VISUALIZATION AND STATISTICAL COMPARISONS OF MICROBIAL COMMUNITIES USING R PACKAGES ON PHYLOCHIP DATA

SUSAN HOLMES*
Statistics Department, Stanford University,
Stanford, CA 94305, USA
**E-mail:susan@stat.stanford.edu*
www-stat.stanford.edu/˜susan/

ALEXANDER ALEKSEYENKO
Center for Health Informatics and Bioinformatics,
NYU School of Medicine,
New York, NY USA

ALDEN TIMME
Statistics Department, Stanford University,
Stanford, CA 94305, USA

TYRRELL NELSON
Department of Civil and Environmental Engineering
Stanford University, Clark Center E-250
318 Campus Drive, Stanford CA, 94305, USA

PANKAJ JAY PASRICHA
Division of Gastroenterology and Hepatology
Stanford University Medical Center
Alway Building, Room M211
300 Pasteur Drive,
Stanford, CA 94305, USA

ALFRED SPORMANN
Department of Civil and Environmental Engineering
Stanford University, Clark Center E-250
318 Campus Drive, Stanford CA, 94305, USA

This article explains the statistical and computational methodology used to analyze species abundances collected using the LNBL Phylochip in a study of Irritable Bowel Syndrome (IBS) in rats.

Some tools already available for the analysis of ordinary microarray data are useful in this type of statistical analysis. For instance in correcting for multiple testing we use Family Wise Error rate control and step-down tests (available in the `multtest` package). Once the most significant species are chosen we use the hypergeometric tests familiar for testing GO categories to test specific phyla and families.

We provide examples of normalization, multivariate projections, batch effect detection and integration of phylogenetic covariation, as well as tree equalization and robustification methods.

Keywords: Hypergeometric Test; PhyloChip; projections; Quality Control; R; Phylogenetic Tree

1. Introduction

We present here some examples of using robust multivariate methods for the specific challenges of microbiome studies. We use as a running example a comparative study of microbiological communities in healthy and IBS rats sampled at different locations in the intestine. The results of the biological analysis have been submitted elsewhere,[1] we concentrate here on the statistical and computational challenges involved in such a project.

1.1. *IBS in humans and rats*

It is believed that alterations in the microflora of humans with IBS comes from changes in colonic fermentation patterns as has been described in King et al.[2] Recently, some research groups have been able to use culture-independent methods and deep high throughput 16S ribosomal RNA gene sequencing to demonstrate significant differences in the microbiome of IBS patients.[3,4] The complexity induced by high individual variation of the microbiome suggested that a good starting point in this comparative study would be a rodent model that mimics the human condition. We have as our working hypothesis that the enteric microflora of adult rats with colonic hypersensitivity would differ from that of controls. We use a comprehensive and relatively simple way of studying the microflora using a 16S rRNA gene DNA microarray called the Phylochip.[5] The Phylochip has the advantage over high-throughput sequencing assays in that it is designed to detect presence and abundance of individual species. A major drawback of utilizing the Phylochip platform for this project was that the chip design was not specific to the intestinal microbiome and as a consequence there is a very unequal resolution in certain phyla, representing unequal knowledge about prokaryotic constituents of these phyla.

1.2. *The data and software platform*

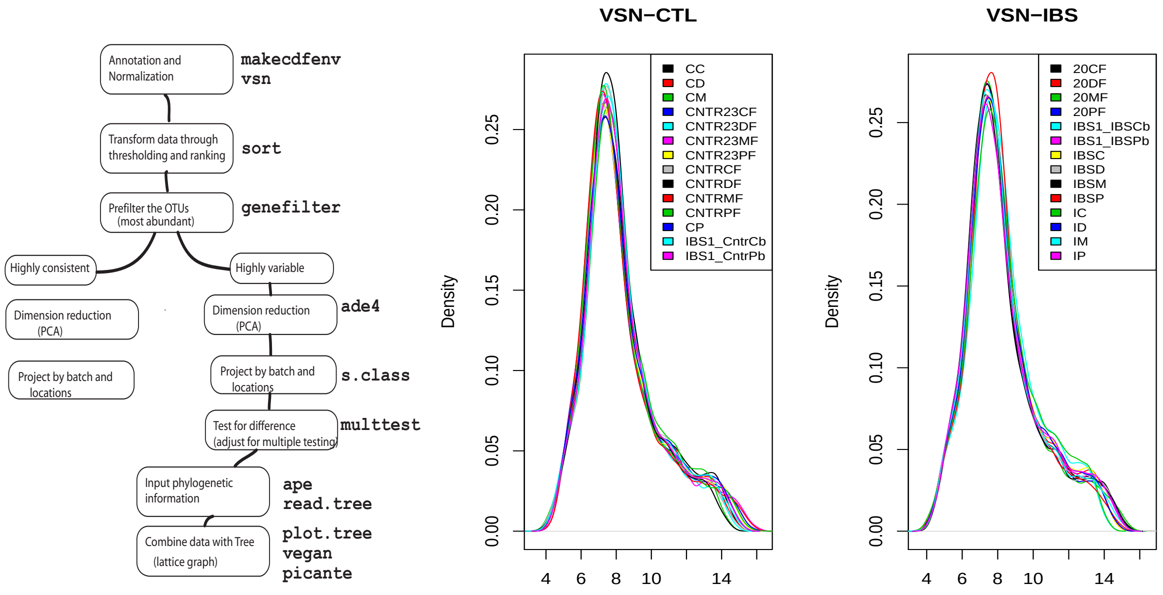
Data were collected on the microbial community of different sections of the large bowel of rats with colonic hypersensitivity induced by neonatal acetic acid irritation. This microarray consists of 500,000 oligonucleotide probes capable of identifying 8743 of bacteria and archaea and provides a comprehensive census for presence and relative abundance of most known prokaryotes in a massive parallel assay. This array uses the the GeneChip (Affymetrix Corporation) technology, thus we could use the Bioconductor[6] suite of tools for annotation[7] and normalization of the data in the same way as is usual for microarray studies.[8] We then used multivariate methods to visualize comparisons between different groupings of the data enabling us to enhance our quality control of the experimental protocol.

