frontiers in Genetics

ORIGINAL RESEARCH
published: 09 April 2021
doi: 10.3389/fgene.2021.644516

G2S: A New Deep Learning Tool for Predicting Stool Microbiome Structure From Oral Microbiome Data

Simone Rampelli[1]*, Marco Fabbrini[1,2], Marco Candela[1], Elena Biagi[1], Patrizia Brigidi[2] and Silvia Turroni[1]

[1] Unit of Microbiome Science and Biotechnology, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [2] Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy

OPEN ACCESS

Edited by:
Saumya Patel,
Gujarat University, India

Reviewed by:
Francesco Asnicar,
University of Trento, Italy
Khanh N. Q. Le,
Taipei Medical University, Taiwan

***Correspondence:**
Simone Rampelli
simone.rampelli@unibo.it

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 21 December 2020
Accepted: 09 March 2021
Published: 09 April 2021

Citation:
Rampelli S, Fabbrini M, Candela M, Biagi E, Brigidi P and Turroni S (2021) G2S: A New Deep Learning Tool for Predicting Stool Microbiome Structure From Oral Microbiome Data.
Front. Genet. 12:644516.
doi: 10.3389/fgene.2021.644516

Deep learning methodologies have revolutionized prediction in many fields and show the potential to do the same in microbial metagenomics. However, deep learning is still unexplored in the field of microbiology, with only a few software designed to work with microbiome data. Within the meta-community theory, we foresee new perspectives for the development and application of deep learning algorithms in the field of the human microbiome. In this context, we developed G2S, a bioinformatic tool for taxonomic prediction of the human fecal microbiome directly from the oral microbiome data of the same individual. The tool uses a deep convolutional neural network trained on paired oral and fecal samples from populations across the globe, which allows inferring the stool microbiome at the family level more accurately than other available approaches. The tool can be used in retrospective studies, where fecal sampling was not performed, and especially in the field of paleomicrobiology, as a unique opportunity to recover data related to ancient gut microbiome configurations. G2S was validated on already characterized oral and fecal sample pairs, and then applied to ancient microbiome data from dental calculi, to derive putative intestinal components in medieval subjects.

Keywords: gut microbiome, oral microbiome, deep learning, microbiome, paleomicrobiology

INTRODUCTION

Deep learning is increasingly being used to make inference on large and complex data. Unlike traditional algorithms, in which the expertise and rules are already coded, deep learning algorithms are built to automatically detect patterns in data (Murphy, 2012; Bishop, 2016), also embedding the computation of variables into the models themselves to yield end-to-end models (Goodfellow et al., 2016). In particular, the construction and training of deep learning algorithms have been enabled by the increasing availability of big data and the rapid growth in the number and size of public available databases. So far, deep neural networks have been key to advances in modern artificial intelligence, with applications such as facial recognition, speech recognition and self-driving vehicles. More recently, new applications have been pioneered in the fields of molecular biology and metagenomics. Indeed, the same deep learning approaches are beginning to be applied to genetics, agriculture and medicine (Alipanahi et al., 2015; Leung et al., 2016; Ching et al., 2018; Demirci et al., 2018; Wainberg et al., 2018; Webb, 2018; Le, 2019; Le and Huynh, 2019; Le et al., 2019; Quang and Xie, 2019). However, deep learning is still unexplored in the field of microbial metagenomics, with only a few approaches suitable for microbiome data (Geman et al., 2016; Reiman et al., 2017; Galkin et al., 2020), and a huge untapped potential yet unexplored.

