# Machine-learning Based Identification of Metagenomic Source Environment

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

A Thesis

Presented to the

Faculty of

San Diego State University

\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

In

Bioinformatics & Medical Informatics

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

by

Jillian Burke

Summer 2019

## Dedication

I dedicate this thesis to my science and mathematics teachers, who have shared with me not only their knowledge but also their love of the natural world and learning. I wish to give special thanks to Drs. Robert Edwards, Rebecca Bart, and Mark Campbell, for offering me guidance and opportunity as well as inspiration.

I also dedicate this work to my earliest and most influential teachers, my parents - Joan and Paul Burke, without whose unwavering support this thesis would not have been possible.

## Abstract

The Sequence Read Archive (SRA) is a publicly-available database that contains over 100,000 whole genome sequence metagenomes. The lack of a standardized classification system for describing the environment each of these genomes was sampled from has created a number of problems in the database, including many datasets with ambiguous or uninformative source information. Here, we develop a quantitative method that can quickly predict the environment from which a metagenome was sampled. A training set of 2176 manually classified genomes from the SRA database was created and the percent abundances of various bacteria from known environments in each of these genomes were predicted using the computational tool FOCUS. These bacterial abundances, as well as several other metrics including percent human DNA and percent viral DNA, were used to train a random forest classifier. I demonstrated a 78.5 percent accuracy in classifying metagenomes based on these metrics, demonstrating that this approach can be used to automatically classify sequences in the SRA.

# 

# Chapter 1

# Introduction

## 1.1 Metagenomics and Microbiome Data

Metagenomics refers to the study of genetic material taken directly from the environment. A metagenome typically contains data from a complex mixture of organisms that coexist within an environment, including microorganisms such as bacteria and viruses and is used to understand the community-wide composition and/or dynamics of the environment1. In contrast to traditional genomic studies, which sequence the genetic material from a single organism at a time using purified samples from a single source rather than direct environmental sampling**.**

As with most fields of biology that rely upon genetic sequencing, short-read based technology is the most commonly used approach in metagenomic studies while third-wave, or long-read sequencing technology, is less commonly used due to the relatively small amount of data retrieved via this technology, increased cost, and higher error rate, which can inhibit the accurate identification of closely related organisms within a community. However, it is possible to obtain metagenomes using this approach 2 and as this technology improves its use in metagenomics is likely to increase. RNA sequencing, sometimes called metatranscriptomics, is also used 3.

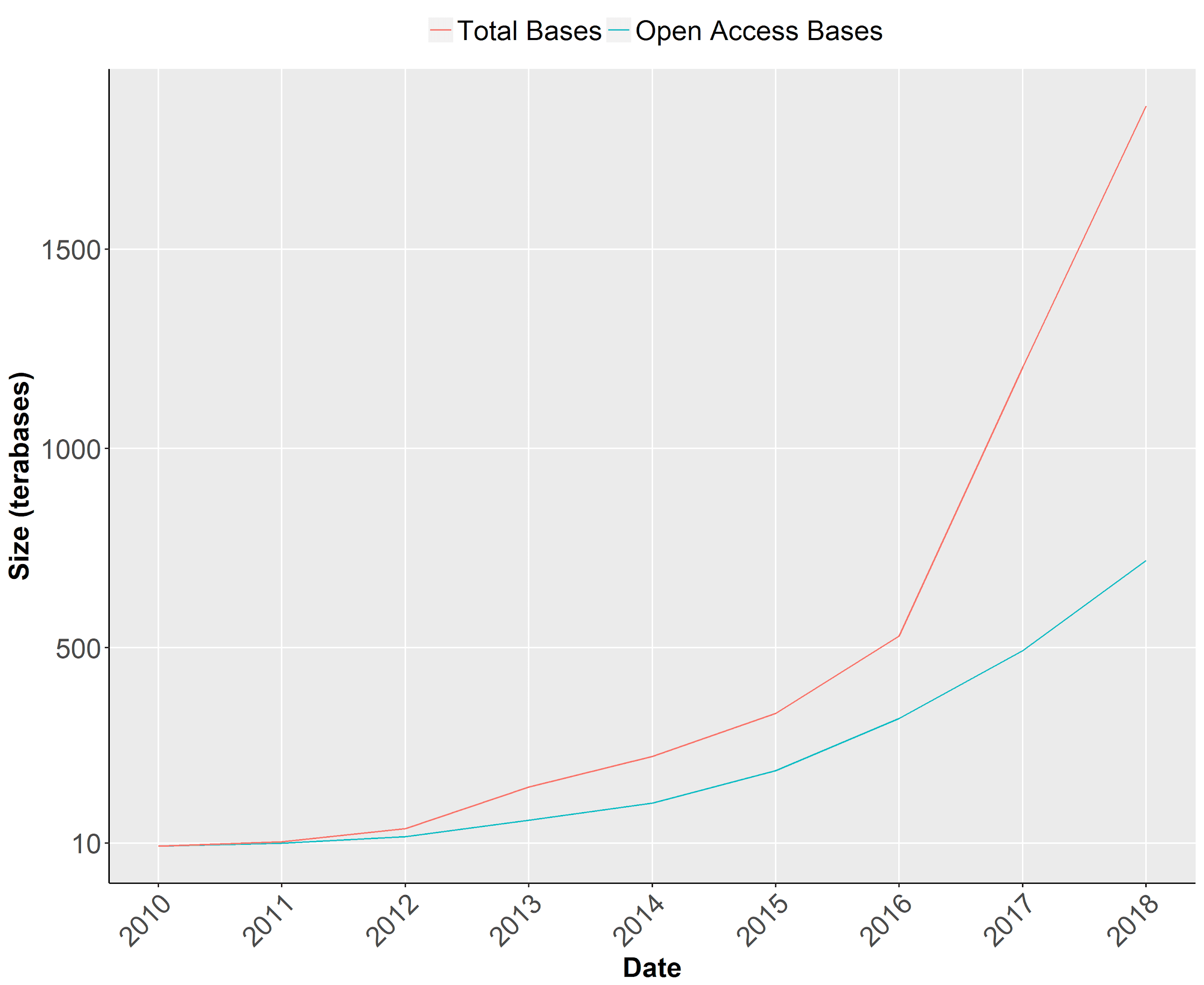
The term metagenome can be used to describe two distinct approaches to sequencing strategies: whole-genome sequencing (WGS) or sequencing of specific regions of the genome, typically the 16s rRNA region of bacteria, also known as amplicon sequencing 4. In the WGS sequencing approach, the entirety of the DNA recovered from the sample is sequenced. The methods discussed here apply only to WGS and whole RNA-sequencing approaches, irrespective of the specific sequencing platform used to obtain the data.

## 1.2 Public Databases in Biological Science

With the increasing focus on sequencing technology, biological science has also become increasingly reliant on computer science to efficiently store and access data, leading to the creation of databases to house this information 5. Much of the information is also made available to researchers through open-access or public databases, designed to give scientists the ability to access as well publish their own data. Yet simply making this data available is not enough. People who wish to use it must also have the knowledge and capability to identify and download the data they seek. This points to a growing need for tools and protocols to help those who would use this wealth of open access data do so efficiently.

## 1.3 The Short Read Archive (SRA) Database

One database designed to make the rapidly expanding collection of genomic sequencing data available to researchers is the Short Read Archive (SRA) created by the National Center for Biotechnology Information (NCBI). This database already houses over 100 terabases of open access next-generation sequencing data and continues to grow exponentially 6. Metagenomic data makes up an estimated 11% of this, or roughly 100,000 individual runs 6.

