The ISME Journal (2010), 1–3

www.nature.com/ismej

COMMENTARY

Supervised classification of microbiota mitigates mislabeling errors

Dan Knights, Justin Kuczynski, Omry Koren, Ruth E Ley, Dawn Field, Rob Knight, Todd Z DeSantis and Scott T Kelley

The ISME Journal advance online publication, 7 October 2010; doi:10.1038/ismej.2010.148

The exponential growth of DNA sequencing technologies and concomitant advances in bioinformatics methods are revolutionizing our understanding of diverse microbial communities (Riesenfeld *et al.*, 2004; Tyson *et al.*, 2004; Hugenholtz and Tyson, 2008; Tringe and Hugenholtz, 2008; Caporaso *et al.*, 2010). Large-scale microbial metagenomics studies have particularly exciting applications in the arena of human health, laying the foundation for the Human Microbiome Project (HMP). In the context of the HMP and related efforts, care has been taken to understand the impact of amplification biases or sequencing errors. However, far less attention has been paid to the impact of errors in metadata on biological interpretations and the mitigation of such errors. During processing and pooling of hundreds of samples, some mislabeling is likely. Figure 1 illustrates a real world example: several 16S rRNA amplifications of bacterial community DNA samples collected along a time series were accidentally mislabeled (late switched to early) (Koenig *et al.*, 2010). Automated detection of such errors will be important as datasets become increasingly large and complex.

Mislabeled metadata are especially problematic in large-scale collaborations where data analysis is far removed from data generation: researchers cannot reconfirm sample labels or resequence questionable samples, and must rely on the accuracy of available

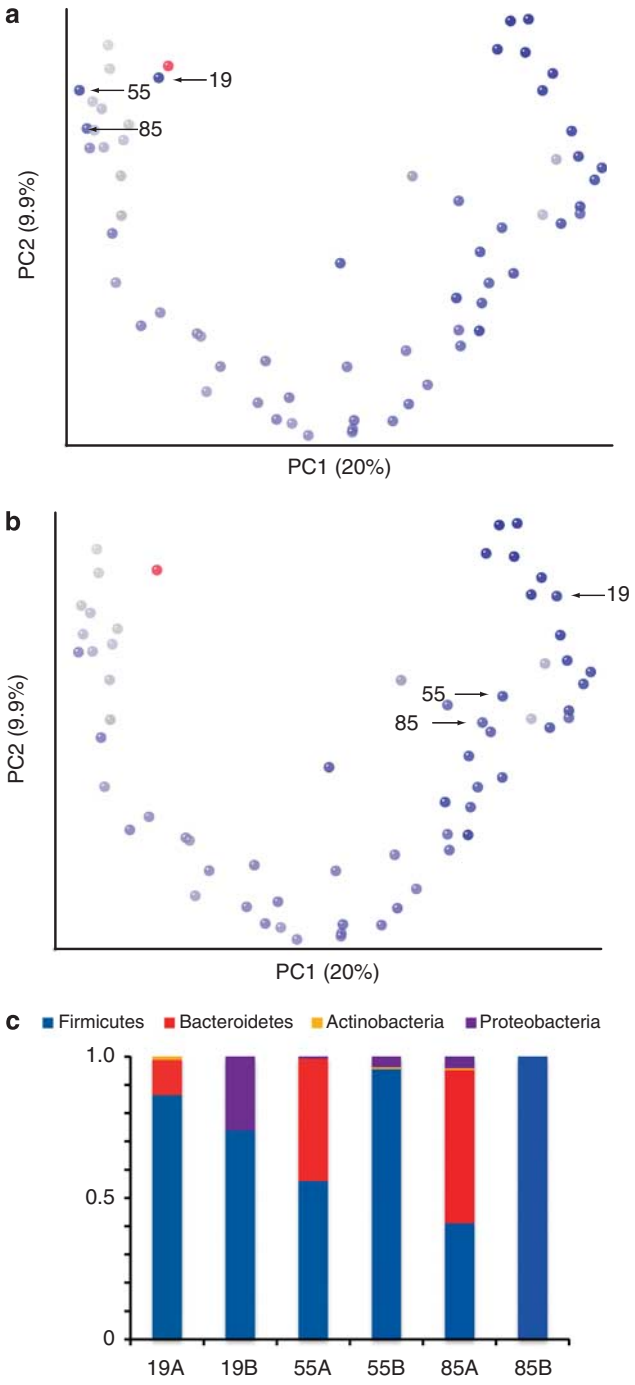

Figure 1 Resequenced 454 16S rRNA genes from infant time series experiment. These data are 60 fecal samples obtained over 2.5 years from a single individual. (**a**) Principal coordinates analysis of unweighted UniFrac distances derived from sequences from the initial sequencing run. (**b**) Corrected data. (**c**) Taxonomic discrepancies between the initial run (**a**) and the corrected run (**b**). Sample points are colored according to collection time where dark blue points represent time points that were collected early during the experiment, whereas the light gray time points represent later samples. Note that time points from days 19, 55 and 85 are misplaced in panel **a** (too dark for their position), and after resequencing, they cluster with other dark blue samples (early time points).

