

Abstract

Numerous studies are currently underway to characterize the microbial communities inhabiting our world. These studies will dramatically expand our understanding of the microbial biosphere and, more importantly, will reveal the secrets of the complex symbiotic relationship between us and our commensal bacterial communities. An important prerequisite for such discoveries are computational tools able to rapidly and accurately compare large datasets generated from complex bacterial communities. We describe a statistical method for detecting differentially abundant features between two populations using count data (e.g. 16S rRNA surveys to find differentially abundant taxa). In high-complexity environments, our method employs the false discovery rate to improve specificity and properly handles low abundance taxa. We demonstrate the use of our tool on several publicly available datasets: 16S rRNA surveys of human and mouse gut microbiomes, and metabolic subsystem data from 85 microbial and viral metagenomes. These methods provide a statistical approach for analyzing frequency data to detect differentially abundant categories between two populations, specifically targeted at clinical studies comprising large numbers of samples.

Availability: A web server implementation of our methods and free source code are available at <http://metastats.cccb.umd.edu>.

Background

Current metagenomics software available for comparing communities

- DOTUR - clusters sequences into OTUs given a distance matrix. Calculates rarefaction curves and nonparametric estimators of richness and diversity (Fig. 1A).
- SONS - extended DOTUR by finding OTUs shared between multiple communities (Fig. 1B).
- J-libshuff - hypothesis test: are these two libraries drawn from the same environment?
- UNIFRAC and TREECLIMBER - analyses to cluster phylogenetically-similar communities.
- MEGAN - graphical interface to compare the taxonomic composition of environments.
- Rodriguez-Brito *et al.* - uses a computationally intensive bootstrapping approach to determine enriched or depleted subsystems between two single samples.

