Gist – an ensemble approach to the taxonomic classification of metatranscriptomic sequence data.

Samantha Halliday[1,2] and John Parkinson[1,3,4,*]

[1] Program in Molecular Structure and Function, Hospital for Sick Children, Toronto, Ontario, M5G 1L7, Canada

[2] Department of Computer Science, University of Toronto, Toronto, Ontario, M5S 1A8, Canada

[3] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, M5S 1A8, Canada

[4] Department of Biochemistry, University of Toronto, Toronto, Ontario, M5S 1A8, Canada

* To whom correspondence should be addressed:

john.parkinson@utoronto.ca

Running Title: **A taxonomic classifier for metatranscriptomic data**

Keywords: metatranscriptomics, taxonomic classification, machine learning

ABSTRACT

The study of whole microbial communities through RNA-seq, or metatranscriptomics, offers a unique view of the relative levels of activity for different genes across a large number of species simultaneously. To make sense of these sequencing data, it is necessary to be able to assign both taxonomic and functional identities to each read. Such assignments allow biochemical pathways to be appropriately allocated to discrete species, enabling the capture of cross-species interactions. Currently, these annotation tasks are commonly performed by looking for long matching subsequences. Such approaches are dependent on homology, and are challenged by highly diverse species. Alternative methods, based on compositional analysis of shorter fragments, have not yet demonstrated comparable performance. Here we introduce a novel program for generating taxonomic assignments, called Gist, which integrates information from a number of machine learning methods and the Burrows-Wheeler Aligner. Uniquely Gist optimizes weightings of methods for individual genomes, facilitating high classification accuracy on next-generation sequencing reads. Further innovations of value to the field include the ability to incorporate prior knowledge about taxon abundances as well as the return of multiple assignments, including to parent taxa. We validate our approach using a synthetic metatranscriptome generator based on Flux Simulator, termed Genepuddle, and on real data. Our results demonstrate the capacity of composition-based techniques to accurately inform on taxonomic origin without resorting to longer scanning windows that mimic alignment-based methods, reducing dependence on reference genomes. Gist is made freely available under the terms of the GNU General Public License at **compsysbio.org/gist**.

INTRODUCTION

Recent advances in high-throughput sequencing are profoundly transforming our understanding of the relationship between complex microbial communities (microbiomes) and their environments. In the context of human health, it is increasingly apparent that the composition of the intestinal microbiome has a significant impact on many diseases including type I diabetes, inflammatory bowel disease (IBD), obesity, and rheumatoid arthritis (Angelakis et al. 2015; Greenblum et al. 2012; Ley et al. 2005; Bervoets et al. 2013; Tong 2015; Loh and Blaut 2012; Kostic et al. 2015; Hara et al. 2013). Typically, studies of complex bacterial communities have largely relied on marker gene (e.g. 16S rRNA) surveys, which yield only limited functional insights (McHardy et al. 2007). With the recognition that multiple combinations of microbial taxa can confer similar functional outputs, efforts have begun to define microbiome function, in addition to the taxa responsible, through untargeted DNA or RNA sequencing (metagenomics and metatranscriptomics respectively) (Xiong et al. 2012; Damon et al. 2012; Lesniewski et al. 2012; Poulsen et al. 2013; Gosalbes et al. 2011). For example, key fermentation products of abnormal bacterial metabolism in the human gut (short-chain fatty acids, especially propionic acid) produced by certain *Clostridia*, *Desulfovibrio*, and *Bacteroidetes* species, have been shown to trigger neuroinflammation in a mouse model, resulting in behavioral changes consistent with autism spectrum disorders (MacFabe 2012; Frye et al. 2016).

A major focus in the analysis of these complex datasets is the accurate determination of the taxa present. Beyond defining taxa responsible for critical functions, taxonomic assignments permit binning of reads that can help with sequence assembly, allowing the generation of longer genomic scaffolds or transcripts while minimizing the generation of chimeras (Kumar and Blaxter 2010; Kumar et al. 2013; Li et al. 2012). However, due to the vast diversity of microbes encountered in microbiomes, taxonomic assignment of sequence reads remains challenging.

Three general categories of techniques for assigning taxonomic labels have been developed: phylogenetic, alignment- or similarity-based, and compositional (Bazinet and Cummings 2012). Phylogenetic

strategies, which exploit models of evolutionary relationships, are computationally intensive and rely on the reprocessing of results from alignment and/or composition based methods to quantify the distance between the assigned reads and the reference data (Berger et al. 2011; Munch et al. 2008). In alignment-based strategies, the results of a sequence similarity search method such as BWA (Li and Durbin 2009) are used to map reads directly onto known reference sequences (e.g. genomes or sets of known transcripts). Due to the reliance on databases that represent only a fraction of bacterial diversity, these methods perform poorly for data containing taxa that have not previously been well-sampled and can also be confounded by lateral gene transfer events (MacDonald et al. 2012). Compositional methods offer an alternative. Typically, such methods count the frequencies of short fragments of reads, called *k*-mers, using a sliding window of some preset length *n* to scan either the nucleotide or amino acid content of a given reference sequence. These counts are then used to generate *k*-mer profile distributions yielding a position-independent summary of sequence content against which *k*-mer distributions of sequence reads can be compared and used for assignment. This process is generally more robust, as it can detect short motifs of diagnostic relevance out of context, such as pathogenicity markers (Rosen et al. 2008). A number of machine learning algorithms have been applied to perform composition-based assignment including: naïve Bayes (NB; (Rosen et al. 2008)), *k*-means clustering (Kelley and Salzberg 2010), hidden Markov models (HMM; (Brady and Salzberg 2009)), support vector machines (SVM; (McHardy et al. 2007; Patil et al. 2012)) and Gaussian-kernelized *k*-nearest neighbors (*k*NN; (Diaz et al. 2009)).

Despite their ability to overcome limitations in taxonomic sampling, current taxonomic classifiers based on compositional approaches are still limited in their ability to deal with problems involving large numbers of species and can be sensitive to sequencing errors (Vervier et al. 2016). A further issue with current implementations of these algorithms lies in their objective of assigning reads to distinct taxa, which becomes challenging when a read cannot be unambiguously assigned to a single genome. Such ambiguity typically arises when a phylogenetic branch is under sampled relative to the rest of the taxa in the database, a significant challenge with microbiomes. Finally, genome-specific biases in sequence com-

position suggest that no single algorithm will yield optimal results across all genomes, with different algorithms likely to perform better for certain taxonomic groups (Brady and Salzberg 2009). To overcome these challenges, ensemble methods that combine several methods offer the potential for improved classification performance. For example, WEVOTE combines predictions from several tools to predict taxon assignments (Metwally et al. 2016), and Phymm (Brady and Salzberg 2009) use combinations of one statistical model under several different weightings.

Here we present a new ensemble classifier, Gist, which uses a predominantly Bayesian framework to integrate predictions from a number of complementary methods. Uniquely, Gist establishes an initial set of weights for each method, specific to each genome in its reference dataset. This weighting optimizes the ability of the combined set of methods to associate sequence reads to a specific genome, achieving high-quality results with much smaller *k*-mers than in previous implementations of composition-based approaches. To avoid errors due to ambiguities in classification, probabilistic output generated by Gist allow reads to be assigned to appropriate taxonomic ranks. The use of a Bayesian model, additionally allows Gist to incorporate prior knowledge about the distribution of data (e.g. based on known 16S rRNA abundance information). We validate our approach using synthetic data, as well as metatranscriptomic data from a previous study of the intestinal microbiome associated with a non-obese diabetic (NOD) mouse model. The performance of Gist is compared against five established short-read classifiers, all developed for use with metagenomic data: Naïve Bayes Classifier, which introduces a Bayesian framework to model genomic *k*-mers (Rosen et al. 2008) with a high-dimensionality Bernoulli distribution; Kraken, which modifies this model by implementing a novel root-to-leaf approach based on mapping *k*-mers in the context of a taxonomic tree (Wood and Salzberg 2014); CLARK, which iterates on the NBC model by instead introducing reduced sets of *k*-mers (Kumar et al. 2013); Centrifuge (Kim et al. 2016) and Kaiju (Menzel et al. 2016), which are primarily alignment-based algorithms which look for maximum-length, exact-sequence matches (MEM) in nucleotide and peptide space respectively, both employing a Burrows-Wheeler transform with Ferragina and Manzini's FM–index (Ferragina and Manzini

2005) for efficiency, although Kaiju additionally features a more traditional, more fault-tolerant 'Greedy' alignment mode. While our focus is on metatranscriptomic data, with additional modification we propose that Gist, with its improved performance and reduced dependence on exhaustive genome databases, represents an effective solution to the taxonomic classification of short read metagenomic datasets, either by binning prior to assembly, or where low coverage limits assembly options.

