Host Phenotype Prediction from Differentially Abundant Microbes Using RoDEO

Anna Paola Carrieri[1], Niina Haiminen[2], and Laxmi Parida[2](✉)

[1] IBM Research UK, Warrington WA4 4AD, UK
[2] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
parida@us.ibm.com

Abstract. Metagenomics is the study of metagenomes which are mixtures of genetic material from several organisms. Metagenomic sequencing is increasingly used in human and animal health, food safety, and environmental studies. In these high-dimensional (metagenomic) data, the phenotype of the host organism, e.g., human, may not be obvious to detect and then the ability to predict it becomes a powerful analytic tool. For example, consider predicting the disease status of an individual from their gut microbiome.

In this study, we compare various normalization methods for metagenomic count data and their impact on phenotype prediction. The methods include RoDEO, Robust Differential Expression Operator, originally developed for gene expression studies. The best prediction accuracy is observed for RoDEO-processed count data with linear kernel support vector machines in most cases, for a variety of real datasets including human, mouse, and environmental samples.

We also address the problem of identifying the most relevant microbial features that could give insight into the structure and function of the differential communities observed between phenotypes. Interestingly, we obtain similar or better phenotype prediction accuracy with a small subset of features as with the complete set of sequenced features.

Keywords: Metagenomics · Phenotype prediction · Differential abundance · Feature selection

1 Scientific Background

Technological advances in high-throughput sequencing and annotation now allow the characterization of genomes, transcriptomes, and most recently metagenomes as part of everyday research in many fields. While single-gene, usually 16S ribosomal RNA (rRNA), sequencing can be used to infer bacterial community members, whole-genome shotgun sequencing can reveal details of the activity and function of the microbial community. Meta-transcriptomic sequencing can be applied to further investigate the actively transcribed sequences. One of the

A.P. Carrieri and N. Haiminen contributed equally to this work.

A. Bracciali et al. (Eds.): CIBB 2016, LNBI 10477, pp. 27–41, 2017.
DOI: 10.1007/978-3-319-67834-4_3

major research challenges of the current decade is gaining insight into the structure, organization, and function of microbial communities which will be enabled by both experimental and computational metagenomic analyses [1].

Since the sequencing methods yield relative rather than absolute gene or species counts, a fundamental methodological question of appropriate normalization and scaling of the counts arises. Approaches such as using the raw counts, log-transformed counts, length-normalized counts, and other normalization methods have been investigated [2–4]. We propose applying RoDEO projection as a pre-processing method for metagenomic counts.

RoDEO (Robust Differential Expression Operator, http://researcher.watson.ibm.com/group/5513) [5] was originally designed for detecting differentially expressed genes from single species RNA-sequencing data. The underlying non-parametric model and ranking-based ordering of genes can be applied in the context of various count data, including species counts from metagenomic samples. We apply RoDEO on metagenomic count data due to its robust design that does not rely on any assumptions regarding the underlying count distributions, and its applicability even in the absence of replicate samples, a common characteristic of metagenomic data.

In this paper we investigate the task of predicting the phenotype of the host organism (or environment) starting from OTU (Operational Taxonomical Unit, e.g., species or genus) counts. This question is relevant, for example, if we aim to predict the disease state from gut or fecal microbiome samples of humans and animals [7,8]. A recent related work on the topic includes a study of approaches to metagenomics-based prediction tasks and potential strength of microbiome-phenotype associations [9].

We investigate the effect of RoDEO projection on the prediction accuracy, and contrast it with existing normalization methods, namely Log-transformation, DESeq2 [10], and CSS (Cumulative Sum Scaling) [2]. We compare several kernel options for SVM (Support Vector Machine) prediction. SVMs are well established fundamental machine learning methods that have been applied in genomic, transcriptomic, and recently also in the microbial phenotype prediction context [11]. We find that the linear kernel SVM yields the best accuracy values across all the datasets and normalization methods. We also consider Random Forests (RF) [12] as they are state-of-the-art classification approaches and are appropriate for this type of data [13].

Furthermore, we investigate the problem of identifying a subset of OTUs that are important for differentiating the phenotypes. The process of selecting a subset of features consists of reducing the size of an high-dimensional dataset to retain only relevant, differentiating features [14]. We apply feature selection by identifying the most differentially abundant OTUs between the phenotype sample groups, and use them for predicting the host phenotype. The top differentiating OTUs are selected using two differential gene expression methods RoDEO DE and DESeq2.

