Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond,[1,2,6,9] Samir Wadhawan,[3,6,7] Francesca Chiaromonte,[4] Guruprasad Ananda,[1,3] Wen-Yu Chung,[1,3,8] James Taylor,[1,5,9] Anton Nekrutenko,[1,3,9] and The Galaxy Team[1]

[1]http://galaxyproject.org; [2]Division of Infectious Diseases, Division of Biomedical Informatics, School of Medicine University of California San Diego, San Diego, California 92103, USA; [3]Huck Institute for the Life Sciences, Penn State University, University Park, Pennsylvania 16803, USA; [4]Department of Statistics, Penn State University, University Park, Pennsylvania 16803, USA; [5]Departments of Biology and Mathematics & Computer Science, Emory University, Atlanta, Georgia 30322, USA

How many species inhabit our immediate surroundings? A straightforward collection technique suitable for answering this question is known to anyone who has ever driven a car at highway speeds. The windshield of a moving vehicle is subjected to numerous insect strikes and can be used as a collection device for representative sampling. Unfortunately the analysis of biological material collected in that manner, as with most metagenomic studies, proves to be rather demanding due to the large number of required tools and considerable computational infrastructure. In this study, we use organic matter collected by a moving vehicle to design and test a comprehensive pipeline for phylogenetic profiling of metagenomic samples that includes all steps from processing and quality control of data generated by next-generation sequencing technologies to statistical analyses and data visualization. To the best of our knowledge, this is also the first publication that features a live online supplement providing access to exact analyses and workflows used in the article.

[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aunl/p/windshield-splatter.]

Metagenomics is often thought of as an exclusively microbial enterprise, as one of the field's seminal papers was titled "Metagenomics: application of genomics to uncultured microorganisms" (Handelsman 2004). Because we simply do not know the number of bacterial taxa, the major motivation behind metagenomic studies was the need to estimate the biodiversity of various environments by direct sampling of potentially unculturable organisms (Beja et al. 2000, 2001; Tyson et al. 2004; Venter et al. 2004; DeLong 2005; Tringe et al. 2005; Gill et al. 2006; Poinar et al. 2006; von Mering et al. 2007). However, our understanding of eukaryotic diversity may not be much more advanced. Although the number of distinct eukaryotic (and, in particular, insect) taxa is likely far below microbial, the existing confusion about the species number is as striking. For example, Erwin (1982) obtained an estimate of 30 million insect species via extrapolation. This figure was fiercely debated, and the latest calculations converge on an educated guess on the order of 10 million (May 1988; Erwin 1991; Mayr 1998; Odegaard 2000). If we assume that these estimates are correct, then only a minute number of insect species have been described to date. For example, as of February 2009 the taxonomy database at the National Center for Biotechnology Information (NCBI) lists 318,068 species from all branches of life. In this study we apply existing metagenomic methodologies to directly determine the taxonomic composition of biological matter collected by the front end of a moving vehicle. Although our specimen collection strategy is straightforward, we set ourselves the nontrivial task of taxonomic identification of collected species. Because morphological identification is precluded by the destructive nature of the collection procedure, only DNA sequence analysis is feasible making this study de facto metagenomic.

Metagenomic methodology has been evolving rapidly in the past 5 yr, and now includes a diverse array of approaches for profiling (binning) of complex samples (for excellent reviews, see McHardy and Rigoutsos 2007; Raes et al. 2007; Kunin et al. 2008; Pop and Salzberg 2008). Classification procedures make use of multiple sequence features including GC content (Foerstner et al. 2005), oligonucleotide composition (McHardy et al. 2007; McHardy and Rigoutsos 2007; Chatterji et al. 2008), and codon usage bias (Noguchi et al. 2006). Homology-based methods compare sequence reads against existing protein markers (Baldauf et al. 2000; Ludwig and Klenk 2001; Rusch et al. 2007; Wu and Eisen 2008) or genomic data (Angly et al. 2006; DeLong et al. 2006; Poinar et al. 2006; Huson et al. 2007). For our study (a eukaryotic metagenome survey), a homology-based approach is more suitable, as we do not expect compositional properties (i.e., GC content) to be informative for, say, a particular family of insects. In addition, because we expect high taxonomic complexity within our samples, the coverage of individual eukaryotic genomes will likely be small, rendering protein (gene)-based approaches useless. Hence our best chance for successful phylogenetic profiling of windshield samples is the approach used by Poinar et al. (2006) and Huson et al. (2007), which relies on the comparison of metagenomic reads against existing sequence databases.

[6]These authors contributed equally to this work.
Present addresses: [7]Department of Genetics, University of Pennsylvania Medical School, 415 Curie Blvd., Philadelphia, PA 19104, USA; [8]Cold Spring Harbor Laboratory, One Bungtown Rd., Cold Spring Harbor, NY 11724, USA.
[9]Corresponding authors.
E-mail spond@ucsd.edu; fax (619) 543-5094.
E-mail james.taylor@emory.edu; fax (404) 727-2880.
E-mail anton@bx.psu.edu; fax (814) 863-6699.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.094508.109. Freely available online through the *Genome Research* Open Access option.

19:2144–2153 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

Because metagenomics is such a recent addition to life sciences, a well-designed software solution implementing the aforementioned methodologies is lacking, rendering metagenomic analyses too cumbersome for experimental biologists to perform. As an example, consider homology-based binning approaches exemplified by Poinar et al. (2006). While well-engineered systems such as CAMERA (Seshadri et al. 2007) or MG-RAST (Meyer et al. 2008) provide a powerful computational infrastructure, and visualization tools such as MEGAN (Huson et al. 2007) allow researchers to analyze the phylogenetic makeup of their metagenomic samples, metagenomic studies remain quite challenging. Indeed, homology searches (provided, e.g., by CAMERA) or taxonomic visualization are just two components of a multistep metagenomic pipeline. As an example, suppose a researcher has generated a collection of short sequencing reads from two metagenomic samples. He or she wants to identify the taxonomic representation of the reads and contrast species abundance between the two samples. The starting point of this analysis is a collection of sequencing reads and associated base quality scores. Next, the researcher would like to do the following:

1. Evaluate the quality of sequencing reads and select high-quality segments;
2. Search for high-scoring hits in existing databases;
3. Assign taxonomic labels to sequencing reads based on their database matches;
4. Visualize the taxonomic composition of metagenomic samples; and
5. Perform a comparison of taxonomic composition between the two samples.

