BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 6 2010, pages 715–721
doi:10.1093/bioinformatics/btq041

Genome analysis

Advance Access publication February 3 2010

Identifying biologically relevant differences between metagenomic communities

Donovan H. Parks and Robert G. Beiko*

Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, Nova Scotia, Canada B3H 1W5

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Metagenomics is the study of genetic material recovered directly from environmental samples. Taxonomic and functional differences between metagenomic samples can highlight the influence of ecological factors on patterns of microbial life in a wide range of habitats. Statistical hypothesis tests can help us distinguish ecological influences from sampling artifacts, but knowledge of only the P-value from a statistical hypothesis test is insufficient to make inferences about biological relevance. Current reporting practices for pairwise comparative metagenomics are inadequate, and better tools are needed for comparative metagenomic analysis.

Results: We have developed a new software package, STAMP, for comparative metagenomics that supports best practices in analysis and reporting. Examination of a pair of iron mine metagenomes demonstrates that deeper biological insights can be gained using statistical techniques available in our software. An analysis of the functional potential of '*Candidatus* Accumulibacter phosphatis' in two enhanced biological phosphorus removal metagenomes identified several subsystems that differ between the *A.phosphatis* stains in these related communities, including phosphate metabolism, secretion and metal transport.

Availability: Python source code and binaries are freely available from our website at http://kiwi.cs.dal.ca/Software/STAMP

Contact: beiko@cs.dal.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 11, 2009; revised on January 14, 2010; accepted on January 26, 2010

1 INTRODUCTION

With its focus on ecologically relevant assemblages of micro-organisms, metagenomics—the assessment of randomly collected DNA fragments from a microbial community—has revealed a new world of microbial diversity and complexity. Metagenomic studies can be motivated by several goals, including the discovery of novel genes of interest (Béjà, *et al.* 2000; Yooseph *et al.*, 2007), validation of metabolic hypotheses (García Martín *et al.*, 2006; Hallam *et al.*, 2004; Mou *et al.*, 2008), profiling of the relationship between microbial community composition and variation in environmental or geographic parameters (Dinsdale *et al.*, 2008a; Ley *et al.*, 2006) and assessment and comparison of the global metabolic complement found in one or more habitats (Brulc *et al.*, 2009; Edwards *et al.*, 2006; Tringe *et al.*, 2005; Turnbaugh *et al.*, 2009). Detailed assessment of communities requires intensive sequencing effort and careful experimental design to ensure that experimental questions can be adequately addressed (Hamady and Knight, 2009).

A common outcome of a metagenome project is a set of functional predictions for the collected and assembled DNA reads, assigned via homology search and classified using a scheme such as the SEED (Overbeek *et al.*, 2005) or KEGG (Kanehisa *et al.*, 2004). When multiple samples are collected, functional classes can be organized into contingency tables that allow the assessment of statistically significant differences in the relative abundance of these classes. For example, Edwards *et al.* (2006) sampled communities from an oxidized and anoxic environment within the Soudan iron mines and found evidence that these two communities preferentially utilize different respiratory pathways. Several dedicated bioinformatics applications have been developed for the statistical analysis of metagenome pairs, including XIPE-TOTEC (Rodriguez-Brito *et al.*, 2006), ShotgunFunctionalizeR (Kristiansson *et al.*, 2009) and MEGAN (Mitra *et al*., 2009). Packages focusing on the complementary problems of assessing differences between groups of metagenomes have also been developed (Gianoulis *et al.*, 2009; Lozupone and Knight, 2005; Schloss *et al.*, 2009; White *et al.*, 2009).

While such comparisons are informative, it is vital to distinguish results that are *biologically* relevant (i.e. truly different in frequency between two or more sites, due to some underlying taxonomic or ecological phenomenon) from those that are merely *statistically* significant. Statistical significance is neither a necessary nor a sufficient criterion for biological relevance (Nakagawa and Cuthill, 2007) but is typically used as a filter to remove uninteresting features where the observed difference can reasonably be attributed to being a sampling artifact. Reasoning about the biological relevance of a feature requires consideration of effect sizes and their associated confidence intervals. Interpretation of statistical results can also benefit from transforming raw P-values to alternative measures with possibly superior interpretations (Storey and Tibshirani, 2003) and by allowing interactive filtering that permits focusing on features with specific statistical properties.

Here, we present a discussion of the key statistical concerns in metagenomic data analysis and introduce a new application, STAMP (STatistical Analysis of Metagenomic Profiles), which provides a user-friendly graphical environment for performing the statistical techniques discussed in this article.

*To whom correspondence should be addressed.