The human microbiome, i.e., the sum of the different microbial ecosystems that colonize the niches of the human body, plays an important role in human physiology and its dysbiotic variations can severely impact our health (Kau et al., 2011). For example, shifts in the composition of microbial communities inhabiting the oral cavity and gastrointestinal tract have been associated with the onset and/or progression of various conditions, such as periodontitis (Griffen et al., 2012) and other modern chronic disorders, including inflammatory bowel disease (Glassner et al., 2020), obesity (Rampelli et al., 2018), cardiovascular disease (Pietiäinen et al., 2018) and some forms of cancer (Helmink et al., 2019; Karpiński, 2019; Wong and Yu, 2019). The importance of the human microbiome in health and disease makes it imperative to understand the drivers of its variation. In this context, a new frontier is represented by the meta-community theory, according to which human symbiont microbial ecosystems are in intimate connection, showing reciprocal influences and exchanges (Koskella et al., 2017; Miller et al., 2018). Supporting a meta-community view of human microbial ecology, a close link between oral and intestinal microbiomes has recently been hypothesized, with the former reflecting changes in the latter, in both healthy and diseased individuals (Bajaj et al., 2015; Iwauchi et al., 2019; Prodan et al., 2019; Schmidt et al., 2019). Another scale of human microbiome variation is represented by its change across the evolutionary timeline. In particular, a large body of literature indicates that the current human gut microbiome has evolved toward at least two different configurations, rural and urban, both associated with the corresponding subsistence strategy. Compared to the first, generally considered as the pristine human gut microbiome, the urban configuration is characterized by an overall compression of microbial biodiversity, a wholescale loss of commensal microbial groups, and an increased presence of genes related to antibiotic resistance and xenobiotics metabolism (Yatsunenko et al., 2012; Schnorr et al., 2014; Obregon-Tito et al., 2015; Rampelli et al., 2015; Ayeni et al., 2018; Jha et al., 2018). These changes, collectively referred to as "microbiota insufficiency syndrome" (Sonnenburg and Sonnenburg, 2019), have been identified as contributing factors to the rise in chronic inflammatory non-communicable diseases. However, mainly due to the paucity of ancient stool samples, the truly ancestral human gut microbiome is still unknown and the evolutionary trajectories and drivers leading to its contemporary configurations have yet to be described, leaving important gaps in knowledge of the gut microbiome-human host co-evolutionary trajectories. Contrary to ancient fecal samples, dental ones are more common and well preserved, allowing for the extraction of the ancient oral microbiome from ancient DNA preserved in dental tartar. Consistent with the meta-community vision, the ancient configuration of the oral microbiome can somehow mirror the structural features of the intestinal one due to the intrinsic connections between the two ecosystems. In this scenario, here we developed a new deep learning-based tool, G2S, which infers the gut microbiome configuration from the oral microbiome data of a given individual. G2S is based on a model trained and tested on a total of 305 and 79 paired samples of oral and stool microbiome, respectively, retrieved from multiple studies with individuals of various geographical origins, including United States, Fiji, United Kingdom, and European countries (The Human Microbiome Project Consortium, 2012; Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). Our approach may be relevant for predicting the gut microbiome configuration when fecal data are not available, and particularly suitable for human archeological records, where coprolites and fecal sediments are indeed rare compared to dental calculi and other human remains.

MATERIALS AND METHODS

G2S software is built in an R environment, using the R packages "base," "stats," and "keras," containing "tensorflow." The G2S source code is available on the website https://github.com/simonerampelli/g2s and it can be run using a command line interface on computer with Windows, Linux and OS X as the operating system.

The G2S tool was trained and tested on a total of 768 paired samples (i.e., oral and stool samples from the same 384 individuals), including samples from 171 healthy adults from United States, 7 from Italy, 29 from Sweden, 37 from United Kingdom, and 140 from Fiji (The Human Microbiome Project Consortium, 2012; Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). Eighty% of the subjects were used for the training dataset and 20% for the test dataset, without overlapping to avoid overfitting. Both 16S rRNA gene reads and shotgun metagenomics sequences were used, analyzed by the QIIME 2 pipeline (Bolyen et al., 2019) or the MetaPhlAn2 software (Truong et al., 2015), respectively.

The performance of G2S in predicting fecal microbiome configuration from the same individual's oral microbiome sample was compared with that of other available approaches, including Random Forest (Breiman, 2001) and a stochastic algorithm, i.e., a customized method that generates mock profiles of the stool microbiome by randomly imputing the abundances of bacterial families in the range of the training dataset (see **Supplementary File 1** for script source).

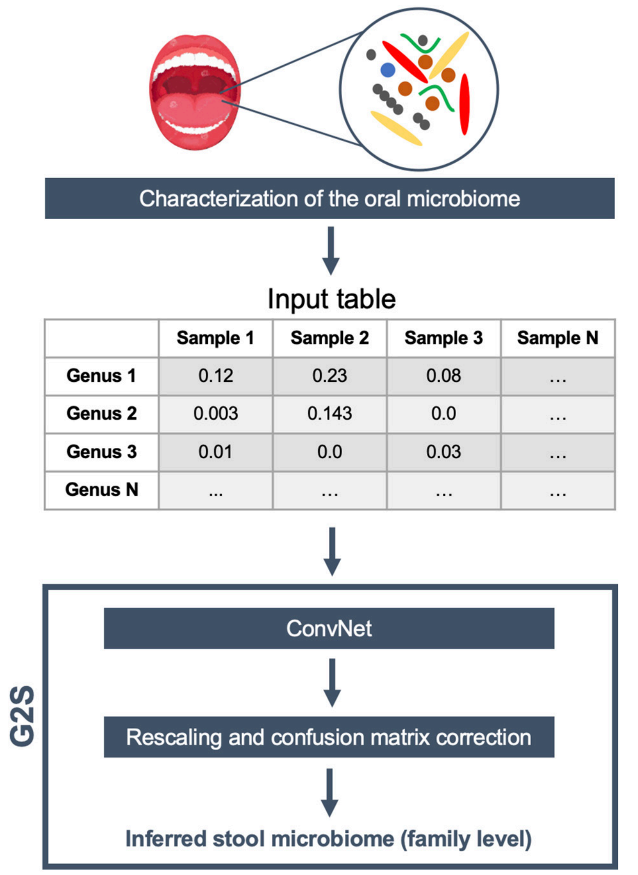
Microbiome data from dental calculi of 4 adult human skeletons (G12, B17, B61, and B78), characterized by sequencing the V5 and V6 regions of the 16S rRNA gene (8 samples in total) (Warinner et al., 2014), were used to illustrate the potential and results of G2S. No ethics committee approval was required to perform the analysis included in this study.

