Title
Pseudoalignment for metagenomic and metatranscriptomic read assignment

Author
Schaeffer, Lorian

Publication Date
2016

Peer reviewed|Thesis/dissertation

Pseudoalignment for metagenomic and metatranscriptomic read assignment

By

Lorian Victoria Schaeffer

A dissertation submitted in partial satisfaction of

the requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Lior Pachter, Chair
Professor Gregory Barton
Professor Nicholas Ingolia
Professor Kimmen Sjolander

Fall 2016

Abstract

Pseudoalignment for metagenomic and metatranscriptomic read assignment

By

Lorian Victoria Schaeffer

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Lior Pachter, Chair

The first step in many metagenomic and metatranscriptomic analysis workflows is assigning high-throughput sequencing reads to specific strains or transcripts, providing the basis for identification and later quantification. However, the high degree of similarity between the sequences of many strains and genes makes it difficult to assign reads at the lowest level of taxonomy, and reads are typically assigned to more general taxonomic levels where they are unambiguous. Recent developments in RNA-Seq analysis have found direct-match k-mer based methods to be extremely accurate and fast when comparing sequenced RNA-Seq reads to transcriptomes. While similar methods have been used in metagenomics before now, none are highly accurate at distinguishing similar strains, and none have been applied to metatranscriptomic data. We explore connections between metagenomic and metatranscriptomic read assignment and the quantification of transcripts from RNA-Seq data to develop novel methods for rapid and accurate quantification of microbiome strains and transcripts.

We find that the recent idea of pseudoalignment introduced in the RNA-Seq context is highly applicable in the metagenomics and metatranscriptomics settings as well. When coupled with the Expectation-Maximization (EM) algorithm, reads can be assigned far more accurately and quickly than is currently possible with state of the art software, making it possible and practical for the first time to analyze abundances of individual genomes in metagenomics and metatranscriptomics projects.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thanks to my friends and family, whose support has made the completion of this dissertation possible.

Special thanks to my advisor, Lior Pachter, for his encouragement and boundless enthusiasm throughout my research, whether I had results or not. While everyone in the Pachter lab has been a delight, I am especially grateful for Shannon Hateley, who gave me considerable support during my wetlab trials, and Isaac Joseph, who helped keep me moving forward when nothing was working. I also enormously appreciate the help of Harold Pimentel, Nicolas Bray, and Páll Melsted, who were ever helpful in getting kallisto to do what I needed it to do.

Outside the Pachter lab, thanks go to Carolyn Elya, whose experience with Drosophila gut microbes was invaluable to my own investigations. Thanks also to Andrew Toseland, without whose simulated metatranscriptome data I would have missed a lot of interesting results. Finally, thanks to Diya Das for her significant help with editing everything, ever.

Chapter 1: History of abundance estimation in metagenomics and metatranscriptomics

1.1 Metagenomics

Microbial ecology, the study of microorganisms and their environmental roles, has been revolutionized by developments in sequencing technology over the last decade. Until the advent of Sanger DNA sequencing in the 1970s, the primary method to identify the composition of a microbiome was through culturing techniques, but it has become evident that culturing fails to reveal all microbial members (Dunbar et al., 2002). Sanger sequencing revealed a much larger microbial world than we had previously known about; next-generation sequencing has shown us even more detail. The ability to sequence a wide variety of microbes has opened up the ability to analyze microbial communities in great depth, and discover far more about microbial interactions than we were ever able to before.

To understand and compare microbial communities, we determine which taxa are present, and then how much of each is present. These two procedures are known as taxa identification and abundance estimation, respectively. Both are complex problems and are frequently performed together as a single step, since taxa identification can be thought of as a special case of abundance estimation where abundance is determined to be either zero or non-zero.

Taxa identification has been an important problem which has been studied long before next-generation sequencing. Before sequencing, identification was carried out by culturing microbiomes, then isolating individual microbes by dilution and identifying them by morphology. This had obvious weaknesses, most especially the fact that most microbes aren't easily culturable (Lagier et al., 2015). The net result of culture-based identification was a very limited understanding of microbiomes, focusing on the species that grow well under typical culture conditions. This led us to vastly underestimate the complexity of microbial communities, especially those from environments particularly different from culture media or those with a very low density of microbes (Dunbar, et al., 2002).