We then separated the data into consistently present species and those presenting higher variability. Previous computational approaches include the use of the weighted `unifrac` (Wasserstein distance[9]) between communities.[10] Here we take a geometrical approach to the visualization and detection of various multidimensional biases and changes in variability, as well as the combination of phylogenetic and low rank information. This is more akin to Purdom[11] who also combines phylogenetic and abundance data, but for PCR sequenced phylotypes. Figure 1(a) shows a diagram of the data analysis workflow we chose to follow.

2. Details of the Data Analysis Procedures

2.1. *Prefiltering and Normalization of the Microarray Data*

We created and used a custom-tailored package containing the annotation of all the probes on the Phylochip using the `makecdfenv`[7] package. As with standard expression data, the data need to be preprocessed to ensure that the variance was independent of the level of abundance as described in Durbin et al[12] and implemented in the `vsn` package[8] in the Bioconductor[6] suite of `R`[13] packages. Figure 1 (b) shows the densities of each of the arrays in the two groups after variance stabilizing normalization.



(a) Different stages of Data Analysis (b) Density after variance stabilizing transformations.

Fig. 1: Tools were transposed from the standard microarray analyses

3. Batch Effect Detection using projections on Principal Planes

A standard principal component analysis was done on the centered and scaled abundance data. In the first set of data, we had originally 24 samples, 12 from IBS, 12 from healthy controls that we wanted to compare, the 12 samples for each group came from 4 locations in the large intestine, however the first apparent differences came from batch groups. We had a first batch of samples corresponding to analyses that were done on day 1 consisted of 6 arrays (3 IBS/3CTL), a second batch 18 arrays (9IBS and 9CTL), done on a second date with a different protocol and array batch. We used the additional ability provided by the projection of supplementary group means and variance as in the function `s.class` in the `ade4`[14] package to explore these batch effects in the laboratory methods used to generate the data. The ellipses are computed using the means, variances and covariance of each group of points on both axes, and are drawn with these parameters: the center of the ellipse is centered on the means, its width and height are given by the variances, and the covariance sets the slope of the main axis of the ellipse. In Figure 2, on the left we see the first two batches although both balanced with regards to IBS and healthy rats were extremely different in variability and overall multivariate location. In order to explore this further, a third batch was generated with the same arrays as batch 2 but the same experimental protocol as batch 1. We see that the third group faithfully overlaps with batch 1 thus showing that the batch effect was not due to a difference in arrays but to the experimental protocol. This shows the utility of PCA in quality control. After finding this particular effect we redid part of the data collection procedure, using only the protocol used in batches 1 and 3, we analyzed 24 samples. We also added 8 samples from mucosal linings, 4 from IBS , 4 for control in each of the 4 intestinal locations. We combined the data into a 32 column matrix of abundance of 8364 species. Since the abundance data were extremely variable and we had seen the sensitivity of the data to varying conditions and protocols we decided to pair the data by location and type. For each pair we had an IBS and a CTL rat, for a sample collected in the location and in the same way, we used the pairing design to minimize the biases from experimental artifacts.

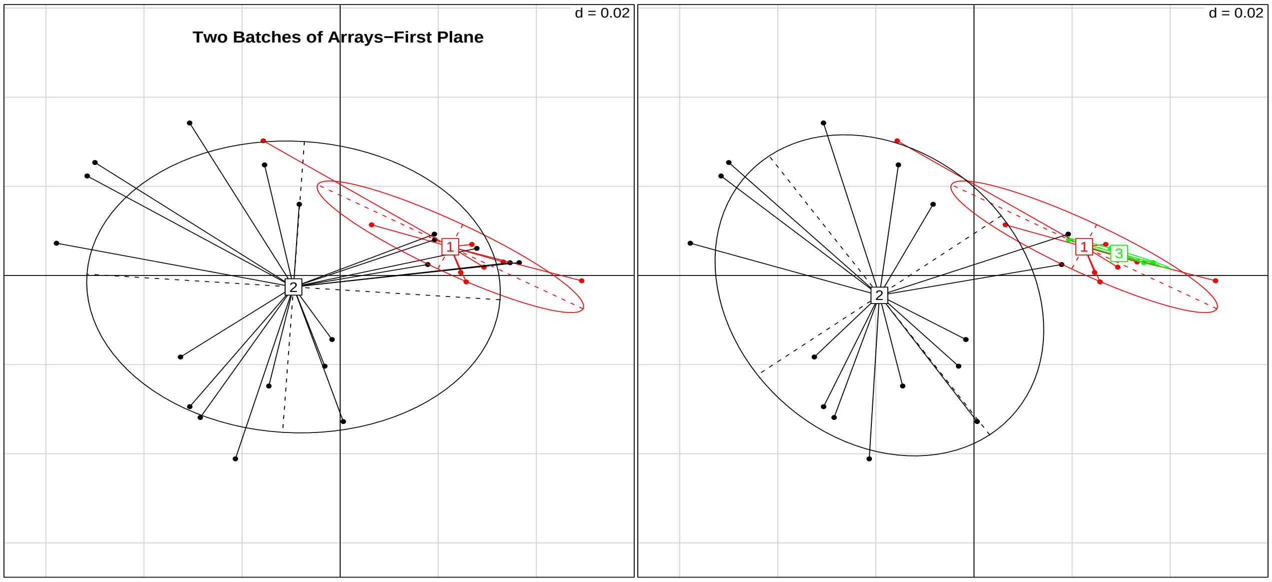


Fig. 2: On the left the first plane of the PCA shows the first set of data with two batches and on the right the third set of arrays was added.