Table 1. Contingency table summarizing data for a feature of interest

	Sample 1	Sample 2	
Sequences in feature	x_1	x_2	$R_1 = x_1 + x_2$
Sequences in other features	y_1	y_2	$R_2 = y_1 + y_2$
Total assigned sequences	$C_1 = x_1 + y_1$	$C_2 = x_2 + y_2$	$N = C_1 + C_2$

2 METHODS

2.1 Input data

The statistical methods and software discussed here can be applied to any count data obtained from a pair of metagenomic samples. Typically, we are interested in a collection of related features which define a profile (e.g. a functional profile indicating the number of sequences assigned to different biological subsystems or pathways). The statistical assessment of a particular feature can be carried out using a contingency table as shown in Table 1. The table entries x_1 and x_2 are the number of sequences in the two samples assigned to the feature of interest, while y_1 and y_2 are the numbers assigned to other features. The total number of sequences in the profile is given by the column sums C_1 and C_2. For hierarchical classification schemes, such as a taxonomy or the SEED functional assignments (Overbeek *et al.*, 2005), we can also investigate the number of sequences within a feature relative to the total number of sequences assigned to a parental category in the hierarchy. In this case, the column sums express the number of sequences assigned to the parental category rather than the total number of sequences in the sample.

2.2 Statistical hypothesis tests

The *P*-value produced by a statistical hypothesis test indicates the probability of an observed difference occurring simply by chance. Features in a profile with *P*-values below a nominally chosen threshold (e.g. 0.05) are termed statistically significant and can reasonably be assumed to be enriched in one of the metagenomes due to ecological or taxonomic differences as opposed to being the result of a sampling artifact. Rodriguez-Brito *et al.* (2006) recently addressed the need to assess statistical significance by introducing a non-parametric, bootstrap test, XIPE-TOTEC. This test has been applied in several recent studies (Brulc *et al.*, 2009; Dinsdale *et al.*, 2008b; Edwards *et al.*, 2006; Mou *et al.*, 2008; Poretsky *et al.*, 2009; Qu *et al.*, 2008; Turnbaugh *et al.*, 2006; Turnbaugh *et al.*, 2009; Urich *et al.*, 2008; Willner *et al.*, 2009).

XIPE-TOTEC builds a null distribution for a given feature by drawing sequences randomly with replacement from a set consisting of all sequences from the two metagenomes. Two samples of M sequences are drawn from this pooled set, and the difference in the number of sequences from the feature of interest used as a test statistic. This process is repeated multiple times in order to estimate the null distribution accurately. The difficulty with this approach is that M is a free parameter. Increasing M reduces the width of the null distribution which in turn increases the number of features identified as being significant (Supplementary Table S1). A similar approach with the same requirement was recently proposed by Allen *et al.* (2009).

To assess the probability of any observed difference being a sampling artifact, we must consider the number of sequences originally obtained from each sample (i.e. C_1 and C_2). Several 'classical' tests meet this requirement and can be classified according to whether they assume sampling with or without replacement. Although these tests aim to produce the same results, in practice there can be considerable variation in the *P*-values they produce (Supplementary Tables S2, S3 and S4).

2.2.1 Sampling without replacement Monte Carlo permutation tests are a widely used non-parametric technique for modeling the distribution of a test statistic under a given null hypothesis [see Manly (2007) for a comprehensive treatment]. The null distribution is approximated by calculating values of the test statistic under a sufficiently large number of random permutations of the sample labels. A *P*-value can then be calculated as the proportion of this approximate null distribution that is equal to or more extreme than the observed data.

To exactly model the null distribution, all possible permutations must be considered. A permutation of the sequences within a pair of metagenomic samples can be viewed as drawing sequences without replacement from a finite population consisting of two types of sequences (i.e. those from the current feature of interest and those from other features). This is the definition of a hypergeometric distribution. Fisher's exact test uses this distribution to efficiently calculate the exact *P*-value without having to exhaustively enumerate all permutations (Agresti, 1990).

The chi-square test and G-test are well-known large sample approximations to Fisher's exact test (Agresti, 1990). Although these approximations are accurate for equal sample sizes, for unequal sample sizes they can produce *P*-values substantially smaller than those given by Fisher's exact test (Supplementary Tables S2–S4). Yates' continuity correction is often recommended for these approximation methods and has the benefit of making them conservative at the expense of being less accurate (Supplementary Tables S2–S4). For contingency tables with 'small' entries, traditionally defined as 5 or 10, these approximations are not recommended (Supplementary Table S3; Cochran, 1952) and have been shown to produce unsatisfactory results under a range of other conditions (Agresti, 1992 and references within).

The execution time of Fisher's exact test calculated under the commonly used 'minimum likelihood' methodology (Supplementary Table S5) is a linear function of the number of sequences assigned to a feature in either sample (i.e. R_1). Our implementation of this approach takes ∼1 s to compute when the number of sequences assigned to a feature is 10 000 (Supplementary Fig. S1). As such, we recommend Fisher's exact test be used instead of a large sample approximation test because metagenomic profiles rarely contain features with more than a few thousand sequences, typically of unequal size and often have many features resulting in 'small' table entries. Rivals *et al.* (2007) also recommended the use of Fisher's exact test after performing a similar analysis of potential hypothesis tests for the enrichment or depletion of gene ontology (GO) categories.