RESULTS

Implementation of the G2S Software

G2S adapted a deep convolutional neural network (ConvNet) to predict gut microbiome configurations from oral microbiome data. Several model architectures were tested in order to find the best performing algorithm, either by testing hidden layers with different number of units, and/or by adding a weight regularization step or a dropout procedure (data not shown). The final ConvNet was structured with two hidden layers, each with 50 units, and a final linear layer with 13 units and no activation function. We selected mean square error as the loss function, and mean absolute error as the metric to evaluate the differences between predictions and targets during training. In order to minimize overfitting problems due to the small number of samples within the dataset, we also included a weight regularization step, by adding to the loss function a cost associated with having high weights. The cost was proportional to the square of the weight coefficient value (L2 regularization or weight decay). Finally, to further prevent overfitting, dropout was applied to the first two layers, obtaining a better prediction and a significant reduction in losses and minimum absolute errors with a rate value of 0.5.

For ConvNet training and testing, we downloaded all available paired samples (i.e., gingival and stool samples from the same individual) from the HMP project (The Human Microbiome Project Consortium, 2012). In order to increase the generalization capability of our ConvNets, while minimizing geography-related bias (He et al., 2018*)*, we integrated our dataset with all available paired samples (i.e., oral and fecal samples) from healthy adults from other literature studies (Zaura et al., 2015*;* Brito et al., 2016*;* Russo et al., 2018), selecting both 16S rRNA gene and shotgun metagenomic datasets (see also **Supplementary Table 1**). Our final dataset included paired samples of 171 individuals from United States, 7 from Italy, 29 from Sweden, 37 from United Kingdom, and 140 from Fiji, for a total of 384 oral and 384 stool samples, divided into 528 16S rRNA gene and 240 shotgun fastq files. Specifically, 16S rRNA gene sequences were analyzed using the QIIME 2 pipeline (Bolyen et al., 2019) and the Greengenes database (DeSantis et al., 2006) in order to obtain the microbiome classification at different taxonomic levels. On the other hand, the shotgun metagenomic samples were analyzed by MetaPhlAn2 (Truong et al., 2015) using the default parameters. The genus-level abundance table of 384 oral microbiome samples was normalized feature-wise prior to its usage for deep learning. In particular, the data were centered on the mean of each specific genus and scaled according to their standard deviation. Only 50 genera present in more than 4 samples with relative abundance greater than 0.1% were retained for the analysis. The 12 bacterial families of the stool microbiome dataset with the highest contribution in terms of median relative abundance, including *Bacteroidaceae, Porphyromonadaceae, Lachnospiraceae, Ruminococcaceae, Veillonellaceae, Rikenellaceae, Alcaligenaceae, Streptococcaceae, Bifidobacteriaceae, Clostridiaceae, Prevotellaceae*, and *Erysipelotrichaceae*, were selected as features to be predicted by ConvNet analysis. An additional variable, called "Other" (i.e., the percentage remaining to reach 100%), was also considered a feature to be inferred. The training and test datasets were separated to contain 80 and 20% of all profiles, i.e., 305 and 79 paired oral and fecal samples, respectively. In order to better evaluate the model, we used a k-fold cross-validation approach with 4 partitions and 500 epochs. We got the best performance after the 151st epoch, with a mean absolute error of 4.1%. To increase the predictive performance of ConvNet, the results were then transformed as follows: (i) negative predictions were set to 0, and (ii) the sum of the value for each sample was rescaled to 100%. Finally, based on the results of the training dataset, we also built a confusion matrix to adjust the predictions of those families with recurring over- or underestimation. G2S includes all of these steps in a single R script, and requires only a relative abundance table of the oral microbiome (between 0 and 1) at the genus level with samples in the columns and


Characterization of the oral microbiome
Input table
G2S
ConvNet
Rescaling and confusion matrix correction
Inferred stool microbiome (family level)

	Sample 1	Sample 2	Sample 3	Sample N
Genus 1	0.12	0.23	0.08	...
Genus 2	0.003	0.143	0.0	...
Genus 3	0.01	0.0	0.03	...
Genus N	...	...	...	...