Although this example outlines just one of many possible metagenomic analyses, even in this simplified case only a few steps can be performed with a collection of disjoint resources. While excellent software tools for performing individual steps (CAMERA or MEGAN) exist and are freely available to the scientific community, we lack a truly integrated solution in which analyses can be easily concatenated, converted to workflows, shared among colleagues, and published in a readily reproducible form. Returning to our example analysis: Step 1 presents a great difficulty for most experimentalists as sequence and quality files are extremely large and exist in many different formats. Step 2 requires a powerful computational infrastructure that allows very large sequence data sets to be searched against even larger databases. CAMERA and MG-RAST provide a public BLAST search and annotation service, enabling large-scale comparisons against a predefined set of databases. Steps 3 and 4 can be performed with MEGAN. But because MEGAN is distributed as a standalone package, it may be challenging to process the results of large BLAST searches on a desktop computer. Step 5 may be performed with a spreadsheet application (provided no data set exceeds the upper row limit for popular spreadsheet applications). The comparison between samples will require novel statistical approaches such as tag counting (Robinson and Smyth 2007; Marioni et al. 2008) that are yet to be implemented in biologist-friendly applications.

We set out to address these challenges by implementing a homology-based workflow for the analysis of metagenomic samples. As our example data set, we used sequencing reads generated by the 454 Life Sciences (Roche) FLX instrument using DNA obtained from organic matter collected by the front-end (windshield and bumper) of a moving vehicle from two geographic locations (see Methods). First, we built a complete pipeline, in which a user uploads reads generated by the sequencing machine (alternatively, reads can be obtained directly from the sequencer), performs quality control (QC), generates alignments, and conducts full taxonomic representation analysis entirely within a web browser. Because we designed our system using our existing Galaxy platform (http://galaxyproject.org), analyses described here can be easily shared among colleagues or referenced in supplementary materials to publications in a way that is completely transparent and reproducible. Next, we use the pipeline to answer the following two questions: (1) Is modern technology combined with available sequence data sufficient to identify eukaryotic taxa from low coverage random sequence samples? (2) Can "eukaryotic metagenomics" be used to contrast the species composition of distinct geographic locations?

Results and Discussion

A metagenomic toolkit and comparison with existing tools

Galaxy (http://galaxyproject.org) contains close to 100 tools enabling users to perform diverse tasks from simple text manipulation to analysis of next-generation sequencing data. As most metagenomic projects begin by generating massive numbers of sequencing reads, Galaxy provides a convenient environment for metagenomic tools where the user can take advantage of the already existing comprehensive functionality. The toolkit we designed contains six components described in Table 1. A casual user can apply this toolkit to very large data sets without leaving his or her web browser. The best way to describe the functionality of our toolkit is to apply it to a well-known, published data set and provide a step-by-step explanation, which can be found as a part of the on-line supplement to this manuscript. In that way we not only offer a detailed description of our implementation but also prove that it performs as expected. As our demonstration data set we have chosen the sample 1 and sample 2–4 data sets from Huson et al. (2007), constituting a metagenomic survey of the Sargasso Sea (Venter et al. 2004). This choice also allowed us to compare our approach with that of MEGAN, which, until now, has been the single comprehensive software package for taxonomic profiling of sequencing reads. The results produced by our pipeline were nearly identical to those returned by MEGAN, suggesting that our approach works correctly (see the online supplement for detailed comparison and full explanation of parameters used and live supplement at http://usegalaxy.org/u/aun1/p/windshield-splatter). All analyses were conducted entirely within Galaxy without any external dependencies. The same applies to all tables and figures in this manuscript: No external software packages were used for statistical analyses or data visualization.

Nucleotide databases are useful for metagenomic analyses

We tested whether the phylogenetic makeup of metagenomic samples can be estimated from nucleotide sequences. There are several compelling reasons for doing this. First, the rate of growth of nucleotide databases, such as whole-genome shotgun (WGS) databases and trace archives, greatly surpasses that of protein databases. As a result, a comparison with nucleotide databases employs a much larger space of reference sequences that can be used for taxonomic labeling of metagenomic samples. Second, because metagenomic sampling is performed at the nucleotide level (i.e., sequencing reads are derived from genomic, plasmid, or other sources of nucleic acids), a comparison against protein databases may not always be appropriate. For example, metagenomic

Table 1. Metagenomic tools in Galaxy framework

Tool	Description
Fetch taxonomic representation	Fetches taxonomic information for a list of GI numbers (sequences identifiers used by the National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov).
Summarize taxonomy	Given taxonomy representation (produced by "fetch taxonomic ranks" described above). This utility computes a summary of all taxonomic ranks.
Draw phylogeny	Given taxonomy representation (produced by "fetch taxonomic ranks"). This tool produces a graphical representation of phylogenetic tree in PDF format (e.g., see Fig. 2).
Find diagnostic hits	Identifies sequence reads corresponding to a particular taxonomic group or, in other words, diagnostic of a particular taxonomic rank. It takes data generated by "fetch taxonomic ranks" as input and outputs for either a list of sequence reads unique to a particular taxonomic rank or a list of taxonomic ranks and the count of unique reads corresponding to each rank.
Find lowest diagnostic rank	Identifies the lowest taxonomic rank for which a mategenomic sequencing read is diagnostic. It takes data sets produced by "fetch taxonomic ranks" as the input.
Poisson two-sample text	Tests if the number of reads between two taxa is significantly different. Assumes that the data comes from a Poisson process and calculates two *Z*-scores (Z1 and Z2) based on the work by Huffman (1984), computes corresponding *P*-values, and performs multiple test correction.