RESULTS

Gist – an ensemble taxonomic classifier for analyzing metatranscriptomic sequence datasets

We present Gist (Genome Identification of Short Transcripts), an ensemble classifier that combines the output of the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) and four classification methods used for examining *k*-mer composition: Gaussian naïve Bayes (NB), nearest neighbor search (1NN), a Gaussian mixture model (GMM), and a novel technique, the expected co-delta correlation (ECC) to assign taxonomic labels to metatranscriptomic read data. Each of the four classifiers we implemented is run with both amino acid and nucleotide information, and can be configured independently to use a different *k*-mer length from the other models, resulting in 8 adaptive elements, plus input from BWA and up to two per-strain priors, for a total of 11 components.

Once models have been built, labeled training data is used to determine the reliability of each of the methods for each of the N classes, yielding a $9 \times N$ table of weights. These are learned using a single-layer neural network, in a technique called ensemble averaging. Of the components, only the outputs of NB and GMM models are truly generative probabilities, so this training process is important in re-scaling the outputs of the other models into comparable ranges. The optimal orientation of each fragment in nucleotide space (forward vs. reverse complement) is determined by scoring sequences in both directions against all classes, and keeping scores for the direction that yields the higher overall score, analogous to selecting for maximum likelihood in a fully Bayesian framework.

To classify reads, data is analyzed in two passes; the first pass uses only the fastest methods to reduce the number of candidate genomes to a manageable size, and then the second pass uses all methods to determine final scores for the most likely hits. A user-configurable system of quotas and thresholds can be used to produce multiple results if so desired.

These final scores are expanded taxonomically before output. After the most likely genomes have been identified for a given read, their scores are compared to the scores found in neighboring strains using a one-tailed Student's *t*-test relative to the mean distribution of other members of the species. If it is found that the *t*-test for the selected taxonomic label's score is insufficiently distinct from that of its siblings, then the parent taxonomic unit, e.g. species, will be returned instead. This process can repeat recursively all the way up to the level of order depending on user-configurable thresholds, sacrificing exactness for improved confidence.

See Figure 1 for an overview of the pipeline. More detail about the ensemble components is provided in the Materials and Methods section.

Comparison of Gist against other taxonomic classifiers using simulated datasets

To assess Gist's performance, we first considered its ability to accurately assign taxonomic labels to simulated metatranscriptomic datasets consisting of short sequence reads relative to five state of the art classifiers, NBC (Rosen et al. 2008), Kraken (Wood and Salzberg 2014), CLARK (Ounit et al. 2015), Centrifuge (Kim et al. 2016), and Kaiju (Menzel et al. 2016). To evaluate classifier performance, we used a previously published mouse gut microbiome (Xiong et al. 2012) to produce simulated training and test datasets based on taxonomic profiles derived from 16S rRNA survey data. In this study, non-obese, diabetic (NOD) mice were reared under germ-free conditions and initially colonized with altered Schaedler flora (ASF). ASF is considered to consist of a community of 8 strains of bacteria commonly found in the

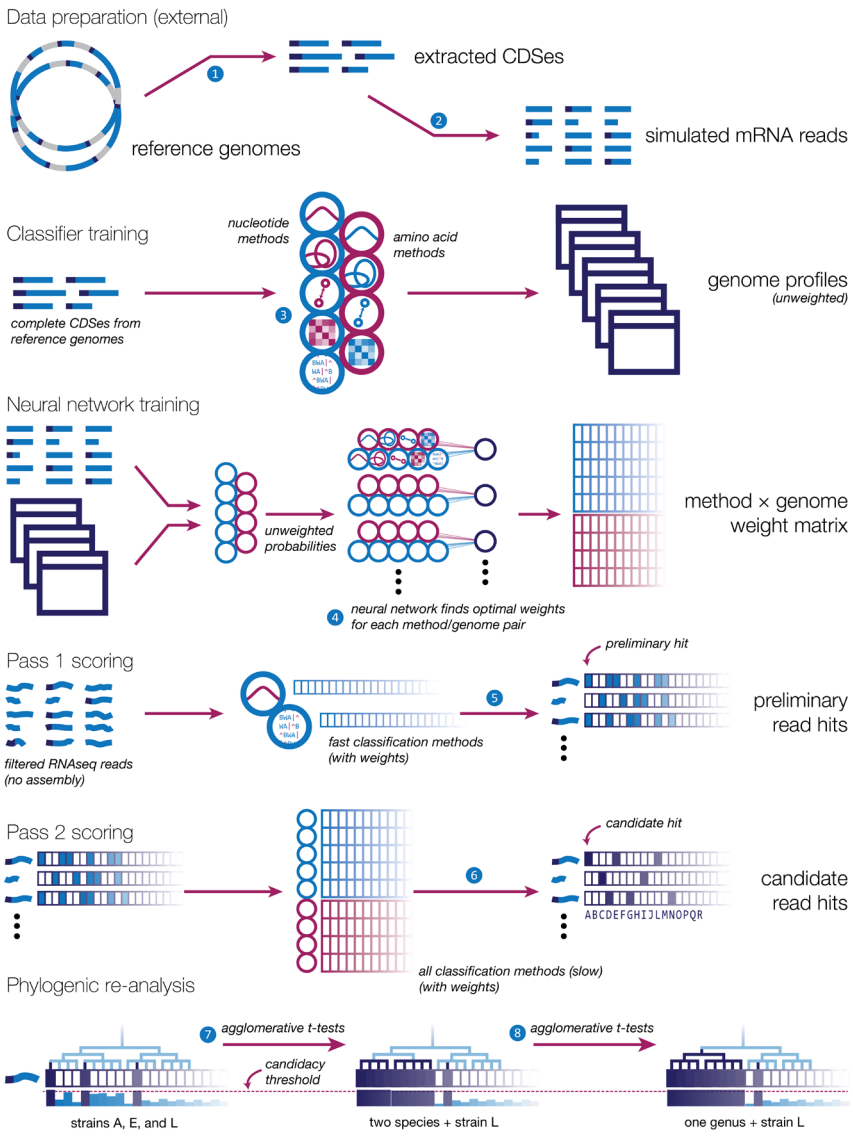


Figure 1. Program overview. (1) In the initial data preparation step, coding sequences from reference genomes are sourced to (2) generate simulated metatranscriptomes using an external tool, such as Genepuddle. (3) During the classifier training step, model components are trained to recognize the original genomic sequences, producing a database of genome profiles, each corresponding to one class. (4) The neural network training step uses the simulated mRNA reads from the first data preparation step to determine the optimal weights for a single-layer perceptron, resulting M×C weights, i.e. one for each pair of genome and method. The algorithm is now ready to process data. (5) In the scoring steps, each model component estimates the likelihood of each read being drawn from each genome, creating an M×C×R table of raw scores, which is multiplied by the M×C weights generated by the perceptron, and summed along the M dimension, resulting in a C×R table of likelihoods that each read comes from each genome. Two scoring passes are used to reduce the amount of time spent calculating scores for classes which are unlikely to represent a particular read; after the first pass, a 'shortlist' of high-scoring preliminaries is produced. (6) The second pass then generates scores for the genomes on the shortlist using slower methods, refining the estimate. (7) Finally, during the phylogenetic re-analysis step, a series of one-tailed Student's *t*-tests with respect to mean scores is used to broaden the assignment, where the null hypothesis indicates that the top-scoring classes within a given taxon are not significantly higher-scoring than other members of the taxon, and hence the read should be assigned to the whole taxon, instead. (8) This is repeated until the *t*-test fails, subject to a series of user-configurable thresholds. The final output is one or more high-scoring taxa, according to a quota, which is also specified by the user.

murine gut (Table 1; (Wymore Brand et al. 2015)), for which genomic sequences are available (Wannemuehler et al. 2014), providing a useful dataset for benchmarking purposes.

Table 1. Taxa in the Altered Schaedler Flora (ASF).

Taxon ID*	Name
97138	**Clostridium** *sp*. ASF356
97137	**Lactobacillus** *sp*. ASF360
1235801	**Lactobacillus murinus** ASF361
1379858	**Mucispirillum schaedleri** ASF457
1235802	**Eubacterium plexicaudatum** ASF492
1378168	**Firmicutes bacterium** ASF500
84086	*unclassified* **Firmicutes sensu stricto** (*miscellaneous*)
97139	**Clostridium** *sp*. ASF502
1235803	**Parabacteroides** *sp*. ASF519

* As defined by the National Center for Biotechnology Information (NCBI)

Analysis of the 16S rRNA survey data previously generated for this mouse gut dataset produced a large inventory of candidate matches of diverse species within the orders of the expected strains; samples from the top 25 genera, plus one clade known to be present but not represented in the 16S data, were included in the pool of genomes used. The result was a dataset consisting of a total of 295 genomes (representing individual strains) taken from the NCBI FTP server, many of which were still in the draft or assembled contig stage of processing at the time of collection (Table 2 and Supplemental Table S1). The rRNA removal treatment appeared to have been biased against the phylum *Bacteroidetes*, resulting in the complete removal of *Parabacteroides* from the data. This resulted in a deficiency in the constructed database which was amended by manually adding a single strain of *Parabacteroides*. This illustrates the importance of careful 16S curation, and how rRNA removal products can be biased towards certain taxa.