We show that the prediction accuracy obtained selecting the top differential 20 OTUs is comparable, if not higher, to using the entire set of OTUs across all the datasets we consider in our experiments. Although RoDEO DE and DESeq2 yield different sets of top differentiating OTUs, the prediction accuracy values obtained using the different OTUs subsets are very close. While the prediction accuracy obtained using RF is often higher or comparable to the one obtained using SVM with linear kernel, RF is more resource consuming especially for a large number of features, i.e. when we use the entire set of OTUs for the prediction.

2 Materials and Methods

In this section we describe the various normalization, differential abundance, and phenotype prediction methods, as well as the datasets used in this study.

2.1 RoDEO Normalization

RoDEO sequence count data normalization, called *projection*, is not focused on the relative counts of reads for each OTU, but on the relative order of the counts within a sample. The count values of all OTUs in an experiment are utilized in a re-sampling approach, to determine robust relative ranks of the genes in several re-sampled instances of the sequencing experiments. A global parameter P determines the number of possible output values of the projection, ensuring that samples processed with the same P are comparable.

The *projection* process of RoDEO takes as input count data, such as the number of reads mapping to a OTU, and performs repeated re-sampling of the reads falling on the OTUs. In this way RoDEO projection process obtains a distribution which represents several randomized draws of sequencing reads from the input sample, according to the initial OTU abundances. In each re-sampling, the reads falling onto each OTU are counted, the OTUs are ranked by decreasing count, and the cumulative curve of the counts is optimally divided into segments $1, ..., P$. The number of segments P defines the resolution at which DE genes are discovered. We choose P for each dataset according to the number of (non-zero) entries per sample. In the RoDEO publication [5] we use 15–20 segments for human and plant data with tens of thousands of genes. Thus the dimensionality of the sequence count data is reduced from thousands of distinct values onto a small number of P possible values.

The projection and re-sampling makes RoDEO resilient in the presence of noisy and sparse count data with a large value range, such as observed in metagenomic sequencing data, and on a previous application on plant gene expression data [5].

2.2 DESeq2 Normalization

DESeq2 [10] is a well known method designed for differential analysis of count data using shrinkage estimation for sequence count dispersions. In a recent work which evaluates several methods for the identification of differentially abundant genes between metagenomes [4], DESeq2 was found to be among the best approaches for the task.

2.3 Other Normalization Methods

The baseline for comparing RoDEO to other methods of processing the counts is using the raw sequence counts per OTU. Log-transformation is a standard pre-processing step for sequence count data applied in many studies, including the respective studies for the datasets analysed in our paper [2,7,15]. Therefore we take the log of the count data (after adding 1 to all the counts we use the log function in R to compute the natural logarithm).

In addition, we evaluate prediction results on the CSS method as implemented in QIIME. CSS [2] was introduced in conjunction of the mouse microbiome dataset that is included in our study. According to the authors, CSS corrects the bias in the assessment of differential abundance. We include this method in our evaluation since it appears better than DESeq (previous version of DESeq2) for the class separation task studied on the mouse dataset.

2.4 RoDEO Differential Abundance

Differential Abundance of an OTU between two groups is computed as a DA score (analogously to differential expression, DE, in the gene expression context). This score takes into account the projected distributions for each sample in the two groups. In this work we use the *mean* distance between the projected distributions instead of mode used in the original paper. The final score for an OTU is the mean distance between the phenotype group projected distributions for this OTU multiplied by the max. norm distance (measuring overlap) between the distributions.

In order to evaluate datasets at different scales, with different numbers of non-zero OTU counts and total counts, we apply *scaling* [6]. The main idea is, we use a different value for the number of projected values P, depending on the count distributions in the samples. Details on this process on the studied datasets are provided in the Appendix.

2.5 DESeq2 Differential Abundance

DESeq2 provides both a normalization function, and a DE score computation function; we use the resulting DE values as the DA per OTU, obtained from the QIIME [16] microbiome analysis pipeline (version 1.9.1).

2.6 Phenotype Prediction

Support Vector Machines (SVMs) are among the most powerful and versatile binary classifiers used in a myriad of applications. We evaluate SVMs with linear, polynomial, radial and sigmoid kernels for phenotype prediction on three different metagenomic datasets described in Sect. 2.7.

We conduct 10-fold Cross Validation (CV), repeating the process 100 times, on the four different trained SVM kernels on RoDEO projected counts, log-transformed counts, as well as the CSS and DESeq2 processed counts. We report the accuracy of each prediction as the percentage of correct phenotype calls for the test set and we include the Matthews Correlation Coefficient (MCC). The latter coefficient is a measure of the quality of binary classification that can be used even when the two classes are of very different sizes. MCC can assume values between +1 and −1, where +1 indicates a perfect prediction, 0 no better than random and −1 represents total disagreement between predictions and observations.