For a full explanation of tool parameters and examples of input and output formats, see help pages within Galaxy at http://usegalaxy.org (click "use now" and then choose "Metagenomic analyses" within the left side of Galaxy interface).

sequences representing eukaryotic taxa (the focus of our windshield study) are derived mostly from genomic DNA, of which only a small fraction codes for proteins. Finally, nucleotide-level alignments are substantially faster to compute, resulting in more sequence space covered in a unit of time. To see how well nucleotide-level comparisons perform in practice, we repeated the Huson et al. (2007) analysis at the nucleotide level using MegaBLAST (with word size = 16 and all other parameters set to default values) (Zhang et al. 2000) to align the reads against the contents of NT and WGS databases obtained from the NCBI. For direct MEGAN to Galaxy comparison, see Supplemental material. For sample 1, this comparison yielded 639,122 alignments representing 8434 reads. For sample 2–4, the numbers were 292,511 and 5678, respectively. Next, we devised a strategy for purging spurious hits, as the 5% alignment score cutoff used by Huson et al. (2007) is only relevant for protein-level alignments and is not applicable here. First, we examined the properties of alignments by building distributions of alignment lengths and alignable fractions (the proportion of each read that aligns with a database entry). In Supplemental Figure S2, one can see that although the reads were long (median length ~800 bp), most alignments were short. We chose to be conservative and selected only those alignments where the alignable fraction was equal or greater to that of the Q_3 value (top 25%) of the corresponding distribution (0.39 and 0.25 for samples 1 and 2–4, respectively). This filter decreased the number of hits significantly, to 161,556 and 75,567 for samples 1 and 2–4, respectively. From the remaining hits we identified 5847 and 1923 reads that were diagnostic below the Kingdom level. Although this approach is clearly less sensitive (yet more specific) than the protein-level comparison described in the Supplemental material and produced far fewer taxonomically labeled reads, its performance was essentially identical vis-à-vis capturing the phylogenetic makeup of samples 1 and 2–4. Supplemental Figure S3 compares phylogenetic profiles up to the family rank for Gammaproteobacteria reads obtained from nucleotide-level comparisons. Protein-level analysis (Supplemental Fig. S1) of sample 1 identified 22 families of which the Shewanellaceae, Preudomonadaceae, Aeromonadaceae, and Enterobacteriacaea were the most prominent with 1807, 172, 130, and 108 reads, respectively, accounting for >80% of all Gammaproteobacteria reads. The same families are also overrepresented in the nucleotide-level analysis. Notably they account for ~95% of all Gammaproteobacteria, implying that the stringent cutoffs used by us "amplified" the signal by removing ranks with a small number of reads. The same trend is apparent for the sample 2–4 reads (cf. Supplemental Figs. S1 and S3). This experiment therefore suggests that the nucleotide analysis is appropriate for the recovery of the phylogenetic makeup of metagenomic samples, allowing us to use much larger databases with only a moderate increase in computational cost since nucleotide-level comparisons can be performed much faster than protein-level comparisons. Again, this is especially relevant for identifying eukaryotic reads as they have only a minute probability of representing protein-coding regions.

Analysis of the windshield splatter

After ensuring that our approach performed correctly on previously analyzed samples and showing that nucleotide sequence databases can be used for phylogenetic profiling, we turned to the analysis of the windshield data set, for which we were primarily interested in the identification of eukaryotic taxa. We started with two collections of 454 FLX reads representing trip A and trip B (Fig. 1; Supplemental Fig. S9). All data described in the manuscript are publicly available through the Galaxy Library System (see Methods).

The first step of the analysis is to assess the quality of sequencing reads. For 454 FLX data, one is primarily concerned with the overall base quality and the extent of read segmentation due to low quality base calls. The first QC task—overall quality assessment—was performed in Galaxy with the "build distribution of base quality" tool. This analysis showed that both data sets (trip A and B) were of suitable quality with the median *phred* value well above 20 (Supplemental Fig. S4). The second QC task—fragmentation analysis—was performed with the "select high-quality segments" tool. This tool allows the user to select read segments above a specified length that are not interrupted by bases of quality lower than a set threshold. Setting the minimal length of a continuous segment to 50 nucleotides (nt) and the minimum phred quality score to 20 produced only 82,917 (trip A) and 45,354 (trip B) contiguous segments. The length distributions for the remaining reads are shown in Supplemental Figure S5. The dramatic reduction in the number of reads (from 183,241 and 183,169, respectively), suggests that low-quality base calls are very common in sequence reads produced by the 454 base caller. Indeed, bases immediately following homopolymer repeats are known to have high error rates for the 454 platform (Margulies