3.1. *Ranking and Thresholding*

In order to deliver a more robust statistical analysis, we ranked the species abundances within each array: the ranks go from 1 (small) to 8364 (large). This is a standard non parametric statistical procedure that enhances the stability of the results because a few outliers cannot bias the analyses. We considered that there were not more than 2000 species present so we set a threshold at 6000 (this is conservative as for instance a recent study in humans places the estimate of numbers of species in the human gut at between 1,000 and 1,200[15]). We thus suppose that all ranks smaller than 6000 were just noise and set them all to be equal to 6000. This avoids finding large differences in ranks for species that are only present at the noise level. We restrict the first part of our analysis here to the species that appeared present in almost all 32 arrays, ie those that had a ranking larger than 6000 in all but one of the arrays. We can see the distribution patterns with varying thresholds from 5000 to 8400 in Table 1. As in microarray studies, it is important to prefilter the species so that only those yielding consistent

#Arrays	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
$\# > 5000$	3997	241	144	91	64	60	55	43	43	37	45	46	41	28	28	38	23	26	26	38	42	41	33	40	47	56	54	47	59	80	88	167	2766
$\# > 6000$	5180	207	136	71	62	48	32	39	38	31	34	25	25	24	24	24	22	24	20	22	20	26	28	46	43	41	41	45	40	46	72	109	1989
$\# > 7000$	6492	120	82	40	32	27	31	13	33	22	22	18	19	12	13	20	17	14	18	18	25	18	20	26	22	18	21	19	26	30	59	83	1204
$\# > 8000$	7737	70	35	25	9	9	10	12	13	12	15	14	9	11	9	11	7	11	7	11	11	11	10	9	16	9	12	13	12	18	23	38	415
$\# > 8400$	8235	36	24	21	14	6	6	11	10	5	8	5	5	5	6	2	6	7	2	8	4	5	7	6	13	10	5	5	5	6	16	18	112

Table 1: Tables showing the number of species present at a given level of abundance as measured by ranks in 0,1,2,. . . ,32 arrays. We can see in particular that there are about 2,000 species present at least at the rank 6000 in all 32 arrays and about 415 which are highly abundant (> 8000) in all arrays.

signals enter the analysis. In particular, this is important for various testing procedures we will use later (testing for differences between IBS and CTL), where having extra non-meaningful species costs us extra power requiring us to perform more tests than necessary. Table 1 is the basis of most of the prefiltering presented in the paper.

4. Incorporating and adjusting the phylogenetic information

4.1. *Difficulty with the Original Tree: heterogeneous levels of resolution*

We entered the complete phylogeny of 16sRNA provided by GreenGenes into the **R**[13] package ape.[16] We can see in the left tree of Figure 3 that the phylogenetic tree of all the bacteria tested for on the microarrays is not ultra-metric. That is, not every species is at the same distance from the root. When looking at the phylogenetic tree (Figure 3), it is evident that some areas of the tree have much greater resolution than others. The problem with this is that some species of bacteria are probed multiple times by the array. Therefore, they have more chances than other bacteria of showing significance under the null hypothesis. For example, of the 158 bacteria found to be significantly over- or under-abundant in IBS rats at the $\alpha = 0.05$ level in the first dataset (excluding the mucosal samples), nine are C. leptum, ten are R. hansenii, and ten are P. ruminicola. One of the questions that must be answered is whether or not higher resolution in certain areas of the phylogenetic tree caused these species to be over-represented among the bacteria of interest. We will see below that the hypergeometric test provides a way to control the phylogenetic bias at the higher-order level, but there is a lot of information lost when we look only at phyla. In an attempt to conserve information while correcting for this oversampling in certain regions, we also propose a method for collapsing the tree by merging the tips of related species with similar microarray intensities.

