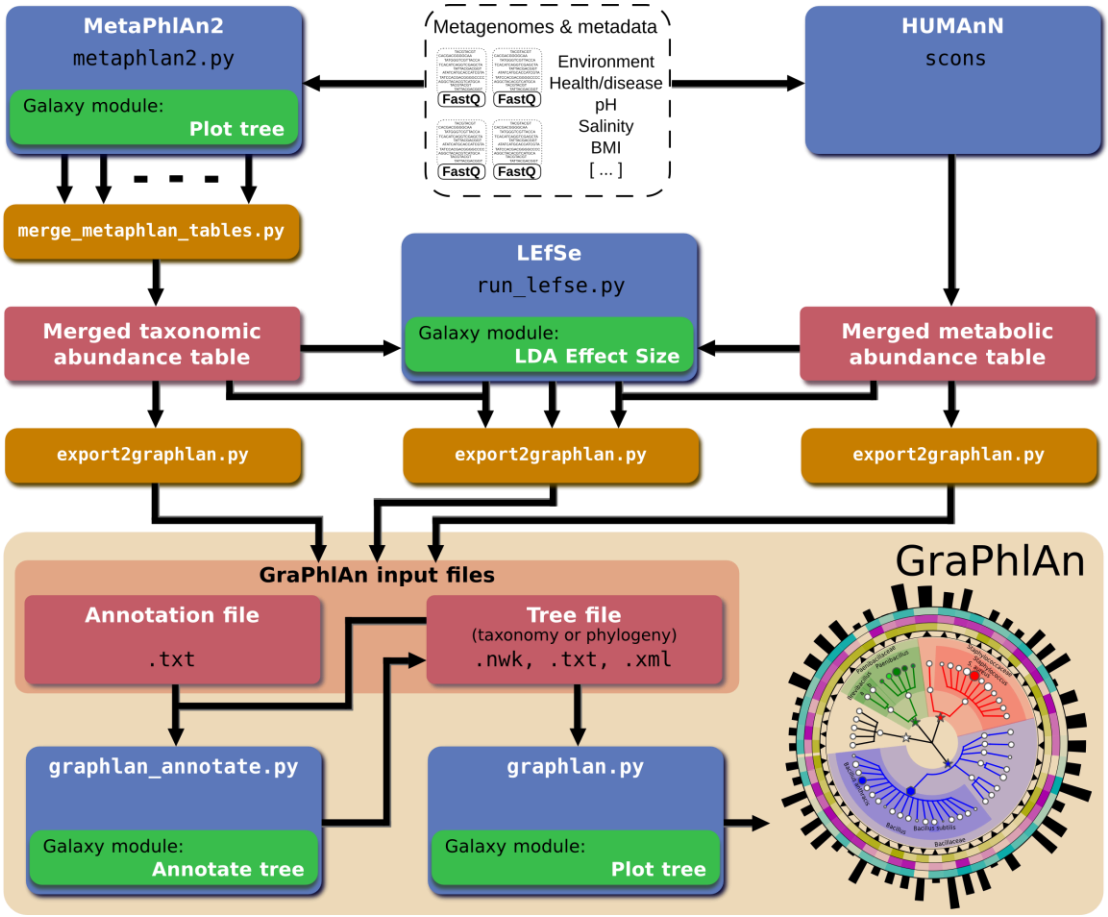

Figure 5. Integration of GraPhlAn into existing analyses pipelines. *We developed a conversion framework called "export2graphlan" that can deal with several output formats from different analysis pipelines, generating the necessary input files for GraPhlAn. Export2graphlan directly supports MetaPhlAn2, LEfSe, and HUMAnN output files. In addition, it can also accept BIOM files (both version 1 and 2), making GraPhlAn available for tools supporting this format including the QIIME and mothur systems. The tools can be ran on local machine as well as through the Galaxy web system using the modules reported in green boxes.*