metadata. However, we have found that supervised classifiers are able to detect and even correct erroneous metadata in some datasets. We have demonstrated the efficacy of this approach by applying two common classifiers to Monte-Carlo simulations of mislabeled metadata from published 16S rRNA microbial community datasets examining (1) variation of bacterial communities among human body habitats (Costello *et al.*, 2009), and (2) the relationship of bacteria on computer keyboards to the users' hands (Fierer *et al.*, 2010).

When the 'alleged' data labels (that is, the metadata supplied by the experimenter) contain errors, we refer to this as 'metadata error'. Normally when we build a supervised classifier we can only compare its predictions to the alleged labels, and so the classifier's 'reported error' may be inaccurate. Here we work through an example in which all metadata error is simulated, and so we can compare a classifier's predicted labels to the true labels to measure the classifier's 'true error'. When all or most of the samples are mislabeled, we expect the classifier to be useless, but what if only a few of the labels are wrong?

After intentionally mislabeling samples at various rates of error, we tried to recover the correct groupings using the taxon relative abundance vectors as input features to two different classification models that have been successful in other high-dimensional classification problems (random forests (Breiman, 2001) and nearest shrunken centroids (Tibshirani *et al.*, 2002)) on two easy classification tasks (classifying general body habitats like skin vs gut, and classifying hand/keyboard samples by individual) and one hard task (classifying specific sites within the skin habitat like palm vs forearm).