Culturing can also be used for abundance estimation via dilution and plating, although there are many additional caveats to this technique: not only are many microbes non-culturable, but some may grow too slowly to be seen and counted, and many clump together, showing only a single colony where multiple individuals were. There is also the basic problem of microbes with incompatible culturing conditions: growing conditions and media nutrients unavoidably select for some bacteria at the expense of others. All of this means that culturing-based

abundance estimation can lead to an underestimation of actual bacterial counts by several orders of magnitude (Pepper & Gerba, 2016).

The first significant step forward in identifying the full variety of microbes in microbial communities was amplicon sequencing, most commonly of the 16S ribosomal RNA gene (Fox et al., 1977; Olsen et al., 1986). This gene is convenient because it has a common structure across all prokaryotes and a number of variable regions that are more sequence similar across closely related taxa. It can be easily amplified through PCR (due to common flanking sequences across bacteria) and sequenced with Sanger sequencing (Hugenholtz & Pace, 1996). This method was the first able to effectively investigate unculturable microbiomes.

Unfortunately, while 16S sequencing is cheap and fast, it suffers from insufficient resolution when applied to taxa identification. Amplicon sequences are initially clustered based on either their similarity to other sequences in the same dataset, or their similarity to 16S gene sequences in a reference database, such as Greengenes (DeSantis et al., 2006), the Ribosomal Database Project (Cole et al., 2014), or SILVA (Quast et al., 2013). Widely adopted pipelines to handle this task currently include QIIME (Caporaso et al., 2010; Navas-Molina et al., 2013) and mothur (Schloss et al., 2009). An arbitrary threshold of sequence similarity is used to distinguish clusters. Often, 97% similarity is used for low-resolution classification; 99% similarity is considered appropriate for species-level clustering, but this is still too broad to distinguish between closely related species, like members of the *Clostridiaceae* or *Enterobacteriaceae* families (Jovel et al., 2016).

Basing taxa identification on 16S rRNA gene sequence similarity is complicated due to both sequencing error rates and the high degree of 16S conservation between some taxa, especially when using a single variable region (Figure 1.1). Analysis of existing databases of 16S sequences has demonstrated that 42% of bacterial genera contain pairs of 16S rRNA gene sequences that can't be distinguished at the 97% similarity level (Vetrovsky and Baldrian, 2013). Overall, 16S-based taxa identification is only reliable at the genus level, due to these issues. Some recent programs have been able to apply machine learning techniques to increase ability to detect small sequence differences (Callahan et al., 2015), but even in the cases where species can be determined, important functional differences between strains of the same species are entirely lost.

Even beyond the specificity issues, 16S sequencing is suboptimal for abundance estimation because 16S genetic copy number varies across taxa, sometimes inconsistently with sequence variation. Some species have single copies of their 16S rRNA gene, while others such as *Photobacterium profundum* have up to fifteen copies (Lee, Bussema, & Schmidt, 2009). Most 16S studies fail to account for this, and implicitly assume that 16S sequence abundance is equivalent to taxa abundance, skewing their analysis of community abundance and diversity. Recent programs have been developed to take copy number into account (Lee, Bussema, &

Schmidt, 2009; Angly et al., 2014; Kembel et al., 2012), but suffer from the fact that the same 16S sequence may be present in different copy numbers in different taxa, as well as the fact that 16S copy number is not known for many taxa. These issues mean that the correlation between true abundances of simulated datasets and abundances estimated using copy-number-corrected 16S sequences is often 0.80 or less. Additionally, biases in primer binding and slight variations in common primer binding sites in the 16S gene can lead to unequal amplification of different taxa, skewing abundance estimations at the sequence generation stage (Kembel et al., 2012).