****

**Figure 1. The Growth of the SRA Database Over Time.** The SRA database has more than doubled in size in the past three years and continues to expand to house additional sequencing datasets**.** Source: National Center for Biotechnology Information, Sequence Read Archive Documentation. https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

When uploading data, users provide “free text” descriptions of their data. Each individual run is assigned a unique ID. There is currently no ontological approach to classify the sources or environments from which these samples were taken. As a result, datasets are frequently labeled ambiguously. For instance, one sample (SRR2187142) is described as a “gut metagenome” but it is not clear if this sample was taken from a human gut or the gut of another animal. Other entries have descriptions that conflict with the actual contents of the data suggesting they may be either mislabeled or contaminated. Multiple samples from the TARA Oceans Project, for example, are described as containing seawater but analysis of the contents shows that the vast majority of the bacterial DNA in the sample are bacteria found on humans and these samples contain less than 2% DNA from microorganisms typically present in ocean water (ERR1711977, ERR1712061, ERR1739844), leading to speculation that these are contaminated in some way.

## 1.4 Project Aims

I created a quantitative method to predict the source environment of a WGS metagenome, relying on the content of the data set rather than user-supplied information. I built a classifier for metagenomes from the SRA database, however, it’s application is not limited to data from the SRA. This approach and the algorithms described here can be used to classify any metagenome, or for example, to ensure that samples have not been mislabeled or contaminated.

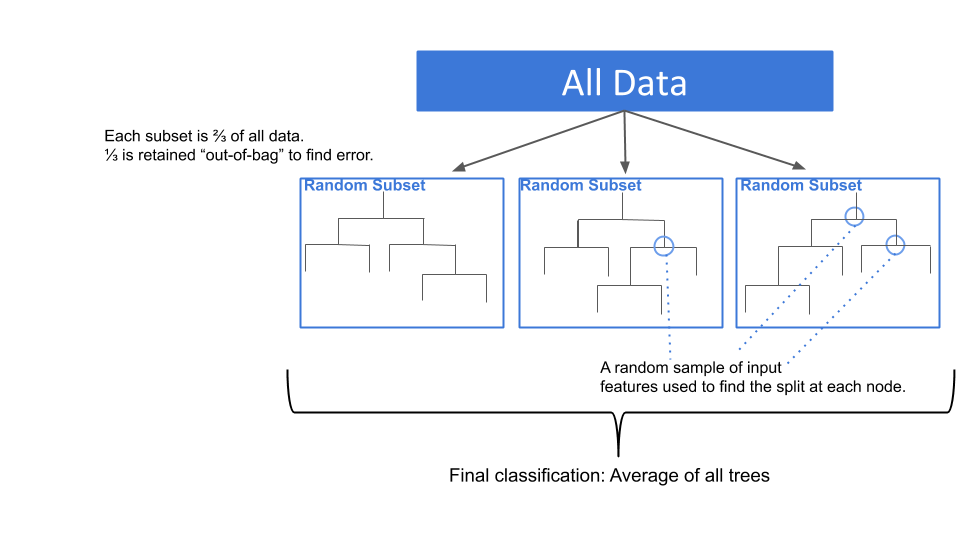
# 

# Chapter 2

# Methods

## 2.1 The Random Forest Algorithm

A random forest method, implemented using the R package randomForest, was used to create a classification model 7. The random forest method is based on bagging of decision trees. The algorithm creates multiple trees using a random subset of data points from the training data in each tree. The exact number created is defined by the user. At each node in a decision tree, the algorithm selects a random subset of the input variables from which to base the split at that node. The user can also define the exact number of variables to select from the total set of input variables. Bagging is then used to assign a final classification to each data point, which is the mode (most common) of the results from all the trees 8**.**

****

**Figure 2. The Random Forest Algorithm.**

The random forest algorithm has several key advantages that make it well suited to this application. Most notably, it is relatively robust to noisy data and outliers due to use of multiple random subsets of the training data. This is particularly valuable in this application as only a small sample of the total number of metagenomes from the SRA database are included in the training set and the data was not assessed for the presence of outliers before beginning. This method is also faster than other bagging methods but at least as accurate 8. Random forests also perform better handling different data sets such as percentages, read counts, and categories, than most machine learning techniques 9.

Finally, the random forest algorithm contains an internal assessment of accuracy called out-of-bag error. As each tree is computed using only a sample, roughly a third, of the entire data set, the remaining two-thirds of the data can be passed through that tree and the accuracy of the tree at predicting the class of each of these “out-of-bag” data points can be determined. After all the trees have been created, the accuracy of predicting each data point is average to create the overall out-of-bag error estimate 8. The random forest algorithm computes this estimate as it runs, which is an advantage as it means that neither a testing set nor any additional computation is not needed to determine the accuracy of the classifier.

## 2.2 Categorical Outcome Variable: Environmental Sources

To be able to classify metagenomes into environments, a categorical outcome variable that is essentially a list of all possible environmental sources is needed. In order to create such a variable, 7,023 bacterial genomes from the PATRIC database were manually curated based on their free-text field, isolation source, and assigned a source location. In other words, the user-supplied description of each of these genomes was individually read and then placed into a category based on where it was sampled from. This curation process resulted in 49 distinct environmental sources. This set of environmental sources was used as the outcome variable in the random forest classifier. (See Supplemental Fig. 1, Appendix A, for a complete list of source categories.)

There is overlap between certain categories and to some extent, this problem is unavoidable when developing any list of environmental sources. A metagenome that was sampled from a bag of raw spinach, for instance, could be correctly included in the category “food” or in the category “plant” and both would be accurate. There is some unavoidable level of noise that is inherent in any system of environmental source classifications that attempts to use a limited ontology to describe complex environments. As described in Section 2.1, however, the random forest tends to be robust to noise within data.

The classification system is intended to provide a baseline for predicting the sampling environment of a metagenome and is not an exhaustive list of all possible environments. It is intended to be as inclusive as possible, however, so that as many genomes as possible may be meaningfully represented by one of these environmental categories. As a result, some of the categories are quite broad, such as “animal,” “human other,” and “water other.” Categories containing the word “other” indicated that similar environments exist within the list of sources but that they do not accurately describe entries from this environment. For example, multiple categories exist to describe data derived from water samples, including freshwater, marine benthic, marine coastal, hot spring, wastewater, etc. Yet some samples, such as those samples from brackish water (a mixture of saltwater and freshwater) are not accurately described by these sources and so the “water other” category was included. For this reason, this classification system is best viewed as a preliminary list that can be subsequently refined to provide better accuracy. Moreover, the very broad category “animal” is a starting point, and those samples can be further refined using text mining and other techniques to sub-classify them.

## 2.2.1 Refinement of the Outcome Variable

Some of the original categories, including human heart, non-human blood, and astronomical body, were removed because less than three entries for them could be found in the PATRIC database. It was assumed that the presence of only one or two bacteria from a particular source would not be enough to accurately classify a metagenome. Outside of this manual refinement, a computational analysis was used to further refine the list of sources.

After creating a preliminary random forest classification, the confusion matrix of the initial results, which shows the accuracy of each level of the source categories and what into what category that mislabeled data were places, was assessed. From this matrix, it was possible to see that certain categories were most often mislabeled as one other category in particular. For example, the groups “human lungs,” “human mouth,” and “human nose” were commonly misclassified as one another. This makes sense intuitive sense as well, considering that these three environments are actually connected physically by air and fluid. PCA analysis (not shown) of the input data also showed that entries from these three groups had the lowest values of the first and second principal components. As a result, these categories were combined into a single group, “human respiratory,” in the final set of source environments.

From the confusion matrix of the initial classifier, it was also possible to see that microbial mat metagenomes were most frequently being misclassified as hot spring samples. and so both the training data and PATRIC source databases were changed so that microbial mats from hot springs were included in the hot spring category rather than the microbial mat category in order to allow for better separation of these two categories. This same process led to the creation of the “biogas material” category after inspection of the confusion matrix revealed that many misclassified samples from various categories including “chemical vessel” and “solid environmental material” were related to biogas production. The random forest classifier was then regenerated using the updated source categories.