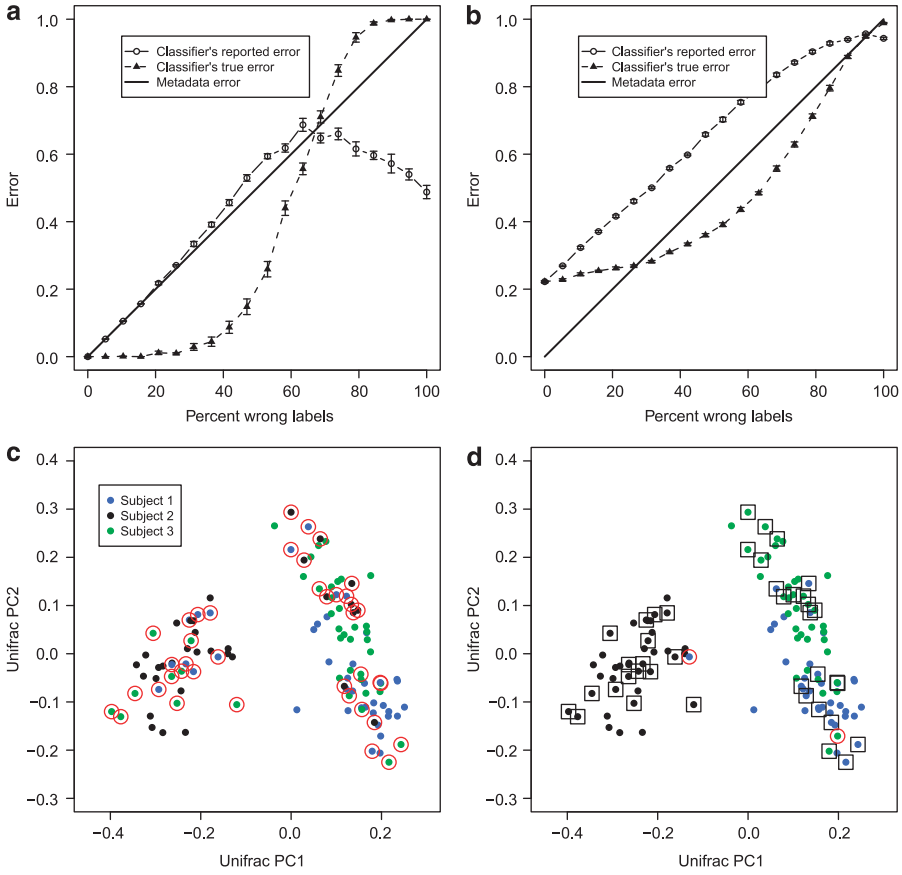

Figure 2 (**a**, **b**) Metadata error correction using random forests for the forensic identification task (**a**) and the general body habitat classification task (**b**). The horizontal axes show the proportion of labels that has been intentionally perturbed, and the vertical axes show the proportion of error in the prediction of the random forest classifier when trained on the full dataset with the perturbed labels. Each point represents the average error for 10 random perturbations of the metadata, with standard error bars. The solid black line simply shows the amount of error in the metadata, and is a useful reference for the other curves. The 'Classifier's reported error' reflects how well the model 'thinks' it is doing based on the partially incorrect metadata, whereas the 'Classifier's true error' reflects a 'god's-eye view' of how well the model is actually doing based on the true metadata. If the model does a good job of learning the differences between categories, it will often discover the true category for a mislabeled sample, although it will still report such a classification as an error. Hence the true error is generally lower than the reported error. (**c**, **d**) Principal coordinates analysis plots of the UniFrac distances between samples in the Fierer *et al.* (2010) dataset; the first two axes (shown) explain 18.0 and 6.3% of the total variation. Panel **c** Shows the data with 40 randomly chosen intentionally confused labels circled in red, and **d** shows the labels predicted by the random forest classifier (trained with 2000 trees and otherwise default settings using the confused labels). This classifier recovered all of the true class labels for those samples, while introducing only two new incorrect labels. Confused labels that were corrected by the model are indicated with a black square; remaining errors are indicated with a red circle.

We first ran the models using the original (correct) metadata for each classification problem, then using perturbed metadata with an increasing proportion of incorrect labels. A detailed description of our methods is provided as Supplementary Information.

We found that both random forests and nearest shrunken centroids are able to recover many of the true data labels when the differences between classes are large. Figure 2a shows that for keyboard user classification, the random forest classifier maintains near-zero true error even with 40% incorrect class labels. Results for the relatively easy body habitat classification were similar. These results imply that in easy classification tasks some classifiers can recover far more of the true data labels than were provided in the alleged metadata (Figures 2c–d and Supplementary Table 1 show an example using the keyboard data). However, in the harder skin site classification task (Figure 2b), random forest's true error rate is only better than the metadata error when the metadata contains between 40 and 80% errors, which is more metadata error than we expect to find in any real dataset.

Conclusions

Supervised classifiers can in some cases be used to detect or correct errata in metadata for microbial communities before the data are subjected to biological interpretation, although they should be used in addition to, and not in place of, careful labeling and data management. In two real datasets, the random forest classifier and the nearest shrunken centroids classifier maintain consistent accuracy until more than 30–40% of samples are mislabeled. Furthermore, when the data categories are highly distinct, the output from the classifier often matches the true data labels much better than the erroneous metadata. In a harder classification task in which the data categories are more subtle, supervised classifiers may not be useful for realistic amounts of metadata error. We encourage future research into this approach, specifically studies of which classifiers are most effective for this purpose, which kinds of metadata are most amenable to this technique, and how best to extend this approach to mislabeling in continuous metadata such as a time series. Although we focused on mislabeled 16S rRNA surveys, similar principles likely hold for metagenomic and other characterizations of microbial communities.

Acknowledgements

We thank Jennifer Wortman, Owen White and Jeremy E Koenig for their input. This work was funded in part by grants from the National Institutes of Health, Howard Hughes Medical Institute, the Crohn's and Colitis Foundation of America, The Hartwell Foundation, and the Arnold and Mabel Beckman Foundation.

D Knights is at Department of Computer Science, University of Colorado, Boulder, CO, USA;
J Kuczynski is at Department of Molecular, Cellular and Developmental Biology, UCB 347, University of Colorado, Boulder, CO, USA;
O Koren is at Department of Microbiology, Cornell University, Ithaca, NY, USA;
RE Ley is at Department of Microbiology, Cornell University, Ithaca, NY, USA;
D Field is at NERC Centre for Ecology and Hydrology, Oxford, UK;
R Knight is at Department of Chemistry and Biochemistry, UCB 215, University of Colorado, Boulder, CO, USA
R Knight is at Howard Hughes Medical Institute, Boulder, CO, USA;
Todd Z DeSantis is at Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA and
ST Kelley is at Department of Biology, San Diego State University, San Diego, CA, USA
E-mail: skelley@sciences.sdsu.edu

References

Breiman L. (2001). Random forests. *Machine Learning* **45**: 5–32.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.

Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. (2010). Forensic identification using skin bacterial communities. *Proc Natl Acad Sci USA* **107**: 6477–6481.

Hugenholtz P, Tyson GW. (2008). Microbiology: metagenomics. *Nature* **455**: 481–483.

Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R *et al.* (2010). Microbes and Health Sackler Colloquium: succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci USA* [e-pub ahead of print].

Riesenfeld CS, Schloss PD, Handelsman J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* **38**: 525–552.

Tibshirani R, Hastie T, Narasimhan B, Chu G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS* **99**: 6567–6572.

Tringe SG, Hugenholtz P. (2008). A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11**: 442–446.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)