Using phenotype prediction to explore the functional capabilities of metagenomically assembled microbial genomes

Fred Farrell[1*], Orkun S. Soyer[1], Christopher Quince[2],

1 School of Life Sciences, University of Warwick, Coventry, United Kingdom
2 Medical School, University of Warwick, Coventry, United Kingdom

* f.farrell@warwick.ac.uk

Abstract

Using machine learning, we associate bacterial and archaeal genomes with functional and metabolic traits. To train these models, we use a database of 84 phenotypic traits associated with 9407 prokaryotic genomes downloaded from the NCBI genome database, and are able to classify organisms as possessing these functions with greater than 90% AUROC score for 65 of these functions. We then use these models to make predictions about the phenotypes of novel microbial genomes assembled from metagenomic studies from a variety of environments.

Introduction

The increasing ease of genetic sequencing has led to an explosion in the amount of such data generated. In the context of microbial ecology, large-scale metagenomic studies such as the Human Microbiome Project [1], the Earth Microbiome Project [2] and the Tara Oceans Project [3] have systematically sequenced the microbial communities in a huge variety of environments at great depth. Amplicon sequencing, such as of the 16S rRNA gene, allows detailed study of the taxonomic makeup of these communities, while shotgun metagenomic sequencing allows characterisation of all genes present in an environment. Increasing depth of coverage and improvements in genome binning algorithms for clustering contigs into genomes, in particular the use of differential coverage across different samples [4, 5], are allowing more and more full and partial genomes to be assembled from shotgun metagenomic studies. Many of these organisms are novel and uncultured, having never been studied in a lab. A recent metagenomic study on aquifer systems [6], for example, reconstructed 2540 separate high-quality, near-complete genomes, and claimed to have discovered an astonishing 47 new phylum-level lineages among them.

Interpretation of these results is time-consuming and difficult, and cannot keep up with the rate of data being generated. In particular, it is difficult to characterise the traits and ecological roles of novel assembled genomes, which may not be very closely related to any known organisms. As such, automated ways to associate genetic data with phenotypic features are required. There are a variety of tools and databases used for this purpose. Genomic databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [7], the Pfam database of protein families [8] and the NCBI COG database of orthologous genes [9] can be used to look up genes and associate them with metabolic pathways and other functions. There are automated tools for interfacing with these databases, for example PICRUSt [10], which performs ancestral state

reconstruction to link 16S sequences with full genomes using the KEGG orthology. However, most analysis of microbial communities from metagenomic studies still relies on manually studying particular genes of interest.

In the interest of developing a way to automatically gain an overview of phenotypic and functional characteristics associated with the results of metagenomic studies, we here present a method of using existing knowledge of phenotypes associated with microbial taxa to make predictions about new organisms. To do this, we train machine learning models to predict phenotype based on the presence or absence of gene families within an organism.

To train these models, we utilised a recently published database, the FAPROTAX database [11, 12], of phenotypes known to be associated with certain microbial taxa, based on an extensive survey of the scientific literature. The bulk of the classifications in this database come from *Bergey's Manual of Systematic Bacteriology* [13]. FAPROTAX currently contains 84 phenotypic traits associated with 4600 microbial taxonomic groups. We train models of the link between genotype and phenotype by combining this database with all of the full genomes available in the NCBI genome database. We show that such models have a good accuracy at determining the phenotype of an unknown species over many functions, with 65 functions out of 84 achieving an AUROC score above 90%. In particular, in many cases the classification accuracy is superior to using manually-curated 'KEGG modules', which are small groups of genes in the KEGG database known to be associated with a particular function.

We then go on to use these models to predict the phenotypes of so-called MAGs (metagenome assembled genomes), genomes assembled from metagenomic studies and otherwise unstudied, from three diverse environments: anaerobic digesters, the human gut and the ocean.

Materials and methods

Databases and preparation of training data

To train our models, we utilized the combination of the recently-published FAPROTAX database of microbial phenotypes and the NCBI genome database. We downloaded all prokaryotic genomes classified as 'full' from the NCBI Genome database. We then used the taxonomic information available from NCBI to assign them phenotypes using the script 'collapse_table.py' which comes as part of the FAPROTAX database [11]. We then called genes in these genomes using Prodigal [14]. Finally, we used Diamond BLASTP [15] to to align these genes against KEGG orthologs and find genes mapping to the orthologs, finally giving a matrix of organisms and their copy numbers of each KEGG ortholog. We have found that using gene copy number rather than simple presence/absence significantly increases classifier performance. The scripts we used to download and process the genomes are available at ...