FIGURE 1 | G2S workflow. The input file is a genus-level relative abundance table (.tsv format), obtained from the characterization of human oral microbiome samples. The stool microbiome is inferred using a deep convolutional neural network (ConvNet) adjusted by a confusion matrix and rescaled to 100%. The results are tabulated as relative abundance.

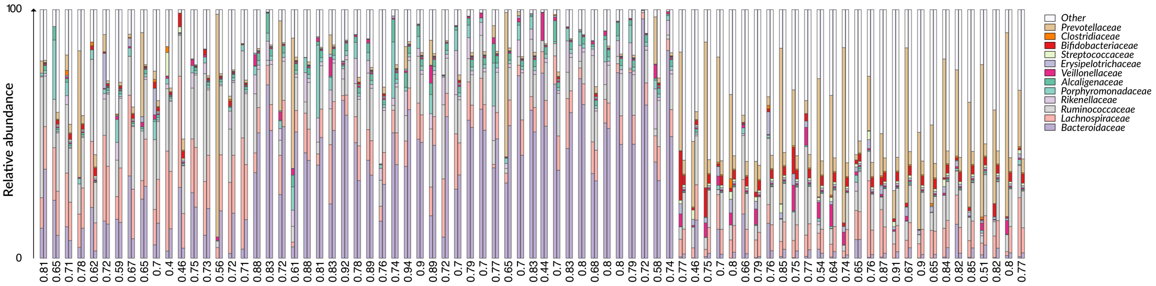


FIGURE 2 | Comparison between G2S predictions and real data from the test dataset. The family level bar plots of the 79 stool samples of the test dataset are visualized next to their inferred configurations obtained by G2S. Spearman correlation coefficients (r) are provided below each pair of bar plots. Samples are derived from the following studies: The Human Microbiome Project Consortium (2012), Zaura et al. (2015); Brito et al. (2016), Russo et al. (2018).

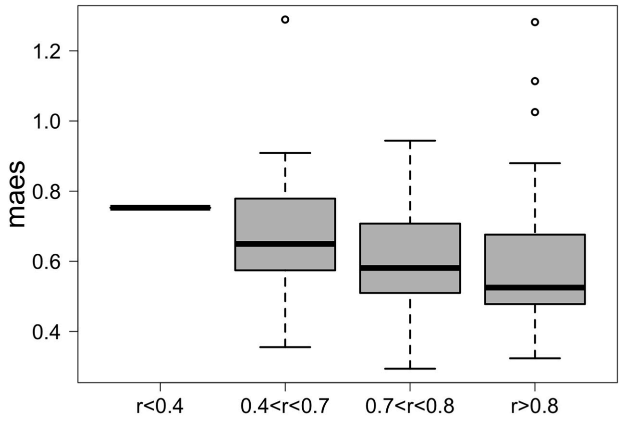


FIGURE 3 | G2S predictions are more accurate when the configurations to be inferred fall within the plane of variation of the training dataset. Box plots of the mean absolute error scaled to one standard deviation (maes) between the real stool microbiome configuration of the samples in the test dataset and the median configuration of the training dataset. Samples were divided into four groups based on the quality of the G2S predictions (i.e., the Spearman correlation coefficients between the real values and the inferred configurations).

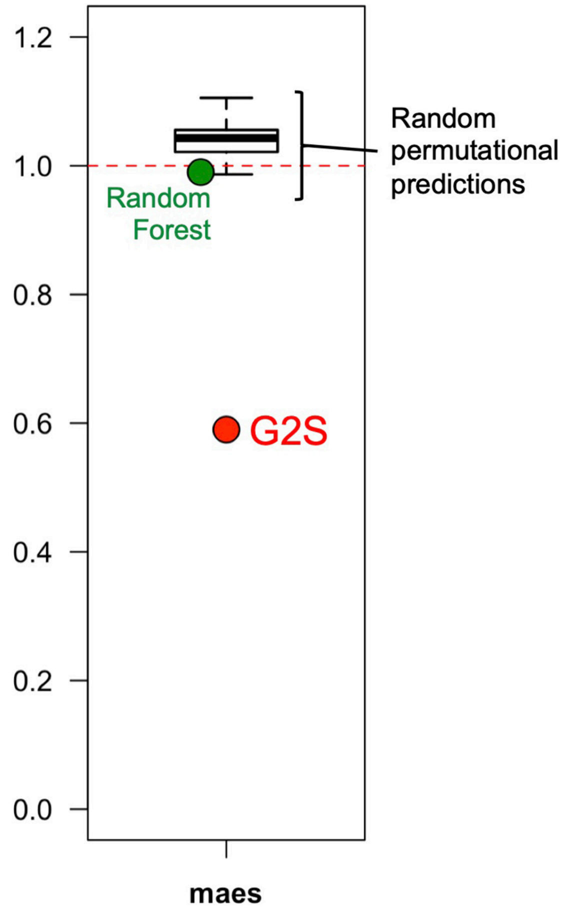


FIGURE 4 | G2S predicts the stool microbiome configuration with better performance than other methods. The mean absolute errors scaled to one standard deviation (maes) between the real data of the samples from the test dataset and the configurations inferred by G2S, Random Forest and a stochastic permutational method (100 predictions), are reported in the dot plot.

the full taxonomy following the Greengenes_05_2013 style in the rows as input file. For each sample analyzed, the predicted microbiome is summarized in a table as the relative abundance of the most abundant bacterial families. Additionally, histograms of the same families are provided, using the "graphics" and "base" R packages. The schematic overview of the G2S framework is provided in **Figure 1**.