Conclusions

We present GraPhlAn, a new method for generating high-quality circular phylogenies potentially integrated with diverse, high-dimensional metadata. We provided several examples showing the application of GraPhlAn to phylogenetic, functional, and taxonomic summaries. The system has already been used for a variety of additional visualization tasks, including highlighting the taxonomic origins of metagenomic biomarkers (Segata et al. 2012a; Segata et al. 2011; Shogan et al. 2014; Xu et al. 2014), exposing specific microbiome metabolic enrichments within a functional ontology (Abubucker et al. 2012; Sczesnak et al. 2011), and representing 16S rRNA sequencing results (Ramirez et al. 2014). GraPhlAn is, however, not limited to microbiome data and has additionally been applied to animal and plant taxonomies (Tree of Sex Consortium 2014) and to large prokaryotic phylogenies built using reference genomes (Baldini et al. 2014; Chai et al. 2014; Langille et al. 2013; Segata et al. 2013).

Compared to the other existing state-of-the-art approaches such as Krona (Ondov et al. 2011) and iTOL (Letunic & Bork 2007; Letunic & Bork 2011), GraPhlAn provides greater flexibility, configuration, customization, and automation for publication reproducibility. It is both easily integrable into automated computational pipelines and can be used conveniently online through the Galaxy-based web interface. The software is available open-source, and the features highlighted here illustrate a number of ways in which its visualization capabilities can be integrated into microbial and community genomics to display large tree structures and corresponding metadata.

Materials & Methods

Implementation strategy

GraPhlAn is composed by two Python modules: one for drawing the image and one for adding annotations to the tree. GraPhlAn exploits the annotation file to highlight and personalize the appearance of the tree and of the associated information. The annotation file does not perform any modifications to the structure of the tree, but it just changes the way in which nodes and branches are displayed. Internally, GraPhlAn uses the matplotlib library (Hunter 2007) to perform the drawing functions.

The export2graphlan module

Export2graphlan is a framework to easily integrate GraPhlAn into already existing bioinformatics pipelines. Export2graphlan makes use of two external libraries: the pandas python library (McKinney 2012) and the BIOM library, only when BIOM files are given as input.
Export2graphlan can take as input two files: the result of the analysis of MetaPhlAn (either version 1 or 2) or HUMAnN, and the result of the analysis of LEfSe. At least one of these two input files is mandatory. Export2graphlan will then produce a tree file and an annotation file that can be used with GraPhlAn. In addition, export2graphlan can take as input a BIOM file (either version 1 or 2). Export2graphlan performs an analysis on the abundance values and, if present, on the LDA score assigned by LEfSe, to annotate and highlight the most abundant clades and the ones found to be biomarkers. Through a number of parameters the user can control the annotations produced by export2graphlan.

Datasets used

The genomic data used for the Tree of Life in Figure 2 was obtained from the Integrated Microbial Genomes (IMG) data management system of the U.S. Department of Energy Joint Genome Institute (DOE JGI) 2.0 dataset (http://jgi.doe.gov/news_12_1_06/). From the KEGG database (Kanehisa & Goto 2000; Kanehisa et al. 2014) we focused on the following modules: M00082, M00083, M00086, M00087, M00088, M00157, and M00159.

In Figure 3, to comprehensively characterize the asymptomatic human gut microbiota, we combined 224 fecal samples (>17 million reads) from the Human Microbiome Project (HMP) (Human Microbiome Project 2012a; Human Microbiome Project 2012b) and the MetaHIT (Qin et al. 2010) projects, two of the largest gut metagenomic collections available. The taxonomic profiles were

obtained by applying MetaPhlAn2. The 139 fecal samples from the HMP can be accessed at http://hmpdacc.org/HMASM/, whereas the 85 fecal samples from MetaHIT were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/, study accession number ERP000108).

The functional profiles used in Figure 4 are the reconstruction of the metabolic activities of microbiome communities. The HUMAnN pipeline (Abubucker et al. 2012) infers community function directly from short metagenomic reads, using the KEGG ortholog (KO) groups. HUMAnN was run on the same samples of Figure 3. The dataset is available on-line at http://www.hmpdacc.org/HMMRC/

The dataset of supplementary Figure S1 refer to a 16S rRNA amplicon experiment. Specifically, it consists of 454 FLX Titanium sequences spanning the V3 to V5 variable regions, obtained from 24 healthy samples (12 male and 12 female) for a total of 301 samples. Detailed protocols used for enrollment, sampling, DNA extraction, 16S amplification and sequencing are available on the Human Microbiome Project Data Analysis and Coordination Center website HMP Data Analysis and Coordination Center (http://www.hmpdacc.org/tools_protocols/tools_protocols.php). This data are pilot samples from the HMP project (Segata et al. 2011).