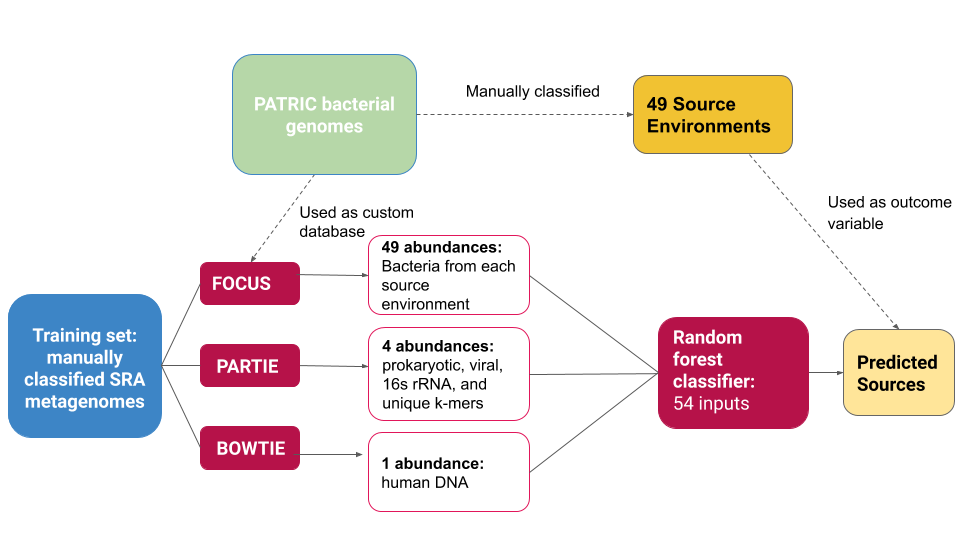
## 2.3 Input Features for the Random Forest Classifier

The input features for creating the random forest classifier are percent read abundances. The majority (49 of 54) of these features are derived from the output of the FOCUS tool, which compares a metagenome to a database of k-mers from known organisms in order to predict which of these organisms are present in the metagenome and at what abundance 10. For this application, a custom database containing only bacteria from the PATRIC database was used. To create this custom database, the 7,024 entries from PATRIC had been previously classified and assigned one of the 49 possible source environments were first included (see section 2.2), and then the remainder of the PATRIC database was searched for entries with a source description identical to one of the manually classified entries. These genomes were given the same environmental category as the matching entry and also included in the FOCUS database, for a total of 102,129 bacterial genomes.

FOCUS predicts only the bacteria present and their relative abundance, so these read abundances were then added according to their assigned environment to obtain the estimated read abundance for each of the 49 possible sources. These abundance values were each used as input features. For example, if FOCUS finds reads in a metagenome that contains 10 different bacterial species in the database of PATRIC genomes with the label “freshwater”, and the total abundance of reads from these 10 bacteria combined is 13%, then the value of the input feature “freshwater” would be 13.0 for that metagenome. If no reads from a particular source environment were found in a metagenome, a value of 0 was used.

In addition to these 49 abundances, the PARTIE tool was used to generate four additional input features. By default, PARTIE computes four statistics about a metagenome: the percentage of k-mers that are unique within the data set, the percentage of reads that are from prokaryotic organisms, from 16s rRNA, and from virii 11.

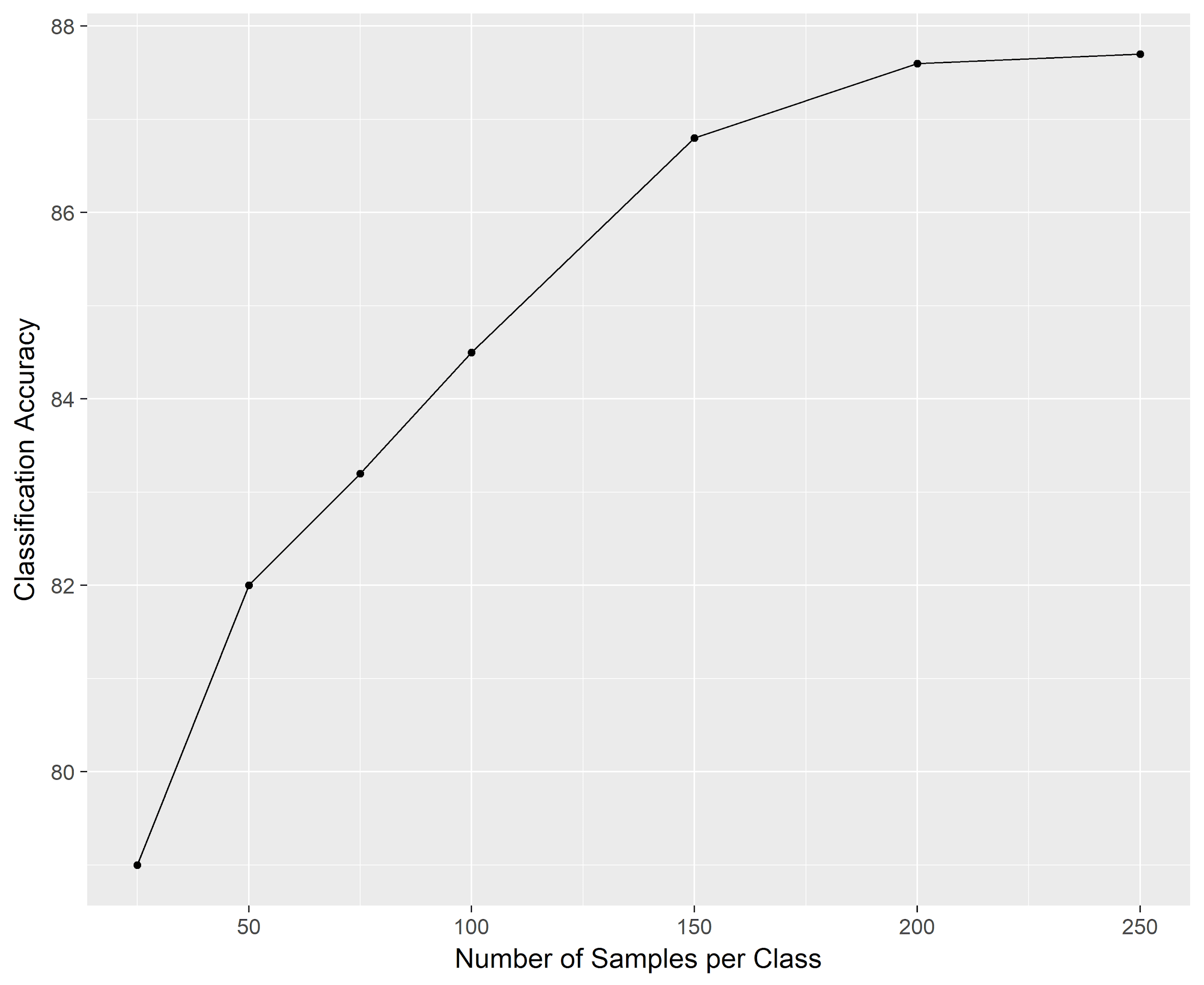
Finally, the percent of the reads in each metagenome that are derived from human DNA was estimated and used an input feature. This was done by using Bowtie2 to map the reads from each metagenome to the human reference genome GRCh38/hg38 12.



**Figure 3. Schematic Overview of the Development of the Random Forest Classifier.** There are 57 input features, all abundance values derived from the output of two existing tools PARTIE and FOCUS. The output of the random forest classifier is a 49-level categorical variable of source environments. This variable was developed by manual curation of a large set of bacterial genomes from the PATRIC database. A training set of SRA genomes was used.

