microorganisms

MDPI

Article

Pet-Human Gut Microbiome Host Classifier Using Data from Different Studies

Nadia Bykova *, Nikita Litovka, Anna Popenko and Sergey Musienko

Atlas LLC, Malaya Nikitskaya 31, 121069 Moscow, Russia; litovka@atlas.ru (N.L.); popenko@atlasbiomed.com (A.P.); musienko@atlasbiomed.com (S.M.)

* Correspondence: bykova@atlasbiomed.com

Received: 18 August 2020; Accepted: 9 October 2020; Published: 15 October 2020

Abstract: (1) Background: microbiome host classification can be used to identify sources of contamination in environmental data. However, there is no ready-to-use host classifier. Here, we aimed to build a model that would be able to discriminate between pet and human microbiomes samples. The challenge of the study was to build a classifier using data solely from publicly available studies that normally contain sequencing data for only one type of host. (2) Results: we have developed a random forest model that distinguishes human microbiota from domestic pet microbiota (cats and dogs) with 97% accuracy. In order to prevent overfitting, samples from several (at least four) different projects were necessary. Feature importance analysis revealed that the model relied on several taxa known to be key components in domestic cat and dog microbiomes (such as Fusobacteriaceae and Peptostreptococcaceae), as well as on some taxa exclusively found in humans (as Akkermansiaceae). (3) Conclusion: we have shown that it is possible to make a reliable pet/human gut microbiome classifier on the basis of the data collected from different studies.

Keywords: gut microbiome; host classification; random forest

1. Introduction

A microbiome is a complex ecological structure that is unique to each environment. Microbiota inhabiting living organism sites, such as the human gut, are of particular interest. Even though metagenomic approaches have made revealing microbiome compositions routine, their characterization and the identification of unique traits is still a challenge.

In the field of microbiome classification, there are several classification settings. One is the classification of the microbiome as a whole. There are also models for disease prediction, and some other individual trait predictions, such as age [1] or specific owner identification (skin microbiome), [2] for example. The gut microbiota, despite its complexity and great variation between individuals, was shown to be predictive of various intestinal diseases and conditions, such as irritable bowel syndrome (IBS) [3], Crohn's disease [4,5], and colorectal cancer [6]. Interestingly, the composition of the gut microbiome also predicts some non-intestinal illnesses, such as coronary artery disease [7], liver fibrosis [8], metabolic diseases/obesity [9], insomnia [10], and bipolar depression [11]. Another classification setting is the detection of contamination in samples. This task mostly arises in the case of water contamination with sewage or animal faeces. In this setting, the fraction of sequences coming from another host can be minimal. Moreover, contamination might come from several different sources making this task quite challenging [12–14]. In our study, we had a specific goal: host discrimination of the whole gut microbiome. This task arose from our need to filter out faecal samples of pets mistakenly or intentionally sent for commercial microbiome analysis in guise of human ones.

The mammalian gut microbiome evolves together with the host. As a general rule, the distance between microbiomes increases with the evolutionary distance between hosts [15]. At the same time,

microbiome composition is also a reflection of the host's dietary category. For example, while sharing the same main bacterial phyla, herbivores and carnivores harbor different families of Clostridiales (Ruminococcaceae and Peptostreptococcaceae, respectively) [15]. Herbivores also show greater diversity at all taxonomic levels than carnivores. It is also generally accepted that species with the same diet are similar at a higher taxonomic level, while host phylogeny reveals itself more at a lower taxonomic level (species and strains) [15,16]. Accordingly, this information was used in host prediction (contamination prediction) models. The oligotyping technique harnesses the existence of the host-specific/preferred species and strains, showing that it is possible to detect sequences from a specific host using just one or several genera [12,13]. This analysis is also considered to be more robust to high fluctuations in microbiome taxonomic composition caused by different sequencing techniques. The total microbiome taxonomy groups can be used as well, for example, as an input to the state-of-art program for contamination tracking called Source Tracker [14]. A disadvantage of Source Tracker for host classification is that it requires host data training from the same experiment to build a prediction model. Here, we checked if it is possible to use regular taxonomy features for accurate microbiome classification of pets and humans.

It is a widely acknowledged problem that next-generation sequencing (NGS) data in general, and microbiome 16S rRNA sequencing data specifically, vary from data center to data center. This introduces strong batch effects that sometimes make meta-analysis a rather sophisticated statistical task [17,18]. This is caused by different protocols for sample collection and storage [19], different DNA purification [20] and amplification protocols [21], and the use of different sequencing platforms [22]. Moreover, sequencing of different regions of the 16S rRNA gene influences the abundance of specific taxa in the resulting data [23]. A recent meta-analysis of human microbiomes found that differences in experimental protocols can affect microbiome composition more than the biological variance of some traits [17]. Another meta-analysis, this time of colorectal cancer studies, demonstrated that samples clustered primarily by study [18]. On the other hand, the incorporation of several studies helped increase the overall accuracy in this study. It should be noted that, in this meta-analysis, each study still had its own set of control (healthy) samples. Here, we aimed to build a classifier using public data where each class (host) was sequenced in a separate study, as there were no studies where human and pet samples were sequenced in the same experiment.

Random forest (RF) algorithm [24] was chosen for the classification task. RFs are highly used machine learning algorithms for microbiome classification [3–11] due to the limited number of model parameters and simple results interpretation.

2. Materials and Methods

2.1. Data

Data with publicly available cat, dog, or human faecal samples used in our study are listed in Table 1. Only projects that performed Illumina sequencing of the V4 region of 16S RNA were included. Overall, data from five cat projects and seven dog projects were collected, providing 321 pet samples in total. Ten human projects provided 1242 samples. Note that from the specified human projects, we used only healthy control (HC) subjects, and from the pet projects we used all subjects (i.e., not only healthy samples) to provide a bigger dataset. After the models were trained, we further tested them on five independent pet and two independent human projects. These additional projects contained 432 animal and 358 human samples.

Table 1. Data used in the study.