After performing 10-fold CV process 100 times, we compute the average of the 100 accuracy and MCC values for each combination of kernel and dataset. The average accuracy and MCC values are summarised in Table 1, while the distribution of accuracy values and their average are visualized in Fig. 1.

Furthermore in Sect. 3.2, we apply, to the whole set of OTUs and to selected subsets of OTUs, SVM with linear kernel together with another prediction method, Random Forests, in order to compare their respective prediction accuracy, MCC and F1 score values.

The F1 measure is widely applied in information retrieval for measuring document classification. F1 score has an intuitive meaning: it tells how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). In statistical analysis of binary classification, the F1 score (which reaches its best value at 1 and worst at 0) is a measure of test accuracy and can be interpreted as a weighted average of the precision and recall.

The SVM and RF prediction is computed using the svm() and rf() R functions (e1071 package). All phenotype prediction results and figures have been produced using R (version 3.2.3).

2.7 Datasets

We investigate the accuracy of phenotype prediction starting from three different available metagenomic datasets: human, mouse, and corpse decomposition data. The human metagenome sequences originate from genome-wide shotgun sequencing, while the mouse and corpse data result from targeted rRNA sequencing. We obtained directly the read counts per OTU in each sample. For more details on the datasets please see the original publications.

Human dataset [7] consists of fecal metagenome of 70-year-old European women with either Normal Glucose Tolerance (NGT) or Type 2 Diabetes (T2D). Though T2D is caused by a complex combination of lifestyle and genetic factors, an altered gut microbiome has been linked to metabolic diseases including obesity, diabetes and cardiovascular disease. All microbiome samples were sequenced with Illumina HiSeq2000, and aligned to 2,382 non-redundant reference genomes (from the National Center for Biotechnology Information (NCBI) and Human Microbiome Project (HMP databases) in order to determine the composition of the gut microbiota. In our study we consider 43 NGT and 53 T2D samples described by a total of 134 OTUs at the family level. The phenotypes for the human dataset are healthy (NGT, 43 samples) and sick (T2D, 53 samples).

Mouse microbiome data [2] consist of mice fecal samples. Mice were fed with either Western (W) or Low-Fat, Plant Polysaccharide-rich (LF-PP) diet. Fecal samples for each mouse went through Polymerase Chain Reaction PCR amplification of the bacterial 16S rRNA gene V2 region. OTUs were classified by RDP11 and annotated. We analyze the dataset composed of 139 samples and 10,172 OTUs. The phenotypes for this dataset are W diet (54 samples) and LF-PP diet (85 samples).

Corpse microbiome data [15] consist of time-series samples from donated human bodies exposed to all natural elements. Two corpses were placed during the spring for 82 days and two corpses were placed during the winter for 143 days. Samples from multiple skin and soil locations were taken at different time points, daily or every other day the first month and less frequently thereafter. 16S rRNA gene (archaeal and bacterial community), 18S rRNA gene (microbial eukaryotic community), and ITS region (fungal community) were sequenced with high-throughput amplicon-based sequencing technology to characterize the full microbial diversity associated with decomposition. Sequence reads were classified into OTUs on the basis of sequence similarity using QIIME. We examine the read counts of 213 samples, having sum of counts above 10, taken from the left knee (skin and soil) at all the time points. There are a total of 17,803 OTUs observed in these samples. We choose this particular body site as it is sampled for both spring and winter conditions with sufficient detail, and there are many non-zero OTUs shared between the two conditions. The phenotypes for the corpse dataset are spring (79 samples) and winter (134 samples).

3 Results

In this section we summarize the phenotype prediction results on full datasets and on selected top differentially abundant features.

3.1 Phenotype Prediction on Full Datasets

Figure 1 summarizes visually the average prediction accuracy for each dataset and kernel, while Table 1 shows in more detail the differences in average prediction accuracy and MCC across the methods and highlights the best results per dataset. The results show that average accuracy and MCC consistently indicate the same combination of normalization and kernel as best for a particular dataset.

Human dataset has the lowest prediction accuracy and the lowest Matthews correlation coefficient. On this data RoDEO is best for nearly every kernel, and especially clearly improves the linear kernel prediction, yielding the best overall accuracy of 67.38% and the best MCC of 0.34.