In the supplementary Figure S2 we used the saliva microbiome profiles obtained by 16S rRNA sequencing on the IonTorrent platform (amplifying the hypervariable region V3). The dataset comprises a total of 13 saliva samples from healthy subjects as described in (Dassi et al. 2014) and it is available in the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/sra).

For the supplementary Figure S3 data represent the temporal dynamics of the human vaginal microbiota, and were taken from the study of (Gajer et al. 2012). Data were obtained by 16S rRNA using the 454 pyrosequencing technology (sequecing the V1 and V2 hypervariable regions). The dataset is composed of samples from 32 women that self-collected samples twice a week for 16 weeks.

Availability, dependences, and user support

GraPhlAn is freely available and released open-source in Bitbucket (https://bitbucket.org/nsegata/graphlan) with a set of working examples and a complete tutorial that guides users throughout its functionality (https://bitbucket.org/nsegata/graphlan/wiki/Home). GraPhlAn uses the matplotlib library (Hunter 2007). GraPhlAn is also available via a public Galaxy instance at http://huttenhower.sph.harvard.edu/galaxy/

Export2graphlan is freely available and released open-source in Bitbucket (https://bitbucket.org/CibioCM/export2graphlan) along with a number of examples helpful for testing if everything is correctly configured and installed. The export2graphlan repository is also present as a sub-repository inside the GraPhlAn repository. The export2graphlan module exploits the pandas library (McKinney 2012) and the BIOM library (McDonald et al. 2012).

Both GraPhlAn and export2graphlan are supported through the Google group "GraPhlAn-users" (https://groups.google.com/forum/#!forum/graphlan-users), available also as a mailing list at: graphlan-users@googlegroups.com.

Acknowledgements

We would like to thank the members of the Segata and Huttenhower labs for helpful suggestions, the WebValley team and participants for inspiring comments and tests, and the users that tried the alpha version of GraPhlAn providing invaluable feedback to improve the software.

References

Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methe B, Schloss PD, Gevers D, Mitreva M, and Huttenhower C. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 8:e1002358.

Baldini F, Segata N, Pompon J, Marcenac P, Robert Shaw W, Dabire RK, Diabate A, Levashina EA, and Catteruccia F. 2014. Evidence of natural Wolbachia infections in field populations of Anopheles gambiae. *Nat Commun* 5:3985.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, and Taylor J. 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* Chapter 19:Unit 19 10 11-21.

Caporaso J, Kuczynski J, Stombaugh J, Bittinger K, Bushman F, Costello E, Fierer N, Pena A, Goodrich J, Gordon J, Huttley G, Kelley S, Knights D, Koenig J, Ley R, Lozupone C, McDonald D, Muegge B, Pirrung M, Reeder J, Sevinsky J, Turnbaugh P, Walters W, Widmann J, Yatsunenko T, Zaneveld J, and Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335 - 336.

Chai J, Kora G, Ahn T-H, Hyatt D, and Pan C. 2014. Functional phylogenomics analysis of bacteria and archaea using consistent genome annotation with UniFam. *BMC evolutionary biology* 14:207.

Cross RL, and Müller V. 2004. The evolution of A-, F-, and V-type ATP synthases and ATPases: reversals in function and changes in the H+/ATP coupling ratio. *FEBS Letters* 576:1-4.

Dassi E, Ballarini A, Covello G, Quattrone A, Jousson O, De Sanctis V, Bertorelli R, Denti MA, and Segata N. 2014. Enhanced microbial diversity in the saliva microbiome induced by short-term probiotic intake revealed by 16S rRNA sequencing on the IonTorrent PGM platform. *Journal of Biotechnology*.

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, and Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505:559-563.