Logistic regression

To model the link between genotype, in the form of KEGG ortholog copy numbers, we used machine learning, and in particular logistic regression, a commonly used linear model used for classification problems [16, 17]. The genome dataset was split into a training set and a testing set, for testing the performance of the algorithm on unseen data. Additionally, we scaled all input features to have mean zero and variance 1 before performing the regression. Since the number of KEGG orthologs (features) was somewhat larger than the number of training examples, overfitting, whereby the model classifies on features of the training set which are very specific to it, was a serious

problem. To alleviate this, we used logistic regression with an $\ell 1$ penalty term, also known as LASSO logistic regression [18], whereby large parameters are penalized in such a way that only a few of the features have a nonzero weight. In detail, the method involves adding a penalty term equal to the $\ell 1$-norm of all of the coefficients of the regressor, thereby penalising nonzero terms, so that the optimization problem becomes:

$$\min_w \left[||w||_1 + C \sum_{i=1}^{n} \log\left(\exp(-y_i(X_i^T w)) + 1\right) \right] \quad (1)$$

where w is the vector of regression weights, X_i are the feature vectors of each example, y_i the classification targets, and C is a parameter defining the (inverse) strength of the regularization. This method of regularization is often useful in cases where the number of features is large (similar to or larger than the number of training examples), as most of the features are not used in the classification task. For example, a recent study used $\ell 1$-regularized regression to predict complex human traits such as height and heel-bone density from a large array of SNPs (around 100000), significantly improving on previous estimates of heritability based on individual SNPs [19]. In our case, we found that this method significantly outperformed other commonly-used and somewhat more complex classification algorithms, such as random forests and support vector machines.

Metrics and classifier performance

Since many of the classes which we are attempting to predict are highly unbalanced (e.g. of the 9407 unique species with full genomes in the NCBI database only 83 are hydrogentotrophic methanogens), simple classification accuracy is not a very useful measure of classifier performance. Predicting all labels as negative in the above example would give an accuracy of 99.1% despite not being a useful classifier. We therefore need a metric which can take into account class imbalance. We use the area under the ROC (Receiver Operating Characteristic) curve, which is a graph of true postive rate against false positive rate as one varies the cutoff in probability for making a positive prediction [20, 21]. An AUROC score much greater than 0.5 (the score for random predictions) indicates a good classifier. In particular, a score of 1 indicates that all positive cases have been assigned a higher probability than all negative cases.

Prediction of MAG phenotypes

Once classifiers have been trained on the NCBI data, it is possible to use them to make predictions about unseen genomes, such as MAGs generated from shotgun sequencing studies. The MAGs must first be processed to give a matrix of the KEGG ortholog copy numbers associated with them, using the same pipeline applied above the NCBI genomes. These matrices are then used as input into the classifiers to produce a matrix of MAGs and their predicted functions, which can be either presence/absence predictions or porbabilities.

Results

Classification accuracy

Figure 1 shows the performance of the algorithm in the classification task on the test set in terms of AUROC (area under the Receiver Operating Characteristic, see Materials and Methods) score. The results are shown for two classification algorithms, $\ell 1$-regularized logistic regression and the random forest. The regularized logistic regression performs significantly better for many, though not all, functions. The average

score over all functions for LR is 90.1% (versus 84.5% for the random forest), and 65 functions have a score greater than 90%, with 45 higher than 95%.