2.2.2 Sampling with replacement The most appropriate method for calculating exact *P*-values for 2×2 contingency tables has long been debated (Barnard, 1947; Barnard, 1989; Haber, 1987; Ludbrook, 2008). Opponents of Fisher's exact test argue that the supposition of fixed row and column totals causes the test to be overly conservative (Agresti, 1990; Ludbrook, 2008; Mehta and Senchaudhuri, 2003). From the perspective of metagenomics, this amounts to assuming that if we resample our two communities, we would obtain the same number of sequences from each community as in our original dataset (i.e. C_1 and C_2 are fixed) *and* that the total number of sequences assigned to a feature across the two samples would be unchanged (i.e. R_1 is fixed).

We can relax fixing the row totals by performing a bootstrap test, where random samples are generated by sampling with replacement as in the method proposed by Rodriguez-Brito *et al.* (2006). Under this model, we assume that the number of sequences drawn for our two samples always remains the same, but the number of sequences assigned to each feature is free to vary. Drawing sequences with replacement from a finite population consisting of two types of sequences gives rise to the binomial distribution. For large sample sizes, we can approximate these binomial distributions as normal distributions, N_1 and N_2. The normal distribution that results from the difference between N_1 and N_2 forms the basis for the well-known 'difference between proportions' z-test. Despite the stark difference in their formulation, this test is equivalent to the chi-square test (Rivals *et al.*, 2007) and shares its limitations.

To exactly model the null distribution, when sampling with replacement, requires knowledge of the true proportion of sequences assigned to a feature, p, within the two microbial populations being considered. The bootstrap test *estimates* this population parameter as the proportion of sequences

Table 2. Effect size statistics commonly applied to 2 × 2 contingency tables

Effect size statistic	Equation
Difference between proportions	$\mathrm{DP} = p_1 - p_2$
Ratio of proportions	$\mathrm{RP} = p_1/p_2$
OR	$\mathrm{OR} = (x_1/y_1)/(x_2/y_2)$

$p_1 = x_1/C_1$, $p_2 = x_2/C_2$; RP is often referred to as relative risk.

sampled from the feature of interest (i.e. $p_{\mathrm{hat}} = R_1/N$). Alternatively, an exact test can be performed by setting p_{hat} to the value that maximizes our *P*-value as suggested by Barnard (1947). Unfortunately, such an approach is computationally prohibitive even for modest sample sizes despite efforts to optimize the method (Mato and Andres, 1997). As the bootstrap test relies on an estimate of *p*, it is not an exact test, and when this estimate is poor can produce *P*-values that are extremely liberal compared with Barnard's or Fisher's exact test (Supplementary Table S6).

Given that Barnard's test is computationally prohibitive for the majority of features in a typical metagenomic profile, we must decide between an approximation to Barnard's exact test (e.g. bootstrapping) and Fisher's exact test. Our recommendation is to use Fisher's exact test as it is generally conservative compared with Barnard's exact test (Supplementary Table S6; Mehta and Senchaudhuri, 2003), computationally tractable for metagenomic profiles and familiar to the majority of researchers.

2.3 Effect size

To assess if a feature is of biological relevance, we must consider the magnitude of the observed difference (i.e. an effect size statistic). An arbitrarily small effect can be statistically significant if the sample sizes are sufficiently large (Supplementary Fig. S2), so biological significance must be supported by effect size statistics as well as *P*-values.

Three common effect size statistics are given in Table 2 (Sistrom and Garvan, 2004). The most intuitive statistic is the difference in proportions (DPs) of sequences assigned to a given feature in the two samples. Interpreting the ratio of proportions (RPs) is also natural and provides complementary information to the DP. Consideration of multiple effect size statistics is often essential while assessing biological relevance as features can have a small (large) DP, but a large (small) RP. The odds ratio (OR) is widely used and has many desirable mathematical properties (Bland and Altman, 2000). However, it is often criticized as being difficult to interpret (Agrawal, 2005; Sackett *et al.*, 1996) and we recommend that the *RP* be preferred to the OR while interpreting and reporting results.

2.4 Confidence intervals

A confidence interval (CI) indicates the range of effect size values that have a specified probability of being compatible with the observed data. For example, a 95% CI gives a lower and upper bound in which the true effect size will be contained 19 times out of 20. Knowledge of these bounds is often an important aid in assessing biological relevance. Despite this, we are unaware of any comparative metagenomic studies that report effect size CIs (Supplementary Table S7).

There is a close relationship between *P*-values and CIs. A *P*-value indicates the probability of observing a given contingency table under the assumption that samples come from identical microbial communities. CIs make no such assumption. As such, a CI that encompasses the 'identity' effect size (e.g. $\mathrm{DP} = 0$ or $\mathrm{RP} = \mathrm{OR} = 1$) will have a *P*-value >1 minus the coverage of the CI (i.e. a *P*-value ≥0.05 for a 95% CI). If the 'identity' effect size is outside the CI, the *P*-value will be <1 minus the coverage of the CI. CIs are more informative than *P*-values, and many proponents of CIs have suggested they make *P*-values unnecessary (Nakagawa and Cuthill, 2007). Nonetheless, *P*-values are a useful summary statistic and provide a natural way to rank and filter results when multiple hypothesis tests are performed.