Table 2. Taxa in the simulated mouse dataset.

Genus	Strains
Actinomadura	3
Aerococcus	2
Anaerostipes	3
Bacteroides	9
Basfia	1
Blautia	2
Brevibacillus	6
Brevibacterium	2

Butyrivibrio	2
Dorea	3
Enterococcus	11
Eubacterium	7
Exiguobacterium	3
Glaciibacter	1
Idiomarina	5
Lachnoanaerobaculum	2
Lactobacillus	52
Leifsonia	2
Mannheimia	3
Microlunatus	1
Moorella	1
Mucispirillum	1
Paenibacillus	9
Parabacteroides	1
Pelotomaculum	1
Peptostreptococcus	3
Propionibacterium	14
Roseburia	5
Staphylococcus	53
Streptococcus	79
Thermaerobacter	2
Thermomonospora	1

Simulated datasets of 100 bp reads were generated using a novel pipeline Genepuddle from this pool of 295 strains. Genepuddle is more suitable for metatranscriptomic experiments than the widely-used synthetic metagenomic read generator, MetaSim (Richter et al. 2008), because it is based on Flux Simulator (Griebel et al. 2012), a tool that accounts for biases in sequencing errors, read lengths and abundance distributions associated with RNA sequencing, rather than DNA sequencing. Two synthetic datasets were produced: an *unbiased* dataset, with equal counts for each strain, and a *biased* dataset, with abundances derived from the 16S rRNA count data described above, to reflect differences expected in real data. Unbiased and biased test datasets consisted of 737,500 reads (2,500 per strain) and 85,990 reads, respectively. A replicate of the unbiased dataset was used as training data for the ensemble's weights according to the method described previously; although genomic sequences can be used directly, this bootstrapped data better approximates real, unassembled input data.

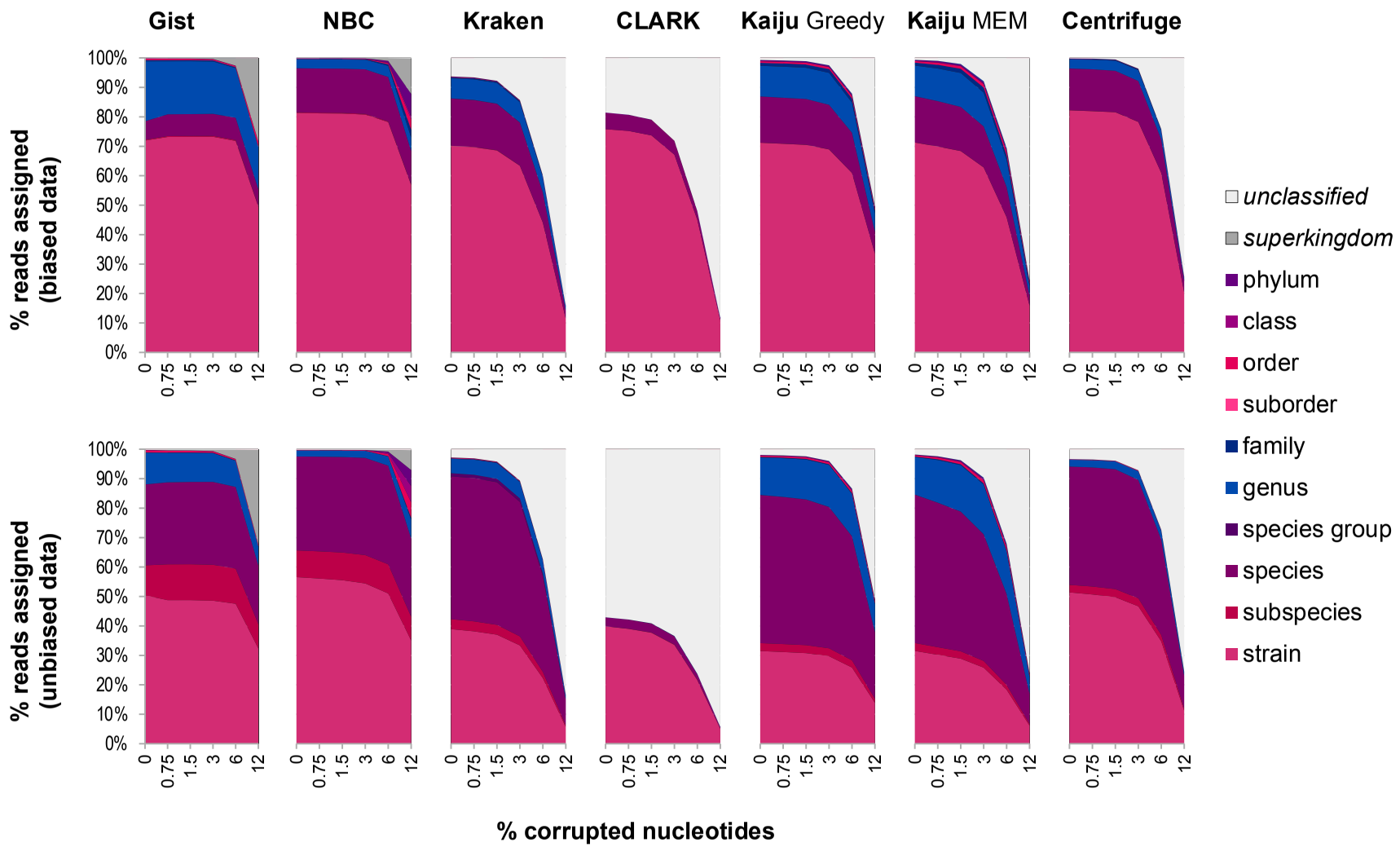


Figure 2. Simulated datasets. Performance on simulated mouse datasets in the presence of isotropic noise. Top: performance of different classifiers on data where each of the 295 strains is represented by an equal number of reads. Bottom: performance of the same when strain abundance is determined by counts from 16S data. The horizontal axis indicates the number of corrupted nucleotides in each 100 nt read, reflecting how each program deals with single-point mutations.

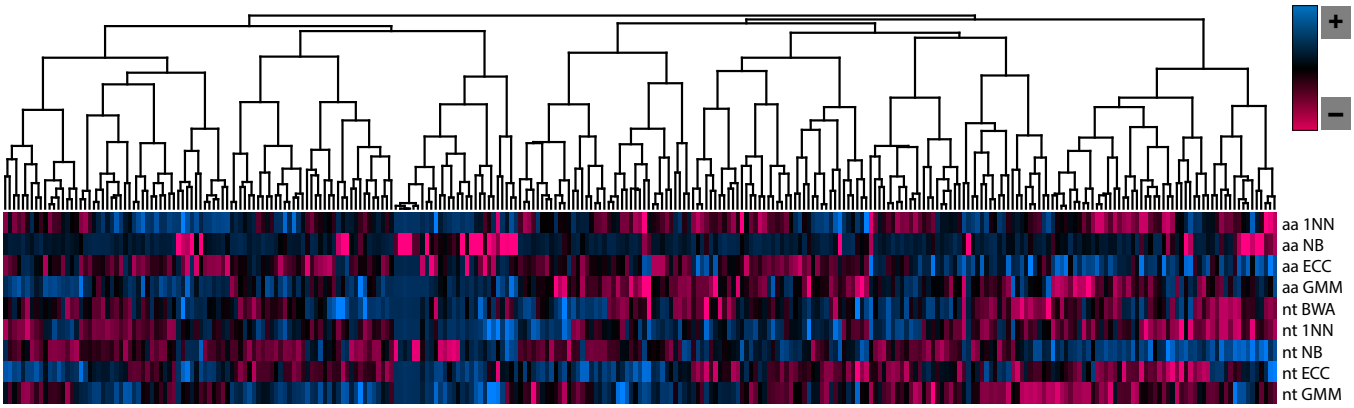
To evaluate each classifier's output on the synthetic data, we used a novel tool, that we called *Lincomp*, which uses NCBI-defined phylogenetic relationships to report accuracy at the lowest supported taxonomic rank. Thus if a classifier assigns a read to the wrong species but the correct genus, it is considered reliable at the genus level for that read. For both unbiased and biased datasets, Gist consistently outperformed CLARK and Kraken across all levels of noise in terms of being able to accurately assign reads at the level of genus or higher (Figure 2). Kaiju and Centrifuge performed more strongly below the genus level on the biased dataset. Kraken and CLARK both aim to improve on NBC's running time by database pruning and through the use of hashing methods. While CLARK maintains high precision, in terms of strain assignments, both of these approaches decline in performance proportionately to the level of noise present in the data, as their comparatively long and error-intolerant *k*-mers (both default to 31 nt) are unable to overcome sequence polymorphisms introduced by our error model, a significant concern described by CLARK's authors as limited sensitivity (Ounit and Lonardi 2016). Centrifuge and Kaiju show more resilience to small levels of noise, when the interval between corrupted nucleotides is high, but like CLARK and Kraken, they face a steep drop-off when the sequences have diverged by 12% or more, at which point all four methods identify a significant number of reads as 'unclassifiable.'