The mouse data prediction is nearly perfect for most kernels and normalization methods. Only the Log data with sigmoid and radial kernels, as well as DESeq2 and CSS with polynomial kernel have lower accuracy.

On the corpse data, different kernels have quite different behavior. The worst seems to be sigmoid kernel and again the best is the linear kernel, where CSS slightly improves over RoDEO and yields 96.3% accuracy and 0.92 MCC, compared to 96.0% accuracy and 0.91 MCC of RoDEO.

Human prediction accuracy is not as high as for the other datasets studied here; in the original study they improve it by assembling novel entities from the unmapped reads and using them as additional features for prediction. This demonstrates there is still significant relevant content in the microbiomes that have not been encountered and annotated before. Still, in the mouse and corpse datasets using sequences mapped against existing databases yield highly accurate separation of phenotypes.

Most importantly, the best prediction accuracy is observed for RoDEO processed data in most cases and for the linear kernel. CSS is the second best method, followed by DESeq2 and Log. Also note that RoDEO clearly improves prediction accuracy on the clinically relevant human dataset, improving the chances of correctly diagnosing Type 2 Diabetes based on the gut microbiome.

Table 1. Accuracy as the average percentage of correct phenotype predictions in the cross validation results using linear, polynomial, radial, and sigmoid kernels. The values in the accuracy table correspond to the rightmost plots in Fig. 1. On the right, Matthews correlation coefficient (MCC) values are reported for each dataset and method. The best accuracy and MCC values are reported in black bold text.

		Accuracy (%)				MCC			
		Lin	Pol	Rad	Sig	Lin	Pol	Rad	Sig
Human	RoDEO	**67.38**	67.00	62.40	55.72	**0.34**	0.33	0.26	0.00
	Log	56.38	55.25	63.08	56.70	0.12	0.10	0.24	0.03
	DESeq2	56.00	57.60	63.00	55.71	0.12	0.15	0.24	0.0
	CSS	58.40	60.31	55.70	55.71	0.17	0.20	0.06	0.0
Mouse	RoDEO	**100.0**	99.97	**100.0**	98.55	**0.999**	0.998	**0.999**	0.968
	Log	99.99	99.90	76.64	61.86	0.998	0.997	0.514	0.087
	DESeq2	**100.0**	61.15	**100.0**	99.99	**0.999**	0.0	**0.999**	0.998
	CSS	**100.0**	94.11	**100.0**	99.98	**0.999**	0.883	**0.999**	0.998
Corpse	RoDEO	96.0	93.90	94.47	49.33	0.91	0.86	0.88	–0.01
	Log	82.7	75.75	62.9	56.87	0.63	0.51	0.0	0.01
	DESeq2	94.8	83.4	93.7	65.56	0.88	0.65	0.87	0.27
	CSS	**96.3**	81.27	93.6	93.7	**0.92**	0.60	0.86	0.86

3.2 Phenotype Prediction on Selected Features

In order to establish a baseline on the de-duplicated datasets we use for feature selection, as discussed in the Appendix, we first evaluate prediction accuracy