Project Name	Host	Host Type	Number of Samples	Number of Samples (Train Dataset)	PMID	Author/Year	Ref.
PRJNA504021	Felis catus	pet	65	65	31844119	*Marsilio* et al. *(2019)*	[25]
PRJNA349988	Felis catus	pet	44	44	27912797	*Duarte* et al. *(2016)*	[26]
PRJNA248757	Felis catus	pet	30	30	25279695	*Bell* et al. *(2014)*	[27]
PRJNA338653	Felis catus	pet	19	19	30709324	*Whittemore* et al. *(2019)*	[28]
PRJNA350163	Felis catus	pet	6	6	28278278	*Vientós-Plotts* et al. *(2017)*	[29]
PRJNA488105	Canis familiaris	pet	34	34	no paper		
PRJNA525542	Canis familiaris	pet	32	32	31565574	*Jarett* et al. *(2019)*	[30]
PRJNA358232	Canis familiaris	pet	30	30	no paper		
PRJNA391562	Canis familiaris	pet	23	23	29852000	*Herstad* et al. *(2018)*	[31]
PRJNA493249	Canis familiaris	pet	19	19	32027665	*Fujishiro* et al. *(2020)*	[32]
PRJDB5398	Canis familiaris	pet	13	13	29643280	*Omatsu* et al. *(2018)*	[33]
PRJNA492898	Canis familiaris	pet	6	6	32027665	*Fujishiro* et al. *(2020)*	[32]
PMID29795809	Homo sapiens	human	681	46	29795809	*McDonald* et al. *(2018)*	[34]
PMID25417156	Homo sapiens	human	200	45	25417156	*Goodrich* et al. *(2014)*	[35]
PMID28195358	Homo sapiens	human	115	45	28195358	*Hill-Burns* et al. *(2017)*	[36]
PMID28179361	Homo sapiens	human	102	45	28179361	*Pascal* et al. *(2017)*	[37]
PMID31027508	Homo sapiens	human	49	45	31027508	*Liu* et al. *(2019)*	[38]
PMID26179554	Homo sapiens	human	31	31	26179554	*Keshavarzian* et al. *(2015)*	[39]
PMID28429209	Homo sapiens	human	22	22	28429209	*Petrov* et al. *(2017)*	[40]
qiita_10928	Homo sapiens	human	21	21	no paper		
PMID29404425	Homo sapiens	human	12	12	29404425	*Zhou* et al. *(2018)*	[41]
PMID28191884	Homo sapiens	human	9	9	28191884	*Halfvarson* et al. *(2017)*	[42]
Additional Projects							
PRJNA470724	Felis catus	pet	74		29971046	*Bermingham* et al. *(2018)*	[43]
PMID32078625	Felis catus	pet	46		32078625	*Jha* et al. *(2020)*	[44]
PMID32078625	Canis familiaris	pet	192		32078625	*Jha* et al. *(2020)*	[44]
PRJNA401442	Canis familiaris	pet	56		no paper		
PRJNA589580	Canis familiaris	pet	35		no paper		
PRJNA592436	Canis familiaris	pet	29		no paper		
PRJNA385551	Homo sapiens	human	284		28959739	*Bian* et al. *(2017)*	[45]
PRJNA493726	Homo sapiens	human	74		30872359	*Li* et al. *(2019)*	[46]

2.2. QIIME2

All the raw data fastq files were processed with QIIME 2 (Flagstaff, AZ, USA) [47] to obtain Chao diversity estimations [48] and feature tables. The following parameters were employed in the QIIME2 microbiome analysis pipeline:

- DADA2 [49] denoising quality parameter value was set to 10 (–p-trunc-q 10),
- taxonomy assignment using QIIME2 feature-classifier [50],
- a random subsample of 5000 reads was used to calculate feature tables,
- alpha-diversity was calculated by sampling 5000 random reads five times from the whole sample to decrease the impact of low-abundance bacteria; the resulting chao index is the mean of these iterations results,
- a custom reference database was used; the database is a restricted version of SILVA database [51], aligned to the HITdb database [52] in order to leave mostly gut bacteria.

2.3. Grouping Features at Different Taxonomic Levels

Abundance data at a genus level was used in the analysis. Each genus is represented as its full taxonomy, namely: kingdom, phylum, class, order, family, genus. Therefore, the original genus table can be grouped into higher-level features. During this procedure, the abundances of genera grouped by the same higher taxa were summed. The analysis was then performed for the tables at each taxonomic level.

2.4. Filtering Rare Features

Statistical hypothesis testing and RF training were only performed for the taxa left after filtering out features with low abundance. The taxonomic features where over 90% of samples had zero read for both pet and human data were defined as rare. This procedure reduced the initial 386 genera to 138, 123 families to 55, 54 orders to 29, 26 classes to 18, 15 phyla to 10 (Table S1).

2.5. Mann-Whitney Test

Two-sided Mann-Whitney test was performed on the projects' median abundance values for each feature (i.e., for each feature there were 12 pet values versus 10 human values in the test). We applied both Holm–Bonferroni [53] and FDR Benjamini–Hochberg [54] procedures to correct for the testing of multiple features. Significant features were further used to build restricted versions of the RF models. Two-sided Mann-Whitney test was also applied for the Chao diversity values that characterize each sample.

2.6. t-SNE

t-SNE algorithm from Python3 sklearn library [55] with Bray-Curtis distance was used to visualize the data. The analysis was performed for the balanced class dataset (see below). The input taxa were restricted to 138 most abundant genus (see above).

2.7. Balanced Class Dataset

From the initial dataset that contained different numbers of animal and human samples, and different numbers of samples from different projects, we constructed a dataset balanced by host (321 animal and 321 human samples). It included all our animal data and a subset of human data. The accession numbers of specific samples that fell into the dataset are listed in Table S2. The human subset was balanced by projects (i.e., we aimed to take the same number of samples from each project). The sampling was without replacement. This dataset was also used for data visualization using t-SNE (see above).

2.8. CLR (Center Log Ratio) Transformation

CLR transformation of feature tables was performed with Python3 library scikit-bio version 0.5.6 (http://scikit-bio.org). CLR transformation converts compositional data from Aitchison geometry to the real space [56]. Non-transformed or CLR-transformed data were optionally used to train the RF model.

2.9. RF Implementation

The input data matrix for the model consisted of feature abundance values and a column with mean Chao diversity values. Optionally, the data were restricted to fewer features, or CLR-transformed data was used beforehand. To construct an optimal prediction model, we first performed parameter selection using stratified 5-fold cross-validation. The best parameters were defined by the highest average test accuracy achieved at cross-validation. We varied the following model parameters:

- max_features in the range from 2 to the 'number of features',
- max_depth in the range from 2 to 52,
- min_samples_split in the range from 2 to 52,
- n_estimators from the set {1,5,10,50,100,500,1000}.

To speed up the parameter selection process, we performed it in two steps. First, we selected the best set of parameters using all parameter combinations in a smaller range. During the second step, each parameter was refined while all the other parameters were fixed to the values obtained on the first step (Text S1). The model was then fit on the whole dataset with the best parameters, and out-of-bag scores were reported as a final performance estimation of the model. The project is realized using Python 3 sk-learn library. The source code of the project is available at GitHub (https://github.com/nadiabykova/microbiota_host_classifier).