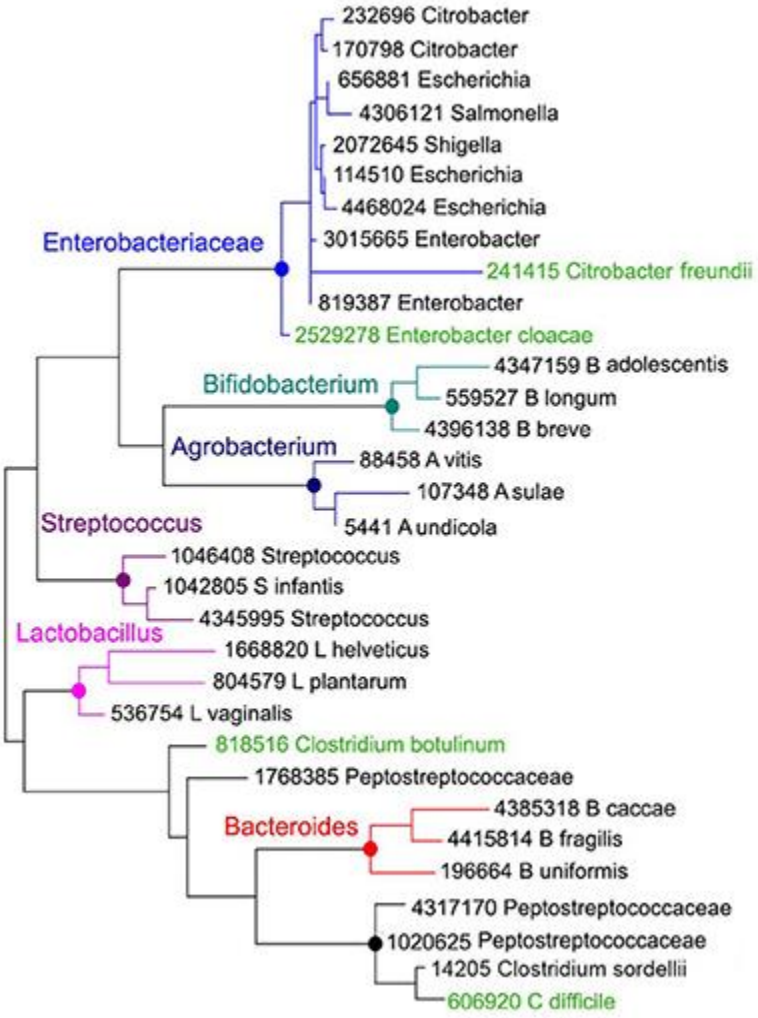

Figure 1.1. A phylogenetic tree of the V4 region of the 16S rRNA gene. Significant shared 16S sequence is present within quite a few genera. Figure from (Jovel et al., 2016).

The current alternative to amplicon sequencing is shotgun metagenomics, in which the entire metagenome is fragmented and sequenced. This gives a much more complete and representative picture of the microbiome in question, at the cost of significantly more sequencing. Using Sanger sequencing, such costs were highly prohibitive, on the order of $1000

per megabase of sequence. As a result, early shotgun metagenomic sequencing was usually utilized in highly novel environments with mostly uncultured microbes, including samples from acid mine biofilm (Tyson et al. 2004), seawater (Venter et al. 2004), deep-sea sediment (Hallam et al. 2004), and soil (Tringe et al. 2005).

Full shotgun metagenomics didn't become widely popular or cost-effective until the development of "sequencing-by-synthesis" in the mid-2000s, when Solexa (later Illumina) was able to both increase the throughput of sequencing and significantly drop the price per base; the tradeoff was reads a tenth of the length of Sanger reads (Margulies et al. 2005; Zhang et al. 2006) (Figure 1.2). In addition, as next-generation sequencing does not require cloning before sequencing, sample generation and library creation became much more straightforward and less prone to failure. While the yield, price, and ease of use were a boon to the field, the short reads led to analysis difficulties: algorithms designed for Sanger reads were inaccurate with these much shorter reads, and the much larger data volumes highlighted their slowness as well. New algorithms were needed to take full advantage of this new technology.