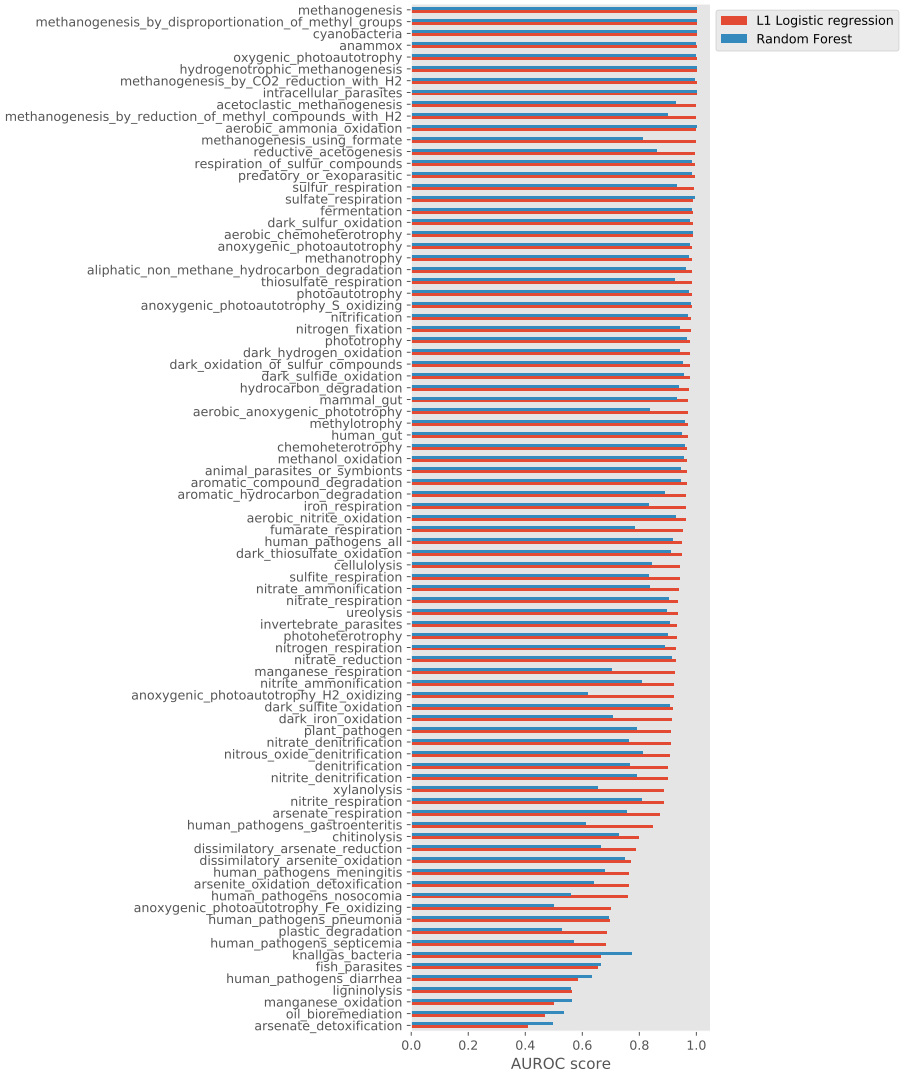

Fig 1. Overall performance of classification algorithms. The AUROC score on each classification task (each function) is shown for two classification algorithms: $\ell 1$-regularized logistic regression and the random forest. Functions are ordered by the LR score.

Gene orthologs used by classifiers

Table 1 shows the KEGG orthologs with non-zero coefficients used by the LR classifiers and their weights for some example functions. Due to the $\ell 1$-regularization, the number of non-zero coefficients is rather low. Three representative functions, all having classifiers with AUROC scores greater than 95%, are shown. Many of the KEGG orthologs picked out by the classifiers are genes known to be involved in these functions, as we might hope. In particular, consider the prediction of methanogens, a relatively easy task since it is known that methanogens must possess the *mcrA* gene, this being a necessary and sufficient condition for methanogenesis [REF]. Indeed, subunits of this gene have the highest weight, and a total of only 9 genes are used by the classifier.

Looking at the some complex traits, for sulfate respiration (i.e. disimmilatory sulfate reduction to H_2S), the model assigns a lot of weight to subunits of a quinone-modifying oxidoreductase, which is indeed associated with sulfur metabolism [REF]. Interestingly, however, none of the genes picked out by the classifier are directly part of the metabolic pathway for this process as described in the KEGG module for dissimilatory sulfate reduction, see Figure 2. The situation is similar with hydrogenotrophic methanogenesis, with classification mostly determined by components of energy-converting hydrogenases which are not directly part of the autotrophic methanogenesis pathway, along with *mcr* genes indicating that the microbe is a methanogen.

Figure 3 shows a scatter plot of AUROC score (i.e. classifier performance) against the number of orthologs used to make the prediction. It can be seen that there is a correlation between these two variables, with some highly accurate classifiers built out of a large number of genes. However, there is also a noticable cluster of functions with high accuracy achieved with only a few genes (less than 100). These functions may be particularly interesting, as it is more likely that these small groups of orthologs are causally associated with the function, rather than just being genes which typically occur in parts of the phylogenetic tree which have the function and may or may not have any direct relation to it. This issue is explored further in the section on performance across taxa, below. Also, note that most of the functions that perform poorly, which typically use very few genes to classify, have very low support in the training data in terms of number of positive examples.