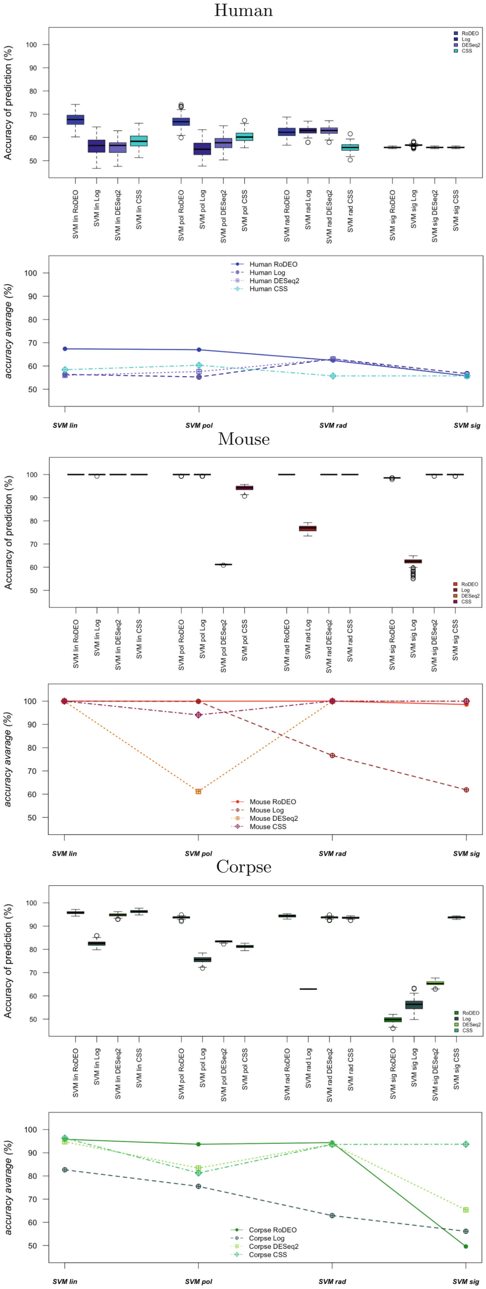

Fig. 1. Phenotype prediction accuracy in the 100 iterations of 10-fold Cross Validation for each SVM kernel (linear, polynomial, radial, sigmoid), RoDEO processed data and other normalization methods for human, mouse, and corpse data. For each dataset, the distribution plot of average accuracy across 100 iterations is followed by the corresponding overall average accuracy plot.

using all the OTUs. We compute 10-Fold Cross Validation using SVM with linear kernel (as we show in Sect. 3.1 that the best prediction accuracy overall is obtained with linear kernel) and Random Forest prediction methods. The average accuracy, MCC and F1 score values obtained for each dataset and each normalization and prediction method are shown in Table 2 in the "All OTU" columns. For the mouse dataset, similarly to the results shown in Table 1, accuracy is near perfect for all prediction methods, thus omitted from this evaluation.

Next we apply RoDEO and DeSeq2 differential expression methods to RoDEO projected data and DeSeq2 normalized data, respectively, and rank the OTUs according to their differential abundance (DA) scores for all three datasets. We select the top X where $X = 2, \ldots, 50$ most differential abundant OTUs and perform 10-fold CV on these subsets of different sizes using SVM linear and RF, to evaluate the prediction accuracy using the selected features only. The results are shown in Fig. 2. The horizontal lines denote the accuracy values reported in Table 2 for all OTUs. Using the most differentially abundant OTUs allows us to achieve similar or even better accuracy, MCC and F1 score compared to using the whole set of OTUs.

Based on the results in Fig. 2, we choose the value $X = 20$ as a representative small number of OTUs that yields phenotype prediction accuracy comparable

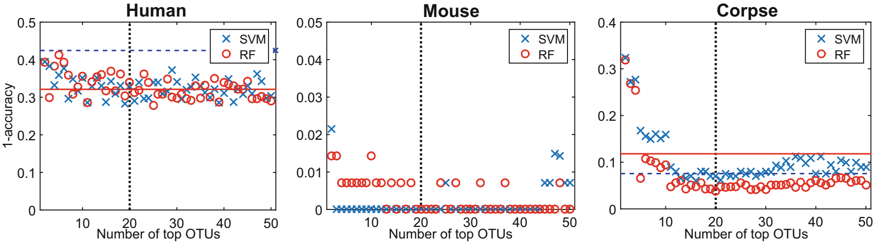

Fig. 2. RoDEO processed SVM and RF 10-fold CV results on varying numbers of top OTUs. Horizontal lines (SVM dashed, RF solid) denote the accuracy values when using all OTUs.

Table 2. Accuracy, MCC and F1 average values of 10 cross-fold validation results using linear kernel SVM and Random Forest prediction methods and considering either the top 20 DA OTUs or the complete set of OTUs. The best accuracy, MCC and F1 values for each dataset is shown in bold text.

		Accuracy (%)				MCC				F1			
		Subset 20		All OTU		Subset 20		All OTU		Subset 20		All OTU	
		RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM
Human	RoDEO	**70.11**	66.22	67.88	57.55	**0.42**	0.32	0.33	0.05	**0.71**	0.69	0.71	0.60
	DESeq2	65.66	58.88	61.00	59.66	0.32	0.19	0.22	0.15	0.68	0.58	0.65	0.61
Corpse	RoDEO	93.93	**94.39**	88.2	92.44	0.86	**0.88**	0.75	0.84	0.89	**0.92**	0.83	0.89
	DESeq2	93.42	89.67	86.47	93.85	0.86	0.79	0.71	0.86	0.90	0.85	0.79	0.89