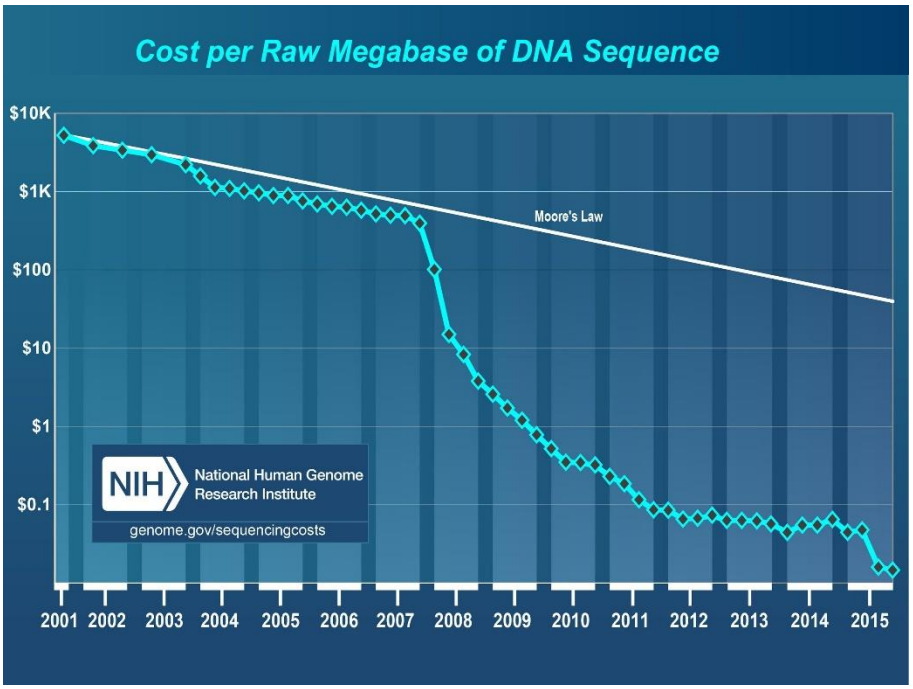

Figure 1.2. Drop in sequencing cost per megabase since 2001. The sudden decrease in sequencing costs in 2008 is due to the shift from Sanger-based to next-generation sequencing technology. The price decrease had a profound effect on the use of high-volume sequencing in a wider range of metagenomic projects. Figure from Wetterstrand, accessed 2016.

The problem of taxa identification and abundance estimation with shotgun metagenomic sequencing can be likened to being given the mixed-together pieces of several hundred similar jigsaw puzzles, and attempting to recreate the source puzzles without access to the boxes. High degrees of similarity between bacteria make this particularly difficult, especially when looking at

the species or strain level. The biggest problem is multimapping reads: reads that could have come from a number of different genomes. The methods used to address ambiguously aligned reads have grown more sophisticated over time, and have greatly improved the accuracy and resolution of taxa identification and abundance estimation.

There are several basic approaches to metagenomic taxa identification via shotgun sequencing. The most straightforward is to look at specific sets of marker genes, and use their presence or absence as a barcode to indicate taxon (Gupta & Sharma, 2015; Nguyen, et al., 2014; Sunagawa et al., 2013); this is an improvement on single amplicon sequencing, but still ignores most of the information present in metagenomic data. Methods that rely on sequence composition characteristics can use features like k-mer frequency and GC% content to identify taxa, but this is only effective for long-read sequencing methods like Sanger and 454, not the short reads of Illumina sequencing. And while assembly is one of the most effective ways to identify taxa, it requires more sequencing depth than many metagenomic studies can afford (Ghurye, Cepeda-Espinoza, & Pop, 2016). So as processing power and the pool of known bacterial genomes increases, most recent algorithms have been based on the idea of aligning sequenced reads directly to databases of reference genomes.

One of the earliest short-read alignment-based metagenomic methods, published in 2007, was MEGAN (Huson et al., 2007). MEGAN uses BLAST (or other aligners) to compare reads against a database of sequenced genomes, accepting alignments passing an E-value threshold. An important aspect of MEGAN, used in many subsequent programs, is its handling of ambiguous reads: they are assigned to the lowest common ancestor (LCA) of all likely sources. So, for instance, if a read mapped to several *E. coli* strains, MEGAN would assign the read simply to *E. coli*, rather than any specific strain (Figure 1.3). This avoids potential misassignment, and also helps account for missing genomes -- a distant LCA assignment may suggest that the actual genome wasn't present in the sample -- but often causes assignment results to be unhelpfully vague.

A subsequent alignment-based tool with improvements in abundance estimation was GAAS (Angly et al., 2009). GAAS improves on MEGAN by iteratively estimating relative genome abundance, rather than accepting the initial abundance suggested by raw read alignment. It also features more complex processing of BLAST results, statistical weighting of similar BLAST hits, and normalizing estimated abundances by genome length (which the program itself can estimate, in the case of de novo genomes). However, it still bases its taxa assignment on alignment E-values directly, with no way to optimize the choice between very similar ambiguous alignments.