Gajer P, Brotman RM, Bai G, Sakamoto J, Schutte UM, Zhong X, Koenig SS, Fu L, Ma ZS, Zhou X, Abdo Z, Forney LJ, and Ravel J. 2012. Temporal dynamics of the human vaginal microbiota. *Sci Transl Med* 4:132ra152.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, and Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 15:1451-1455.

Goecks J, Nekrutenko A, Taylor J, and Galaxy T. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11:R86.

Han MV, and Zmasek CM. 2009. phyloXML: XML for evolutionary biology and comparative genomics. *BMC bioinformatics* 10:356.

Human Microbiome Project C. 2012a. A framework for human microbiome research. *Nature* 486:215-221.

Human Microbiome Project C. 2012b. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207-214.

Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9:90-95.

Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, Fitzgerald MG, and Fulton RS. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207-214.

Kanehisa M, and Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27-30.

Kanehisa M, Goto S, Sato Y, Furumichi M, and Tanabe M. 2011. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*:gkr988.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, and Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42:D199-205.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, and Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31:814-821.

Letunic I, and Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127-128.

Letunic I, and Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39:W475-478.

Maddison DR, Swofford DL, and Maddison WP. 1997. NEXUS: an extensible file format for systematic information. *Systematic Biology* 46:590-621.

McDonald D, Clemente J, Kuczynski J, Rideout J, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, and Caporaso J. 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* 1:7.

McKinney W. 2012. pandas: a Foundational Python Library for Data Analysis and Statistics. *O'Reilly Media, Inc*.

Mulkidjanian AY, Makarova KS, Galperin MY, and Koonin EV. 2007. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat Rev Micro* 5:892-899.

Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, and Li P. 2004. Taverna: a tool for the composition

and enactment of bioinformatics workflows. *Bioinformatics* 20:3045-3054.

Ondov BD, Bergman NH, and Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics* 12:385.

Qin J, Li R, Raes J, Arumugam M, Burgdorf K, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende D, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J, Hansen T, Le Paslier D, Linneberg A, Nielsen H, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Weissenbach J, Bork P, Ehrlich S, Wang J, and Consortium M. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59 - 65.

Ramirez KS, Leff JW, Barberán A, Bates ST, Betley J, Crowther TW, Kelly EF, Oldfield EE, Shaw EA, and Steenbock C. 2014. Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B: Biological Sciences* 281:20141988.

Razin S. 1992. Peculiar properties of mycoplasmas: The smallest self-replicating prokaryotes. *FEMS Microbiology Letters* 100:423-431.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, and Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537-7541.

Sczesnak A, Segata N, Qin X, Gevers D, Petrosino JF, Huttenhower C, Littman DR, and Ivanov, II. 2011. The genome of th17 cell-inducing segmented filamentous bacteria reveals extensive auxotrophy and adaptations to the intestinal environment. *Cell Host Microbe* 10:260-272.

Segata N, Bornigen D, Morgan XC, and Huttenhower C. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 4:2304.

Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, Huttenhower C, and Izard J. 2012a. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol* 13:R42.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, and Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. *Genome Biol* 12:R60.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, and Huttenhower C. 2012b. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* 9:811-814.

Shogan BD, Smith DP, Christley S, Gilbert JA, Zaborina O, and Alverdy JC. 2014. Intestinal anastomotic injury alters spatially defined microbiome composition and function. *Microbiome* 2:35.

Tree of Sex Consortium. 2014. Tree of Sex: A database of sexual systems. *Scientific Data* 1.

Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R,

Nessel L, Li H, Bushman FD, and Lewis JD. 2011. Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334:105-108.

Xu Z, Hansen MA, Hansen LH, Jacquiod S, and Sorensen SJ. 2014. Bioinformatic approaches reveal metagenomic characterization of soil microbial community. *PLoS One* 9:e93445.