CIs can vary considerably with sample size (Supplementary Fig. S3; features 1a and 2a). Critically, CIs provide us with a mean to infer the biological relevance of a feature even when it is marginally statistically significant; features 1c and 2c demonstrate that the difference between a statistically significant and non-significant feature can be minimal in terms of effect size and CI bounds. The CI for example 2c indicates that an effect size of 0 is just as likely as the 'counter-null hypothesis' of a true difference between proportions of ∼8.5% (Rosenthal *et al.*, 2000). This cautions us against making dichotomous decisions about biological relevance based solely on how a *P*-value compares with a nominal level of significance (i.e. 0.05). Inferring biological relevance is best done in the context of all available information, much of which does not lend itself to numerical analysis of a univariate response, such as the *P*-value.

2.5 Multiple test correction

A typical metagenomic profile consists of several hundred features. When performing multiple hypothesis tests, it is useful to modify the *P*-values so that they reflect a particular interpretation. For example, if a profile contains 100 features, the number of features with a *P*-value <0.05 due to chance variation will in general be 5. If we wish to examine a list of features where the probability of observing one or more false positive is less than a specified probability, we can use a correction method that directly controls the family-wise error rate (FWER). Commonly applied FWER methods include Bonferroni, Holm-Bonferrnoi and Šidák (Abdi, 2007). Alternatively, during exploratory analysis, we may be willing to accept a specific percentage of false positives. This can be achieved using the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995) or the Storey FDR approach (Storey and Tibshirani, 2003). These approaches calculate a *q*-value for each feature, indicating the expected proportion of false positives within the set of features with a smaller *q*-value.

These approaches complement each other while performing an exploratory analysis. The list of significant features obtained when no multiple test correction method is applied gives us an initial global look at those features which may be differentially abundant between our samples. An FDR approach can be used to refine this initial list and makes the number of expected false positives explicit. Finally, an FWER technique can be applied to focus our attention to only those features where the observed enrichment or depletion is highly unlikely to be a sampling artifact.

3 IMPLEMENTATION

Here, we introduce our open source software package for performing STatistical Analyses of Metagenomic Profiles (STAMP). Our software provides a user-friendly graphical interface to allow for the easy adoption of the statistical methods discussed in this article.

3.1 Implementation details

STAMP is implemented in Python (http://www.python.org) and can be executed on all major platforms.

3.1.1 Input data STAMP can read the functional and taxonomic profiles produced by MG-RAST (Meyer *et al.*, 2008) and all of the 'abundance profiles' available at IMG/M (Markowitz *et al.*, 2008). Custom profiles can be specified in an accessible tab-separated values file format. We are currently developing other parsers for popular community resources such as the naïve Bayesian rRNA classifier at RDP (Cole *et al.*, 2009).

3.1.2 Statistical hypothesis tests We provide an optimized implementation of Fisher's exact test. Although Fisher's exact test is preferred to asymptotic approximations, we have provided implementations of the chi-square and G-test (with and without Yates' continuity correction) for completeness. Based on the discussion in Section 2.2.2, some users may favor using the non-parametric bootstrap test. We also provide an

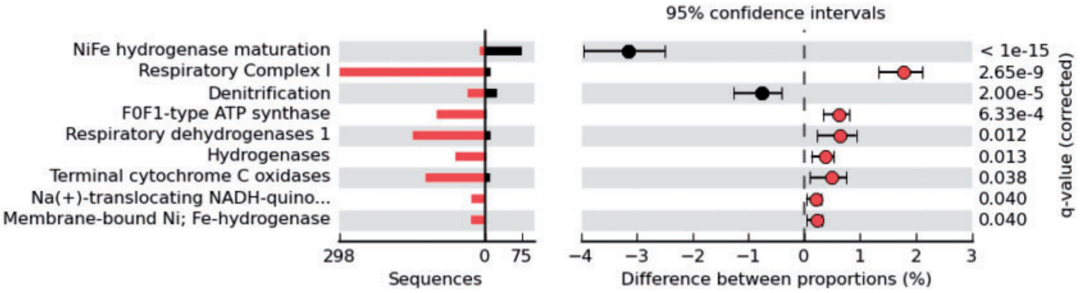

Fig. 1. Respiratory subsystems from the 'red' and 'black' iron mine metagenomes. Corrected P-values were calculated using Storey's FDR approach. Subsystems overrepresented in the 'red' ('black') community have a positive (negative) difference between proportions.