Comparison to KEGG modules

It is instructive to compare the performance of our classifiers to the use of KEGG modules, where an equivalent module exists for that function, i.e. compare the performance to a 'classifier' where an organism is judged capable of a function if it has a complete KEGG module for that function. Table 2 shows the results of this comparison for three FAPROTAX functions with corresponding KEGG modules. Note that the KEGG module method does not require training, so the metrics are over the entire NCBI dataset, whereas for the classifier they are only for the held-out test set. Also, the former method gives only presence/absence of a function rather than a probability, so the AUROC score cannot be calculated, so we use alternative metrics based on classification: the F_1 score and the confusion matrix.

It can be seen that our logisitc regression classifier does significantly better than KEGG modules in assigning these functions as they appear in the FAPROTAX database. This suggests that having the enzymes or proteins described in the KEGG module for a function is not in fact a necessary or sufficient condition for actually performing that function, and that other genes are more predictive. However, it is possiblt that the discrepancy is due instead to inaccuracy in the FAPROTAX database, e.g. species which do perform the functions being missed from the database and therefore getting flagged as false positives with the KEGG method. More work would

Sulfate respiration		
KO	**Weight**	**Description**
K16885	0.664	quinone-modifying oxidoreductase; subunit QmoA
K16887	0.279	quinone-modifying oxidoreductase; subunit QmoC
K05299	0.151	formate dehydrogenase alpha subunit [EC:1.2.1.43]
K06951	0.129	uncharacterized protein
K11261	0.106	formylmethanofuran dehydrogenase subunit E [EC:1.2.99.5]
K10302	0.082	F-box protein 22
K14127	0.070	F420-non-reducing hydrogenase iron-sulfur subunit [EC:1.12.99.-]
K09777	0.052	uncharacterized protein
K11358	0.049	aspartate aminotransferase [EC:2.6.1.1]
K03738	0.047	aldehyde:ferredoxin oxidoreductase [EC:1.2.7.5]
K07550	0.045	benzoylsuccinyl-CoA thiolase BbsB subunit [EC:2.3.1.-]
K09740	0.040	uncharacterized protein
K12598	0.015	ATP-dependent RNA helicase DOB1 [EC:3.6.4.13]
K06873	0.009	uncharacterized protein
K03388	0.008	heterodisulfide reductase subunit A [EC:1.8.98.1]
K09116	0.003	uncharacterized protein
K12309	0.001	beta-galactosidase [EC:3.2.1.23]
K14086	0.001	ech hydrogenase subunit A

Methanogenesis		
KO	**Weight**	**Description**
K03421	0.567	methyl-coenzyme M reductase subunit C
K00400	0.276	methyl coenzyme M reductase system; component A2
K00579	0.160	tetrahydromethanopterin S-methyltransferase subunit C [EC:2.1.1.86]
K00399	0.081	methyl-coenzyme M reductase alpha subunit [EC:2.8.4.1]
K07463	0.023	archaea-specific RecJ-like exonuclease
K17618	0.023	ubiquitin-like domain-containing CTD phosphatase 1 [EC:3.1.3.16]
K00401	0.022	methyl-coenzyme M reductase beta subunit [EC:2.8.4.1]
K09728	0.020	uncharacterized protein
K09613	0.002	COP9 signalosome complex subunit 5 [EC:3.4.-.-]

Hydrogenotrophic methanogenesis		
KO	**Weight**	**Description**
K03421	0.231	methyl-coenzyme M reductase subunit C
K14109	0.201	energy-converting hydrogenase A subunit R
K00401	0.169	methyl-coenzyme M reductase beta subunit [EC:2.8.4.1]
K14098	0.136	energy-converting hydrogenase A subunit G
K14097	0.104	energy-converting hydrogenase A subunit F
K14093	0.058	energy-converting hydrogenase A subunit B
K06862	0.057	energy-converting hydrogenase B subunit Q
K14094	0.049	energy-converting hydrogenase A subunit C
K17278	0.043	membrane-associated progesterone receptor component
K08074	0.042	ADP-dependent glucokinase [EC:2.7.1.147]
K00442	0.032	coenzyme F420 hydrogenase subunit delta
K09613	0.031	COP9 signalosome complex subunit 5 [EC:3.4.-.-]
K09493	0.017	T-complex protein 1 subunit alpha
K14099	0.009	energy-converting hydrogenase A subunit H
K02938	0.005	large subunit ribosomal protein L8e
K00399	0.001	methyl-coenzyme M reductase alpha subunit [EC:2.8.4.1]

Table 1. Details of classifiers for specific functions. Tables showing the all nonzero weights in the logistic regression models trained on three functions from the FAPROTAX database. Note that there are 9647 KEGG orthologs used in our models, so the vast majority of weights are set to zero in these models.