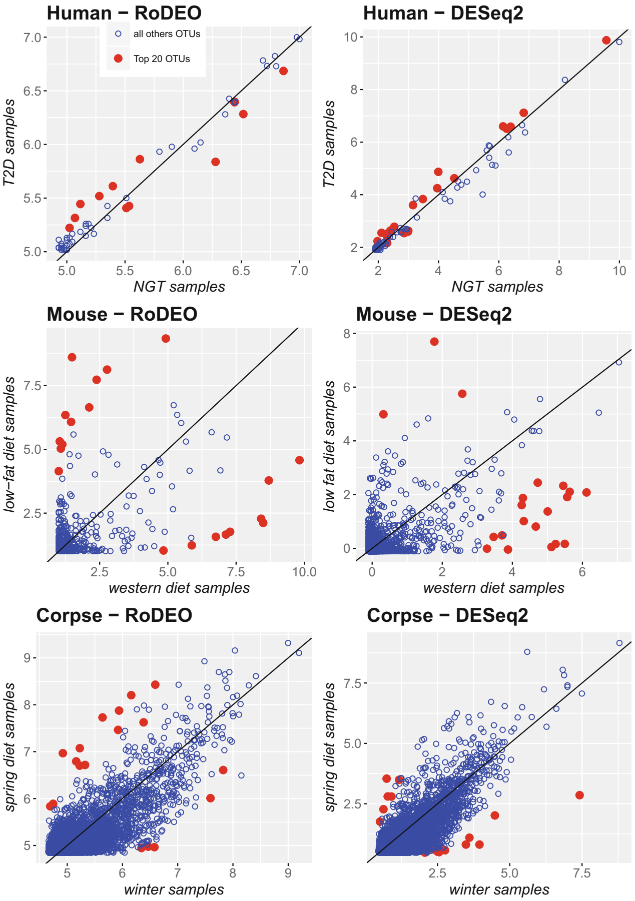

Fig. 3. Visualization of all OTUs (blue) and top 20 differentially abundant OTUs (red), for RoDEO (left) and DESeq2 (right) processed data. Each dot represents the average value of RoDEO projected or DESeq2 normalized samples having one phenotype (x) versus the other phenotype (y). The scale in each plot corresponds to either RoDEO projected values or DESeq2 normalized values. (Color figure online)

to all OTUs for all three datasets. In the following, we study in detail using the top 20 differentially abundant OTUs for phenotype prediction.

Table 2 shows that linear kernel SVM and RF methods using the whole set of OTUs or the top 20 OTUs give overall similar accuracy/MCC and F1 score results over all the three datasets. Furthermore, the results show that accuracy, MCC and F1 score are consistent as they indicate the same best combination of normalization, DA and kernel methods for a particular dataset. For the human dataset the best prediction result is given by RF method using RoDEO projected data and its subset of 20 top DA OTUs. For the corpse dataset the best prediction

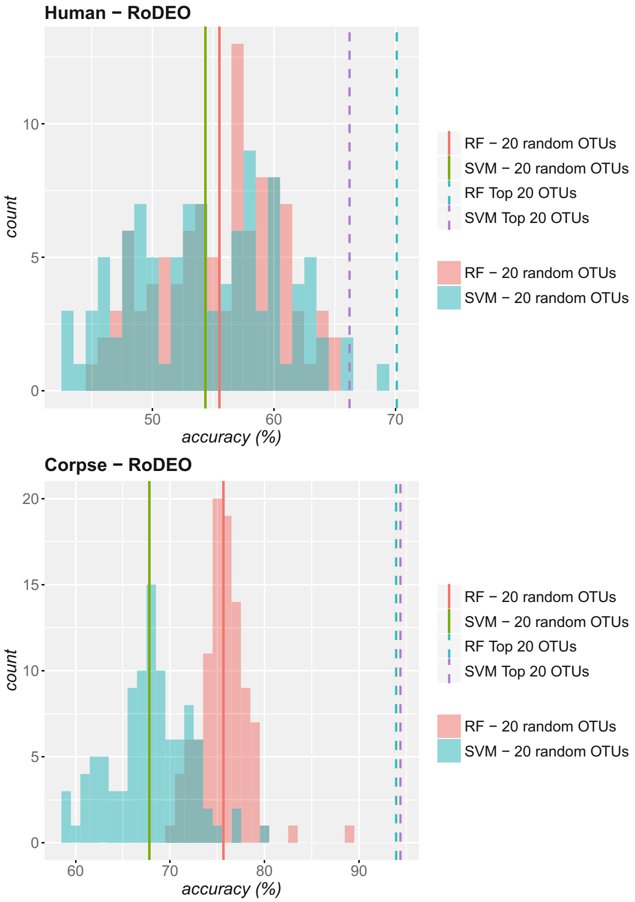

Fig. 4. The histogram shows the distribution of average phenotype prediction accuracy in 10-fold CV of 100 random subsets of 20 OTUs. Solid lines represent the average accuracy of the random OTUs subsets, while dashed lines show the average accuracy of 10-fold CV obtained using only the 20 top DA OTUs.

is obtained with linear kernel SVM on RoDEO projected data using the subset of 20 top DA OTUs.