Per-genome weighting of methods optimizes discrimination between taxa

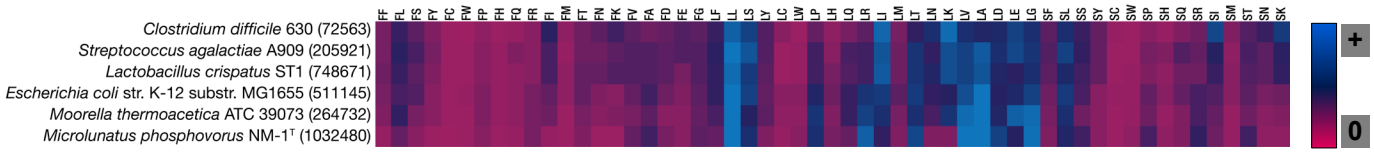
During classification, each of the nine methods shown generates a score, which is weighted and summed to produce the final class-read score. The weights used reflect the contrastive balance necessary to distinguish each genome's reads, as well as a representation of how well each element of the ensemble models each genome. To illustrate how weights help distinguish between genomes, we performed a hierarchical clustering of weightings assigned based on the simulated mouse dataset (Figure 3A).

Focusing on the peptide naïve Bayes classifier, we further show profiles of six representative genomes (*Clostridium difficile* 630, *Streptococcus agalactiae* A909, *Lactobacillus crispatus* ST1, *Esche-*

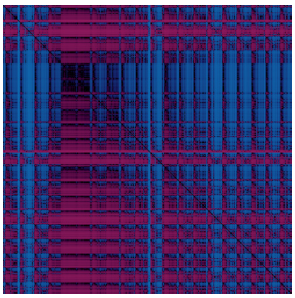
(A) Clustered neural network weights for each genome



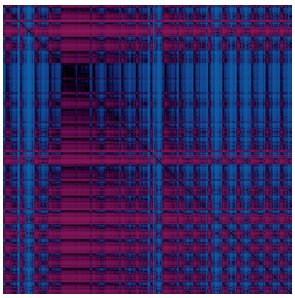
(B) Peptide dimer frequencies in selected genomes illustrate key differences in usage



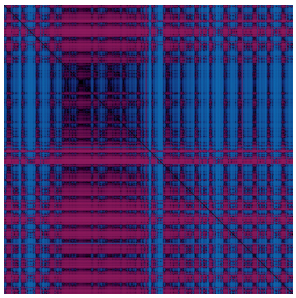
(C) Peptide codelta visually demonstrates taxonomic relationships



Streptococcus agalactiae A909

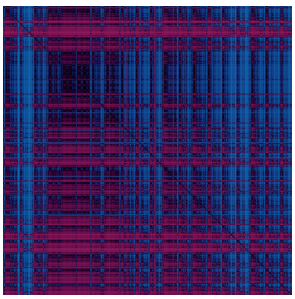


Lactobacillus crispatus ST1

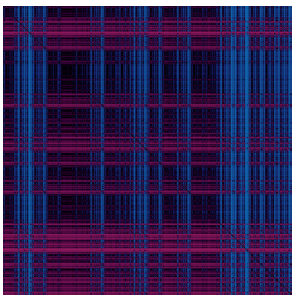


Clostridium difficile 630

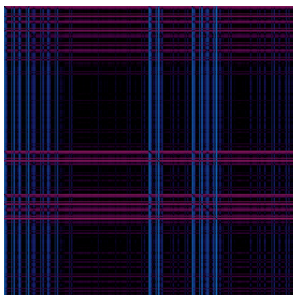
(D) Peptide codelta visually demonstrates sequence entropy



Escherichia coli MG1655

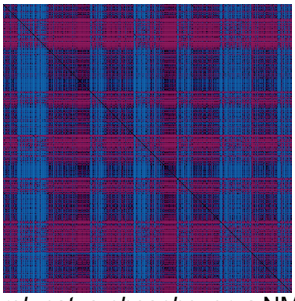


Microlunatus phosphovorus NM-1^{T}

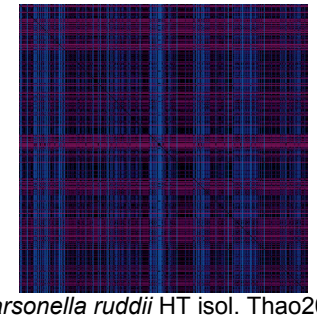


Carsonella ruddii HT isol. Thao2000

Escherichia coli MG1655
(100% noise, same GC%)



Microlunatus phosphovorus NM-1^{T}
(100% noise, same GC%)



Carsonella ruddii HT isol. Thao2000
(100% noise, same GC%)

Figure 3. Genome features. Unique features of genomes picked up in different ways. (A) Weights as learned by the neural network for a human enteric dataset, showing no clear internal structure derived from taxonomy. The data were clustered using Spearman rank correlation with complete linkage. Weights are shown normalized, as the genomes have highly variable total weight (from 10–5 to 10–1.) (B) Comparison of the average peptide pair counts per gene for select strains, demonstrating the variety visible between them. (C) Codelta comparisons. Three firmicute strains, showing strong taxonomic correlation in their similarity. Blue cells show a positive correlation between pairs of amino acid dimers, whereas red cells show a negative correlation. (D) Codelta graph for random genomes with the same GC content as the Candidatus *Carsonella rudii*, *E. coli*, and *M. phosphovorus* genomes, illustrating the relationship between sequence complexity and environment.

richia coli str. K-12 substr MG1655, *Moorella thermoacetica* ATC 39073, *Microluntaus phosphovorus* NM-1^{T}). These indicate the frequency of each dimer averaged across the population of each genome's genes, e.g. the peptide dimer LeuArg is occasionally found in *C. difficile* 630 *M*, but prominent in *Microlunatus phosphovorus* NM-1^{T}.

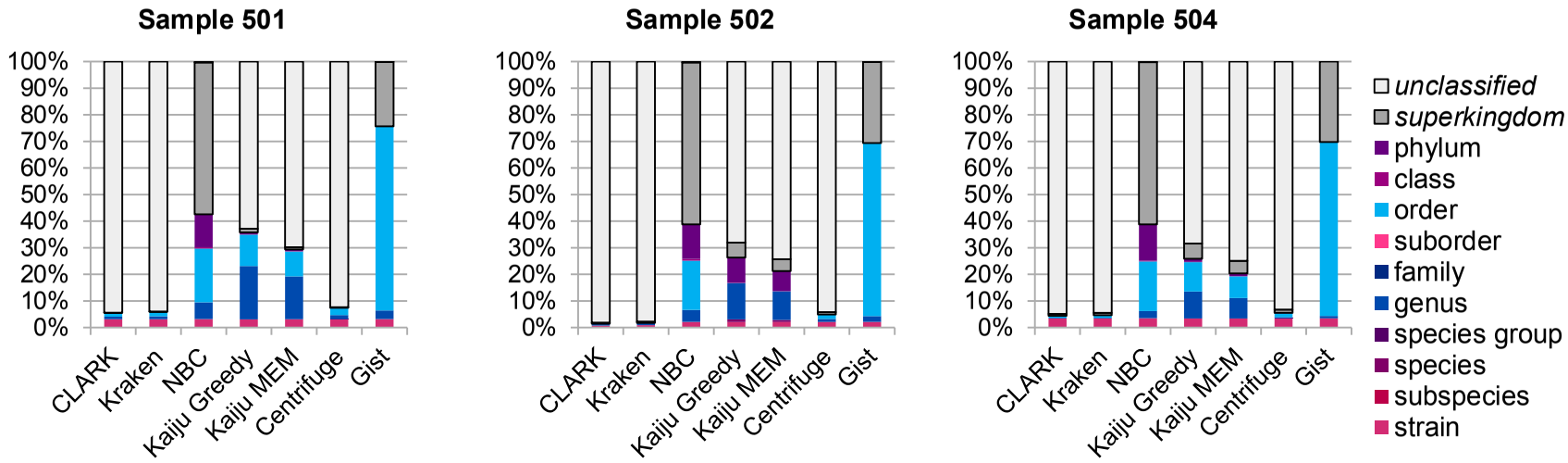
The learned profiles used by other methods also reveal key differences between genomes. Expected codelta correlation (ECC) is a distance metric novel in the Gist model, which complements naïve Bayesian methods by considering the rates of co-occurrence of each pair of *k*-mers in a given genome. To show how peptide dimer correlations change between genomes, we generated codelta tables for six representative genomes. Banding patterns associated with three firmicutes (*S. agalactiae* A909, *L. crispatus* ST1, and *C. difficile* 630) reveal similar patterns of dimer co-occurrence reflecting close taxonomic relationships (Figure 3C). This translates to ECC scores of 1.18, 1.77, and 1.92 between *S. agalactiae* A909 and *L. crispatus* ST1, *S. agalactiae* A909 and *C. difficile* 630, and *L. crispatus* ST1 and *C. difficile* 630 respectively. Conversely, comparisons of *E. coli* MG1655 with two atypical genomes: *Carsonella ruddii*—a symbiont of psyllids (plant lice) with a genome of only 160 kilobases (Nakabachi et al. 2006); and *Microlunatus phosphovorus*—a chemoorganotroph notable for lacking several pathways typical of other Actinobacteria (Kawakoshi et al. 2012) reveal distinct banding patterns of dimer co-occurrence resulting in ECC scores that provide greater discrimination between these three genomes (11.70, 2.72, and 13.76 for *E. coli* MG1655 and *Carsonella ruddii, E. coli* MG1655 and *Microlunatus phosphovorus*, and *Carsonella ruddii* and *Microlunatus phosphovorus* respectively; Figure 3D). Interestingly, further comparisons between these three genomes and random noise matrices generated with the same GC compositions, reveals greatest divergence with *Carsonella ruddii* (ECC = 105.85), suggesting a strong selective pressure for the reduced amino acid complement perhaps associated with more limited functionality as might be expected for a symbiont.