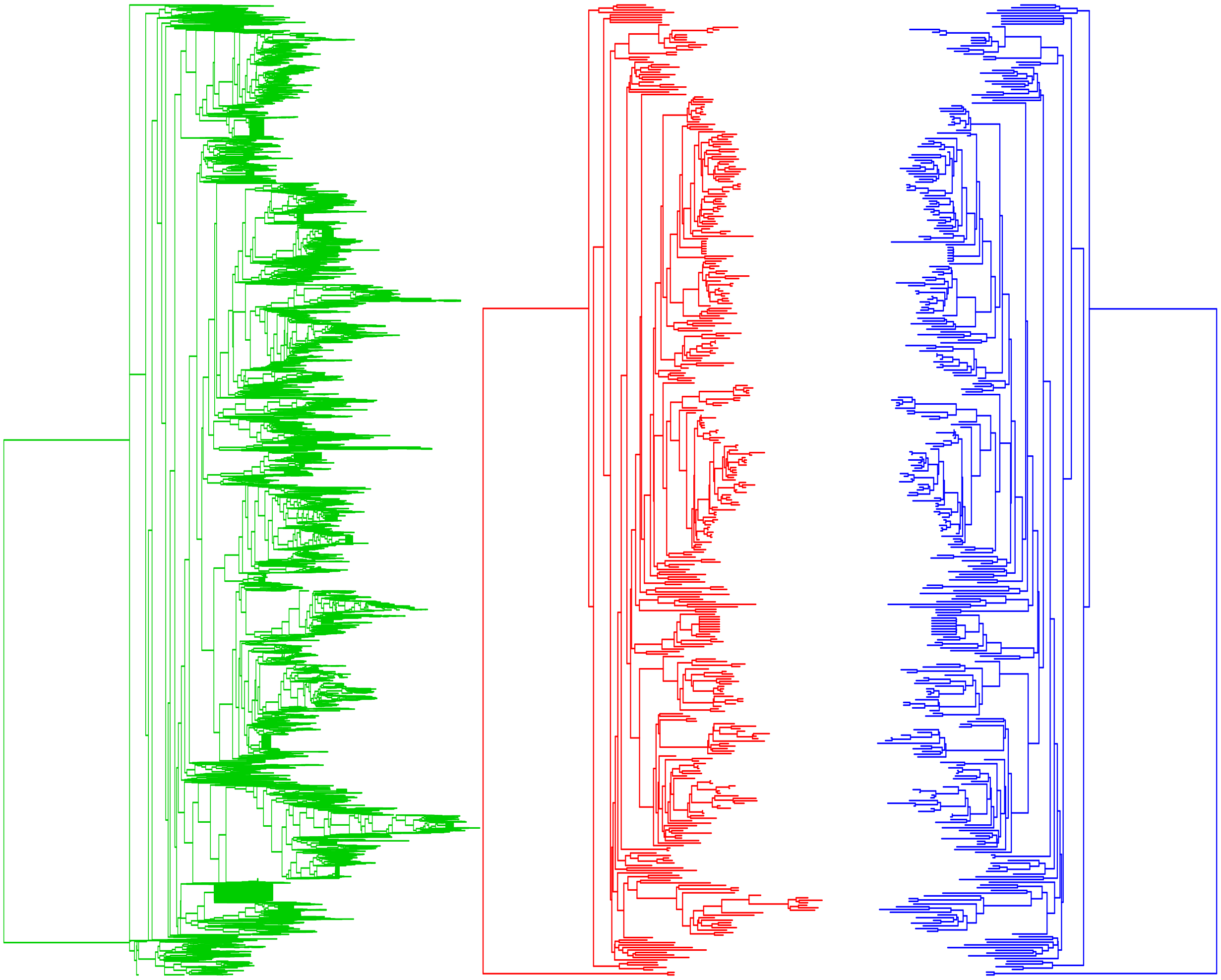


Fig. 3: On the left, we have the tree of all operational taxonomic units (otus) present on the Phylochip, we can observe that the distance to the root of many of the otus is variable, thus indicating a heterogeneous degree of resolution. The two trees on the right are filtered trees representing only the 400 most abundant species. The blue tree on the right was computed by using the collapsing algorithm presented in this section, we see that the long right clade at the bottom of the middle tree has disappeared.

The idea is to control for over-resolution by merging tips of the clades that are more resolved, creating a more level playing field for the multiple testing. We used the length from the root of the tree as the main parameter for collapsing tips. That is, for any two species further from the root than the given maximum distance, we try to merge the two tips. However, merging is only done if the microarray data from the two species are similar enough to be merged. Tips are only merged if there is a low enough variance across the bacteria for each microarray measurement. What is a low enough variance, however, is difficult to define. For the purposes of the analysis here, we used a bootstrap procedure[17] that estimated the $q = 0.9$-quantile for a random collection of groups of size n bacteria. This served as the cutoff of what could be considered a small enough variability within that clade. A collection of n

bacteria is merged only when all their tips are farther than the maximum length from root and $p = 80\%$ of the 32 variances across the collection (one for each microarray) are below the computed thresholds. These were arbitrary thresholds that we have only evaluated empirically by running the algorithm with varying values for n, q and p.

4.2. *Consistently Abundant Species and their place on the Tree*

Here we chose about the top 100 most consistently abundant species following the choice of a threshold of about 8400 as in Table 1. Here we show how we can use the enhanced plotting facilities in R through the Lattice compatible packages, we can plot the complete tree, identifying the part of the tree which is covered by a subset of species.