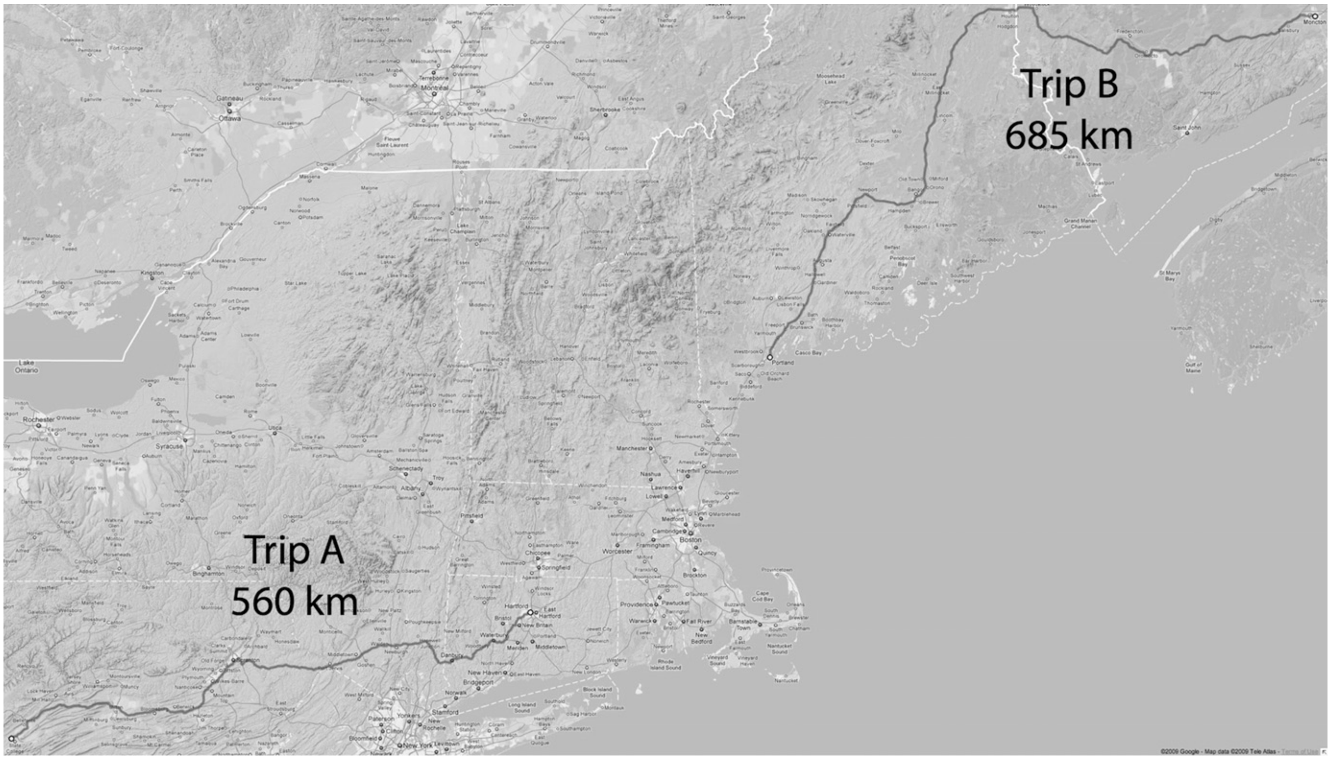

Figure 1. Map of the collection routes. For precipitation and temperature data, see Supplemental Figure 9.

et al. 2005), and efforts have been undertaken to improve error correction for these bases in the software (Brockman et al. 2008). Correspondingly, we have implemented the "DO NOT trigger splitting on homopolymer runs" option in the "select high-quality segments" tool. When this option is selected, the tool identifies low-quality bases immediately after homopolymer runs but does not split reads on them. Running the tool with these settings returned 221,422 (median length, 148 nt) and 212,634 (median length, 127) fragments for the trip A and B data sets, respectively. We used these data sets to perform downstream analyses.

Next, we compared the above trimmed reads with the NT (1,802,654,011 bases) and WGS (102,335,401,932 bases) directly in Galaxy using its MegaBLAST (Zhang et al. 2000) tool with word size set to 16. This analysis yielded 7,488,766 and 3,852,789 hits to 135,317 and 71,730 read fragments from trip A and B, respectively (for overview of alignment length, identity, and alignability distributions, see Supplemental Fig. S6). As with the previously published analysis of the Sargasso Sea data sets, we chose to limit hits to those with alignability exceeding the third quartile in the corresponding distribution (for numeric values, see Supplemental Fig. S6). This conservative filtering restricted the number of hits to 1,873,238 and 920,183, representing 78,219 and 35,419 reads from trips A and B, respectively. From these, 70,546 and 32,966 reads were diagnostic below the Kingdom level. Figure 2 shows a class level overview of trip A and trip B samples. The most prominent difference between the two trips is in the number of reads identified with green plants (Viridiplantae): 10,242 in trip A versus 612 in trip B. It is unlikely that a two orders of magnitude difference reflects a genuine variation in species abundance of such a ubiquitous taxonomic group between the two trips. Because during each trip we collected two samples (left and right sides of the vehicle; see Methods) we were able to trace the majority (9317) of Viridiplantae reads to the left subsample. The most likely explanation for this overabundance is that a piece of plant material (e.g., a leaf or stem fragment) adhered to the collection surface. Aside from green plants, two insect groups, Diptera and Hemiptera, represented the majority of eukaryotic reads in both samples.

The number of sequencing reads can be thought of as a proxy for the relative abundance of taxa in our experiment and therefore can be used to contrast biodiversity estimates between geographic locations. To judge the significance of the difference, we used the approximate two-sample Poisson test (Huffman 1984). Because each set of sequencing reads derived from an insect or a pollen particle colliding with a moving vehicle represents an independent event (much like a phone call arriving to a switchboard or a web page request made to a web server), it can be modeled as a Poisson process. Importantly, the test allows us to correct for differences in read counts between the two trips that result from sample processing and preparation (e.g., differences in DNA concentration between samples). To implement this correction (via a correction factor *d*), one needs to select those species that are expected to be represented equally in compared samples (e.g., are expected to inhabit both collection locations). Because most reads from both trips map to bacterial species, a prokaryotic taxon or group of taxa would be the most appropriate candidate for implementing such a correction. Fierer et al. (2008) report relative abundances of airborne bacteria. Gammaproteobacteria is one of the most prominent groups reported by these investigators (although relative abundance varied significantly across samples) and is the top bacterial group in our analysis (Fig. 2). To select taxa suitable for the calculation of *d*, we traversed the Gammaproteobacteria tree further to calculate numbers of reads and their ratio between the two