Comparison of Gist against other taxonomic classifiers in the absence of reference genome datasets

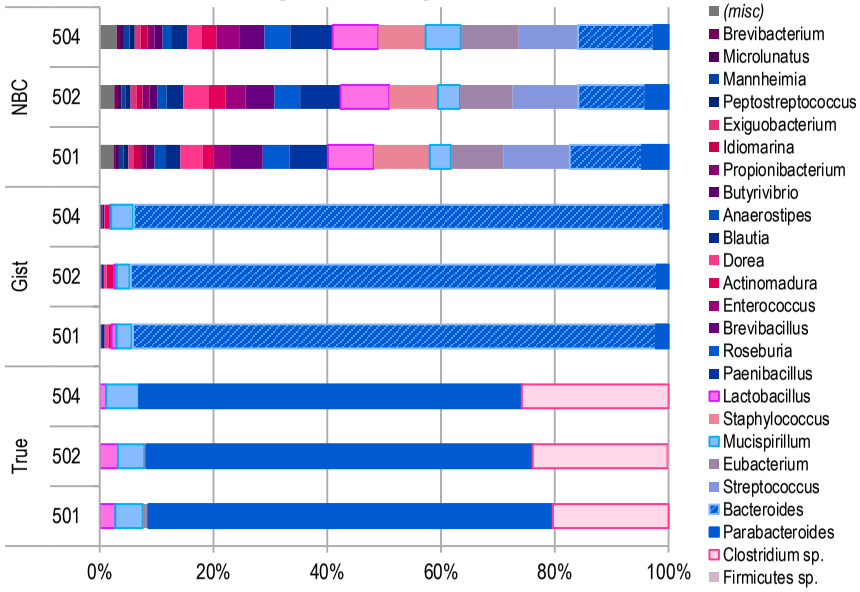
In our initial benchmarking, we noted that Gist's accuracy was similar to NBC. However, given this benchmarking was based on reference genomes that were included in the training data, we might expect that methods dependent on sequence similarity would yield accurate assignments. Here we examine the performance of the methods in the context of environmental samples for which reference genomes may not be available. For our test dataset, we exploited the same mouse colon study as above, however, rather than base this analysis on simulated data from the 295 genomes identified through 16S rRNA analysis, we instead analyzed 175,884 reads of 76 bp generated from three metatranscriptomic datasets, designated *501, 502* and *504* (Xiong et al. 2012), using the classification model trained from the synthetic community experiment, above. The datasets correspond to three biological replicates from three mice with RNA libraries prepared using similar methods. To evaluate algorithm performance, the combined outputs of three aligners (BWA (Li and Durbin 2009), BLAT (Kent 2002), and BLASTN (McGinnis and Madden 2004)) were used to align reads to the more recently sequenced genomes of the ASF community (Wannemuehler et al. 2014), only two strains of which were included in the set of 295 genomes used for training); the results are summarized at the bottom of Figure 4B. Again, we used our in-house tool Lincomp to evaluate accuracy of taxonomic assignments with respect to this gold standard.

Figure 4 reveals that Gist demonstrates a significant improvement in performance relative to CLARK, Kraken, and Centrifuge. Although the latter three classifiers obtained similar levels of performance at the strain level to Gist, Kaiju, and NBC, each was able to classify less than 10% of each dataset, with the vast majority of sequence reads annotated as unclassifiable (Figure 4A). Conversely, NBC, Kaiju and Gist exhibited much greater success in identifying taxonomically related sequences belonging to the 293 strains present in the training set that were not identical to ASF strains. Furthermore, Gist improved over the other methods by correctly annotating ~70% reads at the level of order or better, compared to 25-30% for NBC (with NBC being able to classify an additional 10-12% at the phylum level) and 15-30% for Kaiju in either running mode. Examination of specific taxonomic assignments revealed that Gist at-

(A) Classifier performance on real data



(B) Taxonomic assignments by classifiers on real data



(C) Performance binned by true taxonomic label

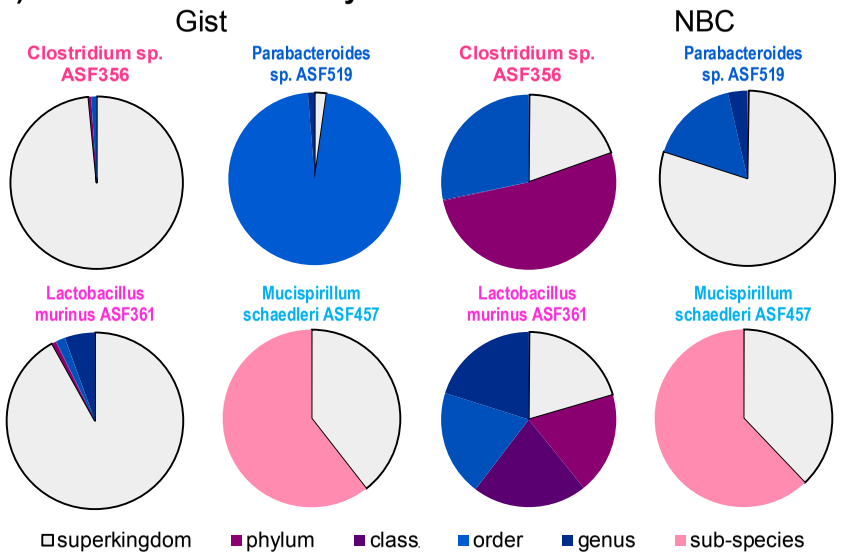


Figure 4. Real data. A) Results of classifying the actual datasets collected from the colons of non-obese diabetic (NOD) mice comparing classification of Gist, Kraken, NBC, CLARK, Kaiju, and Centrifuge against one another. Each community consists of 9 strains (altered Schaedler flora, or ASF). Data analysis was approximated with the same strains from Figure 2, which included only one of the ASF strains. B) Genera found during dataset classification for Gist and NBC, contrasted with the true ASF strain identities. C) Subsets of A) using Gist and NBC for taxa which were known to be most abundant in the NOD 504 dataset.

tributes over 90% of the reads to the genus *Bacteroides*, compared to only ~12-15% for NBC (Figure 4B). While mapping to ASF strains suggests ~70-75% of reads should map to *Parabacteroides sp. ASF519*. Given that *P. distasonis* was included in the training data and that neither NBC nor Gist assigned more than a few hundredths of the data to the genus *Parabacteroides*, these results suggest that *Parabacteroides sp. ASF519* likely has a quite dissimilar gene complement to that of *P. distasonis*, and may be misclassified. These results are reflected in Figure 4C, where we note that both Gist and NBC appear to correctly annotate the majority of the 117,528 reads mapped to the *P. sp. ASF519* genome to either the order or phylum level. Of the sequences belonging to the *M. schaedleri ASF457* genome, both Gist and NBC correctly assign ~3/5 of the 9,564 reads to the sub-species level; however, Gist yields inferior performance to NBC for both the 43,602 reads that map to the *C. sp. ASF 356* genome and the 2,726 reads that map to the *L. murinus ASF 361* genome. For both strains, the majority of reads appear to have been erroneously predicted by Gist as belonging to an unknown member of *Bacteroides*; this likely reflects Gist's use of short *k*-mer signatures (codon bias, GC content etc.) compared to NBC's reliance on closest matching long *k*-mer.