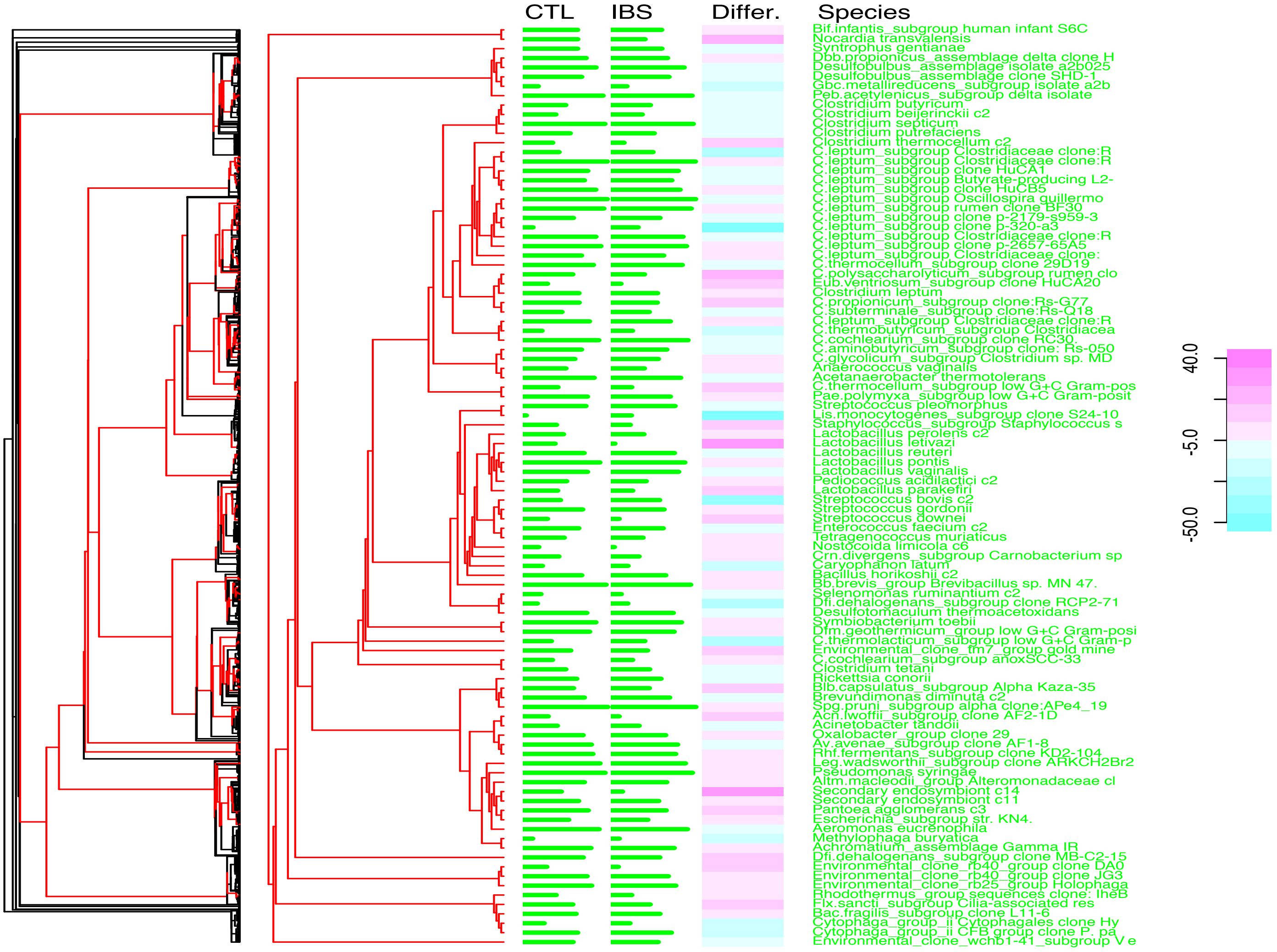


Fig. 4: The left tree shows the complete tree on all species in black with the subtree of most abundant species in red, this subtree is the one plotted on the next panel. Values of abundance in CTL and IBS rats are plotted in the next two columns, the pink/blue scaled variables are the truncated rank differences between the two groups.

4.3. *Highly variable species*

We concentrate now on the species which are abundant enough to be considered consistently present (more than 15 out of 32 arrays over 7800) but that also show high variability (standard deviation above 150). These values were arbitrarily chosen to retain about 100 species. There were actually 99 such species for which we had complete annotation information. We then

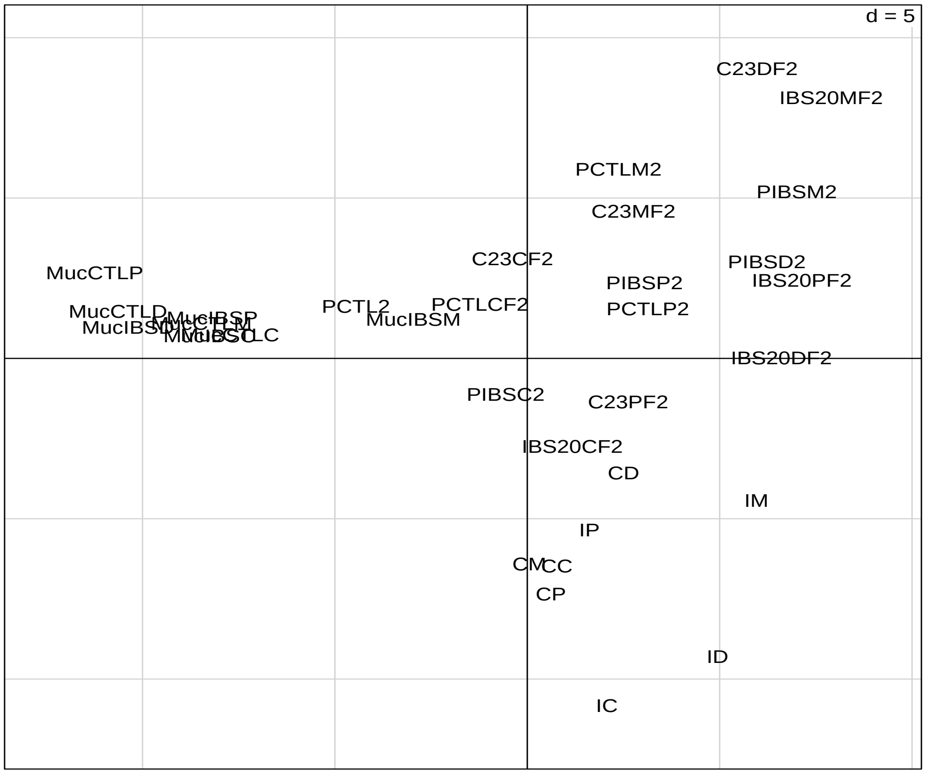


Fig. 5: Principal component analysis of top most abundant and variable species, we see the Mucosal location is the explanation for the first component, all the mucosal samples have negative loadings on this factor.