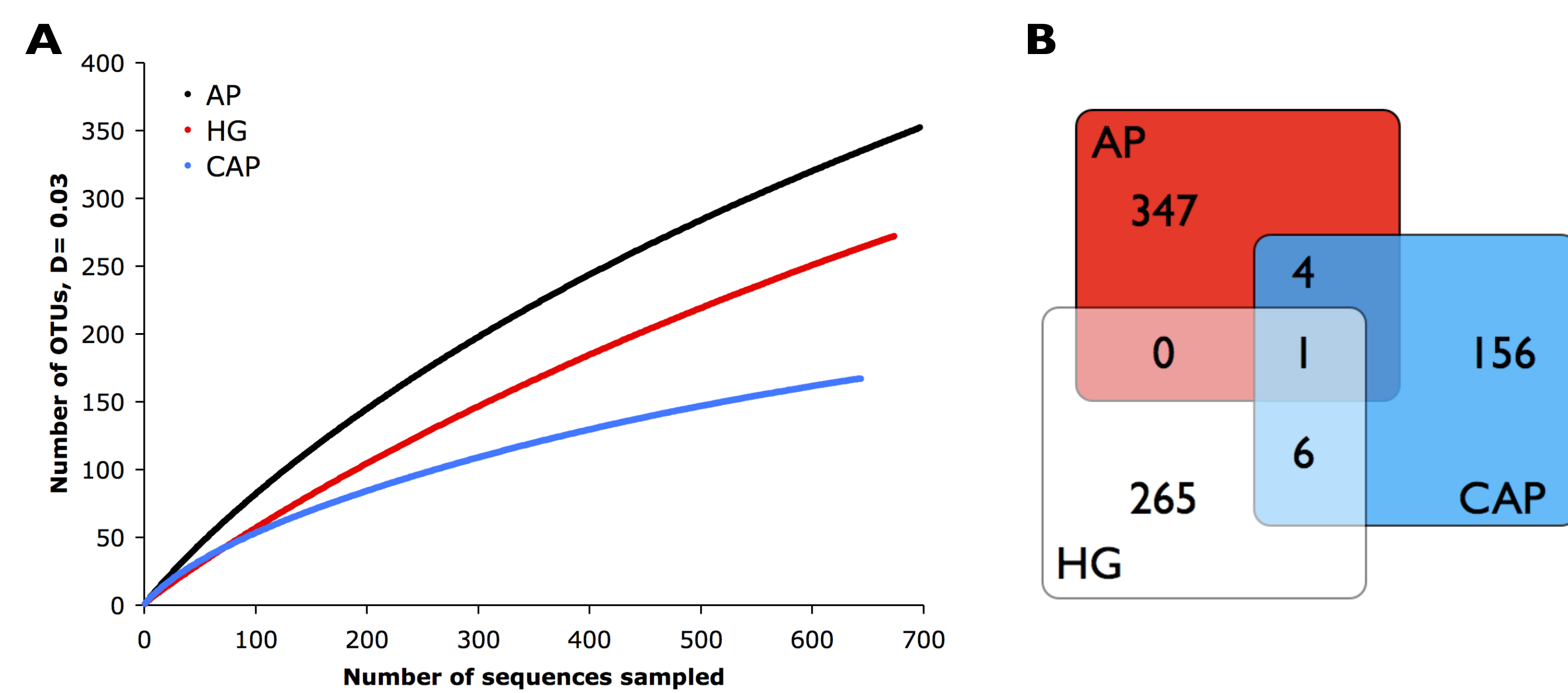


Figure 1

Comparison of three environments using (A) DOTUR and (B) SONS. DOTUR estimates the diversity and sampling coverage of an environment using rarefaction curves. Analyzing the OTU generated by DOTUR, SONS determines how many OTUs are shared between environments.

Most of these programs answer the general question of **whether** two environments differ, rather than exactly **how** they differ. There is an evident need for software capable of detecting differences in microbial communities in studies with multiple subjects. We are specifically interested in differential abundance, e.g. features in a population that are either enriched or depleted compared to another population.

Methods

We designed the following methodology to detect differentially abundant features (e.g. taxa, subsystems, pathways...) between two populations in metagenomic studies with multiple subjects.

Feature abundance matrix

The input to our method can be represented as a feature abundance matrix whose rows correspond to specific features, and whose columns correspond to individual metagenomic samples. The cell in the i^{th} row and j^{th} column is the total number of observations of feature i in sample j (Fig. 2).

	S1	S2	S(N-1)	SN
T1	c(1,1)	c(1,2)	c(1,N-1)	c(1,N)
T2	c(2,1)	c(2,2)			
...					
T(M-1)	c(M-1,1)				
TM	c(M,1)			c(M,N)

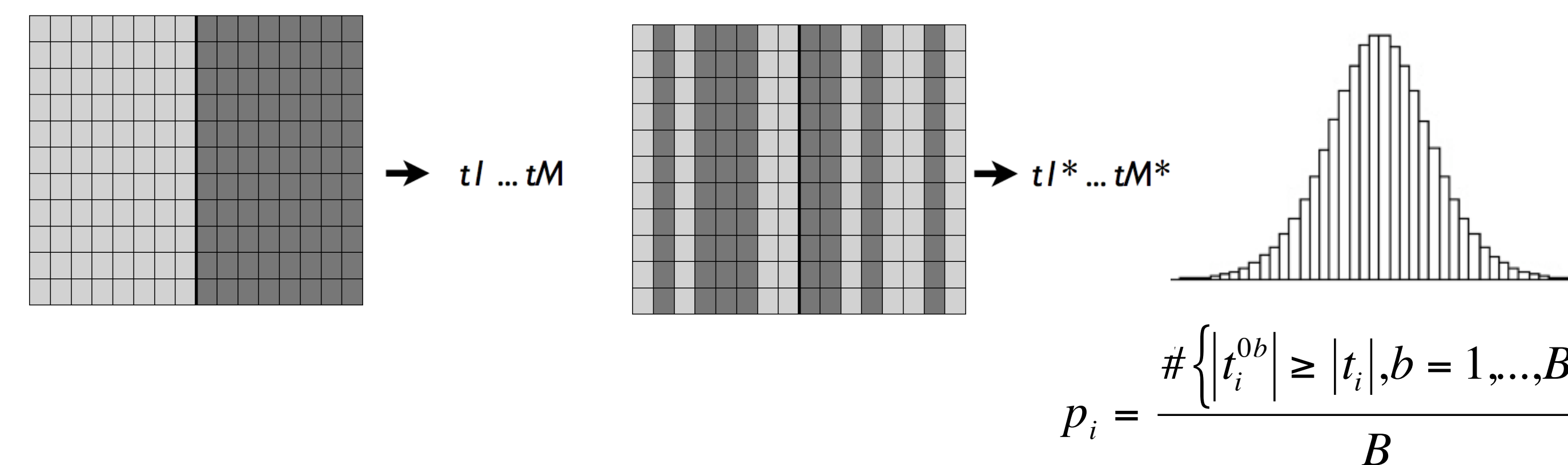
Figure 2 The feature abundance matrix.