Running time

Table 3 lists wall time and CPU time requirements for each method on a single node with two eight-core Intel Xeon (Sandy Bridge) E5-2650 2.0 GHz CPUs and 64 GB of RAM. 100,000 synthetic reads of 100 nt in length were used. Kaiju Greedy is the fastest method, requiring less than a minute of CPU time to achieve performance comparable to the results produced by NBC in 3.0 hours, which does not natively support multithreading. In total, Gist required 2090% of the CPU time used by NBC, but only 130% of actual (wall) time, due to multithreading. With the exception of Kaiju, execution time (Table 3) appears to directly correlate with the performance, in terms of sensitivity, reported in Figure 4A.

***Table 3. Classifier runtimes*.**

Classifier	Wall time (s)	CPU time (s)
Kaiju Greedy	3	48
Kaiju MEM	4	64
Centrifuge	12	192
CLARK	36	576
Kraken	192	3072
NBC	10801	10801
Gist	14089	225424

DISCUSSION

Gist yields precision and sensitivity on-par or surpassing existing composition-based methods at a very short *k*-mer length by combining several sequence features. This enables Gist to obtain unprecedented sensitivity. Previous work has shown that amongst alignment-free techniques, the Naïve Bayes Classifier (NBC) (Rosen et al. 2008) offers robust precision (Bazinet and Cummings 2012) and the best sensitivity (Ounit et al. 2015). NBC uses only a single nucleotide naïve Bayes distribution to classify reads. To function with such accuracy comparatively large *k*-mers are required. This reliance can be problematic where environmental strains feature significant rates of sequence polymorphism with respect to genomes used for reference, which has driven interest in the use of so called gapped *k*-mers, also known as spaced seeds (Břinda et al. 2015; Ounit and Lonardi 2016). Both Kraken and CLARK have received gapped *k*-mer versions of their algorithms, SEED-KRAKEN (Břinda et al. 2015) and CLARK-S (Ounit and Lonardi 2016) with varying results; while SEED-KRAKEN's authors report substantial performance improvement over its antecedent, CLARK-S exhibits a smaller increase in sensitivity compared to CLARK. As another program, LMAT, also showed significant improvements when gapped *k*-mers were added (Břinda et al. 2015), it seems likely that CLARK's novel database pruning methodology may be responsible for the discrepancy, as this pruning eliminates much of the data redundancy that might otherwise serve to make the program resilient to sequence polymorphisms; close orthologs may almost completely cancel out one another. Gist's mixture of short (4–6 nt) and medium (9–15 nt) length *k*-mer methods circumvents the need for gapped *k*-mers, so even in the presence of strong sequence divergence, several of

its methods can identify forms of evidence for relationships between sequences, including non-homologous genes from the same genome or niche, through properties like codon usage.

Of the non-compositional methods we benchmarked against, Centrifuge and Kaiju, we found the results surprisingly divergent considering their closely related algorithms. Past methods that depend exclusively on peptide sequence comparison, such as MetaCV, (Liu et al. 2013) have faced challenges in taxonomic classification because of the loss of clade-specific markers, particularly codon bias. This outcome for Kaiju was anticipated by the results shown in Figure 2, where Centrifuge exhibits higher strain-level sensitivity, but the final performance in Figure 4A was comparable to NBC and Gist, and, at some ranks, superior. It seems likely that Kaiju's Greedy method represents a good compromise between the exactness of pure long *k*-mer methods like CLARK and Kraken, and the more fault-tolerant medium-length approach shown by NBC. Centrifuge's exclusive use of nucleotide sequences, as well as its requirement of a 16 nt starting seed, likely result in a limited ability to make assignments in the absence of closely related reference genomes.

In general, compositional classifiers for metagenomic taxonomy assignments implement a single machine learning technique: RITA (MacDonald et al. 2012), NBC (Rosen et al. 2008), Kraken (Wood and Salzberg 2014), and CLARK (Ounit et al. 2015) all rely on NB; TACOA (Diaz et al. 2009) uses *k*-nearest neighbors (*k*NN); MetaCV (Liu et al. 2013) uses a modified protein-based HMM; and Phymm/PhymmBL (Brady and Salzberg 2009) exploit weighted averages of Markov models, called interpolated Markov models. While RDP, a short-fragment taxonomic classifier (Cole 2004), combines both NB and *k*NN, its use is restricted to classification of ribosomal RNA fragments. To our knowledge, Gist is the first short read classifier that combines many approaches within a single unified model. Uniquely, Gist additionally adopts weights on a per-genome basis to better capture distinct features associated with individual genomes resulting in weighting schemes that optimize the ability of each method to discriminate between genomes. This improves Gist's flexibility over an unweighted model in accommodating the complexity of genetic composition without succumbing to the type of over-fitting one would

expect from a strictly instance-based technique, such as pure *k*NN, while also avoiding the cost of training a powerful discriminatory method such as a support vector machine to accommodate every new classifier category.

A further advantage of Gist over many taxonomic classifiers is the reporting of taxonomic assignments only at the rank that is supported by the underlying model. This design feature was introduced based on an appreciation that different lineages evolve at different rates. Unlike other classifiers, such as RITA, which identifies uniform taxonomic groupings based on user defined thresholds (MacDonald et al. 2012), Gist outputs labels for short reads that may be usefully exploited for the purpose of intuiting taxonomic groupings. Thus, the decision of whether to assign a read to a parent or child taxon is determined from the probability of assignments to each of the child taxa; reads receiving equal probabilities to two child taxa are assigned at the level of the parent. Importantly, such assignments are made without any assumption about relative branch length between each child and its parent; this decreases the likelihood of errors that may arise from instances of horizontal transfer or atypical mutation rates, as associated, for example, with genes under strong selective pressures (Tamames and Moya 2008; Abby and Daubin 2007). Such problems are commonly encountered in the clustering of 16S rRNA data, where taxa of high genetic diversity may be partitioned depending on sequence input order (He et al. 2015; Westcott and Schloss 2015).

The availability of 16S rRNA data that define community members can greatly reduce search space for metatranscriptomic classification, thereby minimizing errors (Rosen et al. 2008). Such data may not always be available. Gist supports the input of strain abundance data from external sources during classification, allowing the operator to guide taxonomic assignments on the basis of available prior knowledge, a key strength of its partially Bayesian framework.

Many recent methods, including Kraken, CLARK, Kaiju, and Centrifuge, emphasize conservative running times. Gist's high running time requirements (Table 3) oppose this trend. With the exception of

Kaiju, however, these comparatively quick methods have not achieved levels of sensitivity on par with NBC, which is also noted as being relatively slow. As Kaiju does not use compositional data to perform classifications, it cannot assign genes to taxa unless such relationships have already been directly evidenced in existing databases, a challenge that is readily addressed by algorithms employing short *k*-mers. At the same time, short *k*-mer algorithms may lack the ability to discriminate between closely related genomes, requiring the supplementation of additional methods. While we acknowledge that the implementation of our ensemble methodology results in long runtimes, given the importance of assigning critical functions to key taxa, we nevertheless consider it imperative to emphasize the need to prioritize accuracy over speed of execution. Improving Gist's efficiency is a primary goal for future development work, with the stipulation that its generalization ability not be sacrificed.

Taxonomic classifiers, like many kinds of bioinformatics programs, often have a limited post-release development cycle, as they are rapidly succeeded by new methods. Many cease development shortly after publication. One unfortunate consequence of this is that the databases offered with these programs eventually become outdated, complicating comparisons in typical usage scenarios. This underscores the importance of confidence metrics; *k*-mer and alignment methods lacking a minimum confidence score may assign data to spurious, unrelated taxa based on very small amounts of evidence. While the authors of CLARK and Kraken have made some efforts to consider the reporting of confidence levels, CLARK, Kraken, and Centrifuge ensure their results primarily by being very cautious about the assignments they make. This comes at the cost of leaving most of the data unconsidered when the database poorly represents the available sequences, as demonstrated in Figure 4A. Additionally, the newer programs considered in this paper (Kraken, CLARK, Centrifuge, and Kaiju) all provide utilities or detailed instructions for downloading new data directly from NCBI, ENSEMBL, or another central repository, although even this is not completely future-proof; on September 20, 2016, the NCBI's FTP version of its repository of bacterial genomes, from which many tools traditionally instructed users to obtain data, underwent reorganization, rendering older database construction methods incompatible.

To ensure that Gist does not become obsolete due to changes in available reference sequences, we are currently developing an efficient pipeline for updating and constructing both databases and training data tuned to the latest available information and the user's needs. In addition, with Gist's Bayesian inference framework for coordinating the outputs of its classifiers, it should be practical to integrate other sequence classification algorithms. Given the noteworthy performance of NBC and Kaiju, we are currently working on integration of these methods into the Gist pipeline. Gist (and associated tools, Lincomp and Genepuddle) is provided as open source software under the GNU General Public License, version 3.0, and is available for download on GitHub at https://github.com/rhetorica/gist, or from its website at http://compsysbio.org/gist.