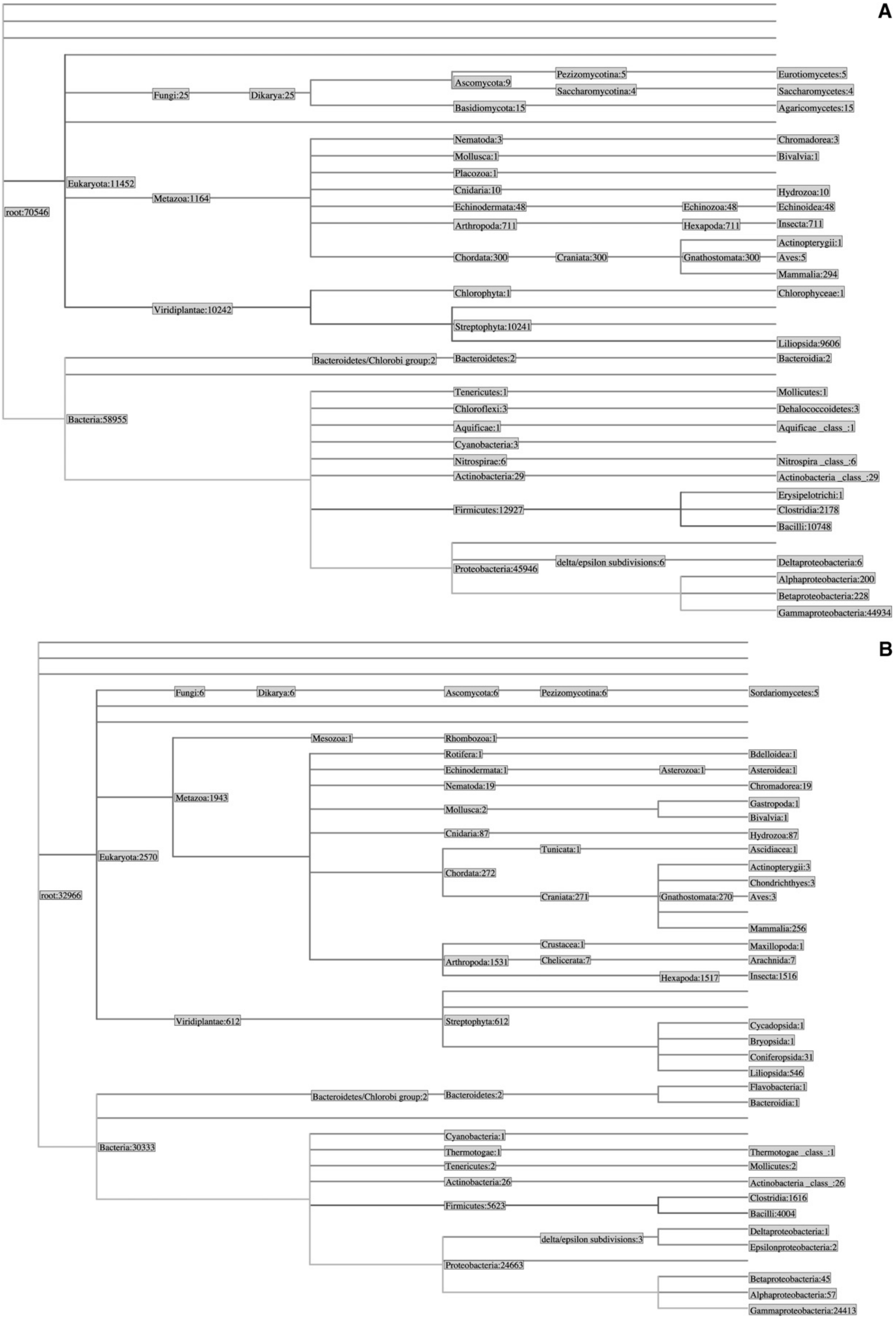

Figure 2. A class-level phylogenetic profile of windshield splatter material. (*A*) Trip A; (*B*) trip B.

trips for individual genera (only taxa for which one of the samples contains at least 50 reads were chosen) (Supplemental Table 1) and selected the median B/A value of 0.32 as *d* for subsequent analyses. We then used this value to perform the approximate two-sample Poisson test described above with the false discovery rate (FDR) correction to generate a list of taxonomic groups with significantly (at 1% level, see Methods) different read abundances between trips A and B (Table 2). The list included unexpected entries such as the genus *Homo* even though the two trips were uneventful. Such matches are likely caused by road debris (which often includes roadkill) adhering to the collecting tape. This illustrates, at least at genus level, that the reliability of taxonomic identification is severely biased by the amount of sequence data available for a particular taxon. Because few entries in NT and WGS databases are derived from, say, white-tailed deer (*Odocoileus virginianus*, a prevalent large mammal roadkill in the northeastern United States), reads truly representing this species are more likely to match abundant human sequences. Similar taxonomic distortions may apply to the insect samples. The most abundant invertebrate genus was *Acyrthosiphon*, which includes the pea aphid *Acyrthosiphon pisum*. This genus provides a unique opportunity to test the reliability of our methodology as pea aphids require a bacterial endosymbiont, *Buchnera aphidicola*, for survival and reproduction (Baumann et al. 1995; Brisson and Stern 2006). Thus we would expect the count of reads identified with *B. aphidicola* to exhibit the same pattern as that observed with *A. pisum* reads. Indeed, the number of B. aphidicola reads is higher along trip B ($X = 9$, $Y = 59$) (Table 2) confirming this expectation.

Table 2. Taxa with significant (at 1% level) differences in read abundance between trip A and trip B

Rank	Name	Trip A	Trip B
Phylum	Arthropoda	711	1531
	Chordata	300	272
	Cnidaria	10	87
	Firmicutes	12,927	5623
	Proteobacteria	45,946	24,663
Class	Bacilli	10,748	4004
	Betaproteobacteria	228	45
	Clostridia	2178	1616
	Gammaproteobacteria	44,934	24,413
	Hydrozoa	10	87
	Insecta	711	1516
	Mammalia	294	256
Order	Aeromonadales	540	21
	Bacillales	83	58
	Clostridiales	2178	1615
	Diptera	296	350
	Enterobacteriales	41,174	23,729
	Hemiptera	383	1027
	Hydroida	10	87
	Lactobacillales	10,643	3943
	Primates	112	10
	Pseudomonadales	1792	408
	Rhodospirillales	56	1
Family	Aeromonadaceae	540	21
	Aphididae	382	1016
	Clostridiaceae	2170	1608
	Culicidae	86	64
	Drosophilidae	32	95
	Enterobacteriaceae	41,172	23,729
	Enterococcaceae	706	1512
	Hominidae	97	6
	Hydridae	10	87
	Lactobacillaceae	5837	209
	Leuconostocaceae	2978	1498
	Pseudomonadaceae	1703	391
	Streptococcaceae	928	545
Genus	*Acyrthosiphon*	381	995
	Aeromonas	540	21
	Anopheles	80	45
	Anopheles	80	1
	Buchnera	9	59
	Clostridium	2170	1607
	Drosophila	31	94
	Enterobacter	4142	5507
	Enterococcus	706	1511
	Erwinia	2	240
	Homo	96	4
	Hydra	10	87
	Klebsiella	15,169	1695
	Lactobacillus	5740	167
	Lactococcus	809	509
	Leuconostoc	2971	1496
	Photorhabdus	57	1
	Providencia	123	3
	Pseudomonas	1648	390
	Salmonella	4044	1870
	Serratia	3242	29
	Shigella	674	376
	Stenotrophomonas	93	9
	Yersinia	1258	196

A complete metagenomic pipeline (Screencasts 1 and 2)