Supplementary materials

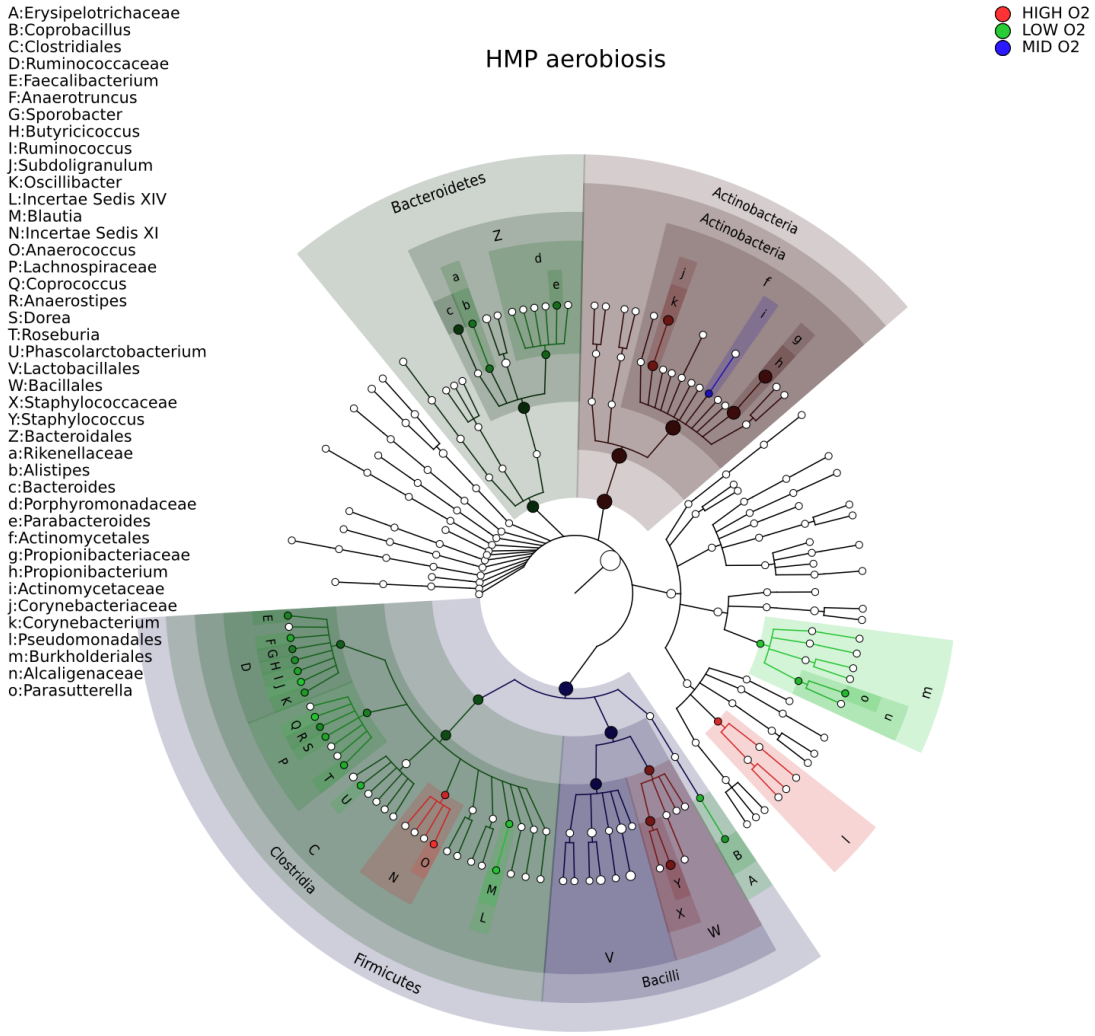

Figure S1. Aerobiosis analysis under aerobic, anaerobic, and microaerobic conditions. *The cladogram shows the aerobiosis analysis of the HMP data in three O_2-dependent classes: aerobic (red), anaerobic (blue), and microaerobic (green). The node size reflects the abundance level of each clades, colors are assigned accordingly to one of the three classes, while the lightness intensity of colors respect the LDA score assigned by LEfSe to biomarkers. Data used for this image is available as indicated under "Datasets used" paragraph in "Materials and Methods" section.*