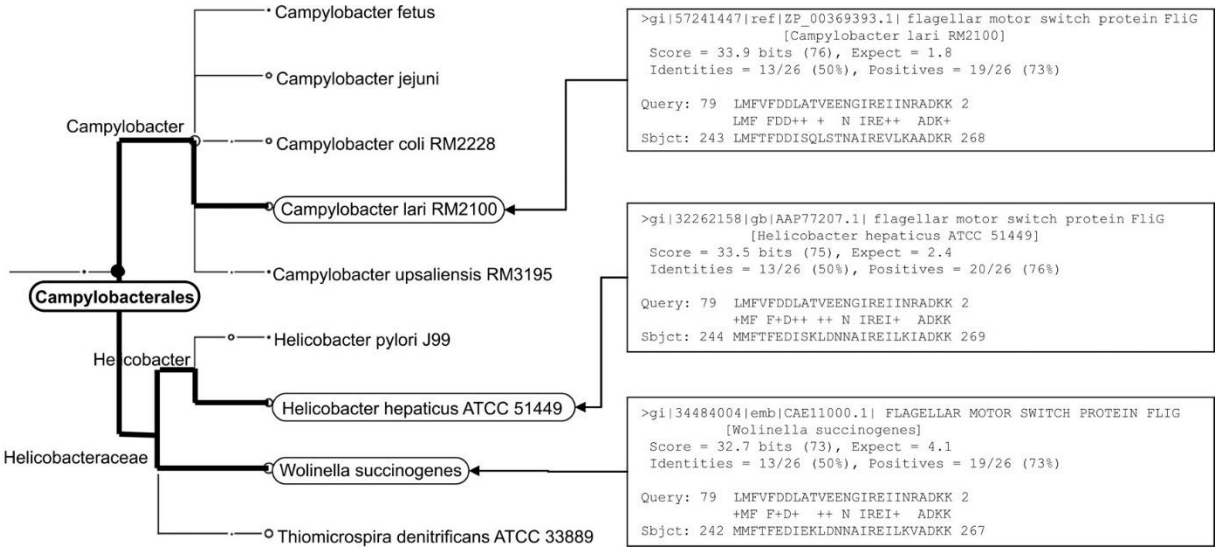

Figure 1.3. MEGAN's use of lowest common ancestor during read assignment. MEGAN traces the specific BLASTX matches (on the right) from a single read through their taxonomy, assigning the read to the taxon Campylobacterales (on the left), as it is the lowest-common taxonomic ancestor of all three matched strains (in the middle). Figure from (Huson et al., 2007).

GRAMMY (Xia et al., 2011), which explicitly models ambiguous alignments using a probability matrix incorporated into its mixture model, was a notable improvement. The mixture parameters are solved via Expectation-Maximization, taking advantage of ambiguous alignments to estimate similarity between reference genomes. While an improvement over simply taking the 'best' ambiguous alignment or the LCA, GRAMMY still has difficulty distinguishing between very similar genomes.

GASiC (Lindner et al., 2013) is one of the most accurate alignment-based algorithms, but pays for its accuracy by being extremely slow. GASiC achieves accuracy by simulating reads from each genome in a sample, then aligning them to their source genome to update basic alignment-based abundance estimation. Thus, it has multiple full alignment steps, as well as a time-consuming simulation step, but it handles very similar genomes much better than previous options.

1.2 k-mer based estimation

While full alignment-based methods can produce highly accurate taxa assignments and abundance estimations, they suffer from the slowness of standard seed-based alignment algorithms. Alignment can be reasonably fast when dealing with single eukaryote genomes, but aligning against the rapidly growing collection of sequenced prokaryotes (50,000 and counting) becomes prohibitively computationally expensive. Thus, most recent metagenomic programs have been exploring the opportunities made available by exact-match k-mer comparisons.