The windshield sample analysis described above was performed entirely in Galaxy without using external tools. It can be distilled to a generic metagenomic analysis pipeline that contains the 16 steps outlined in Figure 3A (an exact ”click through” is shown in Screencast 1). In this example, we start with a typical output of a 454/Roche sequencer: a read file in FASTA format and a base quality file (data sets 1 and 2) (Supplemental Fig. 4A). Next, we select high-quality segments from each read (data set 3). A high-quality segment is defined as a run of nucleotides of length *L* where all bases have the *phred* quality value above *Q* (in this example $L = 50$ and $Q = 20$). We next rename the sequences and use a Galaxy MegaBLAST tool to compare the reads with NT and WGS databases (data sets 7 and 8). While MegaBLAST searches are running, we compute the lengths of sequences (data set 9). Results of the two MegaBLAST runs (NT and WGS) are merged (data set 10) and joined with data set 9, which contains sequence lengths. The resulting data set (11) now contains all columns from data sets 10 and 9, including alignment length (column 5) and sequencing read length (column 15). From this data set, it is simple to filter suboptimal hits by retaining those reads that satisfy the $c5/c15 \leq 0.5$ condition (value in column 5, alignment length divided by the value in column 15, read length) to data set 12. Here, we defined suboptimal hits as those in which an alignment between a 454 read and a database entry covers less than 50% of the read's length (of course, this should be adjusted appropriately in each case as, e.g., was done above). At this point we start a metagenomic analysis sensu stricto by fetching taxonomic information contained in the table of hits using the "fetch taxonomic representation tool" (data set 13). An application of the "find lowest diagnostic rank" tool produces data set 14, which contains a list of reads specific to ranks below Kingdom level. Finally, we build a phylogenetic tree from the list of ranks (data set 16) and tabulate a list of taxonomic groups it contains (data set 15).

What if the user would like to repeat this analysis using a different set of input data sets or different tool settings? The history shown in Figure 3A already contains all the information necessary to build and reproduce this analysis precisely. To convert the history into an executable workflow shown in Figure 3B, the

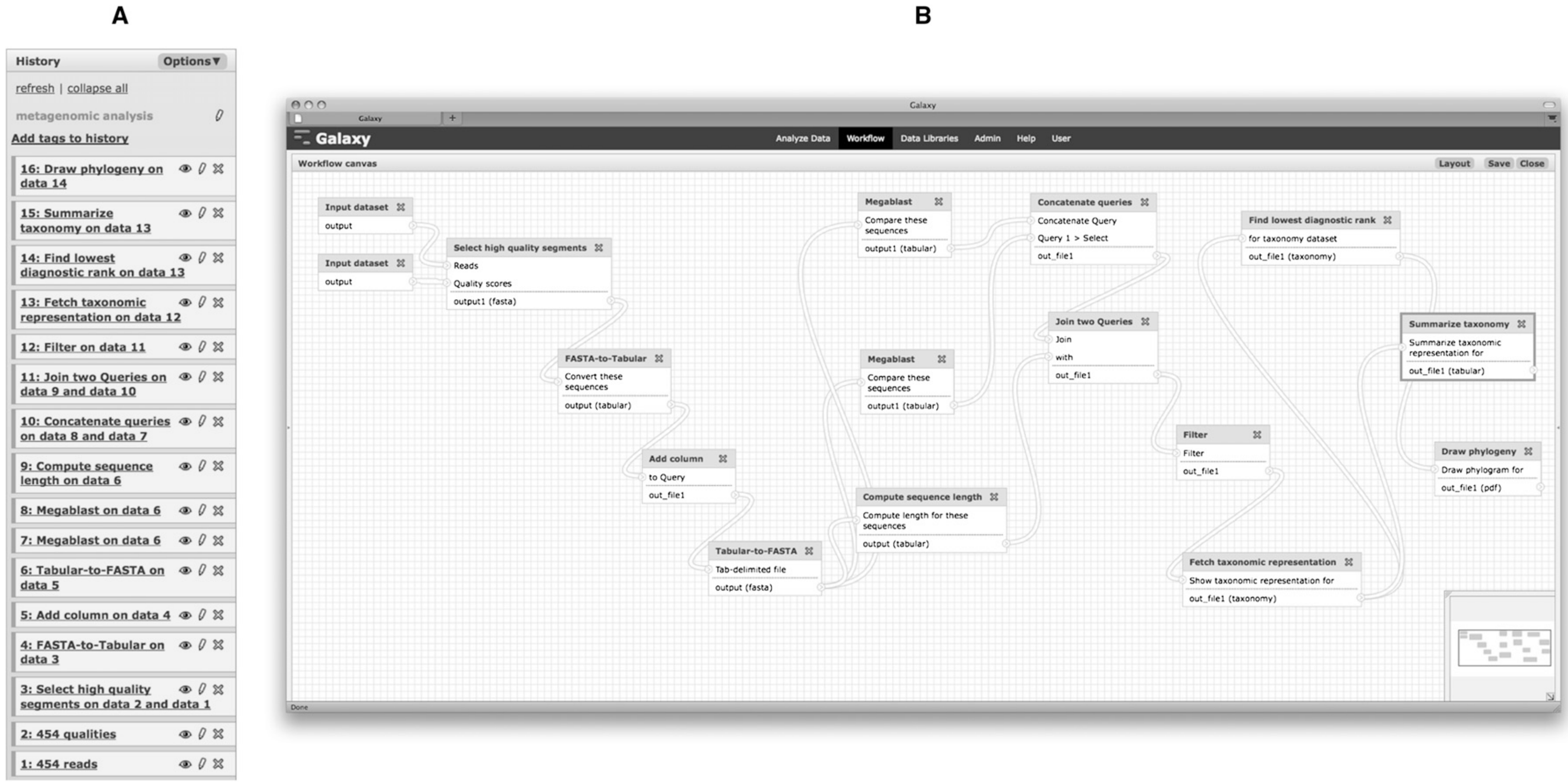

Figure 3. (*A*) Galaxy history pane showing all steps of a metagenomic analysis described in the study. (*B*) Workflow representation of analysis. Using workflow functionality, the user can rerun analyses in their entirety.

user simply needs to click the "construct workflow from the current history" link (see Supplemental Screencast 2). Once the workflow is created, it can be executed, edited, renamed, shared, and deleted. The running, renaming, and deletion options are self explanatory. The editing option allows the modification of all aspects of the workflow, such as all connection and tool parameters. For example, in the framework of this study a user may want to rerun the entire analysis with modified filtering conditions in step 3 and/or 12 or apply it to a different set of sequencing reads.