took the results of the PCA analysis and combined them with the tree information by using the loadings on the first two components (which account for 55% variance) and plotted them alongside the phylogenetic sub tree of the species we had retained as most variable. This plot is much easier to read than the projections of long species names in the two dimensional principal plane. We have colored in red the species that are more abundant in the mucosal samples.

4.4. *Multiple Testing for finding differentially expressed species*

The first set of analyses showed that the main differences were batch effects and differences between the mucosal and other samples, so we decided to proceed by pairing the data by location, batch and mucosal types, thus removing the extra variance due to these factors. Thus we proceed into the testing phase using a paired design and we will use corrections made on the paired t-test rather than the ordinary one. We will use truncated paired differences

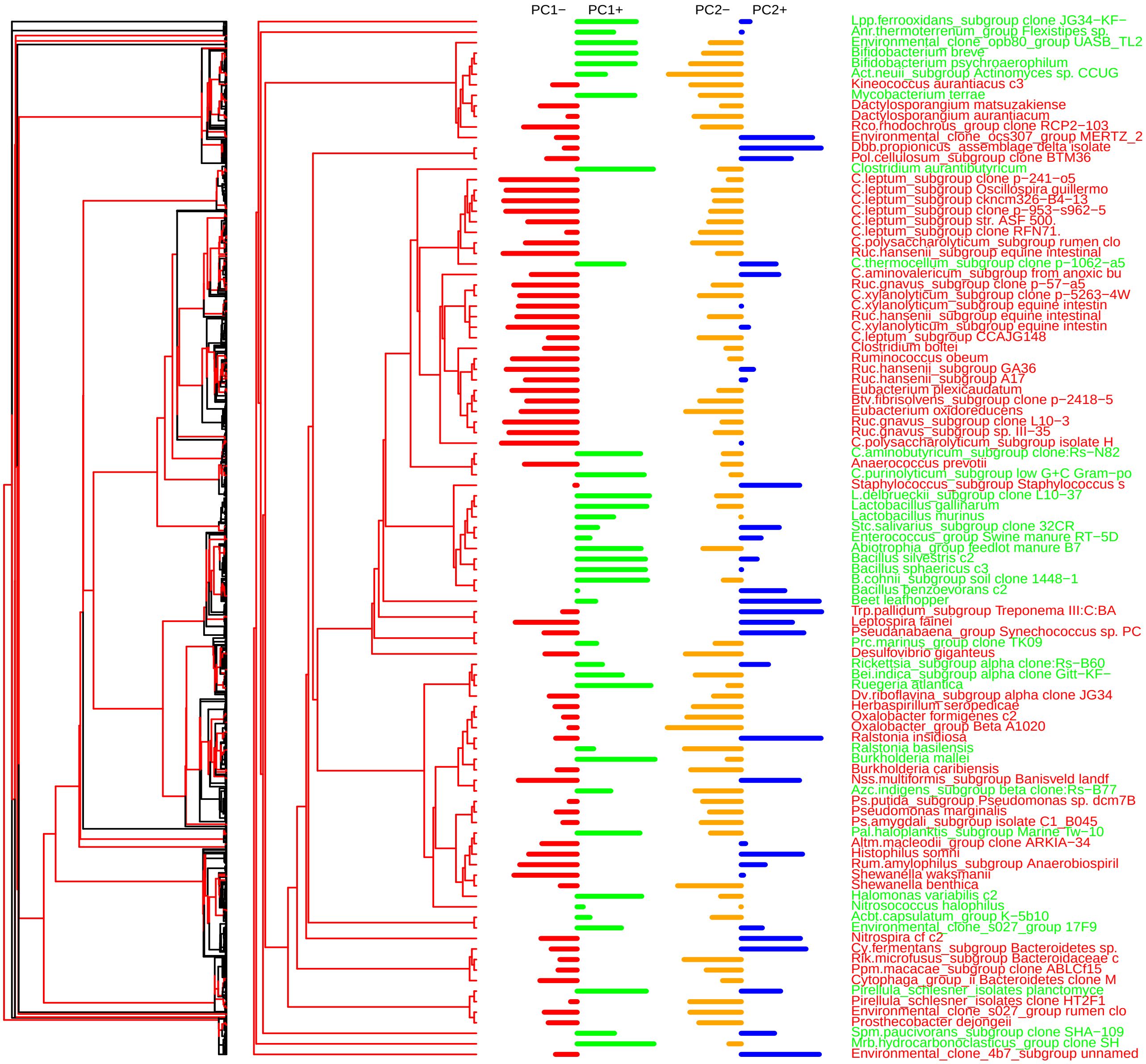


Fig. 6: Complete tree with the subtree of most variable among the consistently abundant species and the loadings on the first two principal components.

in ranks as input to standard multiple testing programs for finding the adjusted p-values. To control for false discovery due to multiple testing, p-values were adjusted according to the Benjamini-Hochberg procedure, which is able to control for FDR given some assumptions on the expression levels of the bacteria on the microarray. We used the `multtest` package from `Bioconductor`.[6]

4.5. *Significant differences projected onto the Tree*

In order to visualize the parts of the phylogenetic tree most influenced by changes in species abundance between groups we retained the most significantly changed species (up in IBS or up in CTL) on the tree and used the facilities available through the `ape`[18] and the `lattice`[19] packages.