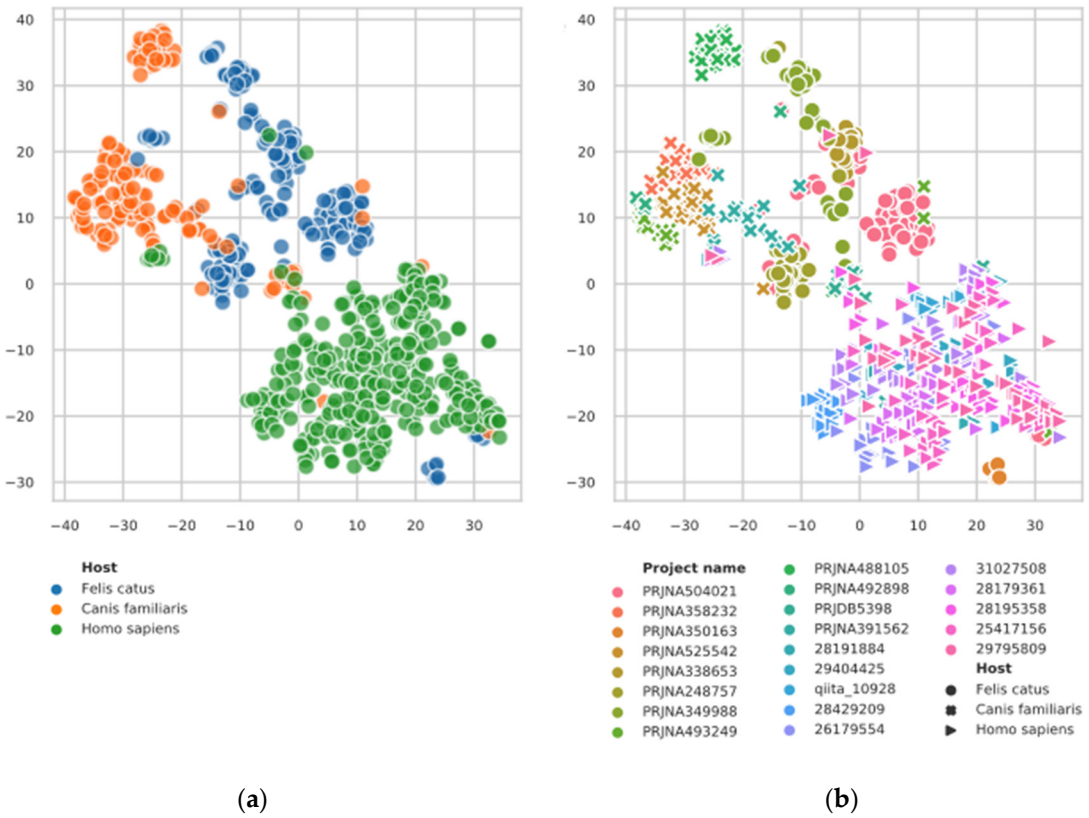
2.10. The Project Learning Curve

Our models are trained on data from specific projects. Therefore, it is possible that the model 'remembers' project features rather than host-specific features. If so, new projects that did not participate in the training would be poorly predicted. To evaluate the ability of these models to overfit to specific projects, we conducted the following experiment: we varied the number of human projects incorporated in the training set and measured the performance on the projects that did not participate in the training. Formally, for n from 1 to $N - 1$, where N is the number of human projects, we formed new balanced training sets consisting of all animal data and samples from n human projects. The number of considered project combinations for each n was set to the minimum value of 200 and $C_{N-1}{}^{n}$. The training sets were balanced by class, and the human part of the sets was balanced by project. Here, we used sampling with replacements to be able to construct the human part of the set with size 321 for each n. We also randomized the sampling from the human projects taking five random samples for each combination of projects. Therefore, in total, approximately 200×5 models were trained for each value of n (less for the cases where $n < 4$). Accuracy on projects that did not participate in the training was then evaluated.

3. Results

3.1. t-SNE Plot

To visualize our data, we first built a t-SNE plot on the balanced class dataset (i.e., a sample of original data that contains equal amounts of pet and human samples (see Materials and Methods)). Figure 1 shows that human, cat, and dog samples cluster within groups, indicating that they can possibly be classified using taxonomic features' abundance values. Samples from specific studies also tend to cluster together, but the host signal is stronger.



(**a**) (**b**)

Figure 1. t-SNE plot for the dataset on the genus level. The t-SNE plot was built using 138 most abundant genera of the balanced dataset (see Materials and Methods). The Bray–Curtis dissimilarity between vectors was used. The samples are colored by host (**a**), or by the study name (**b**). On side b, the host is shown by a marker shape. See the in-plot legend for the specific name-color mapping.

3.2. Taxa Differentially Abundant in Pets and Humans

We studied the distribution of taxa abundances in pets and humans at several taxonomic levels (see Materials and Methods). First of all, the data show substantial variation inside host groups. Even on the phylum level, 'project-outliers' can be noted, illustrating that the batch effect can significantly affect the abundance of specific taxa (Figure S1). To detect the taxa differently abundant in pets and humans, we applied two-sided Mann-Whitney test to the projects' median values of each taxon abundance at each level. The significant taxa detected are listed in Table S3. On the phylum level, Verrucomicrobia and Fusobacteria were significant (Figure S2A), which is in accordance with previous studies describing Fusobacteria as one of the key phyla of domestic pets, and Verrucomicrobia as a taxon characteristic of the human microbiota while being absent in cats and dogs [57]. On the level of class, Verrucomicrobia, Fusobacteria, and Deltaproteobacteria were detected; Deltaproteobacteria was present in almost all human projects, and absent or present in very small amounts in pet projects (Figure S2B). On the level of order, Verrucomicrobiales and Desulfovibrionales were detected (Fusobacteriales were detected only by the FDR correction method, Figure S2C). On the family level, there were more interesting results (Figure 2). While the same effect at the higher levels was detected for Akkermansiaceae and Desulfovibrionaceae, the difference in Bacteroidetes and Firmicutes was first detected at the family level. Namely, the Peptostreptococcaceae family was characteristic of pets and Ruminococcaceae family for humans; this switch in the usage of Clostridia families was previously described as a difference between carnivorous and herbivorous animals [15]. Peptostreptococcaceae is also described as a prominent taxon for cats [43]. The families of the Bacteroidetes phylum, Rikenellaceae and Marinifilaceae, were detected almost exclusively in humans. FDR correction added another eight families from different phyla. On the genus level, Astilipes and Odoribacter from the Rikenellaceae and Marinifilaceae families,

respectively were detected, Ruminococcus_1 and Faecalibacterium (elevated amounts in humans) from Ruminococcaceae family, Bilophila from Desulfovibrionaceae family, Akkermansia, 5 Lachnospiraceae genera, and Erysipelotrichaceae_UCG-003 were detected (Figure S2D). FDR correction yielded an additional 20 significant genera from different taxa. Taken together these data show that, at the level of general composition, pets and humans do not show a great difference with all the main phylums being insignificant (note that the Verrumicrobia and Fusobacteria consist of only one or two genera), while all the main differences lay at the level of families and specific genera. The defined significant taxa were further used to build restricted RF models. Mann-Whitney test for the median values of the Chao diversity index was insignificant.