## 2.4 Selection of Training Data

In order to develop a training set, abstracts from the SRA were individually examined using keyword searches to identify candidates. Those entries with clear, unambiguous source descriptions were assigned a “ground-truth” value consisting of one level of the output variable (source environment). For 12 of 49 the categories, it was possible to identify at least 250 metagenomes from the SRA in this manner, and from this subset of categories a learning curve was created (Fig. 4). The learning curve shows that the rate at which the model gains accuracy by increasing the size of the training begins to taper off around 150-200 samples per environmental category, suggesting that the model would perform best when at least 150-200 samples for each environment are included.



**Figure 4. The Learning Curve for 12 Environmental Categories**

Certain environmental categories were difficult to identify using keyword searches however, and it is possible that the SRA simply does not contain many metagenomes from these sources at this time. For 18 of the 49 of the environments at least 100 genomes could be identified by manual curation and for an additional 6 categories between 50-99 metagenomes were classified. The remaining categories, for which less than 50 genomes could be identified, were excluded from the random forest classifier at this time due to low sample size.

Because the minimum number of samples for some categories is 50, the maximum number of samples in the training set for each class is capped at 100 in order to prevent the data from being highly unbalanced. The highest ratio of samples in the training sets for any two classes is 2:1, representing a modest imbalance in the input data13.The random forest algorithm can be affected by imbalanced training sets as it relies upon building decision trees from a random subset of the data. When one class has a lower number of samples in the training set, that class is less likely to be chosen for inclusion in the decision trees when the model is being computed, causing the accuracy of the underrepresented, or minority, classes to decrease**.** As a rule of thumb, imbalanced datasets tend to impact the performance of the model when a ratio of 10:1 data points between classes is reached, though even smaller imbalances can have measurable effects 13.

As seen in Table 1, with this level of imbalance, the error rates of the underrepresented classes increase by 2% or less. With higher levels of imbalance, such as a 5:1 ratio between classes in the training set, the accuracy of the minority classes begins to suffers significantly. The overall error of the model improves by 4% using a modestly imbalanced design, with a range of 50-100 samples per class, over the balanced design with only 50 samples per category. This suggests that a modestly imbalanced design model increases the overall accuracy of the model, as it allows higher sample sizes to be used, though at the cost of a minor loss in accuracy in the underrepresented classes.

**Table 1. Impact of imbalanced designed on the within-group accuracy of the underrepresented (minority) classes.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Minority Category (sample size in imbalanced designs)** | **Error in balance design (n=50)** | **Error in imbalanced design (max. 2:1 ratio between classes)** | **Error in imbalanced design (max. 5:1 ratio between classes)** |
| Air (58) | 10 % | 12 % | 15 % |
| Chemical Vessel (99) | 26 % | 27 % | 31 % |
| Fungus (53) | 24 % | 26 % | 36 % |
| Marine Benthic (63) | 32 % | 33 % | 47 % |
| Waste water (92) | 40% | 38 % | 54 % |
| Water other (63) | 20 % | 21 % | 38 % |

## 2.5 Evaluation of Performance

As described inSection 2.1, the random forest algorithm calculates an out-of-bag error rate while it develops the classifier. This means that, unlike many machine learning methods, this data does not need to be subdivided into a training and testing set in order to assess the accuracy of the model. This allows for all of the data to be supplied to the algorithm during the development of the classifier, yielding a higher accuracy due to the larger sample size than would be possible if the data were split the data into both a training and testing set.

Plotting the accuracy of the model as a function of the number of trees used, showed that accuracy did not substantially increase when using more than 2,500 trees and so for the final model, 2,500 trees were used. (See Appendix A, Fig. 10.)