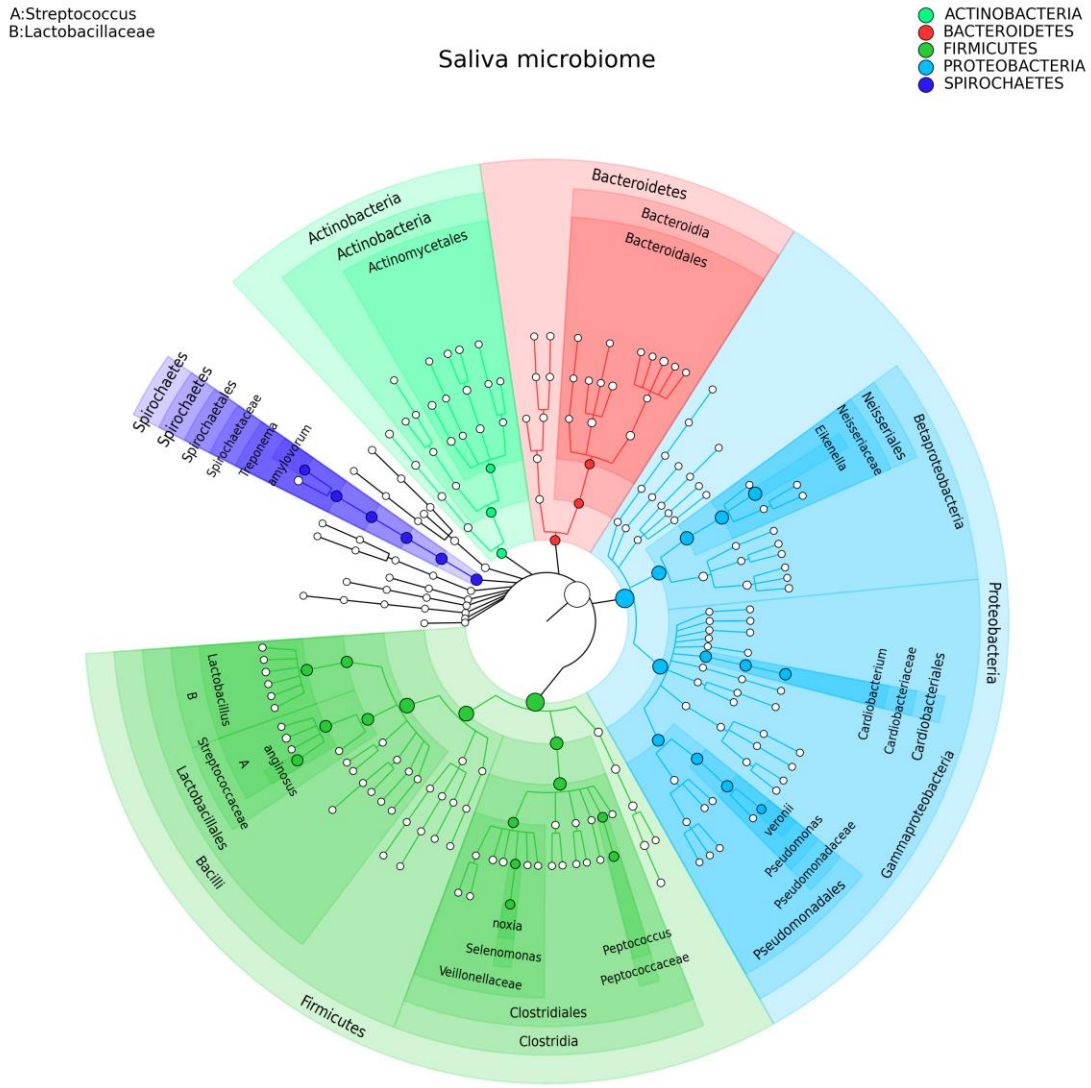

Figure S2. Caracterization of the saliva microbiome. *This image shows the taxonomic enrichment of the first saliva microbiome sequenced using IonTorrent PGM technology. We exploit export2graphlan capability of handle BIOM files to generate the annotation and tree files for GraPhlAn. Data used for this image is available as indicated under "Datasets used" paragraph in "Materials and Methods" section.*