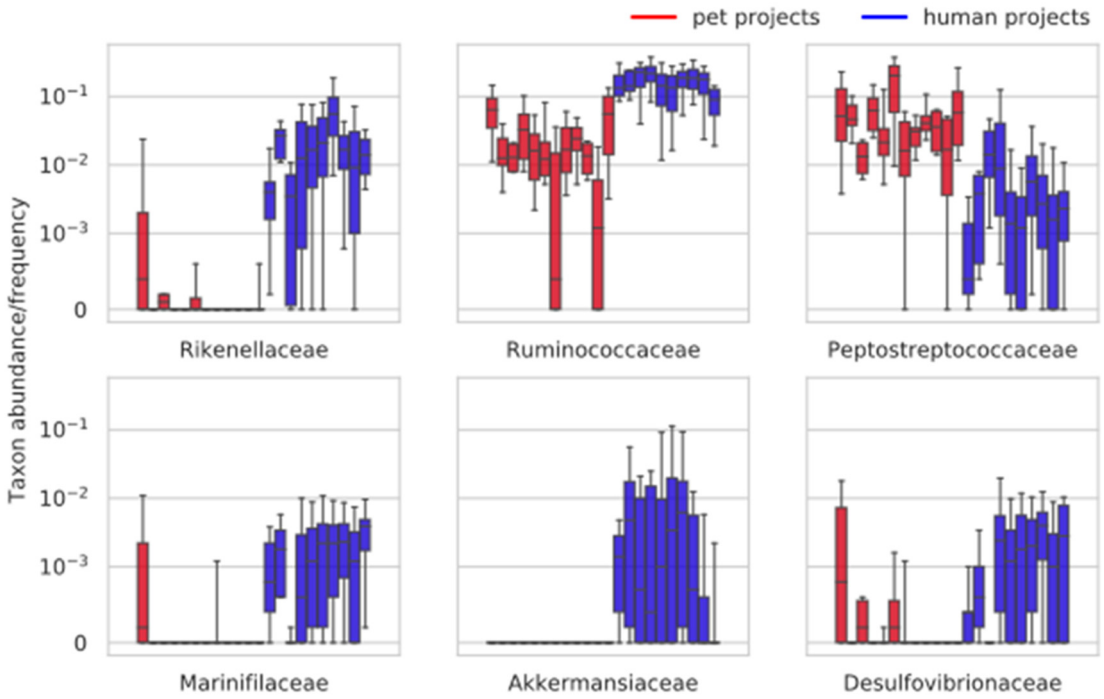


Figure 2. The differentially abundant families between pets and humans. The y-axis shows the fraction of reads from the total (the axis is log-transformed). Red bars correspond to pet projects, and blue bars correspond to human projects. The families significant using the Holm correction method are shown (for families significant according to the FDR correction method, see Table S3).

3.3. Random Forest Models

RF models were trained on the dataset balanced by the host class that was derived from the initial dataset (see Materials and Methods). To define the most appropriate model for host discrimination, we tested several types of models. We tried models on different taxonomic levels, with all or only specific sets of features (as defined by Mann-Whitney test above), we also optionally applied CLR transformation to the initial data (see Materials and Methods). For each model, the best model parameters were first defined using cross-validation. The models with the best parameters were fit on the dataset, and performances were estimated on out-of-bag (OOB) samples. The obtained best parameters, cross-validation and out-of-bag results are summarized in Table S4 and Table 2. Very good results were achieved by all the models. However, genus models clearly outperformed family models, and genus models with more features (Genus_ALL and Genus_MW-FDR) were better than the most restricted Genus_MW-Holm model because a model using more meaningful features allows for better prediction. Remarkably, using only MW-selected FDR features did not lead to a substantial decrease in accuracy. The usage of CLR transformation did not introduce any significant improvement. The fact that the usage of CLR transformation did not significantly improve results in our case might be because the prediction accuracy is good with both methods. The OOB estimation of accuracy of both

Genus_ALL and Genus_MW-FDR models was 0.99 ± 0.002. The process of parameter selection for each model and corresponding ROC curve obtained in cross-validation for each model are presented in Text S1. The ROC curves of all the models together are shown in Figure S3.

Table 2. Out-of-bag estimations of model performances.

Model Name	Level	Features Type	Number of Features	CLR	Accuracy	F1 Score	Precision	Recall
Family_ALL_CLR	Family	all	56	yes	0.981 ± 0.004	0.980 ± 0.004	0.987 ± 0.004	0.974 ± 0.006
Family_ALL	Family	all	56	no	0.983 ± 0.004	0.983 ± 0.004	0.989 ± 0.004	0.977 ± 0.006
Family_MW-FDR_CLR	Family	best_fdr	14	yes	0.966 ± 0.004	0.966 ± 0.004	0.976 ± 0.005	0.955 ± 0.006
Family_MW-FDR	Family	best_fdr	14	no	0.970 ± 0.003	0.970 ± 0.003	0.986 ± 0.004	0.955 ± 0.005
Family_MW-Holm_CLR	Family	best_holm	6	yes	0.954 ± 0.003	0.954 ± 0.003	0.957 ± 0.004	0.951 ± 0.004
Family_MW-Holm	Family	best_holm	6	no	0.953 ± 0.004	0.953 ± 0.003	0.951 ± 0.006	0.955 ± 0.004
Genus_ALL_CLR	Genus	all	139	yes	0.990 ± 0.002	0.990 ± 0.002	0.999 ± 0.001	0.981 ± 0.003
Genus_ALL	Genus	all	139	no	0.992 ± 0.002	0.992 ± 0.002	0.999 ± 0.002	0.985 ± 0.003
Genus_MW-FDR_CLR	Genus	best_fdr	32	yes	0.986 ± 0.002	0.985 ± 0.002	0.997 ± 0.003	0.974 ± 0.004
Genus_MW-FDR	Genus	best_fdr	32	no	0.989 ± 0.002	0.989 ± 0.002	0.998 ± 0.003	0.979 ± 0.003
Genus_MW-Holm_CLR	Genus	best_holm	12	yes	0.972 ± 0.003	0.972 ± 0.003	0.982 ± 0.004	0.963 ± 0.004
Genus_MW-Holm	Genus	best_holm	12	no	0.967 ± 0.003	0.967 ± 0.003	0.981 ± 0.004	0.953 ± 0.005

3.4. Random Forest Feature Importance

RF models are suitable for selecting the most important features that allow us to distinguish between classes. When comparing the most important RF features of full models (Family_ALL and Genus_ALL) with MW-results (FDR correction), we found good correspondence both on the family level (86% intersection between top features, Figure S4A) and genus level (78%, Figure S4B). New families preferred by RF over Mann-Whitney-selected taxa were Enterococcaceae and unclassified bacteria. At the genus level, the RF model brought up Fusobacterium, Collinsella, Anaerobiospirillum, unclassified Fusobacteriaceae, unclassified bacteria, Intestinimonas, Veillonella.