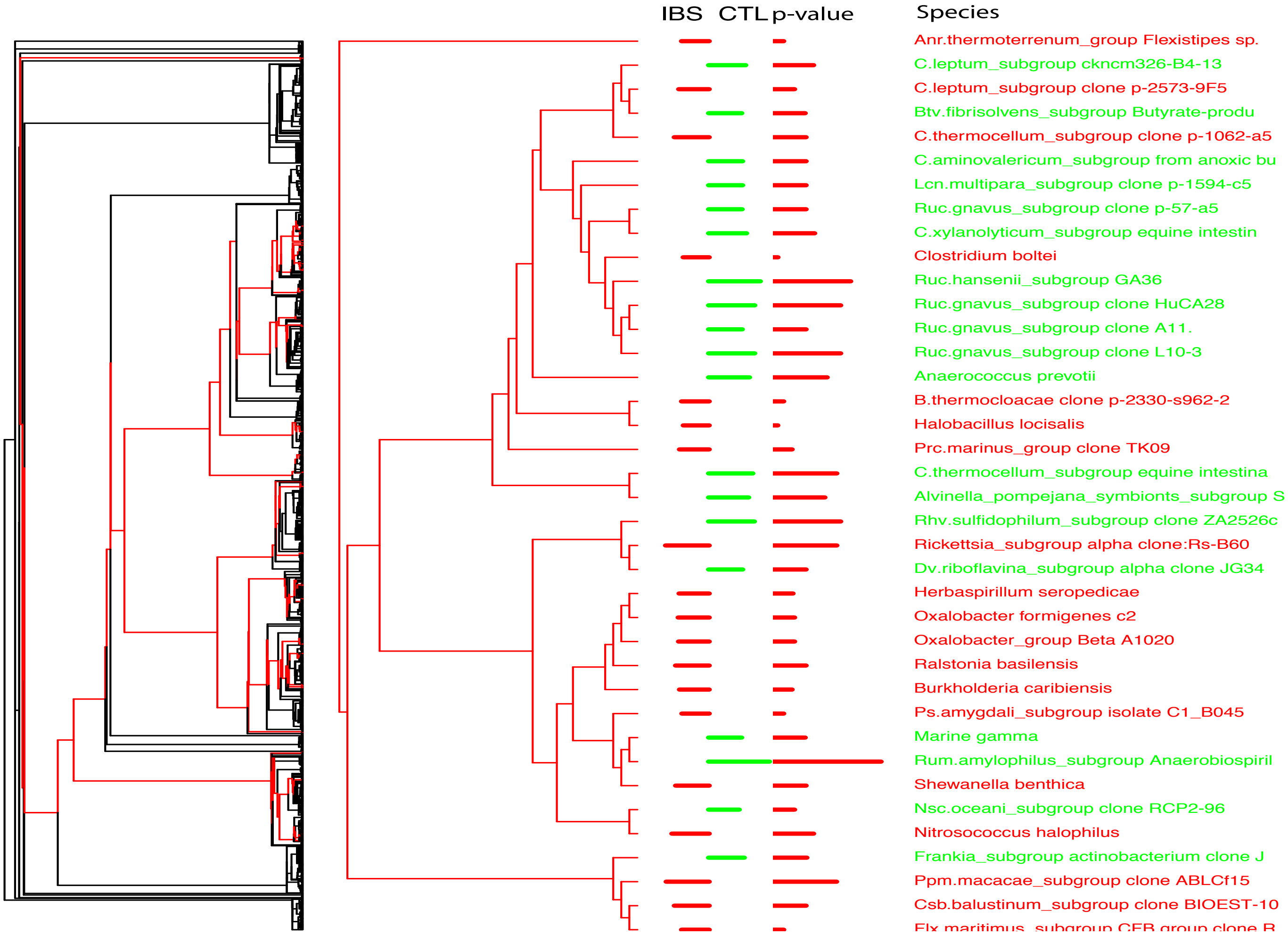


Fig. 7: The left tree shows the complete tree on all species in black with the subtree of set of species that show the most significantly differences between CTL and IBS in red in the second panel. Values of abundance in CTL and IBS rats are plotted in the next two columns, the next column shows the $-\log(pvalue)$, so the largest bars represent the most significantly different species.

4.6. *Category Based Comparisons*

We chose as the list of most significant species those that had adjusted p-values lower than 0.05 in the multiple testing procedure detailed above. We created two lists, one for which the ranked abundances were larger in the IBS, the other for which the ranked abundances were larger in the CTL group. We wanted to find specific families or phyla that are over-represented in either of the lists. This is a similar situation as that of testing significance of Gene Ontology categories for expression studies. We recall that in both situations the relevant test is the hypergeometric and that Fisher's exact test and the hypergeometric test formulation are equivalent.[20] We define the set of prefiltered species (**species universe**) as those that passed the threshold test of being present (> 6000) in at least 31 of the arrays (see Table 1). The chosen species (universe and significant) are then binned by phyla or families, these categories replace the Gene Ontology categories used in microarray studies. We are looking for overrepresentation of certain families or phyla. This method is especially relevant here as the chip does not have equal representation of different families and phyla.