Permutation-based p-values

After normalizing the count data to proportions, we then compute the two-sample t statistic:

$$t_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}$$

To account for the likely non-normal underlying distribution of the data, we perform the following procedure. We randomly permute the treatment labels of the columns of the abundance matrix and recalculate the t statistics. Repeating this procedure for B trials, we obtain B sets of t statistics. For each feature, the p-value associated with the observed t statistic is calculated as the fraction of permuted tests with a higher value:



The false discovery rate (FDR)

Rather than use a standard Bonferroni correction to control false positives, we chose to control the false discovery rate (FDR), which is defined as the proportion of false positives within the set of predictions, in contrast to the false positive rate defined as the proportion of false positives within the entire set of tests. In this context, the significance of a test is measured by a q-value, an individual measure of the FDR for each test. Controlling the FDR is preferable to a Bonferroni correction because statistical power is maintained.

Handling rare features

For low frequency features, e.g. low abundance taxa, the t statistic computation described above is not accurate. We instead compare the differential abundance of sparsely-sampled (rare) features using Fisher's exact test. Fisher's exact test models the sampling process according to a hypergeometric distribution (sampling without replacement), rather than a binomial distribution. The drawback of this approach is that inter-subject variability is ignored by this test.

Results

Gut microbiomes of lean and obese humans (Ley *et al.* 2006)