3.5. Projects Learning Curve

We further tried to estimate if models built this way are robust for project overfitting. To this end, we conducted the following experiment: the pet part of the dataset was fixed, and from the human data we selected various numbers of projects to include in the training set. The remaining human projects were used to control the model's performance (see Materials and Methods). The dependency of model accuracy on the number of human projects in the training set for genus models is shown in Figure 3 (the accuracy is averaged among the combinations of training projects and among the projects used for testing). Figures for the models where accuracy is shown for each test project can be found in Figure S5. Figure 3 shows that models trained using only one human project appear to consistently overfit, leading to decreased accuracy on the test projects. The appropriate number of projects to capture the host differences for our setting should be more than four (as can be identified from the graph). In the models described above, we used more projects (all of them) in the training set, thus we do not expect them to be overfit. For the Genus_MW-FDR model, we do not expect accuracy below 0.9 on other human projects (Figure S5E).

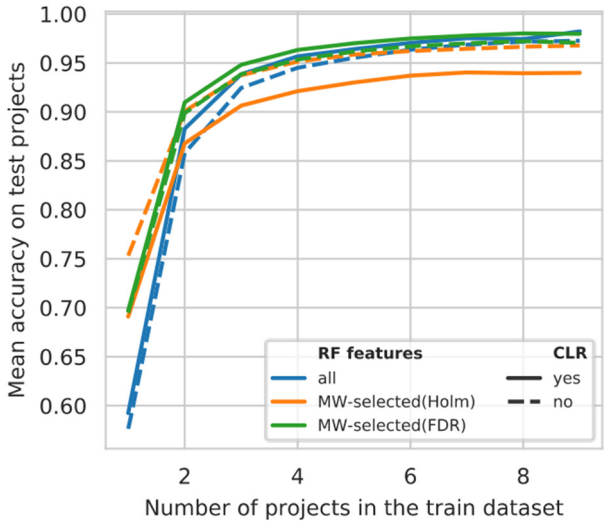


Figure 3. The dependency of accuracy on test human projects from the number of human projects used in the training set (for the genus models). The average value of accuracy among all the models (for the specific *n*) and test projects is shown.

3.6. Model Testing on Additional Projects

To test the performance of our actual models on some new projects, we downloaded several additional human and animal projects (described in Table 1) and applied our models to them. The resulting accuracy for all genus models is summarized in Table S5. The best results were once again shown by Genus_ALL and Genus_MW-FDR models. Genus_MW-FDR model performance is presented in Table 3 and Figure 4. The new data also contained samples from SLE (systemic lupus erythematosus) and RA patients (rheumatoid arthritis); notably, all the models showed better performance on healthy samples, as expected. From animal projects, almost all new samples were dog samples –only two projects also contained cat samples. The dog samples were better recognized as pets by all the models. To obtain the full discriminative characteristics of the models, we constructed 100 balanced animal/human sets of size 200 from these projects (50 cats, 50 dogs, and 100 humans). The average accuracy, precision, recall, and f1 score of the Genus_MW-FDR model on healthy samples is 0.99 ± 0.01, 0.98 ± 0.02, 1.00 ± 0.0, 0.99 ± 0.01 on all samples 0.97 ± 0.01, 0.98 ± 0.02, 0.97 ± 0.02, 0.97 ± 0.01 (Table 4), the estimations for the other genus models are in Table S6.

Table 3. Accuracy of Genus_MW-FDR model on additional projects.

Host Type	Host	Test Project	Accuracy	Number of Samples
human	Homo sapiens	PRJNA385551	1	284
human	Homo sapiens	PRJNA493726	0.932	74
human	Homo sapiens	PRJNA493726_HC	1	19
human	Homo sapiens	PRJNA493726_RA	1	18
human	Homo sapiens	PRJNA493726_SLE	0.865	37
pet	Felis catus + Canis familiaris	PMID32078625	0.983	238
pet	Felis catus	PMID32078625_cat	0.935	46
pet	Canis familiaris	PMID32078625_dog	0.995	192
pet	Canis familiaris	PRJNA401442	1	56
pet	Felis catus	PRJNA470724	0.973	74
pet	Canis familiaris	PRJNA589580	1	35
pet	Canis familiaris	PRJNA592436	1	29

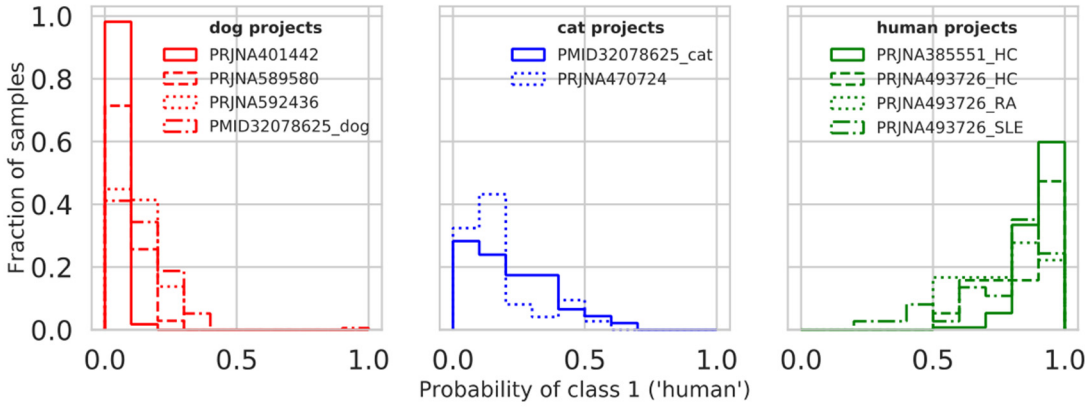


Figure 4. Prediction score distributions of the additional projects. The histograms for the probability of class 1 ('human') for the seven additional projects are present. The dog projects are shown in red, cat projects in blue, and human projects are in green. Specific projects are shown with the combination of color and linetype (see the in-graph legend).

Table 4. Genus_MW-FDR performance on a mixed class dataset.