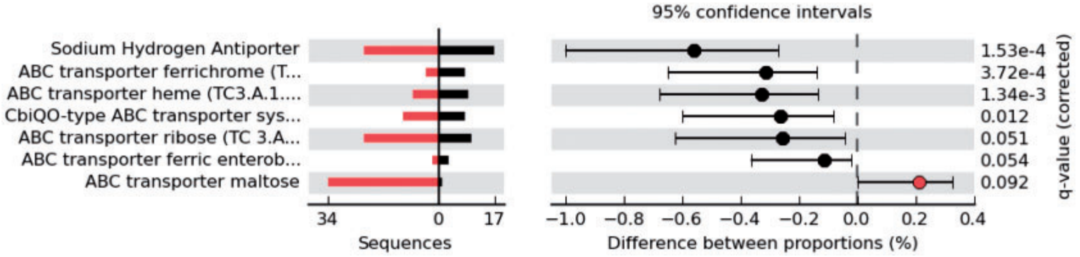

Fig. 2. Membrane transport subsystems from the 'red' and 'black' iron mine metagenomes. Corrected P-values were calculated using Storey's FDR approach. Subsystems overrepresented in the 'red' ('black') community have a positive (negative) difference between proportions.

implementation of Barnard's test for researchers considering profiles resulting in small tables (i.e. $N < 20$). Both one- and two-sided tests are supported, although for metagenomic studies two-sided results should generally be reported for the reasons given in Rivals *et al.* (2007).

3.1.3 Effect size and confidence intervals All the effect size statistics in Table 2 are available within STAMP. CIs for the DP statistic can be constructed using either a standard asymptotic approach, an asymptotic approach with continuity correction, or the Newcombe–Wilson method (Supplementary Table S8; Newcombe, 1998). Standard asymptotic CI approaches are implemented for the RP and OR statistics as these perform well even on tables with small entries (Supplementary Table S8; Agresti, 1999; Lawson, 2004). STAMP also provides a Monte Carlo simulation framework for evaluating the accuracy of a CI method on a particular dataset.

3.1.4 Multiple hypothesis test correction The FWER and FDR methods discussed in Section 2.5 are available within STAMP. Storey's FDR approach is based on the bootstrapping approach discussed in Storey *et al.* (2004).

3.1.5 Filtering of features Features can be filtered based on the number of sequences assigned to each sample or parental category, their observed effect size and their associated P-value. Specific subsets of features can be selected (e.g. all subsystems involved in respiration) and filtering optionally applied to them.

3.1.6 Plots Numerous publication-quality plots can be produced using STAMP. Bar and scatter plots, indicating the relative frequency of all features, permit an initial exploratory analysis of metagenomic profiles (e.g. Supplementary Figs S7 and S8). Extended error bar plots (e.g. Figs 1–3) provide a single figure indicting the number of sequences assigned to a feature along with the P-value, effect size and CI. Several additional plots are also available.

3.1.7 Extensible architecture STAMP uses a plug-in architecture in order to allow new statistical hypothesis tests, effect size statistics, CI methods, multiple comparison procedures or plots to be easily incorporated into the software.

3.1.8 Command-line interface A command-line interface is provided to facilitate batch processing or 'application linking' as recommended by Kumar and Dudley (2007).

3.1.9 Comparing multiple metagenomes Several recent studies have used XIPE-TOTEC to compare multiple metagenomes by performing pairwise tests between all possible pairs of metagenomes (Brulc *et al.*, 2009; Dinsdale *et al.*, 2008b; Turnbaugh *et al.*, 2009; Willner *et al.*, 2009). Such analyses can be performed using the STAMP command-line interface in a manner similar to XIPE-TOTEC or interactively through STAMP's graphical interface. We are currently extending STAMP with additional plots and exploratory tools specifically focused on multiple pairwise tests.

3.2 Comparison with available software

Here and in Supplementary Table S9, we summarize the differences among STAMP, XIPE-TOTEC (Rodriguez-Brito *et al.*, 2006), ShotgunFunctionalizeR (Kristiansson *et al.*, 2009), MEGAN (Mitra *et al.*, 2009) and IMG/M (Markowitz *et al.*, 2008). XIPE-TOTEC reports only statistically significant features without indicating their calculated P-value. ShotgunFunctionalizeR runs within the R statistical computing environment (http://www.r-project.org) and provides statistical tests for comparing groups consisting of multiple metagenomes. MEGAN is a cross-platform tool with a graphical interface that focuses primarily on taxonomic profiles. The taxonomic hierarchy is displayed as a tree structure, and statistical results are textually and graphically reported for each node in the hierarchy. IMG/M is a web portal that provides a number of tools for comparative metagenomics, including pairwise statistical hypothesis testing. Notably, none of these software packages reports effect sizes or CIs.