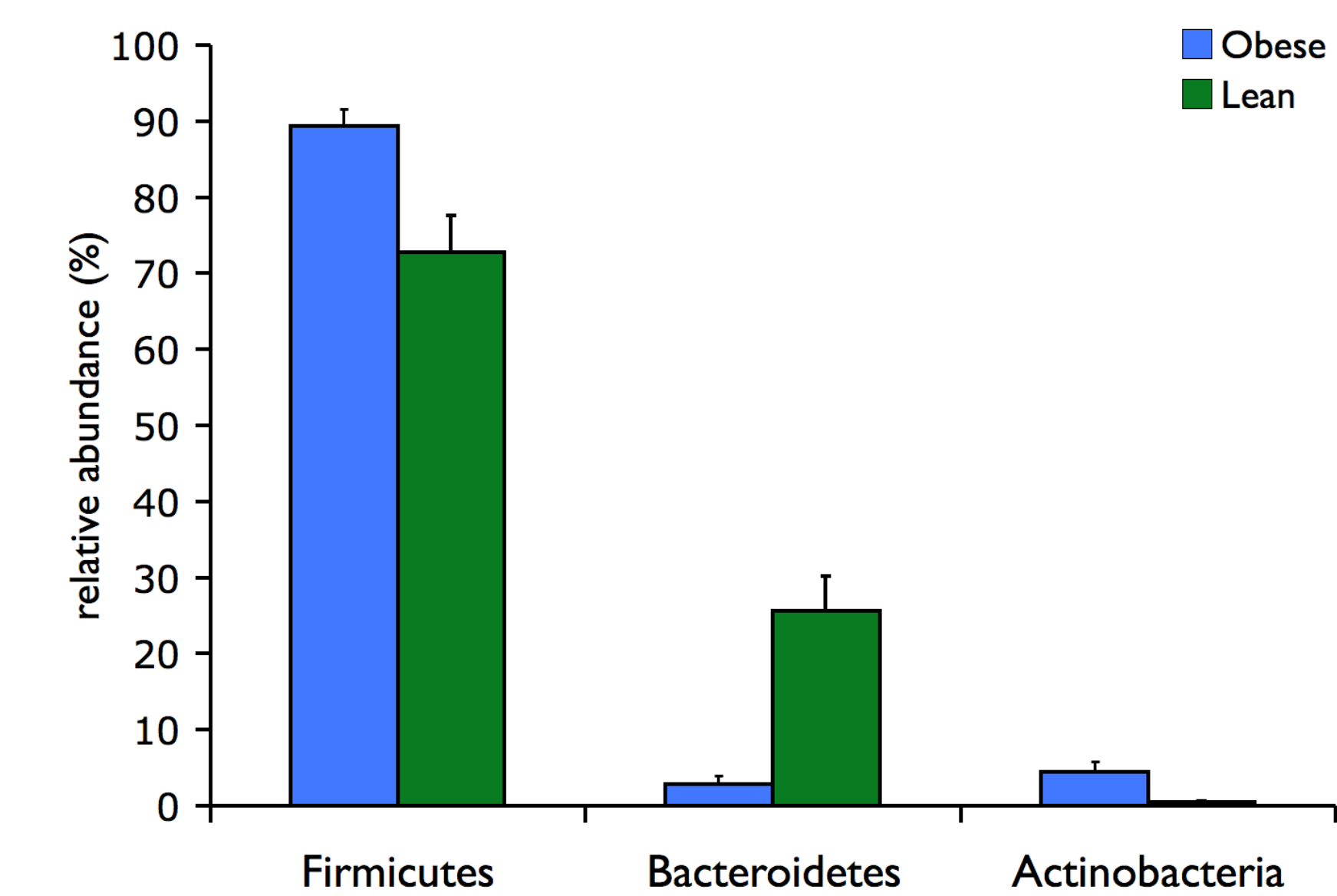


Figure 3 Differentially abundant phyla detected using our method (mean percentage \pm s.e., p-value \leq 0.05). No p-value correction for multiple hypothesis tests was employed. We successfully re-established the major result of Ley *et al.*, and discovered that Actinobacteria are also differentially abundant ($P = 0.004$). Both Firmicutes and Actinobacteria have significantly higher relative abundances in obese people than lean people. In the lean population, Bacteroidetes make up a higher proportion of the gut microbiota than in the obese population.

Microbial and viral functional capacities (Dinsdale *et al.* 2008)