Metric Name	Total Dataset	Only Healthy Controls
Accuracy	0.971 ± 0.010	0.988 ± 0.008
Precision	0.976 ± 0.015	0.977 ± 0.015
Recall	0.966 ± 0.016	1.000 ± 0.000
F1 score	0.971 ± 0.010	0.988 ± 0.008

4. Discussion

One result of our work was identifying the list of taxa differently abundant in pets and humans that is statistically supported by a set of different studies. We show that the main differences do not occur in the general composition of the microbiome, but rather in the usage of specific genera, even though the species stick to very different diets. The main, if not to say the only, Verrumicrobia member in a gut microbiome is Akkermansia genus. Akkermansia bacteria are known to live in a mucin layer and degrade it. Previous studies show that indeed, feline and canine gut microbiome lack Akkermansia because this ecological niche is occupied with other species, specifically members of Bacteroidaceae, Prevotellaceae, Clostridiales, Faecalibacterium, and Fusobacteria phylum [58]. We also show that, even when overfitting occurs (as observed in a few projects in the training set), the RF model appears to be able to dissect host-specific features from the project-specific features, and the joint usage of these features makes it possible to successfully classify samples from new projects. On the other hand, even though we collected data from a substantial number of projects, we cannot guarantee that all the possible sequences techniques were covered and that our model will not fail at some special new project due to skewed taxa abundances. Moreover, samples from people with health conditions are more likely to be mistaken as pet samples.

5. Conclusions

We have shown that it is possible to make a reliable pet/human classifier on the basis of taxonomic feature abundance tables collected from different studies. Our study demonstrates that a classifier with good performance can be built if at least four different studies are included in the training set. We also provide a list of taxa that discriminate between hosts, these results are in line with previous studies.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-2607/8/10/1591/s1. Figure S1: The boxplot of the phylum abundance for the projects used in the study; Figure S2A–D: The boxplot of the taxa significant in the MW test; Figure S3: ROC-curves for family and genus random forest models obtained;

Figure S4: Top families and genera important for human/pet host discrimination as suggested by random forest models; Figure S5: The dependency of accuracy on test human projects from the number of human projects used in the training set; Table S1: Taxonomic feature abundances in the data, features with low abundance; Table S2: The accession numbers of the specific samples used in the training dataset; Table S3: Differentially abundant taxa, results of Mann-Whitney test; Table S4: Summary of models performance; Table S5: The accuracy of the genus models on additional datasets; Table S6: Estimated accuracy, precision, recall, and f1 score of the genus models; Text S1: Parameters selection of random forest models.

Author Contributions: Conceptualization, A.P. and S.M.; methodology, N.B., N.L. and A.P.; software, N.B.; validation, N.B.; formal analysis, N.B., N.L. and A.P.; data curation, N.B. and N.L.; writing—original draft preparation, N.B.; writing—review and editing, A.P.; supervision, A.P.; project administration, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: Atlas LLC funded the study design, collection, analysis, and interpretation of the data; writing of the paper; and decision to submit for publication.

Acknowledgments: We are grateful to Grigory Sapunov and Ancha Baranova for discussion and revision of the manuscript.

Conflicts of Interest: All of the authors were employees of Atlas LLC at the time of this work and received compensation.

References

1. Huang, S.; Haiminen, N.; Carrieri, A.-P.; Hu, R.; Jiang, L.; Parida, L.; Russell, B.; Allaband, C.; Zarrinpar, A.; Vázquez-Baeza, Y.; et al. Human Skin, Oral, and Gut Microbiomes Predict Chronological Age. *mSystems* **2020**, *5*. [CrossRef] [PubMed]
2. Yang, J.; Tsukimi, T.; Yoshikawa, M.; Suzuki, K.; Takeda, T.; Tomita, M.; Fukuda, S. Cutibacterium acnes (*Propionibacterium acnes*) 16S rRNA Genotyping of Microbial Samples from Possessions Contributes to Owner Identification. *mSystems* **2019**, *4*. [CrossRef] [PubMed]
3. Labus, J.S.; Hollister, E.B.; Jacobs, J.; Kirbach, K.; Oezguen, N.; Gupta, A.; Acosta, J.; Luna, R.A.; Aagaard, K.M.; Versalovic, J.; et al. Differences in gut microbial composition correlate with regional brain volumes in irritable bowel syndrome. *Microbiome* **2017**, *5*, 49. [CrossRef] [PubMed]
4. Sprockett, D.; Fischer, N.; Boneh, R.S.; Turner, D.; Kierkus, J.; Sladek, M.; Escher, J.C.; Wine, E.; Yerushalmi, B.; Dias, J.A.; et al. Treatment-Specific Composition of the Gut Microbiota Is Associated With Disease Remission in a Pediatric Crohn's Disease Cohort. *Inflamm. Bowel Dis.* **2019**, *25*, 1927–1938. [CrossRef] [PubMed]
5. Jones, C.M.A.; Connors, J.; Dunn, K.A.; Bielawski, J.P.; Comeau, A.M.; Langille, M.G.I.; Van Limbergen, J. Bacterial Taxa and Functions Are Predictive of Sustained Remission Following Exclusive Enteral Nutrition in Pediatric Crohn's Disease. *Inflamm. Bowel Dis.* **2020**, *26*, 1026–1037. [CrossRef]
6. Qu, K.; Gao, F.; Guo, F.; Zou, Q. Taxonomy dimension reduction for colorectal cancer prediction. *Comput. Biol. Chem.* **2019**, *83*, 107160. [CrossRef]
7. Zheng, Y.-Y.; Wu, T.-T.; Liu, Z.-Q.; Li, A.; Guo, Q.-Q.; Ma, Y.-Y.; Zhang, Z.-L.; Xun, Y.-L.; Zhang, J.-C.; Wang, W.-R.; et al. Gut Microbiome-Based Diagnostic Model to Predict Coronary Artery Disease. *J. Agric. Food Chem.* **2020**, *68*, 3548–3557. [CrossRef]
8. Dong, T.S.; Katzka, W.; Lagishetty, V.; Luu, K.; Hauer, M.; Pisegna, J.; Jacobs, J. A Microbial Signature Identifies Advanced Fibrosis in Patients with Chronic Liver Disease Mainly Due to NAFLD. *Sci. Rep.* **2020**, *10*, 1–10. [CrossRef]
9. Zeng, Q.; Li, N.; He, Y.; Li, Y.; Yang, Z.; Zhao, X.; Liu, Y.; Wang, Y.; Sun, J.; Feng, X.; et al. Discrepant gut microbiota markers for the classification of obesity-related metabolic abnormalities. *Sci. Rep.* **2019**, *9*, 13424. [CrossRef]
10. Liu, B.; Lin, W.; Chen, S.; Xiang, T.; Yang, Y.; Yin, Y.; Xu, G.; Liu, Z.; Liu, L.; Pan, J.; et al. Gut Microbiota as an Objective Measurement for Auxiliary Diagnosis of Insomnia Disorder. *Front. Microbiol.* **2019**, *10*, 1770. [CrossRef]
11. Hu, S.; Li, A.; Huang, T.; Lai, J.; Li, J.; Sublette, M.E.; Lu, H.; Lu, Q.; Du, Y.; Hu, Z.; et al. Gut Microbiota Changes in Patients with Bipolar Depression. *Adv. Sci.* **2019**, *6*, 1900752. [CrossRef]
12. Roguet, A.; Eren, A.M.; Newton, R.J.; McLellan, S.L. Fecal source identification using random forest. *Microbiome* **2018**, *6*, 185. [CrossRef]