Ascertaining the Performance of G2S on the Test Dataset

We first applied G2S to the test dataset to evaluate its cross-validated predictions. In particular, mean absolute errors for each family scaled to one standard deviation of real data (maes) < 1 were considered as reference parameters for a good quality of the prediction. As expected, G2S predicts relative abundances with an average maes of 0.59, ranging from the best score for *Bacteroidaceae* and *Erysipelotrichaceae* (maes = 0.46) to the worst case for *Ruminococcaceae* (maes = 0.77). To gain more insights into the predictive performance of G2S, we globally compared, sample by sample, the inferred microbiome configurations with

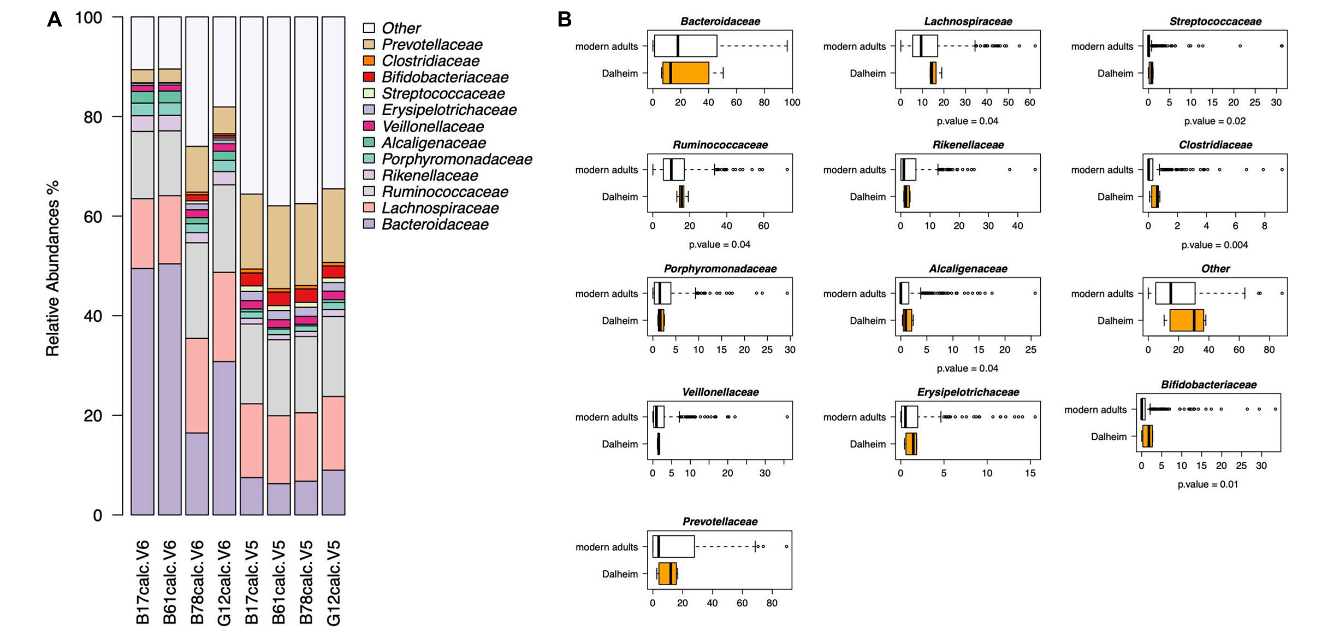


FIGURE 5 | Reconstructing the ancient stool microbiome of adult medieval individuals. **(A)** Bar plots of stool microbiome configurations inferred from 16S rRNA gene (V5 and V6 regions) sequencing data of ancient microbiomes (i.e., dental calculi from the medieval monastic site of Dalheim, Germany [ca. 950–1,200 CE]) (Warinner et al., 2014). **(B)** Comparison between the predicted ancient microbiome configurations and the modern stool microbiome of subjects from the dataset used to implement G2S (The Human Microbiome Project Consortium, 2012; Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). *P*-values were determined by Wilcoxon test.