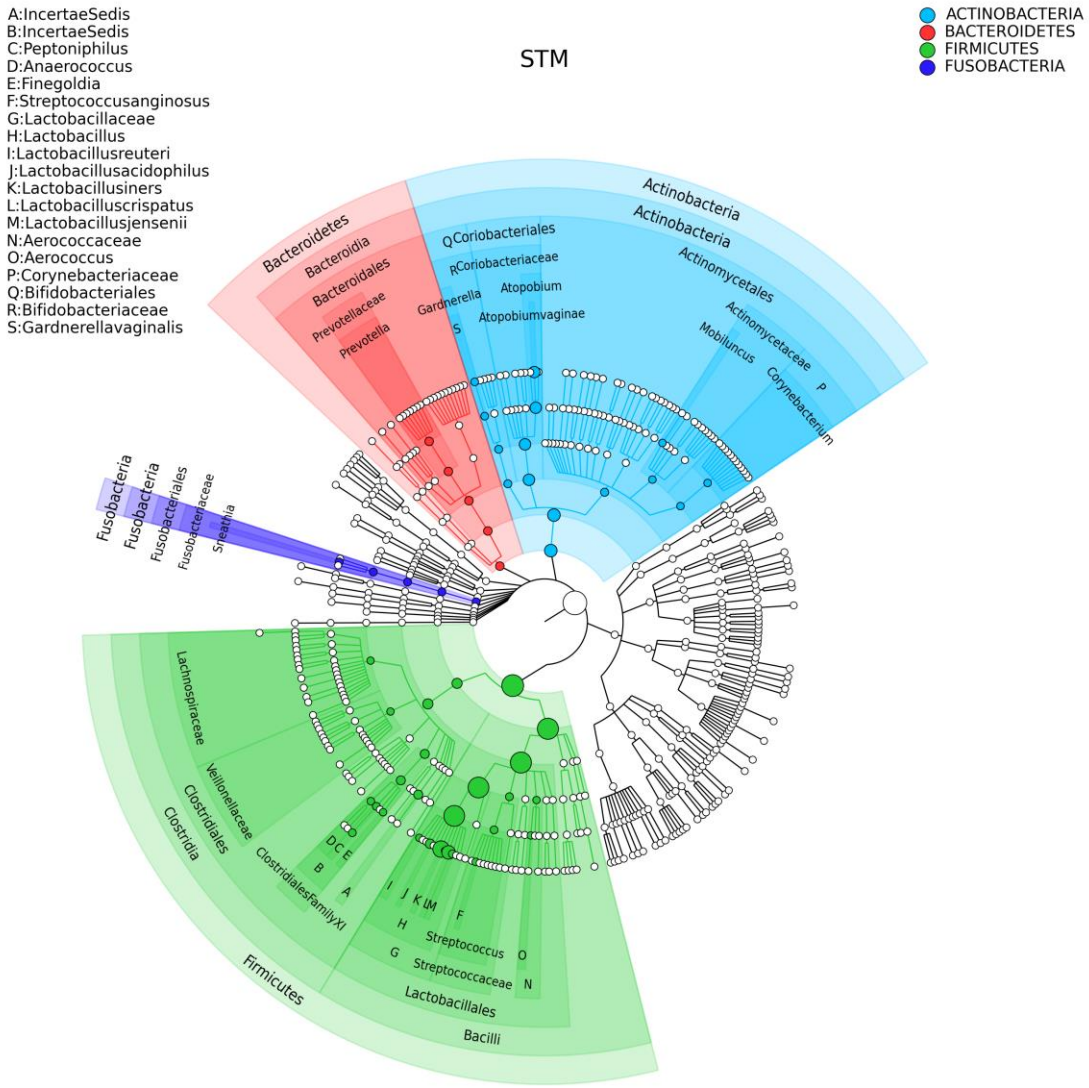

Figure S3. Characterization of temporal dynamics of the human vaginal microbiota. *We take the data as a BIOM file from the (Gajer et al. 2012) study. We use export2graphlan to generate the needed files for plotting the circular tree with GraPhlAn. Data used for this image is available as indicated under "Datasets used" paragraph in "Materials and Methods" section.*