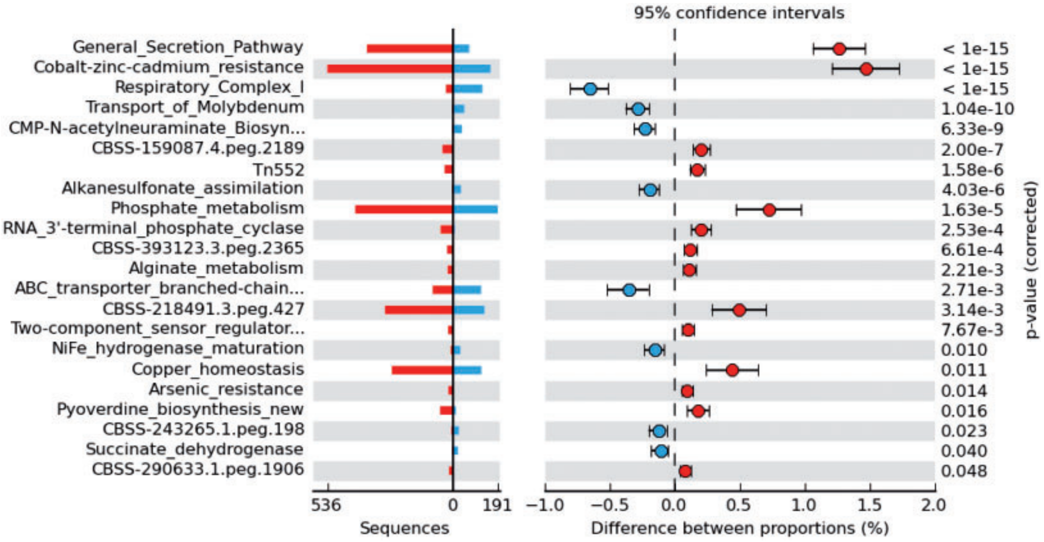

Fig. 3. Subsystems enriched or depleted within the *A.phosphatis* strains of the USA and Australia EBPR communities. Corrected *P*-values were calculated using the Bonferroni correction. Subsystems overrepresented in the USA (Australia) community have a positive (negative) difference between proportions and are indicated by red (blue) coloring.

4 RESULTS

4.1 Soudan iron mine metagenomes

Two samples from distinct habitats within the Soudan Iron Mine in Minnesota, USA, were sequenced and analyzed by Edwards *et al.* (2006). The ‘black’ sample (pH 6.7, redox potential -142 mV) was taken from water within a borehole, and the ‘red’ sample (pH 4.37, redox potential -8 mV) was taken a few centimeters from the mouth of a borehole where oxygen in the passageway has significantly reduced the pH and redox potential of the bore water. Functional profiles for these samples were obtained by Edwards *et al.* (2006) by comparing the unassembled reads obtained from pyrosequencing against the SEED database (Supplementary Methods).

Here, we contrast the statistical results reported by Edwards *et al.* (2006) using XIPE-TOTEC to those obtained using the statistical techniques provided in STAMP. Fisher's exact test identifies ∼11% fewer statistically significant subsystems than XIPE-TOTEC v2.4 with replicate sample sizes of $M = 5000$ (Table 3). This replicate sample size was used in Edwards *et al.* (2006), presumably as a compromise between the size of the two profiles that contain 2319 and 13 221 sequences in the ‘black’ and ‘red’ communities, respectively. Edwards *et al.* (2006) report only 69 statistically significant subsystems as it appears manual editing of subsystems containing few assigned sequences was performed.

Using the filters provided in STAMP, we determined that 15% of the statistically significant subsystems identified by Fisher's exact test have fewer than five sequences assigned to them from each sample (26% have 10 sequences or less). Performing additional filtering with liberal absolute ($\mathrm{DP} \leq 0.5\%$) and relative ($\mathrm{RP} \leq 2$) effect size requirements gives a list of 71 subsystems with sufficient statistical support to warrant further consideration of their biological relevance. Applying Storey's FDR approach with a q-value threshold of 0.05 reduces this list to 60 subsystems and makes it explicit that we should expect three of these to be false positives. Supplementary Figure S4 indicates the number of sequences in each of these 60 subsystems along with their effect size, CI and q-value. Further investigation into the biological relevance of these subsystems will benefit from considering this information.

Table 3. Identifying biologically interesting subsystems

	XIPE-TOTEC	STAMP
Total subsystems	247	
Statistically significant[a]	98	87
Manual filtering	69	–
With ≥5 sequence in at least one sample	NA	74
With RP ≥ 2 or DP ≥ 0.5%, and ≥5 sequences	NA	71
As above with Storey's FDR approach	NA	60

[a]Two-sided test with a significance level of $\alpha = 0.05$.

The metabolic analysis performed in Edwards *et al.* (2006) focused on the ‘respiratory’ and ‘iron uptake and utilization’ metabolic classes. Our analysis supports the hypothesis that the ‘red’ aerobic and ‘black’ anaerobic communities predominantly utilize different respiratory pathways (Fig. 1 and Supplementary Fig. S5). Edwards *et al.* (2006) also proposed that the ‘black’ community has a greater abundance of genes involved in iron uptake and utilization because ferric iron (Fe^{3+}) is limited in this community compared with the ‘red’ sample. Although this hypothesis is supported by the number of different subsystems that are overrepresented in the ‘black’ community (Fig. 2 and Supplementary Fig. S6), caution is warranted as the number of sequences assigned to these subsystems is extremely small.