real data by means of bar plots (**Figure 2**). Spearman correlations between predicted and actual microbiome profiles were used to evaluate predictions for each subject. In particular, we considered as excellent those predictions with r > 0.8 (52% of predictions), good those with r between 0.71 and 0.8 (29% of predictions), discrete with *r* between 0.41 and 0.7 (18% of predictions), and incorrect with $r \leq 0.4$ (1% of predictions). When we analyzed the single case in which G2S inferred an incorrect prediction, we found that the stool microbiome configuration was very peculiar, with the relative abundances of the two keystone bacterial families *Bacteroidaceae* and *Lachnospiraceae* not reaching 5% of relative abundance together (while generally dominant in the ecosystem). It is important to note that G2S worked correctly even when the stool microbiome configurations to be predicted were not so close to the median configuration of the training dataset (maes < 1 even when $r < 0.7$) (**Figure 3**). This was likely due to the large variation captured by the pool of microbiome configurations of the samples in the training dataset.

G2S showed a better mimicry of the relative abundance of microbiomes in the test dataset than other methods, including Random Forest and a stochastic method developed specifically for this comparison, which generates mock profiles of the stool microbiome in the range of the training dataset (**Figure 4**). Random Forest under- or overestimated bacterial families with a global maes of 0.99, ranging from 0.77 for *Bacteroidaceae* to 1.74 for *Streptococcaceae*. The performance of our custom predictor was even more inaccurate, with a total of 100 permutational predictions showing maes between 0.98 and 1.11 (mean = 1.05). The best performance of G2S in predicting the stool microbiome structure is probably due to the predictive power of deep learning that automatically detects patterns in the data, by also embedding the computation of variables into the models themselves to yield end-to-end models.

Case Study: Using G2S in Paleomicrobiology to Predict the Stool Microbiome Profile From Ancient Dental Calculi

In the second part of our analysis, we used G2S to infer the stool microbiome from oral microbiome data of four adult human skeletons with evidence of mild to severe periodontal disease, from the medieval monastic site of Dalheim, Germany (ca. 950–1,200 CE) (Warinner et al., 2014). G2S inferred the stool microbiome structure at the family level, estimating the abundance of the 13 features, i.e., the 12 bacterial families and the category "Other" including all other families (**Figure 5A**). Interestingly, *Bacteroidaceae, Lachnospiraceae, Ruminococcaceae*, and *Prevotellaceae* were the predicted dominant components in the feces of the four subjects, using both V5 and V6 regions as targets of the 16S rRNA gene (together their relative abundance ranged from 52 to 80%). On the other hand, the family *Clostridiaceae* showed the lowest relative abundance (<1%) in all eight samples. Significant differences in taxon relative abundance were found with respect to the stool microbiome of modern subjects from the dataset used to implement G2S, including higher relative abundance of *Ruminococcaceae*, *Lachnospiraceae, Streptococcaceae, Alcaligenaceae, Clostridiaceae*, and *Bifidobacteriaceae* in the predicted ancient microbiome configurations (p-value < 0.05, Wilcoxon test) (**Figure 5B**). This is not unexpected given the profoundly different lifestyles of ancient individuals of the Middle Ages and modern people, in terms of diet, contact with the environment and sanitization practices (The Human Microbiome Project Consortium, 2012; Warinner et al., 2014; Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). Future studies in larger worldwide cohorts, including paired samples of oral and intestinal microbiome, are needed to refine the accuracy of the G2S software and predict a higher number of bacterial families as well as possibly taxa at different phylogenetic levels, possibly including genera and species.

DISCUSSION

G2S is specifically designed to predict the structure of the human stool microbiome from oral microbiome data. In particular, it uses relative abundance tables of the oral microbiome generated by next-generation sequencing, and a deep learning approach that allows high-speed prediction of the stool microbiome without any downstream process. It could be used with both modern and ancient samples, providing a good prediction of the fecal microbiome with a net saving of time and costs. This is particularly relevant in the context of paleomicrobiology, where human coprolites and fecal sediments are very rare compared to dental calculi. However, as G2S appears to work best when the input oral microbial composition is close to the average used during training, caution must still be taken in interpreting the prediction data. Furthermore, G2S was implemented using both 16S rRNA gene and shotgun metagenomics data from different populations across the globe (from United States, Italy, Sweden, United Kingdom, and Fiji), with a good generalization of the results as evidenced by the findings on the test dataset. This provides an opportunity for users who can apply the tool on data obtained through different sequencing techniques simply by formatting their abundance tables with a taxonomy congruent with the Greengenes database. It should also be noted that G2S was built and validated using the 768 paired samples currently available in the literature. This stresses the importance of collecting paired samples (i.e., oral and fecal) in future studies from cohorts from different geographic locations, in order to further extend the range of the training dataset and thus the applicability of G2S. Finally, other future implementations could include predictions at different taxonomic levels, as well as functional predictions thanks to the recent expansion of shotgun metagenomics.