CAMERA and Galaxy

The CAMERA resource (Seshadri et al. 2007) provides an extensive computational infrastructure allowing users to compare their own metagenomic samples against a variety of very large data sets. Supplemental Figure S8 illustrates how a user can import results of his/her analysis into Galaxy and take advantage of the tools described above. Here we uploaded a set of 454 reads representing trip A into CAMERA, ran a MegaBLAST comparison against all metagenomic reads stored in CAMERA, and used the "export" feature to save the resulting MegaBLAST XML file (Supplemental Fig. S6A). We then uploaded the compressed XML file into Galaxy, processed it using the built-in BLAST-XML parser (Supplemental Fig. S6B), and then ran the Group tool to show that the 126,032 hits correspond to 123,241 Metagenomic reads. At this point the user can take advantage of any Galaxy tools to further process these data.

Conclusion

The explosive growth of metagenomics in the past 2–3 yr can be mainly attributed to the rapid proliferation of massively parallel sequencing technologies. This trend is almost certain to continue, and as a result, the future of metagenomics and of sequencing technologies are inextricably linked. Yet extant metagenomic applications frequently place the considerable burden of preprocessing raw metagenomic reads onto the user. For example, MEGAN is an excellent tool for phylogenetic typing of metagenomic reads, but to take advantage of it, the user must perform database searches elsewhere: a procedure that is prohibitively tedious and complex for most experimentalists. This is why our objective was to build a complete pipeline for homology-based taxonomic labeling of metagenomic reads that was self-contained and guided the user from data acquisition and QC, to database searches, and, finally, actual metagenomic analyses. We demonstrate that the classification performance of our solution is on par with currently available applications. At the same time, our software is distributed as open-source, can be downloaded and installed locally, and can be easily extended to develop specialized versions of our pipeline (all code base is available from http://galaxyproject.org).

At present, our approach has two main limitations. First, the set of reference databases is currently limited to NT and WGS nucleotide sequence collections from NCBI. As our computational infrastructure is expected to grow significantly during the current year, we will enable protein-level searches on nonredundant protein data sets as well as trace archives. Further, this is solely a limitation of the Galaxy instance maintained at the Nekrutenko laboratory, but not of the Galaxy framework, which supports several popular clustering solutions "out-of-the-box" and can be easily integrated with a powerful computational resources, eliminating this issue. Second, homology-based phylogenetic profiling is just one of the ways metagenomic analyses are performed. Extensibility of the Galaxy framework makes it straightforward to implement alternative annotation pipelines, such as AMPHORA (Wu and Eisen 2008) and others.

Our second goal was to perform a eukaryotic metagenomic study on the organic matter collected on an automobile's windshield. Specifically, we were interested in addressing two questions: Can one identify eukaryotic taxa from random reads generated by the next-generation sequencing technology from environmental samples? and Is it possible to contrast species abundance between geographic locations? While this pilot analysis provides positive answers to both questions, it also raises important issues and limitations. At present, it is difficult to perform species identification using random sequencing of environmental samples, as only a relatively small fraction of all reads (~8%) mapped unambiguously to eukaryotic taxa. This is not entirely surprising as species richness of our sample is expected to be high, and so every species would be represented by a small fraction of the reads. In addition, the present day sequencing coverage of eukaryotic species is patchy, with most nucleotides derived from a handful of model or medically-relevant organisms. For example, GenBank contains 5,737,786 insect sequences, of which 3,202,129 (or ~56%) belong to just two families: Culicidae (mosquitoes) and Drosophilidae (pomace flies). Consequently, the number of species identified in this study is likely grossly underestimated for the following reason. Consider species A, which has a completely sequenced genome, and a closely related species B, for which only a minimal amount of sequence data are available. Most reads derived from species B will match the sequenced genome from species A, creating an impression that there is no species B in the sample. Thus, despite astronomical rates of growth reported by major sequence repositories, we know very little about species diversity from a genomic standpoint. The dramatic drop in sequencing costs and increase in throughput will make thorough cataloging of biodiversity possible. We are optimistic that we will soon be able to sequence genomic DNA from most holotypes available from major museum collections. Having a user-friendly, extensible, and powerful computational framework in place is clearly necessary to take advantage of these data as they become available.

Methods

Data and tool access

All data sets and tools described in this manuscript are available at the test Galaxy server at http://usegalaxy.org. To access the data, point your browser to the Galaxy site, click "data libraries" in the upper pane of the interface, and look for "windshield splatter." To access the tools, click "metagenomic analyses." Clicking on individual tools will bring their interface into the center pane of the browser and will also provide a detailed description of each tool's functionality, including input and output formats and an explanation of options. These steps are highlighted in a supplemental screencast that can be viewed at Galaxy's site and downloaded as a part of our submission. A live supplement at http://usegalaxy.org/u/aun1/p/windshield-splatter provides access to analyses and workflows used in this manuscript.

Sample collection and DNA isolation

The front bumper of a 2006 Dodge Caravan ("The Wanderer") was divided at the license plate into "left" (passenger side) and "right" (driver side), and was taped with a double-sided carpet tape. On top of the carpet tape, a 3M 5414 Water Soluble Wave Solder Tape was affixed, exposing its sticky side. The tapes were applied on June 23, 2007, at 6 am EDT in State College, Pennsylvania, and removed in tubes containing Tris EDTA buffer at 12 pm EDT in Manchester, Connecticut. New tapes were again applied in Portland, Maine, at 5 pm EDT and removed in Moncton, New Brunswick, at 12 pm EDT the following day.

The tubes containing the 3M water soluble tape were incubated for 1 h at 65°C in order to dissolve the tape and were centrifuged at 4000 rpm for 10 min to pellet the sample. Absolute ethanol was then added to sediment small insects and other debris still floating in the buffer, and the residual tape, if any, was carefully removed ensuring no loss of sample. The pellet was then suspended in 10 mL preheated (5 min at 65°C) 2% CTAB buffer (2% [w/v] CTAB, 1 M Tris, 1.4 M NaCl, 0.5 M EDTA, 50 μg/mL proteinase K) with the addition of β-mercaptoethanol (2 μL/mL), followed by homogenization in a tissue homogenizer for 10 min. The pellet was then incubated for 1.5 h at 65°C, mixing the sample every 15 min. The tubes were centrifuged at 4000 rpm for 10 min, and the supernatant was transferred in a new tube. Equal volume of 24:1 chloroform: isoamyl alcohol was added, and the sample was mixed by inverting for 10 min. After centrifugation at 4000 rpm for 10 min, the upper aqueous phase was transferred into a new tube. The DNA was pelleted using an equal volume of cold (−20°C) isopropanol and by mixing (inverting) the sample for 10 min followed by centrifugation at 4000 rpm for 20 min. The pellet was washed twice with 5 mL of 70% ethanol, dried at room temperature for 15 min, and finally resuspended in 150–250 μL of 10 mM Tris-EDTA buffer. The pellets were then submitted for 454 Life Sciences (Roche) FLX sequencing.