13. Eren, A.M.; Sogin, M.L.; Morrison, H.G.; Vineis, J.H.; Fisher, J.C.; Newton, R.J.; McLellan, S.L. A single genus in the gut microbiome reflects host preference and specificity. *ISME J.* **2014**, *9*, 90–100. [CrossRef] [PubMed]
14. Knights, D.; Kuczynski, J.; Charlson, E.S.; Zaneveld, J.; Mozer, M.C.; Collman, R.G.; Bushman, F.D.; Knight, R.; Kelley, S.T. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **2011**, *8*, 761–763. [CrossRef]
15. Nishida, A.H.; Ochman, H. Rates of gut microbiome divergence in mammals. *Mol. Ecol.* **2018**, *27*, 1884–1897. [CrossRef] [PubMed]
16. Groussin, M.; Mazel, F.; Sanders, J.G.; Smillie, C.S.; Lavergne, S.; Thuiller, W.; Alm, E.J. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun.* **2017**, *8*, 14319. [CrossRef] [PubMed]
17. Lozupone, C.A.; Stombaugh, J.; Gonzalez, A.; Ackermann, G.; Wendel, D.; Vázquez-Baeza, Y.; Jansson, J.K.; Gordon, J.I.; Knight, R. Meta-analyses of studies of the human microbiota. *Genome Res.* **2013**, *23*, 1704–1714. [CrossRef]
18. Shah, M.S.; DeSantis, T.Z.; Weinmaier, T.; McMurdie, P.J.; Cope, J.; Altrichter, A.; Yamal, J.-M.; Hollister, E.B. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut* **2017**, *67*, 882–891. [CrossRef]
19. Gorzelak, M.A.; Gill, S.K.; Tasnim, N.; Ahmadi-Vand, Z.; Jay, M.; Gibson, D.L. Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. *PLoS ONE* **2015**, *10*, e0134802. [CrossRef]
20. Yuan, S.; Cohen, D.B.; Ravel, J.; Abdo, Z.; Forney, L.J. Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome. *PLoS ONE* **2012**, *7*, e33865. [CrossRef]
21. Mao, D.-P.; Zhou, Q.; Chen, C.-Y.; Quan, Z.-X. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol.* **2012**, *12*, 66. [CrossRef] [PubMed]
22. Ratan, A.; Miller, W.; Guillory, J.; Stinson, J.; Seshagiri, S.; Schuster, S.C. Comparison of Sequencing Platforms for Single Nucleotide Variant Calls in a Human Sample. *PLoS ONE* **2013**, *8*, e55089. [CrossRef] [PubMed]
23. Rintala, A.; Pietilä, S.; Munukka, E.; Eerola, E.; Pursiheimo, J.-P.; Laiho, A.; Pekkala, S.; Huovinen, P. Gut Microbiota Analysis Results Are Highly Dependent on the 16S rRNA Gene Target Region, Whereas the Impact of DNA Extraction Is Minor. *J. Biomol. Tech. JBT* **2017**, *28*, 19–30. [CrossRef]
24. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. Marsilio, S.; Pilla, R.; Sarawichitr, B.; Chow, B.; Hill, S.L.; Ackermann, M.R.; Estep, J.S.; Lidbury, J.A.; Steiner, J.M.; Suchodolski, J.S. Characterization of the fecal microbiome in cats with inflammatory bowel disease or alimentary small cell lymphoma. *Sci. Rep.* **2019**, *9*, 1–11. [CrossRef] [PubMed]
26. Duarte, A.M.; Jenkins, T.P.; Latrofa, M.S.; Giannelli, A.; Papadopoulos, E.; De Carvalho, L.M.; Nolan, M.J.; Otranto, D.; Cantacessi, C. Helminth infections and gut microbiota—A feline perspective. *Parasites Vectors* **2016**, *9*, 625. [CrossRef]
27. Bell, E.T.; Suchodolski, J.S.; Isaiah, A.; Fleeman, L.M.; Cook, A.K.; Steiner, J.M.; Mansfield, C.S. Faecal Microbiota of Cats with Insulin-Treated Diabetes Mellitus. *PLoS ONE* **2014**, *9*, e108729. [CrossRef]
28. Whittemore, J.C.; Stokes, J.E.; Price, J.M.; Suchodolski, J.S. Effects of a synbiotic on the fecal microbiome and metabolomic profiles of healthy research cats administered clindamycin: A randomized, controlled trial. *Gut Microbes* **2019**, *10*, 521–539. [CrossRef]
29. Vientós-Plotts, A.I.; Ericsson, A.C.; Rindt, H.; Grobman, M.E.; Graham, A.; Bishop, K.; Cohn, L.A.; Reinero, C. Dynamic changes of the respiratory microbiota and its relationship to fecal and blood microbiota in healthy young cats. *PLoS ONE* **2017**, *12*, e0173818. [CrossRef]
30. Jarett, J.K.; Carlson, A.; Serao, M.R.; Strickland, J.; Serfilippi, L.; Ganz, H.H. Diets with and without edible cricket support a similar level of diversity in the gut microbiome of dogs. *PeerJ* **2019**, *7*, e7661. [CrossRef]
31. Herstad, K.M.V.; Moen, A.E.F.; Gaby, J.C.; Moe, L.; Skancke, E. Characterization of the fecal and mucosa-associated microbiota in dogs with colorectal epithelial tumors. *PLoS ONE* **2018**, *13*, e0198342. [CrossRef] [PubMed]
32. Fujishiro, M.A.; Lidbury, J.; Pilla, R.; Steiner, J.M.; Lappin, M.R.; Suchodolski, J.S. Evaluation of the effects of anthelmintic administration on the fecal microbiome of healthy dogs with and without subclinical Giardia spp. and Cryptosporidium canis infections. *PLoS ONE* **2020**, *15*, e0228145. [CrossRef] [PubMed]