In summary, G2S opens up new possibilities in bioinformatics approaches related to metagenomics, extending *in silico* procedures to predict the human stool microbiome from oral microbiome data. Starting from either modern or ancient oral microbiome samples, the tool infers the stool microbiome with family level resolution. Its main field of application is probably paleomicrobiology, as a tool that can help understand how the gut microbiome of the past was structured, and its implications for human evolution. An update of the G2S tool will be periodically performed to incorporate newly released microbiome studies.

DATA AVAILABILITY STATEMENT

The datasets used for setting up G2S are available at the Human Microbiome Project website https://www.hmpdacc.org/HMQCP/ and NCBI SRA as SRP057504 (Zaura et al., 2015), PRJNA217052 (Brito et al., 2016) and PRJNA356414 (Russo et al., 2018). Microbiome data from ancient samples were taken from the study conducted by Warinner and colleagues (Warinner et al., 2014).

AUTHOR CONTRIBUTIONS

SR: conceptualization and software. SR and MF: formal analysis. SR, MC, and ST: writing—original draft preparation. MF, EB, and PB: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.644516/full#supplementary-material

Supplementary File 1 | R script containing the stochastic method that generates mock profiles of the stool microbiome in the range of the training dataset.

Supplementary Table 1 | List of paired fecal and oral samples from the HMP study as well as from other literature studies dealing with healthy adults (Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). Both 16S rRNA gene sequencing and shotgun metagenomics studies were considered. For each sample, the following data are reported: sample ID, subject ID (and visit when available), geographical origin, reference, sequencing method and body site.

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838. doi: 10.1038/nbt.3300

Ayeni, F. A., Biagi, E., Rampelli, S., Fiori, J., Soverini, M., Audu, H. J., et al. (2018). Infant and adult gut microbiome and metabolome in rural Bassa and urban settlers from Nigeria. *Cell Rep.* 23, 3056–3067. doi: 10.1016/j.celrep.2018.05.018

Bajaj, J. S., Betrapally, N. S., Hylemon, P. B., Heuman, D. M., Daita, K., White, M. B., et al. (2015). Salivary microbiota reflects changes in gut microbiota in cirrhosis with hepatic encephalopathy. *Hepatology* 62, 1260–1271. doi: 10.1002/hep.27819

Bishop, C. M. (2016). *Pattern Recognition and Machine Learning.* New York, NY: Springer.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535, 435–439. doi: 10.1038/nature18927

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15:20170387. doi: 10.1098/rsif.2017.0387

Demirci, S., Peters, S. A., de Ridder, D., and Van Dijk, A. D. J. (2018). DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *Plant J.* 95, 13979. doi: 10.1111/tpj.13979

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Galkin, F., Mamoshina, P., Aliper, A., Putin, E., Moskalev, V., Gladyshev, V. N., et al. (2020). Human gut microbiome aging clock based on taxonomic profiling and deep learning. *Iscience* 23:101199. doi: 10.1016/j.isci.2020.101199

Geman, O., Chiuchisan, I., Covasa, M., Doloc, C., Milici, M. R., and Milici, L. D. (2016). "Deep learning tools for human microbiome big data," in *Proceedings of the 7th International Workshop Soft Computing Applications SOFA 2016. Advances in Intelligent Systems and Computing*, Vol. 633, eds V. Balas, L. Jain, and M. Balas (Cham: Springer), 265–275.

Glassner, K. L., Abraham, B. P., and Quigley, E. M. M. (2020). The microbiome and inflammatory bowel disease. *J. Allergy Clin. Immunol.* 145, 16–27. doi: 10.1016/j.jaci.2019.11.003

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Griffen, A. L., Beall, C. J., Campbell, J. H., Firestone, N. D., Kumar, P. S., Yang, Z. K., et al. (2012). Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* 6, 1176–1185. doi: 10.1038/ismej.2011.191

He, Y., Wu, W., Zheng, H. M., Li, P., McDonald, D., Sheng, H. F., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* 24, 1532–1535. doi: 10.1038/s41591-018-0164-x

Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V., and Wargo, J. A. (2019). The microbiome, cancer, and cancer therapy. *Nat. Med.* 25, 377–388. doi: 10.1038/s41591-019-0377-7

Iwauchi, M., Horigome, A., Ishikawa, K., Mikuni, A., Nakano, M., Xiao, J. Z., et al. (2019). Relationship between oral and gut microbiota in elderly people. *Immun. Inflamm. Dis.* 7, 229–236. doi: 10.1002/iid3.266

Jha, A. R., Davenport, E. R., Gautam, Y., Bhandari, D., Tandukar, S., Ng, K. M., et al. (2018). Gut microbiome transition across a lifestyle gradient in Himalaya. *PLoS Biol.* 16:e2005396. doi: 10.1371/journal.pbio.2005396