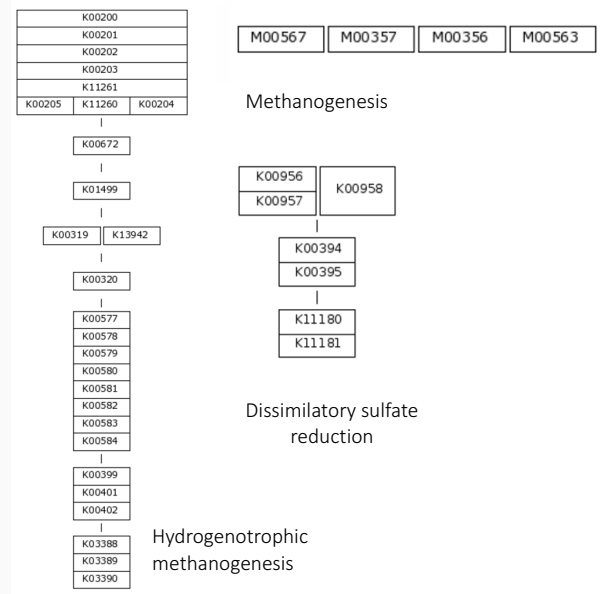

Fig 2. KEGG modules for some functions. Representations of the KEGG modules corresponding to the FAPROTAX functions shown in Table 1. Modules are organized into 'blocks' of orthologs, typically indicating a protein complex. Orthologs positioned next to each other are 'options', i.e. that section of the module is present if any of the adjacent blocks are present.

Fig 3. Classifier performance and complexity. Scatterplot showing the AUROC score of the different classifiers plotted against the number of gene orthologs the classifier uses to make its predictions. Point size is proportional to the number of positive examples in the training set.

	KEGG modules		Classifier	
module	**F1**	**confusion matrix**	**F1**	**confusion matrix**
sulfate respiration	0.84	$\begin{pmatrix} 9197 & 53 \\ 6 & 151 \end{pmatrix}$	0.99	$\begin{pmatrix} 2313 & 0 \\ 1 & 38 \end{pmatrix}$
nitrate respiration	0.14	$\begin{pmatrix} 6821 & 2162 \\ 224 & 200 \end{pmatrix}$	0.622	$\begin{pmatrix} 2237 & 9 \\ 54 & 52 \end{pmatrix}$
hydrogenotrophic methanogensis	0.756	$\begin{pmatrix} 9281 & 43 \\ 6 & 77 \end{pmatrix}$	0.923	$\begin{pmatrix} 2331 & 0 \\ 3 & 18 \end{pmatrix}$

Table 2. Comparison of classifiers to KEGG modules. Table showing the performance of using KEGG module presence/absence against LR classifiers for some functions where equivalent KEGG modules exist. Since the KEGG module approach doesn't give a probability, the AUROC score can't be used, so the F1 score and confusion matrices are comapred.