Contrasting Figures 1 (Supplementary Fig. S5) and 2 (Supplementary Fig. S6) illustrate the benefit of reporting effect sizes and CIs. We should have more confidence in assigning biological relevance to subsystems such as ‘NiFe hydrogenase maturation’, which has a large DP and RP than to a subsystem such as ‘ABC transporter ferrichrome’ where the DP is small and the RP may be relatively small as indicated by the CI for this feature. In broader terms, the evidence for ecological differences causing preferential utilization of alternate respiratory pathways in these communities is far stronger than the evidence for differing concentrations of Fe^{3+} driving an overrepresentation of iron update and utilization genes.

4.2 *Accumulibacter phosphatis* strains in enhanced biological phosphorus removal metagenomes

Enhanced biological phosphorus removal (EBPR) is a treatment process in which micro-organisms are employed to remove excessive inorganic phosphate from wastewater which, if left untreated, would result in eutrophication of outfall ecosystems. Although the economic and environmental benefits of EBPR have led to its global adoption in waste water treatment plants (WWTP), the performance of these systems varies over time and location for reasons that are still poorly understood due to an incomplete understanding of EBPR microbiology. Here, we compare the functional profiles of *A.phoshatis* strains, the dominant EBPR phosphate-accumulating organisms, from two lab-scale WWTP in Australia and the US (García Martín *et al.*, 2006).

The *A.phosphatis* genes from these two communities were assigned to 26 functional classes (Supplementary Fig. S7) that contain a total of 491 SEED subsystems (Supplementary Fig. S8). Of these, 142 (29%) were assigned fewer than 10 sequences from either sample and were not considered further. The number of statistically significant features (P-value ≤ 0.05) identified using Fisher’s exact test was 116 (33%) and varied between 107 and 120 for the other statistical hypothesis tests considered (Supplementary Table S10). Application of Storey’s FDR and the Bonferroni’s FWER approaches reduces this to a list of 77 (22%) and 22 (6%) features, respectively (Supplementary Table S10).

Those features identified as statistically significant after applying the Bonferroni correction are given in Figure 3. While ‘phosphate metabolism’ is significantly overrepresented in the US community, it is not clear that this overrepresentation is related to the key polyphosphate metabolism processes in the EBPR community. Within this group, apolipoprotein N-acyltransferases showed the strongest overrepresentation in the US sample ($P = 8.15 \times 10^{-4}$). Other important systems include ‘general secretion pathway’, which covers a wide range of putative functions including exopolysaccharide biosynthesis and type II (including pilus assembly) and III secretion systems. Several categories of transport proteins related to metals were identified in each of the two samples; metadata about metal ion concentrations in these systems could reveal links between toxic metals and the systems needed to detoxify or exclude them. Overrepresentation of mobile elements (Tn552) and phage-associated proteins (CBSS-159087.4.peg.2189) may indicate endemic strains of phage and mobile elements (Kunin *et al.*, 2008).

5 DISCUSSION

It is essential to consider potential sources of error and how they can influence the results of statistical techniques when making inferences about biological relevance. Comparative metagenomics relies heavily on a ‘guilt by association’ paradigm where the function or taxonomic origin of reads is assigned according to similarity with sequencing from a reference databases. This leads to four notable sources of error in metagenomic profiles: (i) assuming sequence similarity implies functional similarity, (ii) database bias favoring certain functional subsystems or taxonomic units, (iii) misannotations resulting in reads being incorrectly associated with a specific functional or taxonomic category and (iv) reads with no reliable match in a database being effectively discarded. Investigating the seriousness of these sources of errors is an active area of research (Eisen, 1998; Friedberg, 2006; Schnoes *et al.*, 2009). Knowledge of effect size, CI width and number of sequences assigned to a feature is essential when deciding if these sources of error have the potential to explain the observed enrichment or depletion of a statistically significant subsystem or taxonomic unit.

For statistical techniques to aid biological inference, they must be interpreted and reported correctly. STAMP provides a graphical environment for performing statistical analyses and interactively exploring results through publication quality plots with sufficient information to inferring biological relevance. These capabilities make STAMP a valuable tool that will aid researchers in interpreting and communicating the results of their statistical analyses.

ACKNOWLEDGEMENTS

We would like to thank Dennis Wong for insightful discussions and providing us with the EBPR reads from *Candidatus* A.phosphatis, and Michael Porter for help with the Macintosh version of STAMP.

Funding: Killam Trusts (to D.H.P.). R.G.B. acknowledges the support of Genome Atlantic and the Canada Research Chairs program.

Conflict of Interest: none declared.

REFERENCES

Abdi,H. (2007) *Encyclopedia of Measurement and Statistics.* Sage, Thousand Oaks, CA.

Agrawal,D. (2005) Inappropriate interpretation of the odds ratio: oddly not that uncommon. *Pediatrics*, **116**, 1612–1613.

Agresti,A. (1990) *Categorical Data Analysis.* Wiley, New York.

Agresti,A. (1992) A survey of exact inference for contingency tables. *Stat. Sci.*, **7**, 131–153.