MATERIALS AND METHODS

Data

The sequenced reads used in this study are available from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra: SRX134834, SRX134840, SRX134842 for samples 501, 502, and 504, respectively.) For further details of sequence processing, see (Xiong et al. 2012). Genomes for inclusion in the synthetic data were selected by drawing from the 25 most abundant taxa found during 16S rRNA analysis, and then pooling at the genus or family level. As the 16S analysis successfully detected *Mucispirillum schaedleri*, a strain of this species was used directly, for a total of 285 strains. As no strains from the phylum Bacteroidetes were detected by 16S analysis, 9 *Bacteroides* and 1 *Parabacteroides* genomes were added. Flux Simulator (Griebel et al. 2012), an RNA-seq simulator program, was used to produce the synthetic read data, but as Flux Simulator is intended for use with a single genome, it was adapted for synthetic metatranscriptome generation using the GenePuddle frontend, described below. Using Genepuddle, 100 nt artificial mRNA reads were produced from the selected genomes, assuming uniform expression of all genes. Two synthetic datasets, the *biased* and *unbiased* datasets, were produced,

in both test and training versions; the biased datasets distributed the relative abundances of each of the 25 taxa among their proxies, so that e.g. the 80 *Streptococcus* strains, which served as stand-ins for a single, minor, unsequenced *Streptococcus* strain, did not overwhelm the far more important single strain of *Mucispirillum*; the unbiased datasets consisted of precisely 5000 reads per sample.

Algorithm details

Gist was implemented in C++ under GNU/Linux. Scanning for protein sequences was derived from FragGeneScan (Rho et al. 2010), and the *t*-test calculations in the output pass were implemented using ALGLIB (http://www.alglib.net). It incorporates four machine learning methods as classifiers (NB, 1NN, GMM, and ECC), which are used in nucleotide, reverse-complement-nucleotide, and amino acid modes. Scores from the Burrows-Wheeler Aligner (BWA) are also considered, generated separately against each reference genome.

BWA was chosen as an efficient method for identifying close or exact matches, minimizing false positive assignments. BWA was selected over other programs of its type due to its low memory usage and high accuracy when analyzing prokaryotic sequences (Shang et al. 2014). We did not consider protein alignments as they have limited taxonomic resolution (for example, contrast the performance of Centrifuge, a nucleotide-based MEM method, with that of Kaiju, a protein-based MEM method, in Figure 2) due to being unable to detect codon bias. With some normalization, the scores output by BWA are used directly in the ensemble model as if they were probabilities.

Our nearest neighbor search (1NN) implementation is an instance-based method that determines the nearest known gene in each strain and reports the best distance from these genomes with a modified Euclidean metric. This differs from typical nearest neighbor methods, which usually pool all reference data together and return only a single result corresponding to the label of the closest data point among all genes in the reference set. In contrast with BWA, which is also effectively instance-based, 1NN is much more tolerant of rearrangements and short duplications, although it is vulnerable to missense mutations.

Like the BWA scores, the distances returned by the 1NN model are treated as probabilities after some modification; in this case, inversion (i.e. $1/x$).

Naïve Bayes (NB), a popular algorithm for compositional taxonomic classifiers, works by assuming the data is distributed according to a single multivariate distribution. In Gist, a multivariate Gaussian model is used. Gaussian NB is effective at determining the mean of the distribution (i.e. what typical genes from a genome look like), but does not perform well on outliers, and rapidly becomes oversaturated at smaller *k*. It may also fail when a genome has several large, distinctive subpopulations, corresponding to a true data distribution that is multimodal; in these cases, the mode of the Gaussian often falls in a region of low probability.

Expectation–maximization (EM) is a randomized iterative algorithm used to fit model parameters in the presence of latent, or unknown, variables. A Gaussian mixture model (GMM) is used with EM to find the means and variances of multiple subpopulations of genes; the correspondence between each gene and its subpopulation is the latent variable. Consequently, we expect the GMM component to be most effective at modeling a given genome when the genome has recently undergone large-scale horizontal gene transfer from another source, has many genes that do not obey normal codon distributions (e.g. RNA genes), or if contains a large family of proteins with many paralogs.

Expected co-delta correlation (ECC) is a novel technique that provides an efficient trade-off to the calculation of full covariance tables for Gaussian-based methods. It calculates the rates of co-occurrence between pairs of *k*-mers within the read, and then compares this to the average rates for each genome. Because of this two-dimensional relation, ECC can encode motifs of longer lengths by connecting *k*-mers found together in one gene to each other, even if these are discontiguous; for example, applied to translated protein sequences, it is able to identify amino acids most commonly found adjacent to disulfide bridges. For each transcript $\vec{x}$ in the genome, a co-delta table is generated, consisting of the contrasts of the squares of the frequencies of each *k*-mer, i.e.

$$C_{ij} = (x_i - x_j)^2 \cdot \mathrm{sgn}(x_i - x_j)$$

for each cell C_{ij} on the $M \times M$ matrix C, where M is the number of dimensions; 4^k for nucleotides and 20^k for peptides. The result is an anti-symmetric matrix. These codelta tables are then averaged to get the expected codelta matrix for the entire genome (e.g. Figure 3C), and the pairwise difference of a read from the genome mean is then calculated to get the final distance. The average is weighted using the corresponding NB method, so that genes closer to the genome's mean contribute more than outliers.

While each component excels at identifying key features representative of specific genomes, individually they are too simplistic to model the *k*-mer landscapes necessary for accurate classification. For example, ECC does not consider the background rate of each *k*-mer and must therefore be combined with other techniques to be effective. Consequently, a single-layered neural network, similar to logistic regression, is employed to determine the best combination of methods to describe each genome. This ensemble averaging approach significantly improves resolution power at short *k*-mer lengths compared to existing composition-based methods (MacDonald et al. 2012) operating under the same constraint; the resultant joint distribution more accurately represents the shape of each genome's total gene population, and can better predict the expected compositional signatures of unknown genes from related strains, in large part due to the short *k*-mer lengths employed.

Generating the expected weights for each technique is performed during an initial bootstrapping process that is tailored to the dataset provided to the program. This requires synthetic data drawn from a distribution approximating the expected distribution of the real reads. To generate such data, we created a novel prokaryotic metatranscriptome simulation pipeline, Genepuddle. Obtaining the underlying distribution of taxa was accomplished using 16S rRNA counts, substituting genomes from adjacent genera and families when more precise species data were not available, as shown in Supplemental Table S1. While constructing a generic weight and class set incorporating all known genomes is feasible, a reduced dataset improves both speed and results by eliminating the consideration of irrelevant taxa.

Integration of classifier results to yield overall probabilities that a read derives from a given genome is accomplished with Bayesian inference:

$$p(x|G_c) = \exp\left[\sum_{m \in Q} w_{mc} \ln p(x|\theta_{mc})\right] = \prod_{m \in Q} p(x|\theta_{mc})^{w_{mc}}$$

where $p(x|G_c)$ is the overall probability of the read x originating from the genome G_c, w_{mc} is the learned weight from the neural network for method m (from the set of methods Q) and genome c, and $p(x|\theta_{mc})$ is the probability of x originating from genome c modeled with method m using parameters θ_{mc}, referred to as the *raw score*. For a single read, this is repeated for every genome in the dataset.

To improve algorithm efficiency, in the first pass of the algorithm, Q contains only fast methods (BWA and nucleotide naïve Bayes) that are used to prune lower-scoring genomes from consideration. Subsequently, for each read, a reduced set, called the 'shortlist,' of no fewer than D highest-scoring genomes are kept; D is a user-defined quota. More than D genomes may be included at this stage if there are many log-scores that fall within a certain fraction of the highest-scoring hit, e.g. the user may decide that, during the first pass, at least $D = 12$ hits should be considered, or any hit that receives a log-score of at least 98% of the top score.

The scores for these genomes are further refined in a second pass of the algorithm, with Q including all supported methods. The shortlisting process is performed again, this time with more stringent threshold and quota values. The final output report is generated based on the second pass shortlist, using a recursive method which ensures that the program returns larger taxonomic units (i.e., less precise predictions) if the best-scoring taxon appears to be drawn from the same distribution as its immediate relatives when subjected to a one-tailed *t*-test. This ensures that mutations at novel sites are correctly assigned to the closest known reference strain, while mutations in regions known to be highly diverse are placed more conservatively, i.e. being assigned at the species or genus level.

Supporting utilities: Genepuddle and Lincomp

Flux Simulator (Griebel et al. 2012) was adapted to simulate metatranscriptomic datasets through a Python-based pipeline called Genepuddle. Genepuddle disables Flux Simulator's eukaryote-specific features (poly-adenylation) and instructs the program to generate 100 nt unpaired Illumina-like reads with standard parameters for a list of species according to a known count profile. The result is a labeled artificial metatranscriptome in .FASTA format which is suitable for training and testing performance on any metatranscriptomic or metagenomic classifier.