The results and details of the hypergeometric tests can be consulted in Nelson et al, 2010[1] where we conclude in particular that the IBS had significantly more Bacteriodetes and on the other hand there is an overrepresentation of Firmicutes in the healthy controls. At the family level, the results showed that the families of Oxalobacteraceae, Prevotellaceae, Burkholderiaceae, Sphingobacteriaceae were significantly overrepresented in IBS rat. Conversely, the most significantly enriched family in control rats were Lachnospiraceae, including Ruminococcus sp., followed by Erysipelotrichaeceae and Clostridiaceae.

5. Summary

Some methods developed for standard microarray studies can be useful in Phylochip studies, examples shown here include variance stabilization, prefiltering, multiple testing and hypergeometric tests.

Batch effects can be detected through multivariate projections using methods such as PCA complemented with the projections of the relevant means, variance and covariance ellipses on the principal planes. We concluded that the best way to counter batch effects was then to use paired differences between subjects if a comparative design is available.

High between subject variability in bacterial abundances suggests the use of ranks is more effective than the original intensities. This method is known to be robust in the sense that if some of the abundance values are on very different scales, their effect on the overall outcome can be minimized by replacing the original values by the ranks within each array. We have provided an example of such an approach here.

Finally the integration of complex phylogenetic structure is possible through the conjoint use of the many available packages in R for doing phylogenetics and community analysis. We have provided an example of a complex combination of plotting trees and results from PCA.

Acknowledgments

We thank Todd DeSantis and Cindy Wu for enlightening conversations about the Phylochip and all participants in the R and Bioconductor projects for their packages. This work was supported by a grant from the Bio-X Interdisciplinary Initiative Program IIP4-32 to P.J. Pasricha and A. Spormann and grant NIH-5-R01GM086884 to S. Holmes. We thank the reviewers who made useful comments that helped improve the paper.

References

1. T. A. Nelson, S. Holmes, A. V. Alekseyenko, M. Shenoy, T. DeSantis, C. Wu, G. L. Anderson, J. Sonnenburg, P. J. Pasricha and A. Spormann, Phylochip microarray analysis reveals altered gastrointestinal microbial communities in a rat model of colonic hypersensitivity, Submitted.
2. T. King, M. Elia and J. Hunter, *The Lancet* (Jan 1998).
3. E. Malinen, T. Rinttilä, K. Kajander and J. Mättö, *The American journal of Gastroenterology* **100**, 373 (Jan 2005).
4. A. Kassinen, L. Krogius-Kurikka and H. Mäkivuokko, *Gastroenterology* (Jan 2007).
5. K. H. Wilson, W. J. Wilson, J. L. Radosevich, T. Z. DeSantis, V. S. Viswanathan, T. A. Kuczmarski and G. L. Andersen, *Appl. Environ. Microbiol.* **68**, 2535 (2002).
6. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang and J. Zhang, *Genome Biology* **5**, p. R80 (Jan 2004).
7. R. A. Irizarry, L. Gautier, W. Huber and B. Bolstad, makecdfenv_1.18 (2009), `http://cran.r-project.org/doc/packages/makecdfenv.pdf`.
8. W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka and M. Vingron, *Bioinformatics* **18 Suppl 1**, S96 (Jan 2002).
9. S. N. Evans and F. A. Matsen, *arXiv* **q-bio.PE** (Jan 2010).
10. M. Hamady, C. Lozupone and R. Knight, *The ISME Journal* (Jan 2009).
11. E. Purdom, *Annals of Applied Statistics* .
12. B. Durbin, J. Hardin, D. Hawkins and D. Rocke, *Bioinformatics* **19**, 1360 (2003).
13. R. Ihaka and R. Gentleman, *Journal of Computational and Graphical Statistics* **5**, 299 (1996).
14. D. Chessel, A. Dufour and J. Thioulouse, *R News* **4**, 5 (2004).
15. J. Qin, R. Li, J. Raes, M. Arumugam, K. er Solvsten Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. ce Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. ong Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. orben Hansen, D. L. Paslier, A. Linneberg, H. B. Nielsen, E. P. tier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. ang Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. gang Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. D. a nd Francisco Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, M. Consortium, P. Bork, S. D. Ehrlich and J. Wang, *Nature* **464**, 59 (Mar 2010).
16. E. Paradis, Ape (analysis of phylogenetics and evolution) v1.8-2 (2006), `http://cran.r-project.org/doc/packages/ape.pdf`.
17. B. Efron, R. Tibshirani and R. Tibshirani, *An introduction to the bootstrap* (Chapman & Hall/CRC, 1993).
18. E. Paradis, J. Claude and K. Strimmer, *Bioinformatics* **20**, 289 (2004).
19. D. Sarkar, *lattice: Lattice Graphics*, (2009). R package version 0.17-26.
20. I. Rivals, L. Personnaz, L. Taing and M.-C. Potier, *Bioinformatics* **23**, 401 (Feb 2007).
21. S. Kembel, P. Cowan, M. Helmus, W. Cornwell, H. Morlon, D. Ackerly, S. Blomberg and C. Webb, *Bioinformatics* **26**, 1463 (2010).
22. K. S. Pollard, H. N. Gilbert, Y. Ge, S. Taylor and S. Dudoit, *multtest: Resampling-based multiple hypothesis testing*, (2010). R package version 2.4.0.
23. J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, R. G. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens and H. Wagner, *vegan: Community Ecology Package*, (2010). R package version 1.17-0.