Sequence alignment is slow is because it has to allow for some number of mismatches to the reference genome, due to sequencing errors, SNPS, and indels. Allowing for these mismatches requires implementation of the Needleman-Wunsch or related algorithms, all of which are quadratic in the length of the sequences being aligned. However, if you cut sequences into short enough pieces, you can reasonably expect most of them to be free of errors or polymorphisms, allowing one to use very fast exact match alignment instead. Methods based on exact-match k-mers (short subsequences of length k) cut the reference genomes and/or the sequenced reads into overlapping sequences of 21 to 31 bp, and align or compare these k-mers directly.

Table 1.1. Performance of selected metagenomic read assignment tools. Fraction is the average % of simulated reads mapped. Shuffled is the average number of synthetic reads that should not be mapped that mapped. Run time is CPU time in minutes per metagenome. Correlation is average Pearson correlation coefficient between predicted and known relative abundance of phyla in the dataset. Table from (Lindgreen, Adair, & Gardner, 2016).

Analysis tool	Fraction	Shuffled	False positives	Run time	Correlation
CLARK	73.32%	340,607	0.02%	211.50	**0.9922**
EBI	0.08%	**0**	41.74%	~12 days	0.7427
Genometa	39.91%	**0**	0.83%	401	0.9136
GOTTCHA	43.10%	NA	**0.00%**	229.49	0.1777
Kraken	71.98%	19	**0.00%**	60.95	0.9915
LMAT	56.61%	1,486,699	0.63%	981.21	0.9395
MEGAN	42.21%	NA	0.49%	2489.65	0.7728
MetaPhlAn	5.09%	**0**	0.75%	108.51	0.9552
MetaPhyler	0.45%	649	0.05%	26586.15	0.7989
MG-RAST	56.17%	3	0.27%	16881.8	0.9209
mOTU	0.16%	NA	0.10%	45.8	0.9334
One Codex	73.68%	23	**0.00%**	27.77	0.9787
QIIME	58.23%	**0**	0.28%	**8.88**	0.7772
Taxator-tk	45.67%	2	14.07%	9147.92	0.8561

LMAT (Ames et al., 2013) was one of the earliest programs to apply k-mers to metagenomics. The algorithm generates a reference database of k-mers from bacterial genomes, each of which have been assigned to the lowest common ancestor (LCA) of strains containing that k-mer. This database is then simplified into a subset of k-mers, which is compared to the k-mers generated from sequencing reads in a dataset. LMAT then uses the alignment of each k-mer in a single read to determine the most likely assignment of the read as a whole, choosing the genome that contains the most k-mers from the read.

Examining only unique or high-specificity regions of the reference genomes is a popular strategy among k-mer based algorithms, as it solves or at least simplifies the ambiguously-assigned read problem. GOTTCHA (Freitas et al., 2015) uses this approach to limit the size of reference sequence: since it does not split the reference genomes into k-mers, examining only high-specificity genome regions reduces its search space and speeds up alignment. It is still a k-mer based method, as it breaks the sequenced reads into k-mers, and judges read assignment by overall matching of k-mers within those high-specificity regions.

Kraken (Wood & Salzberg, 2014), much like LMAT, assigns reference k-mers to the lowest common ancestor of matching strains. Each read is then broken into k-mers which are compared to the database; the read is assigned to the taxon with the highest number of mapping k-mers to itself and its ancestors. While sometimes this can result in assignment to a specific strain, most often it ends up assigning to higher taxonomic levels.

CLARK (Ounit et al., 2015) also breaks up the reference genomes into k-mers, but then only keeps the unique k-mers for each genome. Reads are assigned to the target with which they share the most k-mers, with a confidence score based on k-mers not shared. Unlike many of the previous programs, CLARK is able to assign most reads at a high taxonomic level, without resorting to lowest common ancestor assignment; however, its overall accuracy at any given taxonomic level is less than Kraken's.

One of the major issues faced by these k-mer based methods is resource intensiveness. While the comparison of read k-mers to reference k-mers is fast, the process of converting reference genomes into k-mers can be extremely memory intensive. Both Kraken and CLARK come in several flavors, depending on the computational resources available. As Kraken loads its entire k-mer database into memory while running, its 'memory-light' version, MiniKraken, uses a much smaller database (4GB instead of 70GB) which drops k-mers from the reference genomes, resulting in significantly less memory requirements, but sensitivity for read identification dropped significantly as well (by 11-25%). Kraken also has a 'fast' version, Kraken-Q, which only looks up one k-mer per read, and simply assigns the read to the source of that k-mer. This significantly speeds up classification, with only small drops in sensitivity and precision; however, it seems likely that this would result in more LCA assignments, and fewer strain-level assignments.