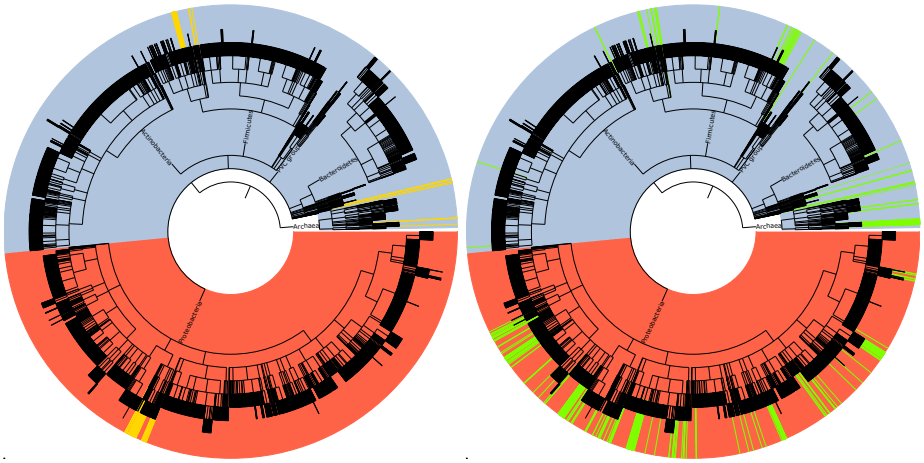

Fig 4. Taxonomic distribution of metabolic traits. Taxonomic trees of all prokaryotic NCBI species with full genomes. For training the cross-taxa verison of the classifier, only the Proteobacteria (red section of the tree) were used, and the models were tested on the rest of the tree. Species capable of a) sulfate respiration and b) nitrate respiration are highlighted on the trees.

be needed to exclude this possibility.

Performance across taxa

As alluded to above, it is not clear how much the genes being used by the classifiers are actually related to the functions being predicted; they must just be genes that happen to be found in a closely-related set of organisms that happen to all perform the function. The way in which functions are spread over the phylogenetic tree of microorganisms varies between functions [22], see Figure 4. As might be expected, closely-related organisms often perform similar functions, with clusters on the tree often sharing the function.

To investigate this phenomenon and attempt to find orthologs with real causal associations with functions, we tried training a model on one part of the phylogenetic tree and test its performance on another. If a classifier can predict phenotype based on genes in a distantly-related, unseen set of organisms, it is likely the genes it is using have a real association with the function. In particular, we tried training our logistic regression models on the Proteobacteria, a large phylum of bacteria, and testing on the rest of the phylogenetic tree of life.

Prediction of MAG phenotypes

Discussion

Conclusion

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

References

1. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–214. doi:10.1038/nature11234.
2. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biology. 2014;12(1):69. doi:10.1186/s12915-014-0069-1.
3. Zhang H, Ning K. The Tara Oceans Project: New Opportunities and Greater Challenges Ahead. Genomics, proteomics & bioinformatics. 2015;13(5):275–7. doi:10.1016/j.gpb.2015.08.003.
4. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nature Methods. 2014;11(11):1144–1146. doi:10.1038/nmeth.3103.
5. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319. doi:10.7717/peerj.1319.
6. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Communications. 2016;7:13219. doi:10.1038/ncomms13219.
7. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017;45(D1):D353–D361. doi:10.1093/nar/gkw1092.
8. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research. 2016;44(D1):D279–D285. doi:10.1093/nar/gkv1344.
9. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science (New York, NY). 1997;278(5338):631–7.
10. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology. 2013;31(9):814–821. doi:10.1038/nbt.2676.
11. Louca S, Jacques SMS, Pires APF, Leal JS, Srivastava DS, Parfrey LW, et al. High taxonomic variability despite stable functional structure across microbial communities. Nature ecology & evolution. 2016;1(1):15. doi:10.1038/s41559-016-0015.

12. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science (New York, NY). 2016;353(6305):1272–7. doi:10.1126/science.aaf4507.

13. Whitman WB, Bergey's Manual Trust, Wiley Online Library (Online service). Bergey's manual of systematics of archaea and bacteria;.

14. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119. doi:10.1186/1471-2105-11-119.

15. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2014;12(1):59–60. doi:10.1038/nmeth.3176.

16. Hastie T, Tibshirani R, Friedman J. Linear Methods for Classification. In: The Elements of Statistical Learning; 2009. p. 101–137.

17. Freedman D. Statistical models : theory and practice. Cambridge University Press; 2009. Available from: http://www.cambridge.org/gb/academic/subjects/statistics-probability/statistical-theory-and-methods/statistical-models-theory-and-practice-2nd-edition?format=HB{&}isbn=9780521112437{#}IqXivBILqUALXPKZ.97.

18. Lee SI, Lee H, Abbeel P, Ng AN. Efficient L1 Regularized Logistic Regression. In: The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference; 2006.

19. Lello L, Avery SG, Tellier L, Vazquez A, de los Campos G, Hsu SDH. Accurate Genomic Prediction Of Human Height. doiorg. 2017; p. 190124. doi:10.1101/190124.

20. Fawcett T, Tom. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861–874. doi:10.1016/j.patrec.2005.10.010.

21. Flach P, Hernández-Orallo J, Ferri C. A coherent interpretation of auc as a measure of aggregated classification performance. Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA. 2011; p. 657–664. doi:10.1145/347090.347126.

22. Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: A phylogenetic perspective. Science. 2015;350(6261).