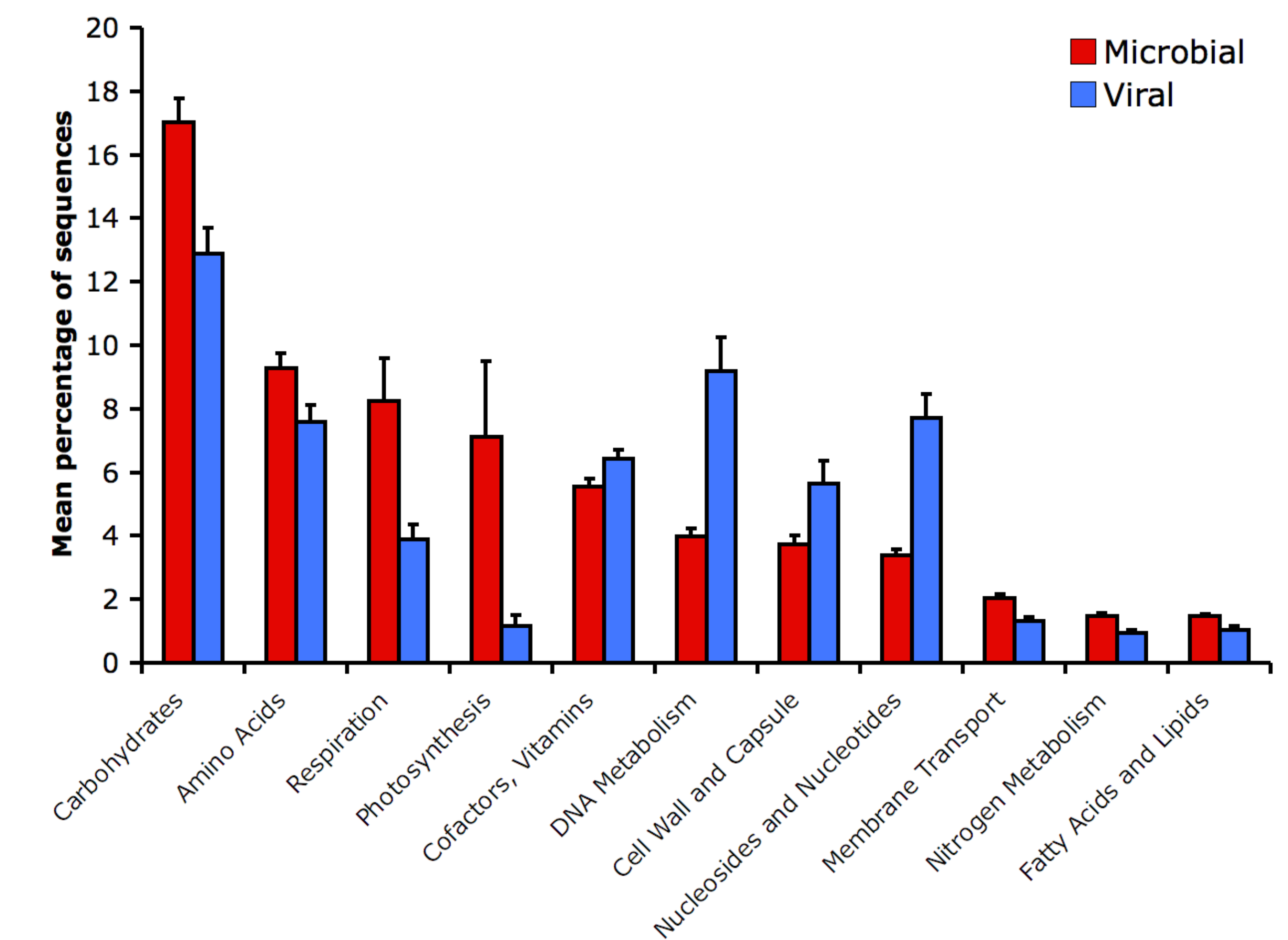


Figure 4 Differentially abundant metabolic subsystems between microbial and viral metagenomes (mean percentage \pm s.e., p-values \leq 0.02). We find that viral metagenomes are significantly enriched for nucleotides and nucleosides ($P < 1e-6$) and DNA metabolism ($P < 1e-4$). Processes for respiration, photosynthesis, and carbohydrates are overrepresented in microbial metagenomes.

Bibliography

1. Dinsdale, E.A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* 452, 629-32 (2008).
2. Ley, R.E., Turnbaugh, P.J., Klein, S. & Gordon, J.I. Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022-3 (2006).
3. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI: Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* 102, 11070-11075 (2005).
4. Rodriguez-Brito, B., Rohwer, F. & Edwards, R.A. An application of statistics to comparative metagenomics. *BMC Bioinformatics* 7, 162 (2006).
5. Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100, 9440-9445 (2003).