33. Omatsu, T.; Omura, M.; Katayama, Y.; Kimura, T.; Okumura, M.; Okumura, A.; Murata, Y.; Mizutani, T. Molecular diversity of the faecal microbiota of Toy Poodles in Japan. *J. Veter Med. Sci.* **2018**, *80*, 749–754. [CrossRef]
34. McDonald, D.; Hyde, E.; Debelius, J.W.; Morton, J.T.; Gonzalez, A.; Ackermann, G.; Aksenov, A.A.; Behsaz, B.; Brennan, C.; Chen, Y.; et al. American Gut: An Open Platform for Citizen Science Microbiome Research. *mSystems* **2018**, *3*, e00031-18. [CrossRef]
35. Goodrich, J.K.; Waters, J.L.; Poole, A.C.; Sutter, J.L.; Koren, O.; Blekhman, R.; Beaumont, M.; Van Treuren, W.; Knight, R.; Bell, J.T.; et al. Human Genetics Shape the Gut Microbiome. *Cell* **2014**, *159*, 789–799. [CrossRef]
36. Hill-Burns, E.M.; Debelius, J.W.; Bs, J.T.M.; Ba, W.T.W.; Ms, M.R.L.; Ms, Z.D.W.; Peddada, S.D.; Do, S.A.F.; Molho, E.; Zabetian, C.P.; et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Mov. Disord.* **2017**, *32*, 739–749. [CrossRef]
37. Pascal, V.; Pozuelo, M.; Borruel, N.; Casellas, F.; Campos, D.; Santiago, A.; Martinez, X.; Varela, E.; Sarrabayrouse, G.; Machiels, K.; et al. A microbial signature for Crohn's disease. *Gut* **2017**, *66*, 813–822. [CrossRef]
38. Liu, H.; Chen, X.; Hu, X.; Niu, H.; Tian, R.; Wang, H.; Pang, H.; Jiang, L.; Qiu, B.; Chen, X.; et al. Alterations in the gut microbiome and metabolism with coronary artery disease severity. *Microbiome* **2019**, *7*, 68. [CrossRef]
39. Keshavarzian, A.; Green, S.J.; Engen, P.; Voigt, R.M.; Naqib, A.; Forsyth, C.B.; Mutlu, E.; Shannon, K.M. Colonic bacterial composition in Parkinson's disease. *Mov. Disord.* **2015**, *30*, 1351–1360. [CrossRef]
40. Petrov, V.A.; Saltykova, I.V.; Zhukova, I.A.; Alifirova, V.M.; Zhukova, N.G.; Dorofeeva, Y.B.; Tyakht, A.V.; Kovarsky, B.A.; Alekseev, D.G.; Kostryukova, E.S.; et al. Analysis of Gut Microbiota in Patients with Parkinson's Disease. *Bull. Exp. Biol. Med.* **2017**, *162*, 734–737. [CrossRef]
41. Zhou, Y.; Xu, Z.Z.; He, Y.; Yang, Y.; Liu, L.; Lin, Q.; Nie, Y.; Li, M.; Zhi, F.; Liu, S.; et al. Gut Microbiota Offers Universal Biomarkers across Ethnicity in Inflammatory Bowel Disease Diagnosis and Infliximab Response Prediction. *mSystems* **2018**, *3*, e00188-17. [CrossRef] [PubMed]
42. Halfvarson, J.; Brislawn, C.J.; Lamendella, R.; Vázquez-Baeza, Y.; Walters, W.A.; Bramer, L.M.; D'Amato, M.; Bonfiglio, F.; McDonald, D.; Gonzalez, A.; et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* **2017**, *2*, 17004. [CrossRef] [PubMed]
43. Bermingham, E.N.; Young, W.; Butowski, C.F.; Moon, C.D.; MacLean, P.H.; Rosendale, D.; Cave, N.J.; Thomas, D.G. The Fecal Microbiota in the Domestic Cat (Felis catus) Is Influenced by Interactions Between Age and Diet; A Five Year Longitudinal Study. *Front. Microbiol.* **2018**, *9*, 1231. [CrossRef] [PubMed]
44. Jha, A.R.; Shmalberg, J.; Tanprasertsuk, J.; Perry, L.; Massey, D.; Honaker, R.W. Characterization of gut microbiomes of household pets in the United States using a direct-to-consumer approach. *PLoS ONE* **2020**, *15*, e0227289. [CrossRef]
45. Bian, G.; Gloor, G.B.; Gong, A.; Jia, C.; Zhang, W.; Hu, J.; Zhang, H.; Zhang, Y.; Zhou, Z.; Zhang, J.; et al. The Gut Microbiota of Healthy Aged Chinese Is Similar to That of the Healthy Young. *mSphere* **2017**, *2*, e00327-17. [CrossRef]
46. Li, Y.; Wang, H.-F.; Li, X.; Li, H.-X.; Zhang, Q.; Zhou, H.-W.; He, Y.; Li, P.; Fu, C.; Zhang, X.-H.; et al. Disordered intestinal microbes are associated with the activity of Systemic Lupus Erythematosus. *Clin. Sci.* **2019**, *133*, 821–838. [CrossRef]
47. Bolyen, E.; Rideout, J.R.; Dillon, M.R.; Bokulich, N.A.; Abnet, C.C.; Al-Ghalith, G.A.; Alexander, H.; Alm, E.J.; Arumugam, M.; Asnicar, F.; et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **2019**, *37*, 852–857. [CrossRef]
48. Chao, A. Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* **1984**, *11*, 265–270.
49. Callahan, B.J.; McMurdie, P.J.; Rosen, M.J.; Han, A.W.; Johnson, A.J.A.; Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **2016**, *13*, 581–583. [CrossRef]
50. Bokulich, N.A.; Kaehler, B.D.; Rideout, J.R.; Dillon, M.; Bolyen, E.; Knight, R.; Huttley, G.; Caporaso, J.G. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **2018**, *6*, 90. [CrossRef]
51. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2012**, *41*, D590–D596. [CrossRef] [PubMed]
52. Ritari, J.; Salojärvi, J.; Lahti, L.; De Vos, W.M. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genom.* **2015**, *16*, 1056. [CrossRef]

53. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
54. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
55. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
56. Gloor, G.B.; Macklaim, J.M.; Pawlowsky-Glahn, V.; Egozcue, J.J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **2017**, *8*, 2224. [CrossRef] [PubMed]
57. Deng, P.; Swanson, K.S. Gut microbiota of humans, dogs and cats: Current knowledge and future opportunities and challenges. *Br. J. Nutr.* **2014**, *113*, S6–S17. [CrossRef] [PubMed]
58. Garcia-Mazcorro, J.F.; Minamoto, Y.; Kawas, J.R.; Suchodolski, J.S.; De Vos, W.M. Akkermansia and Microbial Degradation of Mucus in Cats and Dogs: Implications to the Growing Worldwide Epidemic of Pet Obesity. *Veter Sci.* **2020**, *7*, 44. [CrossRef]