The two methods, RoDEO DE and DeSeq2, yield different sets of top OTUs for all three datasets, but the prediction accuracy, MCC, and F1 scores on them are still quite close. Although the exact OTU names are different, representatives from the same family are selected by both methods, such as *Lachnospiraceae* for mouse.

Figure 3 shows details about the 20 OTUs in each dataset and for both RoDEO and DESeq2 methods. Overall the normalized datasets look quite similar between methods, but there are some differences, also regarding the values for the selected top OTUs. For example, DESeq2 appears to select many OTUs that have high counts in "western diet" compared to "low fat diet", while a more balanced selection is given by RoDEO.

Finally, we validate our method using SVM with linear kernel and RF for RoDEO processed data and considering 100 random subsets of 20 OTUs for each dataset. For each of these 100 random subsets we performed 100 times 10 CF validation and we show in Fig. 4 that using 20 random OTUs yield clearly worse prediction than the one obtained using the top differentiating 20 OTUs computed by RoDEO DE.

4 Conclusion

In this work we evaluate the applicability of the RoDEO projection method for metagenomic sequencing data, applying it on the task of phenotype prediction. We show that RoDEO processing increases the prediction accuracy over current methods when using SVM with a linear kernel, which we find to be the most accurate prediction method overall.

We include metagenomic data across human, mouse, and environmental (corpse decomposition) samples in our evaluation. The human data includes only a handful of OTUs with counts generated by whole-genome shotgun sequencing, while mouse and corpse data include thousands of OTUs sampled by targeted region sequencing. The results suggest that for various types and quantities of metagenomic data, using RoDEO projection of the sequencing counts onto lower dimensional values, together with linear kernel SVM yields the most accurate phenotype prediction results in most cases.

Perhaps surprisingly, in all three real datasets, prediction accuracy using the top few most differentially abundant OTUs is comparable to using all OTUs. This may be explained by random noise in the underlying metagenomic sequencing results, due to the sparse nature of the data and individual variation between the biological samples.

The actual top OTUs selected vary between the RoDEO and DESeq2 methods, but both provide accurate phenotype predictions using the respective OTUs. This indicates potential for accurate disease diagnostics and other phenotype prediction tasks by measuring a handful of most differential features only. RoDEO projection and feature selection, combined with either RF or SVM prediction yields consistently accurate phenotype prediction results.

Appendix: Experimental Details

RoDEO Projection Details on Full Datasets

For each of the 96 human samples with 134 OTUs, we run RoDEO for 100 independent re-sampling simulations, with $P = 7$ number of segments, 10^6 number of reads for the re-sampling and gap parameter equal to 1. For each of the samples we compute the average of projected values for each OTU (average of the 100 iterations), and combine all the obtained values in a single matrix.

Similarly, we apply RoDEO to the 139 mouse samples and 10,172 OTUs for 100 independent re-sampling simulations, with $P = 10$ number of segments, 10^7 number of reads for the re-sampling and gap equal to 1, and we compute the average of projected OTU values.

Finally, we run RoDEO for each of the 213 corpse samples with 17,803 OTUs for 100 independent re-sampling simulations, with $P = 10$ number of segments, 10^7 number of reads for the re-sampling and gap between the samples equal to 2. In the same way as described before, we compute the average of projected OTU values for each sample.

Feature Selection Details

We start the feature selection process deleting duplicated OTUs from each of the three initial raw count datasets described in Sect. 2.7. Removing identical OTUs allow us to deal with smaller datasets and apply Random Forests as an alternative prediction method to SVM. More precisely, for the corpse data we remove about 3000 OTUs passing from an original dataset of 213 samples and 17804 OTUs to a new dataset with 213 samples and 14789 OTUs. For the mouse data we pass from 139 samples described by 10172 OTUs to 139 samples described by only 4411 features. Finally, in the human data we find only 4 OTUs identical in the count and we obtain a new human dataset with 97 samples and 130 OTUs.

We proceed to run DESeq2 on this duplicate-removed data, including the DESeq2 normalization and subsequent DE computation, in order to obtain a ranked list of differentially abundant OTUs. For RoDEO, projection and scaling is required before the DE computation, in order to make the samples directly comparable across phenotypes. Below is a detailed description of the RoDEO scaling process described in Sect. 2.1.