Agresti,A. (1999) On logit confidence intervals for the odds ratio with small samples. *Biometrics*, **55**, 597–602.

Allen,M.A. *et al.* (2009) The genome sequence of the psychrophilic archaeon, Methanococcoides burtonii: the role of genome evolution in cold adaptation. *ISME J.*, **3**, 1012–1035.

Barnard,G.A. (1947) Significance tests for 2 x 2 tables. *Biometrika*, **34**, 123–138.

Barnard,G.A. (1989) On alleged gains in power from lower P-values. *Stat. Med.*, **8**, 1469–1477.

Bland,J.M. and Altman,D.G. (2000) The odds ratio. *BMJ*, **320**, 1468.

Béjà,O. *et al.* (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, **289**, 1902–1906.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B*, **57**, 289–300.

Brulc,J. *et al.* (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc. Natl Acad. Sci. USA*, **106**, 1948–1953.

Cochran,W.G. (1952) The chi-square test of goodness of fit. *Ann. Math. Stat.*, **23**, 315–345.

Cole,J.R. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.

Dinsdale,E. *et al.* (2008a) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.

Dinsdale,E. *et al.* (2008b) Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE*, **3**, e1584.

Edwards,R. *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**, 57.

Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.

Friedberg,I. (2006) Automated protein function prediction–the genomic challenge. *Brief. Bioinform.*, **7**, 225–242.

García,M.H. *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.*, **24**, 1263–1269.

Gianoulis,T.A. *et al.* (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl Acad. Sci. USA*, **106**, 1374–1379.

Haber,M.A. (1987) A comparison of some conditional and unconditional exact tests for 2 × 2 contingency tables. *Commun. Stat. Simul.*, **16**, 999–1013.

Hallam,S.J. *et al.* (2004) Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science*, **305**, 1457–1462.

Hamady,M. and Knight,R. (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.*, **19**, 1141–1152.

Kanehisa,M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kristiansson,E. *et al.* (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics*, **25**, 2737–2738.

Kumar,S. and Dudley,J. (2007) Bioinformatics software for biologists in the genomics era. *Bioinformatics*, **23**, 1713–1717.

Kunin,V. *et al*. (2008) A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res.*, **18**, 293–297.

Lawson,R. (2004) Small sample confidence intervals for odds ratio. *Commun. Stat. Simul.*, **33**, 1095–1113.

Ley,R.E. *et al.* (2006) Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.

Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.

Ludbrook,J. (2008) Analysis of 2 x 2 tables of frequencies: matching test to experimental design. *Int. J. Epidemiol.*, **37**, 1430–1435.

Manly,B.F.J. (2007) *Manly: Randomization, bootstrap and Monte Carlo methods in biology*, 3rd edn. Chapman & Hall/CRC, Taylor & Francis Group 6000 Broken Sound Parkway NW, Suite 3000 Boca Raton, FL, p. 455.

Markowitz,V.M. *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.*, **36**, D534–D538.

Mato,A.S. and Andres,A.M. (1997) Simplifying the calculation of the P-value for Barnard's test and its derivatives. *Stat. Comput.*, **7**, 137–143.

Mehta,C.R. and Senchaudhuri,P. (2003) Conditional versus unconditional exact tests for comparing two binomials. http://www.cytel.com/papers/twobinomials.pdf.

Meyer,F. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Mitra,S. *et al*. (2009) Visual and statistical comparison of metagenomes. *Bioinformatics*, **25**, 1849–1855.

Mou,X. *et al.* (2008) Bacterial carbon processing by generalist species in the coastal ocean. *Nature*, **451**, 708–711.

Nakagawa,S. and Cuthill,I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev. Camb. Phil. Soc.*, **82**, 591–605.

Newcombe,R.G. (1998) Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat. Med.*, **17**, 873–890.

Overbeek,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5891–5702.

Poretsky,R.S. *et al.* (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Env. Microbiol.*, **11**, 1358–1375.

Qu,A. *et al.* (2008) Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. *PLoS ONE*, **3**, e2945.

Rivals,I. *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Rodriguez-Brito,B. *et al.* (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics*, **7**, 162.

Rosenthal,R. *et al.* (2000) *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach.* Cambridge University Press, Cambridge.

Sackett,D.L. *et al.* (1996) Down with odds ratios! *Evidence-based Med.*, **1**, 164–166.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Schnoes,A.M. *et al.* (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamiles. *PLoS Comput. Biol.*, **5**, e1000605.

Sistrom,C.L. and Garvan,C.W. (2004) Proportions, odds, and risk. *Radiology*, **230**, 12–19.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Storey,J.D. *et al.* (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Royal Stat. Soc. B*, **66**, 187–205.

Tringe,S.G. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.

Turnbaugh,P.J. *et al.* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Urich,T. *et al.* (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE*, **3**, e2527.

White,J. R. *et al.* (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.*, **5**, e1000352.

Willner,D. *et al.* (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*, **4**, e7370.

Yooseph,S. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.*, **5**, e16.