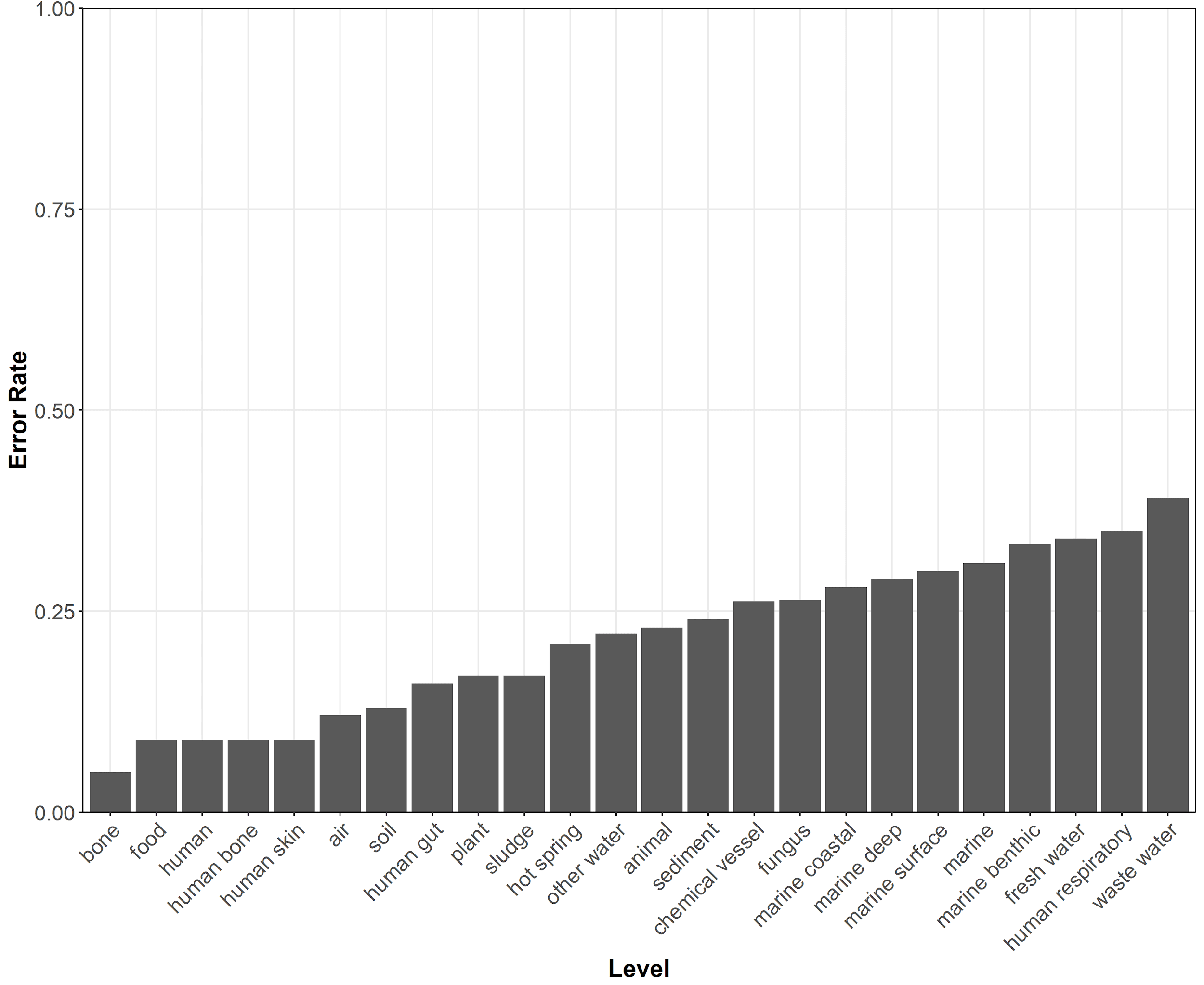
## 

## Chapter 3

## Results

## 3.1 Accuracy of the Random Forest Classifier

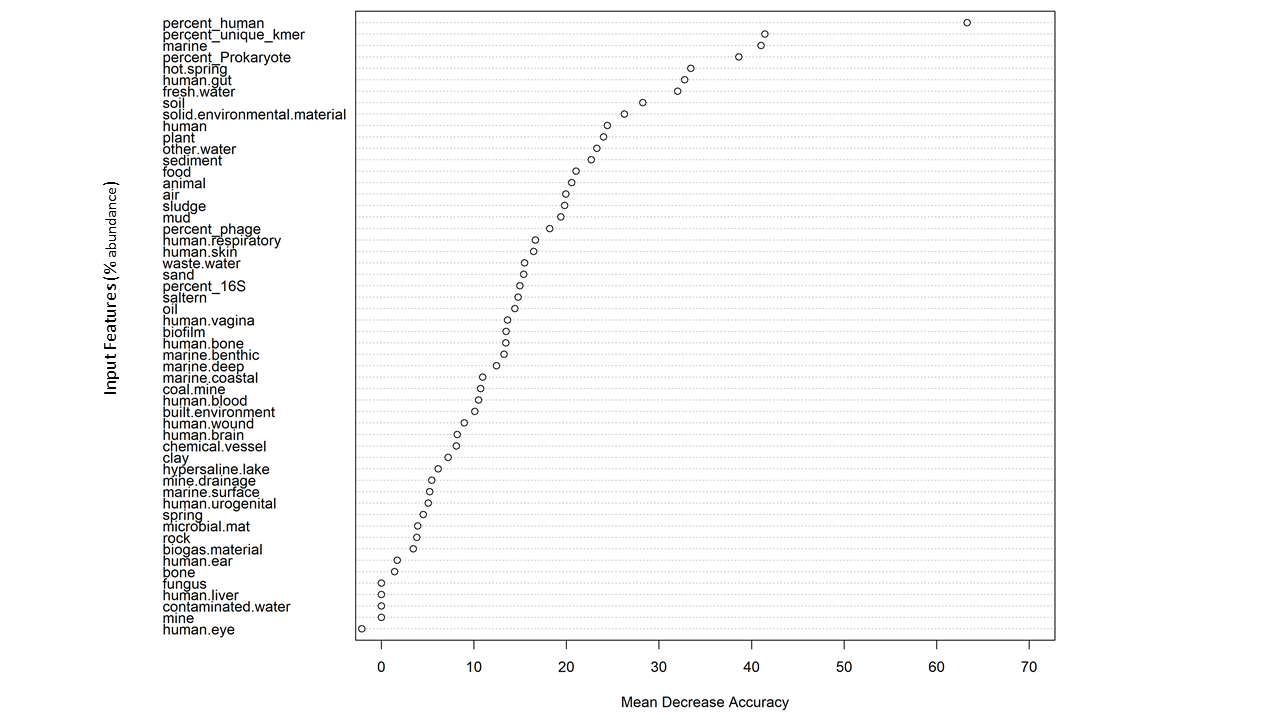
A random forest classifier resulting from 2,500 treesdemonstrated an out-of-bag error rate of 21.5%. There is considerable variation in how the well the model performed at predicting each of the individual levels of the source variable. Fourteen of the twenty-four environmental sources have error rates below 25%, while the rest have relatively high error rates, above 25% (See Fig. 5). The highest error rate seen is 38% in the wastewater category and the lowest is 5% in the bone category. (See Appendix A, Fig. 9 for results of the balanced design model).

****

**Figure 5. Out-of-bag Error Rates by Level of Classifier Variable (Source Environment)**

## 3.2 Feature Importance

A ranking of the importance of the input features, based on the mean decrease in accuracy of the model when this feature is excluded from the analysis was completed (See Fig. 6). Though certain input features contribute little to model, only 24 out of 49 total possible levels of the output variable have been included at this time and so the relative importance of these features may change as additional levels of the output variable are included in the model. For this reason, they have not been excluded from the classifier at this time, despite adding little information to the end result.

****

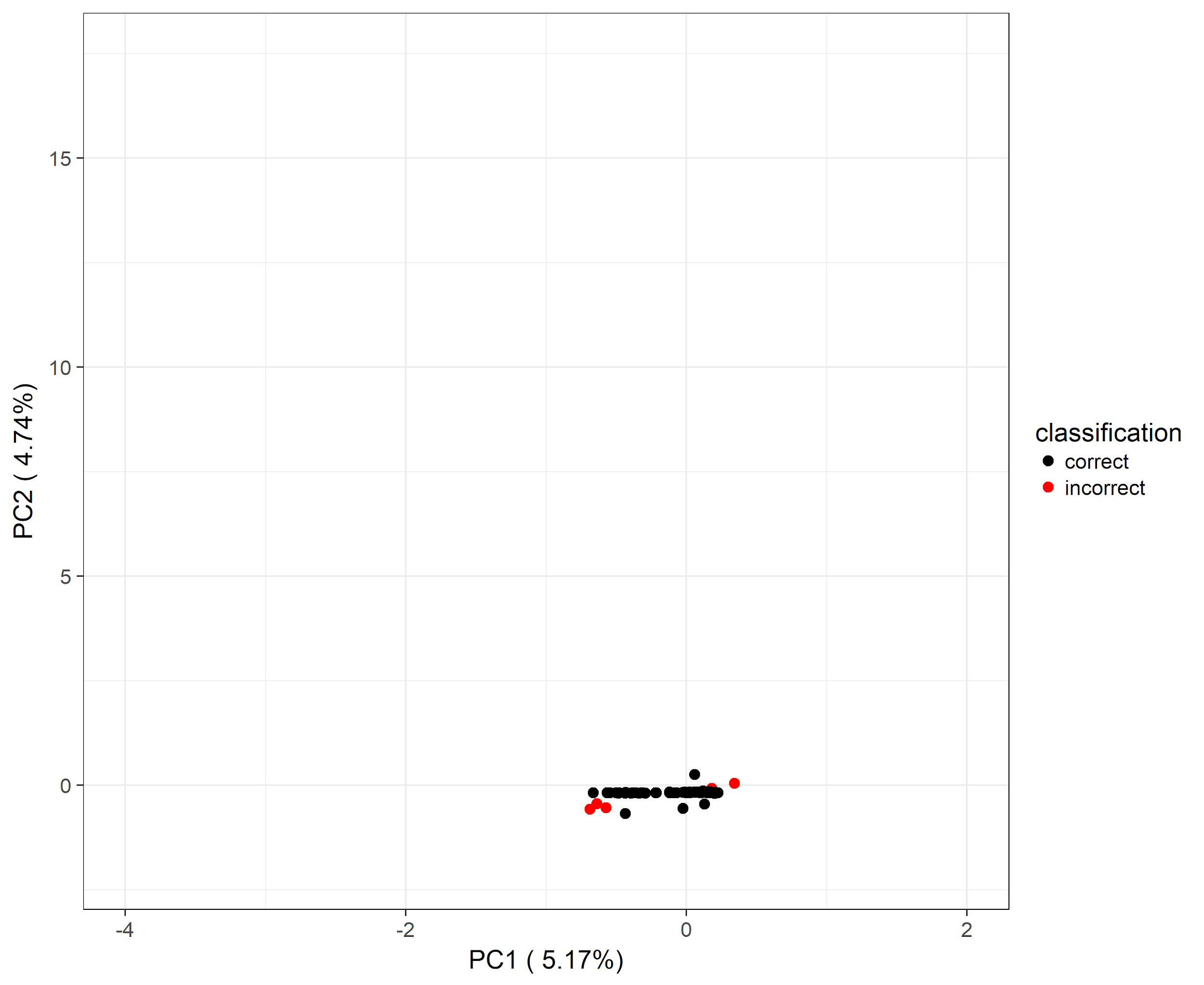
**Figure 6. Ranked Variable Importance.** The mean decrease in accuracy (x-axis) that