CLARK also features several versions. While the default version keeps only completely unique k-mers for each reference genome, the full version keeps all k-mers; this is slightly more accurate, but slower. CLARK-E, on the other hand, is optimized for speed, trading it for a slight drop in precision and sensitivity on most datasets. It does this by only querying non-overlapping k-mers, and assigning the read to the first target hit (possible because CLARK uses unique k-mer sets for each target). CLARK-l ("light") is a version designed to use less memory; it samples only one in 5 consecutive k-mers in each reference database target, leading to a similar amount

of memory used as MiniKraken. Finally, CLARK-S is a new version that attempts to prioritize sensitivity, by allowing specific mismatches between k-mers (Ounit & Lonardi, 2016).

The primary advantage of these k-mer based methods is speed; the more recent entries in particular can analyze a standard-sized metagenomic dataset within half an hour. But the tradeoff is a step back in accuracy, as they estimate abundance by straightforward read assignment, like MEGAN in 2007. Only GOTTCHA adjusts its estimated abundances by genome length. This of course significantly limits abundance accuracy and suggests a route for further improvement.

In one attempt to address this issue, Kraken has an additional layer called Bracken (Lu et al., 2016). Bracken takes Kraken's taxonomy tree for each read and explicitly collapses it to the most likely species-level target, based on the probability that competing reference genomes share reads. Specifically, it uses Bayesian conditional probabilities to adjust Kraken's initial assignments based on the proportion of k-mers that are unique in a genome. This significantly improves Kraken's abundance estimation, and also addresses the issue that lowest common ancestor assignments are not so helpful for abundance estimation.

Despite the progress made in both speed and accuracy of metagenomic analysis, a pervasive problem in methods development has been limited or inadequate benchmarking. Most of the above papers use simulated microbiomes with very few source genomes, ranging from as few as two (Wu & Ye, 2011) to up to 10 (Wood & Salzberg, 2014) or 20 (Ounit et al., 2015) genomes. Needless to say, these limited metagenomes are not representative of the highly complex microbiomes found in nature, containing hundreds of separate strains. This is not due to lack of sufficiently complex microbiomes; a number of artificial metagenomes with over 100 constituent genomes have been constructed from both simulated or sequenced reads. Without applying these more realistic metagenomes to new methods, it is difficult to judge the actual level of improvement over previous methods.

1.3 Metatranscriptomics

Another important contribution of modern sequencing technology to microbial communities has been the application of RNA-Seq in the form of metatranscriptomics, which attempts to do with microbial transcriptomes what metagenomics does with microbial genomes. Sequencing microbial transcripts can help determine the specific functional roles of constituents of a community, by revealing the activity level of genes of known pathways. Since meta'omic sequencing is expensive, it is often infeasible to generate both DNA-Seq and RNA-Seq libraries for a single sample; thus, metatranscriptomic data is often simultaneously used for taxa identification and abundance estimation tasks for which DNA-Seq data would normally be used.

These three tasks -- taxa identification, abundance estimation, and functional analysis -- are made harder by the difficulties of performing RNA-Seq for prokaryotes. The lack of mRNA poly-A tails makes the physical separation of mRNA and rRNA more complex and much less

reliable than in eukaryotes, meaning a much smaller percentage of reads are actually mRNA (Pascault et al., 2015; Mondav, Schmidt, & Tyson, 2010; Carvalhais & Schenk, 2013). Of course, the rRNA reads can be used for 16S taxa identification, as in the case of metagenomic reads; however, this comes with exactly the same issues as was mentioned previously, regarding insufficient resolution and copy number issues (made worse by the transcription from rDNA to rRNA).

The absence of mRNA splicing means that metatranscriptome datasets can be processed by metagenomic programs to determine taxa identification and abundance estimation. However, it is more common to use metatranscriptomic-specific pipelines which do both taxa identification and functional gene identification. Unfortunately, these programs tend to lag significantly behind the state-of-the-art in metagenomic analysis in both accuracy and speed. The ones which do not rely on 16S sequencing almost exclusively use BLAST, which is slow and handles multiple alignments poorly.