For the greatest human sample, i.e. the one with smallest number of zeros, we run RoDEO for 100 independent re-sampling simulations, with $P_g = 7$ number of segments, 10^6 number of reads for the re-sampling and gap parameter 1. The number of segments we use to run RoDEO for all the other 96 human samples varies and depends on the result obtained from the scaling process for a given sample. All the other required parameters are instead equal to the ones used for the greatest sample. We then compute the average of projected values for each OTU (average of the 100 iterations), combine all the obtained values in a single matrix and we add to each row i, representing sample i, the difference between the number of segments P_g used to run RoDEO on the greatest sample g and the number of segment P_i used to run RoDEO on sample i.

Similarly, we apply RoDEO projection and the scaling algorithm to the mouse dataset running 100 independent re-sampling simulations, with $P = 10$ number of segments, 10^7 number of reads for the re-sampling and gap 1, for the greatest mouse sample.

Finally, we run RoDEO on the greatest corpse sample for 100 independent re-sampling simulations, with $P = 10$ number of segments, 10^7 number of reads for the re-sampling and gap between the samples equal to 2. In the same way as described before, we compute the averages of projected OTU values for each sample and we add the difference values from the scaling.

References

1. Anastas, P., et al.: 2020 visions. Nature **463**(7277), 26–32 (2010). https://www.nature.com/nature/journal/v463/n7277/full/463026a.html
2. Paulson, J.N., Stine, O.C., Bravo, H.C., Pop, M.: Robust methods for differential abundance analysis in marker gene surveys. Nat. Methods **10**, 1200–1202 (2013)
3. Parida, L., Haiminen, N., Haws, D., Suchodolski, J.: Host trait prediction of metagenomic data for topology-based visualization. In: Natarajan, R., Barua, G., Patra, M.R. (eds.) ICDCIT 2015. LNCS, vol. 8956, pp. 134–149. Springer, Cham (2015). doi:10.1007/978-3-319-14977-6_8
4. Jonsson, V., Österlund, T., Nerman, O., Kristiansson, E.: Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics **17**(78), 1–14 (2016)
5. Haiminen, N., Klaas, M., Zhou, Z., Utro, F., Cormican, P., Didion, T., Jensen, C., Mason, C.E., Barth, S., Parida, L.: Comparative exomics of Phalaris cultivars under salt stress. BMC Genomics **15**(6), 1–12 (2014)
6. Klaas, M., Haiminen, N., Grant, J., Cormican, P., Finnan, J., Krishna, S., Utro, F., Vellani, T., Parida, L., Barth, S.: Characterizing differentially expressed genes under flooding and drought stress in the biomass grasses Phalaris arundinacea and Dactylis glomerata. Under submission (2017)
7. Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., Bäckhed, F.: Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature **498**, 99–103 (2013)
8. Ross, E.M., Moate, P.J., Marett, L.C., Cocks, B.G., Hayes, B.: Metagenomic predictions: from microbiome to complex health and environmental phenotypes in humans and cattle. PLoS ONE **8**, e73056 (2013)
9. Pasolli, E., Tin, D., Truong, F.K., Waldron, L., Segata, N.: Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput. Biol. **12**(7), e1004977 (2016)
10. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**(12), 550 (2014)
11. Weimann, A., Mooren, K., Frank, J., Pope, P.B., Bremges, A., McHardy, A.C., Segata, N.: From genomes to phenotypes: traitar, the microbial trait analyzer. mSystems **1**(6), 1–19 (2016)
12. Ho, T.K.: Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282 (1995)
13. Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., Pei, Z., Blaser, M.J., Aliferis, C.F., Alekseyenko, A.V.: A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome **1**, 11 (2013)
14. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. JMLR **3**(11), 57–82 (2013)

15. Metcalf, J.L., Xu, Z.Z., Weiss, S., Lax, S., Van Treuren, W., Hyde, E.R., Song, S.J., Amir, A., Larsen, P., Sangwan, N., Haarmann, D., Humphrey, G.C., Ackermann, G., Thompson, L.R., Lauber, C., Bibat, A., Nicholas, C., Gebert, M.J., Petrosino, J.F., Reed, S.C., Gilbert, J.A., Lynne, A.M., Bucheli, S.R., Carter, D.O., Knight, R.: Microbial community assembly and metabolic function during mammalian corpse decomposition. Science **351**(6269), 158–162 (2016)
16. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Gonzalez Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R.: QIIME allows analysis of high-throughput community sequencing data. Nat. Methods **7**(5), 335–336 (2010)