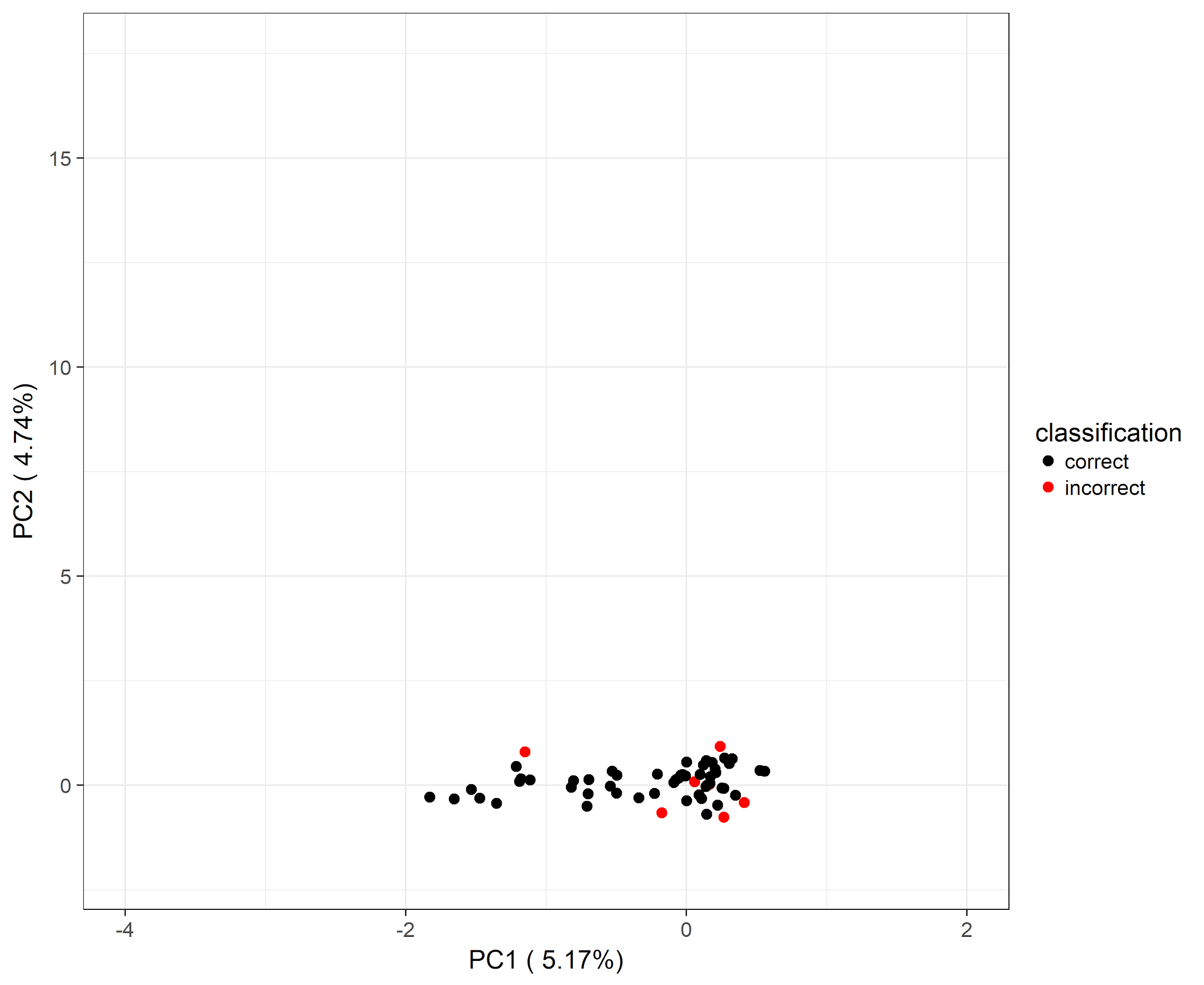
the model experiences when the variable (y-axis) is left out is shown.

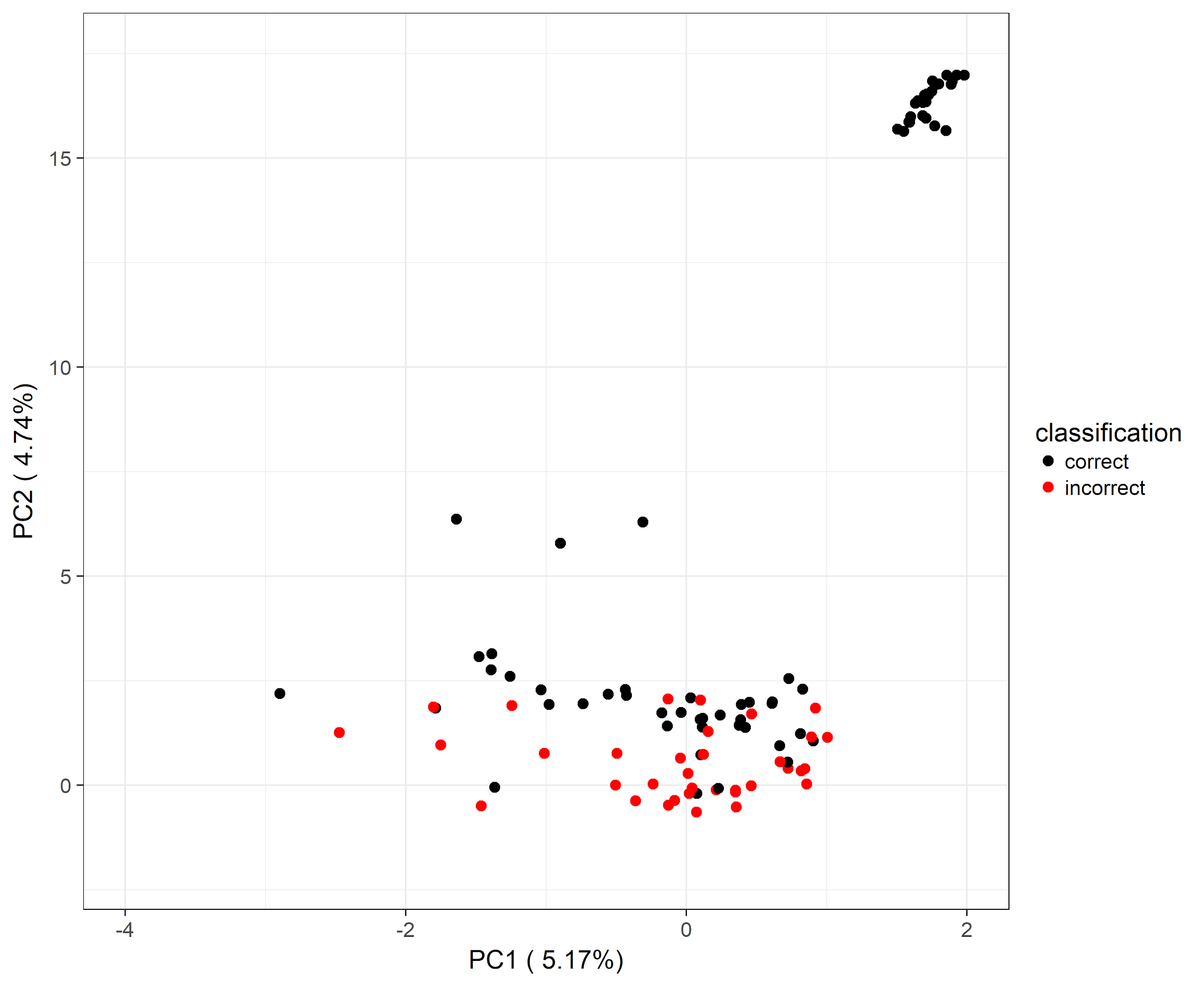
## 3.3 Comparison of High and Low-Accuracy Environments

Principal component analysis (PCA) of the training data computed using R version 3.4, suggests that the accuracy of the levels of the variable may be influenced by the diversity of the training data. A single PCA of the entire dataset was performed. The results of this PCA are shown in Figures 7-8, with selected points displayed so that each plot contains only a single level of the class variable.

Two highly accurate levels, “bone” and “air” are shown in Figures 7A and 7B. In comparison, two groups with lower accuracy, “human respiratory” and “wastewater,” are shown in Figures 8A and 8B.