Karpiński, P. M. (2019). Role of oral microbiota in cancer development. *Microorganisms* 7, 20. doi: 10.3390/microorganisms7010020

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature* 474, 327–336. doi: 10.1038/nature10213

Koskella, B., Hall, L. J., and Metcalf, C. J. E. (2017). The microbiome beyond the horizon of ecological and evolutionary theory. *Nat. Ecol. Evol.* 1, 1606–1615. doi: 10.1038/s41559-017-0340-2

Le, N. Q. K. (2019). Fertility-gru: identifying fertility-related proteins by incorporating deep-gated recurrent units and original position-specific scoring matrix profiles. *J. Proteome Res.* 18, 3503–3511. doi: 10.1021/acs.jproteome.9b00411

Le, N. Q. K., and Huynh, T. T. (2019). Identifying SNAREs by incorporating deep learning architecture and amino acid embedding representation. *Front Physiol.* 10:1501. doi: 10.3389/fphys.2019.01501

Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N., and Yeh, H. Y. (2019). Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext n-grams. *Front Bioeng Biotechnol.* 7:305. doi: 10.3389/fbioe.2019.00305

Leung, M. K. K., Delong, A., Alipanahi, B., and Frey, B. J. (2016). Machine learning in genomic medicine: a review of computational problems and data sets. *Proc. IEEE* 104, 176–197.

Miller, E. T., Svanbäck, R., and Bohannan, B. J. M. (2018). Microbiomes as metacommunities: understanding host-associated microbes through metacommunity ecology. *Trends Ecol. Evol.* 33, 926–935. doi: 10.1016/j.tree.2018.09.002

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* 6:6505. doi: 10.1038/ncomms7505

Pietiäinen, M., Liljestrand, J. M., Kopra, E., and Pussinen, P. J. (2018). Mediators between oral dysbiosis and cardiovascular diseases. *Eur. J. Oral Sci.* 126, 26–36. doi: 10.1111/eos.12423

Prodan, A., Levin, E., and Nieuwdorp, M. (2019). Does disease start in the mouth, the gut or both? *Elife* 8:e45931. doi: 10.7554/eLife.45931

Quang, D., and Xie, X. (2019). FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* 166, 40–47. doi: 10.1016/j.ymeth.2019.03.020

Rampelli, S., Guenther, K., Turroni, S., Wolters, M., Veidebaum, T., Kourides, Y., et al. (2018). Pre-obese children's dysbiotic gut microbiome and unhealthy diets may predict the development of obesity. *Commun. Biol.* 1:222. doi: 10.1038/s42003-018-0221-5

Rampelli, S., Schnorr, S. L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., et al. (2015). Metagenome sequencing of the Hadza hunter-gatherer gut microbiota. *Curr. Biol.* 25, 1682–1693. doi: 10.1016/j.cub.2015.04.055

Reiman, D., Metwally, A., and Dai, Y. (2017). Using convolutional neural networks to explore the microbiome. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2017, 4269–4272. doi: 10.1109/EMBC.2017.8037799

Russo, E., Bacci, G., Chiellini, C., Fagorzi, C., Niccolai, E., Taddei, A., et al. (2018). Preliminary comparison of oral and intestinal human microbiota in patients with colorectal cancer: a pilot study. *Front. Microbiol.* 8:2699. doi: 10.3389/fmicb.2017.02699

Schmidt, T. S. B., Hayward, M. R., Coelho, L. P., Li, S. S., Costea, P. I., Voigt, A. Y., et al. (2019). Extensive transmission of microbes along the gastrointestinal tract. *Elife* 8:e42693. doi: 10.7554/eLife.42693

Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* 5:3654. doi: 10.1038/ncomms4654

Sonnenburg, E. D., and Sonnenburg, J. L. (2019). The ancestral and industrialized gut microbiota and implications for human health. *Nat. Rev. Microbiol.* 17, 383–390. doi: 10.1038/s41579-019-0191-8

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/nmeth.3589

Wainberg, M., Merico, D., Delong, A., and Frey, B. J. (2018). Deep learning in biomedicine. *Nat. Biotechnol.* 36, 829–838. doi: 10.1038/nbt.4233

Warinner, C., Rodrigues, J. F., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., et al. (2014). Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* 46, 336–344. doi: 10.1038/ng.2906

Webb, S. (2018). Deep learning for biology. *Nature* 554, 555–557. doi: 10.1038/d41586-018-02174-z

Wong, S. H., and Yu, J. (2019). Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat. Rev. Gastroenterol. Hepatol.* 16, 690–704. doi: 10.1038/s41575-019-0209-8

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Zaura, E., Brandt, B. W., Teixeira de Mattos, M. J., Buijs, M. J., Caspers, M. P. M., Rashid, M. U., et al. (2015). Same exposure but two radically different responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces. *mBio* 6, e01693–e01695. doi: 10.1128/mBio.01693-15

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.