The Lincomp tool, implemented in C++, produces a rank-specific accuracy report, given two files consisting of sequence labels and taxonomic IDs, with one file serving as the guess and the other as the ground truth. Using the nodes.dmp and names.dmp files from NCBI's Taxonomy database (Federhen 2012) as its reference, or equivalently a collection of Gist profiles for each relevant genome, Lincomp determines the smallest taxonomic unit that the two input files have in common for each label; e.g. a guess of 562 (species *Escherichia coli*) and a true label of 984897 (unranked strain *Shigella dysenteriae* 1) would return 543 (family *Enterobacteriaceae*). It also includes 'taxonomic grep' features, such as the ability to extract or delete a specified taxon and its descendants from an appropriately labeled FASTA file.

ACKNOWLEDGEMENTS

This work was funded by grants from Genome Canada and Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-06664). High performance computing was provided by the University of Toronto SciNet facility. We would to thank Dr. Richard Zemel (University of Toronto Department Computer Science), Dr. Anna Goldenberg (Genetics and Genome Biology, Hospital for Sick Children), Dr. Alan Moses (University of Toronto Departments of Biology, Ecology and Evolutionary Biolo-

gy, and Computer Science) and Dr. Michael Brudno (University of Toronto Department of Computer Science, *et alia*) for valuable discussions in developing the Gist algorithm.

REFERENCES

Abby S, Daubin V. 2007. Comparative genomics and the evolution of prokaryotes. *Trends Microbiol* **15**: 135–141.

Angelakis E, Armougom F, Carrière F, Bachar D, Laugier R, Lagier J-C, Robert C, Michelle C, Henrissat B, Raoult D. 2015. A Metagenomic Investigation of the Duodenal Microbiota Reveals Links with Obesity ed. M. Covasa. *PLOS ONE* **10**: e0137784.

Bazinet AL, Cummings MP. 2012. A comparative evaluation of sequence classification programs. *BMC Bioinformatics* **13**: 92.

Berger SA, Krompass D, Stamatakis A. 2011. Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Syst Biol* **60**: 291–302.

Bervoets L, Hoorenbeeck KV, Kortleven I, Noten CV, Hens N, Vael C, Goossens H, Desager KN, Vankerckhoven V. 2013. Differences in gut microbiota composition between obese and lean children: a cross-sectional study. *Gut Pathog* **5**: 10.

Brady A, Salzberg SL. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6**: 673–676.

Břinda K, Sykulski M, Kucherov G. 2015. Spaced seeds improve *k* -mer-based metagenomic classification. *Bioinformatics* **31**: 3584–3592.

Cole JR. 2004. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33**: D294–D296.

Damon C, Lehembre F, Oger-Desfeux C, Luis P, Ranger J, Fraissinet-Tachet L, Marmeisse R. 2012. Metatranscriptomics Reveals the Diversity of Genes Expressed by Eukaryotes in Forest Soils. *PLoS ONE* **7**: e28967.

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. 2009. TACOA – Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics* **10**: 56.

Federhen S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res* **40**: D136–D143.

Ferragina P, Manzini G. 2005. Indexing compressed text. *J ACM* **52**: 552–581.

Frye RE, Rose S, Chacko J, Wynne R, Bennuri SC, Slattery JC, Tippett M, Delhey L, Melnyk S, Kahler SG, et al. 2016. Modulation of mitochondrial function by the microbiome metabolite propionic acid in autism and control cell lines. *Transl Psychiatry* **6**: e927.

Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, Jiménez-Hernández N, Pérez-Cobas AE, Latorre A, Moya A. 2011. Metatranscriptomic Approach to Analyze the Functional Human Gut Microbiota. *PLoS ONE* **6**: e17447.

Greenblum S, Turnbaugh PJ, Borenstein E. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* **109**: 594–599.

Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, Sammeth M. 2012. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res* **40**: 10073–10083.

Hara N, Alkanani AK, Ir D, Robertson CE, Wagner BD, Frank DN, Zipris D. 2013. The role of the intestinal microbiota in type 1 diabetes. *Clin Immunol* **146**: 112–119.

He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR, Edgar RC, Kopylova E, Walters WA, Knight R, et al. 2015. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* **3**.

Kawakoshi A, Nakazawa H, Fukada J, Sasagawa M, Katano Y, Nakamura S, Hosoyama A, Sasaki H, Ichikawa N, Hanada S, et al. 2012. Deciphering the Genome of Polyphosphate Accumulating Actinobacterium Microlunatus phosphovorus. *DNA Res* **19**: 383–394.

Kelley DR, Salzberg SL. 2010. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics* **11**: 544.

Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. *Genome Res* **12**: 656–664.

Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*.

Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, Peet A, Tillmann V, Pöhö P, Mattila I, et al. 2015. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. *Cell Host Microbe* **17**: 260–273.

Kumar S, Blaxter ML. 2010. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* **11**: 571.

Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts, and parasites using taxon-annotated GC-coverage plots. *Front Bioinforma Comput Biol* **4**: 237.

Lesniewski RA, Jain S, Anantharaman K, Schloss PD, Dick GJ. 2012. The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME J* **6**: 2257–2268.

Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. 2005. Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* **102**: 11070–11075.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, et al. 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* **11**: 25–37.

Liu J, Wang H, Yang H, Zhang Y, Wang J, Zhao F, Qi J. 2013. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic Acids Res* **41**: e3.

Loh G, Blaut M. 2012. Role of commensal gut bacteria in inflammatory bowel diseases. *Gut Microbes* **3**: 544–555.

MacDonald NJ, Parks DH, Beiko RG. 2012. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res* **40**: e111–e111.

MacFabe DF. 2012. Short-chain fatty acid fermentation products of the gut microbiome: implications in autism spectrum disorders. *Microb Ecol Health Dis* **23**.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32**: W20–W25.

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

Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* **7**: ncomms11257.

Metwally AA, Dai Y, Finn PW, Perkins DL. 2016. WEVOTE: Weighted Voting Taxonomic Identification Method of Microbial Sequences ed. H. Tang. *PLoS ONE* **11**: e0163527.

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R. 2008. Statistical Assignment of DNA Sequences Using Bayesian Phylogenetics. *Syst Biol* **57**: 750–757.

Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* **314**: 267.

Ounit R, Lonardi S. 2016. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics* **32**: 3823–3825.

Ounit R, Wanamaker S, Close TJ, Lonardi S. 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**.

Patil KR, Roune L, McHardy AC. 2012. The PhyloPythiaS Web Server for Taxonomic Assignment of Metagenome Sequences ed. S.K. Highlander. *PLoS ONE* **7**: e38581.

Poulsen M, Schwab C, Borg Jensen B, Engberg RM, Spang A, Canibe N, Højberg O, Milinovich G, Fragner L, Schleper C, et al. 2013. Methylotrophic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen. *Nat Commun* **4**: 1428.

Rho M, Tang H, Ye Y. 2010. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* **38**: e191.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH. 2008. MetaSim—A Sequencing Simulator for Genomics and Metagenomics ed. D. Field. *PLoS ONE* **3**: e3373.

Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B. 2008. Metagenome Fragment Classification Using N-Mer Frequency Profiles. *Adv Bioinforma* **2008**.

Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. 2014. Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis. *BioMed Res Int* **2014**: e309650.

Tamames J, Moya A. 2008. Estimating the extent of horizontal gene transfer in metagenomic sequences. *BMC Genomics* **9**: 136.

Tong Y. 2015. Metagenome-wide Association Studies Potentiate Precision Medicine for Rheumatoid Arthritis. *Genomics Proteomics Bioinformatics* **13**: 208–209.

Vervier K, Mahé P, Tournoud M, Veyrieras J-B, Vert J-P. 2016. Large-scale machine learning for metagenomics sequence classification. *Bioinformatics* **32**: 1023–1032.

Wannemuehler MJ, Overstreet A-M, Ward DV, Phillips GJ. 2014. Draft Genome Sequences of the Altered Schaedler Flora, a Defined Bacterial Community from Gnotobiotic Mice. *Genome Announc* **2**: e00287-14-e00287-14.

Westcott SL, Schloss PD. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**: e1487.

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**: R46.

Wymore Brand M, Wannemuehler MJ, Phillips GJ, Proctor A, Overstreet A-M, Jergens AE, Orcutt RP, Fox JG. 2015. The Altered Schaedler Flora: Continued Applications of a Defined Murine Microbial Community. *ILAR J* **56**: 169–178.

Xiong X, Frank DN, Robertson CE, Hung SS, Markle J, Canty AJ, McCoy KD, Macpherson AJ, Poussier P, Danska JS, et al. 2012. Generation and Analysis of a Mouse Intestinal Meta-transcriptome through Illumina Based RNA-Sequencing. *PLoS ONE* **7**: e36009.