**Figure 7A. PCA Results for Selected High-Accuracy Environmental Categories: Bone.**

**Figure 7B. PCA Results for Selected High-Accuracy Environmental Categories: Air.**

****

**Figure 8A. PCA Results for Selected Low-Accuracy Environmental Categories: Human Respiratory.**

# 

**Figure 8B. PCA Results for Selected Low-Accuracy Environmental Categories: Wastewater.**

# Chapter 4

# Discussion

We have shown that it is possible to identify the environmental source of a metagenome using a random forest method with roughly 78% accuracy and using only a relatively limited sample size. The input features used were primarily bacterial read abundances predicted using the FOCUS tool with a custom database of bacteria that had been manually assigned source environments. A ranking of the feature importance supports the decision to include additional features as well, such as the percentages of human DNA content, prokaryotic content, and unique k-mers, which are amongst the most important features to the accuracy of the model.

The method shown here could be used to classify metagenomes from open-source databases such as the SRA and improve their usefulness as a research tool. The levels of the outcome variable, or the environmental sources, could serve as a standardized vocabulary that researchers could use when uploading their data to the SRA or similar databases, improving the organization and ease of use of the sites.

## 4.1 Future Work

The fact that the method relies on a training set means that roughly half of the levels of the outcome variable are still not included in the classifier due to the lack of adequate training data sets. One remaining step of this project, therefore, is to expand the development to including all 49 levels of the source variable. These categories are likely rare within the SRA and thus they could not be located by searching the SRA for keywords related to the environments. It is possible for some of these categories there are not enough datasets deposited in the SRA database at all and that the amount of required data to build a classifier is not publically available at this time.

In addition, to expanding the training set to include the currently unclassified categories, more samples from the six minority categories (i.e. those that do not yet contain 100 data points in the training set) should be included in order to achieve balanced input data, which may increase the accuracy of these categories. The learning curve also shows that the overall accuracy continues to increase until 200 samples per class, whereas this model includes only roughly 100 samples per class. It may be possible to further reduce error by increasing the number of samples in each class from 100 to 200.

The results of a PCA suggest that to further improve accuracy, some adjustment of the outcome variable (environmental sources) may be required. In the higher-accuracy classes, the data appears more tightly clustered and better correlated. In the most accurate category, “bone,” the data forms a single linear cluster along the x-axis, suggesting that the variability in this group can be explained largely first principal component. In comparison, the lower accuracy groups appear relatively diffuse. The group with the lowest accuracy, “human respiratory,” looks to contain at least two distinct clusters. Breaking groups such as this into multiple, more homogenous categories may achieve better predictive accuracy.

For example, all animal samples are placed in one category, meaning that a sample from a domestic farm animal and that of a bee are placed in a single category. While this is technically true, a more accurate classification system would likely result from breaking the animal category into several small ones. Future work should include the refinement of this preliminary classification variable.

Finally, a more balanced representation of the bacteria from each source location within the set of bacterial genomes from the PATRIC database that are used to create the input features may also lead to improved accuracy. Environments that have been more frequently studied. like soil and the human gut, have a higher representation in this database and accordingly, more of these categories were annotated for inclusion.

## 4.2 Limitations

This method requires the creation of a list of discrete environments in order to classify each genome. In nature, environments are rarely truly separate and instead interact and exchange material with one another. For example, the classification system described here includes the environments “marine coastal” and “sand,” however in nature, coastal ocean waters will contain sand and sand may also be wet with ocean water, so these environments physically overlap. It is likely that many samples that will be classified via this standard vocabulary do not neatly fit into one of the source classifications but instead could be seen as being in at least two categories. This is the primary reason that many ontological approaches consider a directed acyclic graph to allow multiple categories to overlap.

As the database is continually growing and novel environments continue to be sequenced, no fixed set of source classification can ever be truly exhaustive. Accordingly, some researchers attempting to use this system to describe or classify their data would also find that their data does not fit accurately into any of the categories contained in this method. Despite these limitations, this framework provides a baseline that may be updated or improved upon in further work.

A further limitation of this approach is that is requires human curation to identify training data and bacterial genomes for inclusion in the FOCUS database. Any manual curation step is prone to human errors and so it is possible that mislabeled entries may exist in either the training data or the classifications of the bacterial samples from PATRIC. This limitation is potentially overcome by the use of the random forest method, which has been demonstrated to perform well even with weakly-correlated input features and errors within the output classifications (up to 5%) of the training set.

## 4.3 Conclusion

This proposed list of source categories and random forest-based classification approach allows researchers a quick method for checking that the data they have downloaded from the SRA or other sources is actually correctly annotated, or to check for contamination of their datasets. It also aims to provide a set of standardized vocabulary or keywords to be used in source descriptions when uploading data to the SRA database. Adopting such a set of standardized source descriptions could substantially increase the ease of use of this and other open-access databases.