Testing for abundance differences

In order to access the significance of differences in read counts corresponding to a particular taxon between trip A and trip B, we used a Poisson two-sample test (Huffman 1984). For each taxon, we used two formulae to calculate *Z*-scores, and corresponding *P*-values from a standard normal distribution:

$$Z_1 = \frac{Y - dX}{\sqrt{d(X+Y)}} \; Z_2 = \frac{\sqrt{Y+\frac{3}{8}} - \sqrt{d(X+\frac{3}{8})}}{\sqrt{1+d}}.$$

Here, *X* represents the number of reads unique to a particular taxon in trip A, whereas *Y* stands for the number of reads unique to the same taxon in trip B. *d* in both equations is a correction factor that accounts for biases in sample collection, DNA concentration, read numbers, etc., between the two trips. Therefore we used *d* as the "expected" ratio of *Y* over *X* that accounts for these differences. *P*-values corresponding to the *Z*-scores were obtained using the R statistical package and corrected for multiple testing using the FDR correction method. Taxa with *P*-values significant at 1% after correction for both formulae were considered as informative of differences between the two trips.

Acknowledgments

This study was inspired by Professor Robert J. Baker. Efforts of Galaxy team members, including Greg Von Kuster, Nate Coraor, and Dan Blankenberg, were critical for development of the metagenomic toolkit. We thank Katherine S. Pollard and Jonathan Eisen for critical suggestions. Stephan Schuster provided advice on comparison of our methodology with that of MEGAN. We also thank members of Nekrutenko and Makova laboratories, who helped in our collection efforts. Kateryna Makova kindly provided her laboratory equipment for our work. Additional equipment was provided by David Gilmore, Ming Tien, and Graham Thomas. This work was funded by Beckman Foundation Young Investigator Award to A.N., NSF grant DBI 0543285 and NIH grant HG004909 to A.N. and J.T., NSF grant DMS-071499 to S.K.P., as well as funds from Penn State University and the Huck Institutes for the Life Sciences to A.N., and from Emory University to J.T. Additional funding is provided, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

References

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA. 2006. The marine viromes of four oceanic regions. *PLoS Biol* **4:** e368. doi: 10.1371/journal.pbio.0040368.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. 2000. A Kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290:** 972–977.

Baumann P, Baumann L, Lai C, Rouhbakhsh D, Moran NA, Clark MA. 1995. Genetics, physiology, and evolutionary relationships of the genus Buchnera: Intracellular symbionts of aphids. *Annu Rev Microbiol* **49:** 55–94.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL. 2000. Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* **289:** 1902–1906.

Beja O, Spudich EN, Spudich JL, Leclerc M, Delong EF. 2001. Proteorhodopsin phototrophy in the ocean. *Nature* **411:** 786–789.

Brisson JA, Stern DL. 2006. The pea aphid, *Acyrthosiphon pisum*: An emerging genomic model system for ecological, developmental and evolutionary studies. *BioEssays* **28:** 747–755.

Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB. 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* **18:** 763–770.

Chatterji S, Yamazaki I, Bai Z, Eisen JA. 2008. CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. *Lect Notes Comput Sci* **4955:** 17.

DeLong EF. 2005. Microbial community genomics in the ocean. *Nat Rev Microbiol* **3:** 459–469.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496–503.

Erwin TL. 1982. Tropical forests: Their richness in Coleoptera and other arthropod species. *Coleopt Bull* **36:** 74–75.

Erwin TL. 1991. How many species are there?: Revisited. *Conserv Biol* **5:** 330–333.

Fierer N, Liu Z, Rodriguez-Hernandez M, Knight R, Henn M, Hernandez MT. 2008. Short-term temporal variability in airborne bacterial and fungal populations. *Appl Environ Microbiol* **74:** 200–207.

Foerstner KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. *EMBO Rep* **6:** 1208.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312:** 1355–1359.

Handelsman J. 2004. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68:** 669.

Huffman MD. 1984. An improved approximate two-sample Poisson test. *Appl Stat* **33:** 224–226.

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res* **17:** 377–386.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. 2008. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* **72:** 557–578.

Ludwig W, Klenk HP. 2001. A phylogenetic backbone and taxonomic framework for prokaryotic systematics. In *The Archaea and the deeply branching and phototrophic bacteria*, pp. 49–65. Springer, New York.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18:** 1509–1517.

May RM. 1988. How many species are there on earth? *Science* **241:** 1441–1449.

Mayr E. 1998. Two empires or three? *Proc Natl Acad Sci* **95:** 9720–9723.

McHardy AC, Rigoutsos I. 2007. What's in the mix: Phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol* **10:** 499–503.

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4:** 63–72.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al. 2008. The metagenomics RAST server: A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386. doi: 10.1186/1471-2105-9-386.

Noguchi H, Park J, Takagi T. 2006. MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34:** 5623–5630.

Odegaard F. 2000. How many species of arthropods? Erwin's estimate revised. *Biol J Linn Soc* **71:** 583–597.

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, et al. 2006. Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* **311:** 392–394.

Pop M, Salzberg SL. 2008. Bioinformatics challenges of new sequencing technology. *Trends Genet* **24:** 142–149.

Raes J, Foerstner KU, Bork P. 2007. Get the most out of your metagenome: Computational analysis of environmental sequence data. *Curr Opin Microbiol* **10:** 490–498.

Robinson MD, Smyth GK. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23:** 2881–2887.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K. 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5:** e77. doi: 10.1371/journal.pbio.0050077.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. 2007. CAMERA: A community resource for metagenomics. *PLoS Biol* **5:** e75. doi: 10.1371/journal.pbio.0050075.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC. 2005. Comparative metagenomics of microbial communities. *Science* **308:** 554–557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram R, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37–43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P. 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315:** 1126.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9:** R151. doi: 10.1186/gb-2008-9-10-r151.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7:** 203–214.

Received March 31, 2009; accepted in revised form August 19, 2009.