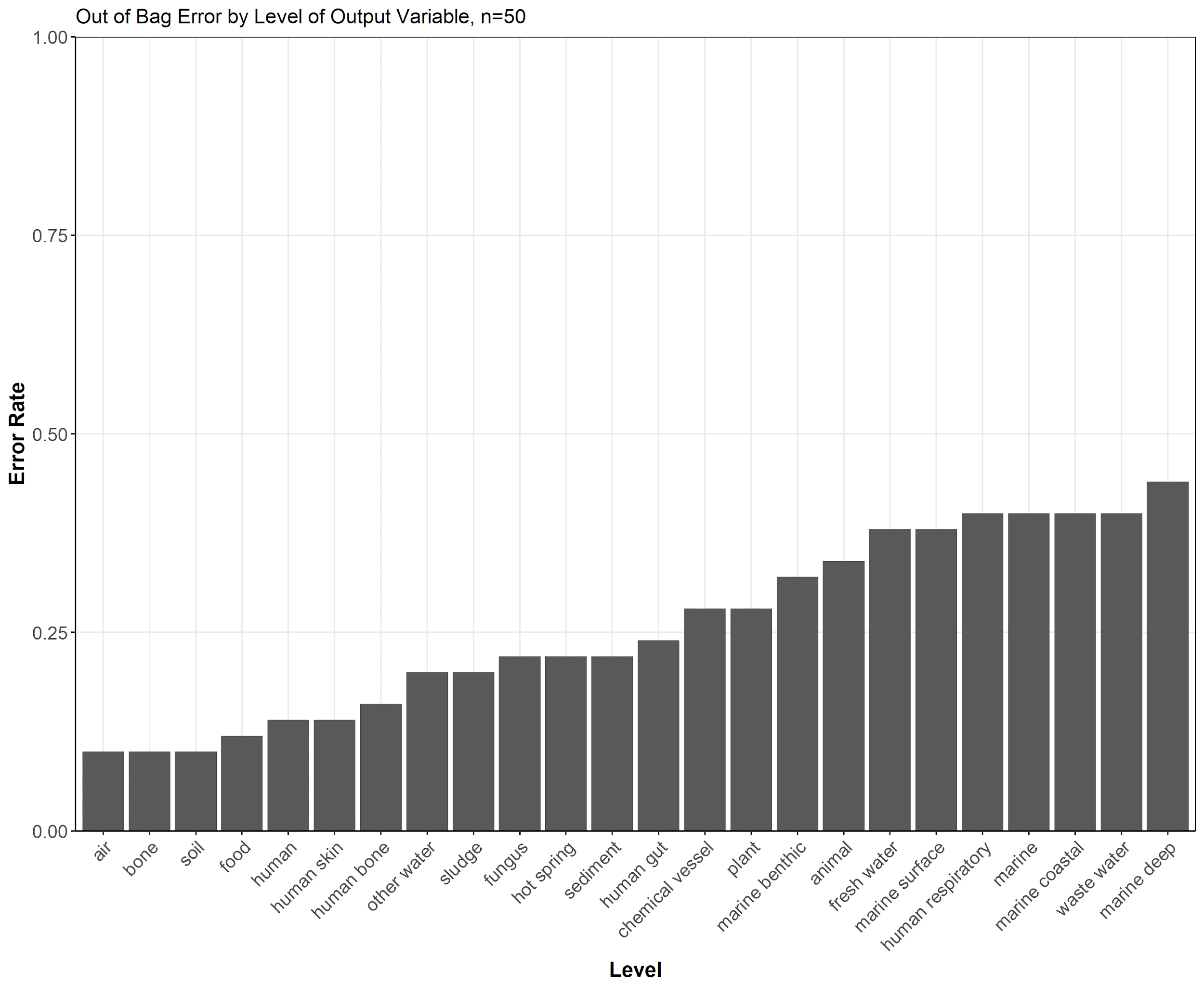
# 

# Appendix A

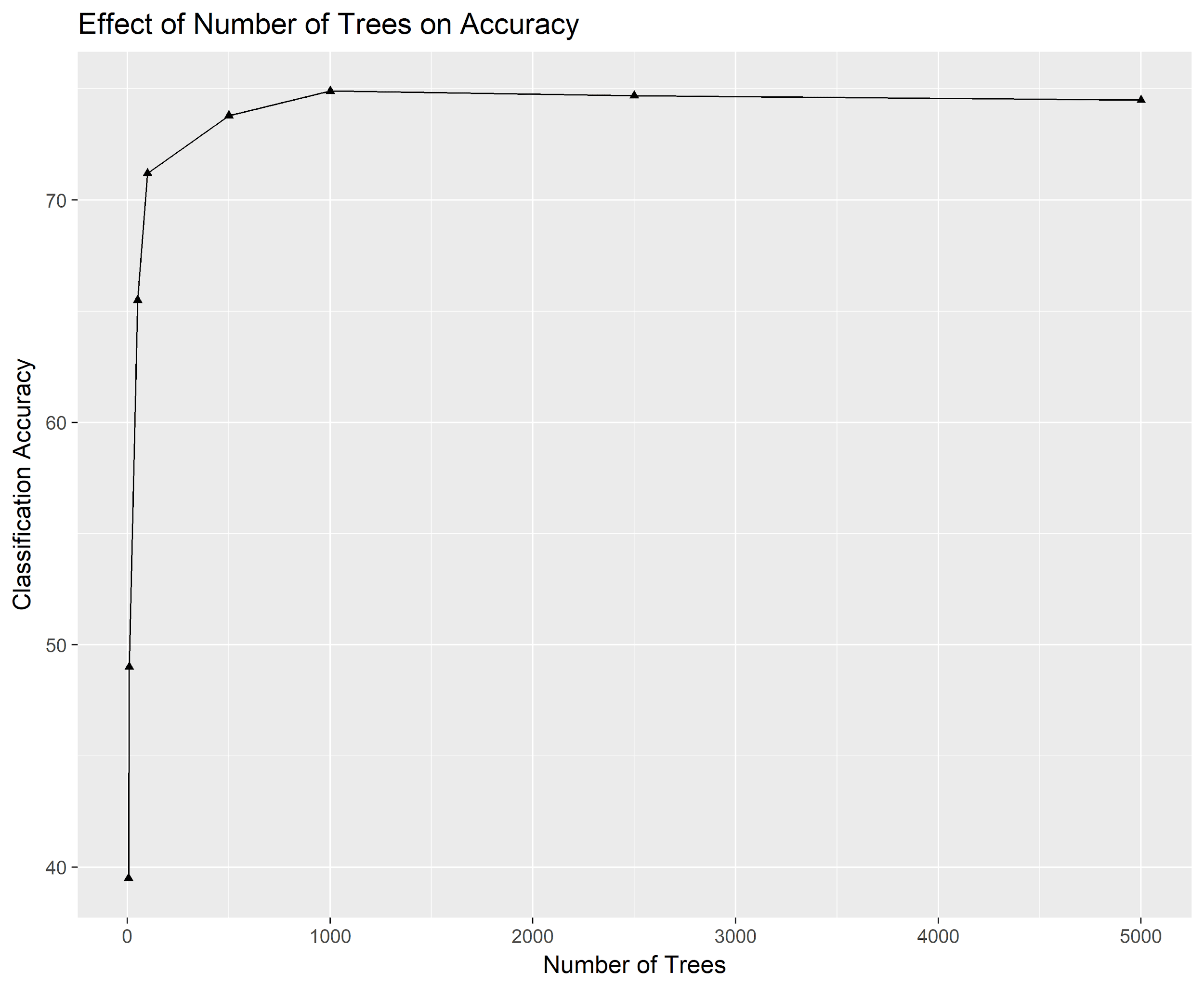
## Supplemental Figures

**Table 2. Full List of Source Environments (outcome variable).**

|  |  |
| --- | --- |
| # | Environmental Source |
| 1 | air |
| 2 | animal |
| 3 | biofilm |
| 4 | biogas material |
| 5 | bone |
| 6 | built environment |
| 7 | chemical vessel |
| 8 | clay |
| 9 | coal mine |
| 10 | contaminated water |
| 11 | food |
| 12 | fresh water |
| 13 | fungus |
| 14 | hot spring |
| 15 | human other |
| 16 | human blood |
| 17 | human bone |
| 18 | human brain |
| 19 | human ear |
| 20 | human eye |
| 21 | human gut |
| 22 | human liver |
| 23 | human respiratory |
| 24 | human skin |
| 25 | human urogenital |
| 26 | human vagina |
| 27 | human wound |
| 28 | hypersaline lake |
| 29 | marine |
| 30 | marine benthic |
| 31 | marine coastal |
| 32 | marine deep |
| 33 | marine surface |
| 34 | microbial mat |
| 35 | mine |
| 36 | mine drainage |
| 37 | mud |
| 38 | oil |
| 39 | other water |
| 40 | plant |
| 41 | rock |
| 42 | saltern |
| 43 | sand |
| 44 | sediment |
| 45 | sludge |
| 46 | soil |
| 47 | solid environmental material |
| 48 | spring |
| 49 | waste water |



**Figure 9. The Out-of-Box Error Rates of the Random Forest Classifier Using a Balanced Design: 50 samples per category. Overall Error: 25.5%**



**Figure 10. The Accuracy of the Random Forest Classifier as a Function of the Number of Trees.**

## 

## References

1. Dinsdale, E. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008)
2. Driscoll, C., Otten, T., Brown, N., & Dreher, T. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Stand. Genomic Science* **12**(9),(2017)
3. Lim, Y. *et al*. Metagenomics and metatranscriptomics: windows on CF-associated viral and microbial communities. *Journal of Cystic Fibrosis* **12**, 154–164 (2013)
4. Jovel, J. *et al*. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology* **7**, e459 (2016)
5. Varmus, H. Genomic empowerment: the importance of public databases. *Nature Genetics* **35**, 3 (2003)
6. Kodama, Y., Shumway, M., & Leinonen, R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, 54-6 (2011)
7. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**(3), 18-22 (2002)
8. Breiman, L. Random forest. *Machine Learning* **45**, 5–32 (2001)
9. Shi, T. & Horvath, S. Unsupervised learning with random forest predictors. *J Comput Graph Stat* **15**, 118–138 (2006)
10. Silva, G., Cuevas, D., Dutilh, B., & Edwards, R. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* **2**, e425 (2014)
11. Torres, P., Edwards, R., & McNair, K. PARTIE: a partition engine to separate metagenomic and amplicon projects in the Sequence Read Archive. *Bioinformatics* **33**(15), 2389–2391 (2017)
12. Langmead, B. & Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4), 357–359. Published (2012)
13. Weiss, G. *Foundations of Imbalanced Learning.* (Fordham University Press, 2012)