Many of the early metatranscriptomic pipelines were simply a series of scripts, such as the one by Hamamura and Meneghin (2010), which is a combination of perl scripts and user-run BLAST queries on various databases. Taken together, they do an admirable job of cleaning the raw reads, identifying 16S rRNA sequences, and assigning functional categories to the mRNA sequences. However, the process is extremely slow (taking many days, even when multithreaded on a powerful server) and error-prone due to the many individual steps. Other similar pipelines are those released by Goncalves et al. (2011), Friedman and Maniatis (2011), and Leimena et al. (2013), all of which suffer from similar problems of ease of use and speed.

Recent improvements in usability include MetaTrans, SAMSA, and COMAN. Both MetaTrans and SAMSA are more coherent pipelines than previous analysis options; while both are script-based, they offer essentially end-to-end coverage of all required steps, bringing together a number of programs for cleaning, filtering, aligning, and annotating metatranscriptomic data, and requiring minimal configuration for standard use cases.

MetaTrans (Martinez et al., 2016) determines functional abundance of transcripts as well as using 16S rRNA sequences present in the dataset to identify taxa. Its functional assignment utilizes gene prediction and clustering, followed by mapping to the MetaHIT database, which is specific for human gut microbiome genes (Qin et al., 2010). Mapping offers the option of using SOAP2 or DIAMOND, an improvement in speed over the standard use of BLAST. It can also perform differential analysis on multiple conditions using DESeq2, a standard R package for RNAseq differential analysis (Love, Huber, & Anders, 2014).

Building on top of MG-RAST, SAMSA (Westreich et al., 2016), a metagenomic taxa identification platform, can break down transcriptional activity by organism or by function. It uses MG-RAST (Meyer et al., 2008) for alignment and annotation, which itself uses translated protein clustering, then uses BLAT to find the closest reference match. While MG-RAST uses e-

values to determine best match, it does keep all matches with equal max e-values, allowing for some amount of multi-aligning. Final transcript abundances, however, are judged solely based on raw counts. While the pipeline does not itself perform differential analysis, it creates output files that can be imported into DESeq. Some caveats are that the pipeline requires FASTQ input, and reads of at least 100bp, although it can join overlapping paired-end reads to achieve this if necessary.

Additional efforts to minimize required setup have been made with online pipelines such as COMAN (Ni, Li, & Panagiotou, 2016). While COMAN only does functional analysis, not taxa identification or abundance estimation, it does judge the contribution of taxa to function if provided with taxa abundances determined elsewhere. In order to speed up processing, COMAN uses the aligner DIAMOND (Buchfink, Xie, & Huson, 2015) rather than BLAST or BLAT. It then uses genome annotations to determine functional contribution, and can infer pathways as well as determine enriched or depleted functions when comparing conditions.

One of the chief problems with judging the performance of metatranscriptomic programs is the lack of commonly-shared simulated metatranscriptomic datasets. Transcriptome simulation is a difficult problem, and most simulated datasets are not made publicly available, so authors of different programs can't easily compare their results against the same dataset. This is compounded by the slowness and complexity of most of the pipelines listed above; installing and running each of them for comparison purposes is prohibitive. This means that while metagenomic programs often compare themselves directly to other options, for both accuracy and speed, none of the above pipelines have any published head-to-head comparisons. The exception is MetaTrans, which compared itself to metagenomic programs MG-RAST and Kraken, but only for biological datasets, not datasets where ground truth was known. That said, it is clear there is definite room for speed and accuracy improvements, especially as the field has not progressed very far past BLAST, conceptually.

Chapter 2: Pseudoalignment for metagenomic read assignment[1]

2.1 Introduction

The analysis of microbial communities via whole-genome shotgun sequencing has led to exceptional bioinformatics challenges (Chen & Pachter, 2005) that remain largely unsolved (Sholz et al., 2012). Most of these challenges can be characterized as "*de novo*" bioinformatics problems: they involve assembly of sequences, binning of reads, and annotation of genes directly from sequenced reads. The emphasis on *de novo* methods a decade ago was the result of a paucity of sequenced reference microbial and archaeal genomes at the time. However, this has begun to change in recent years (Land et al., 2015). As sequencing costs have plummeted, the number of fully sequenced genomes has increased dramatically, and while a large swath of the microbial world remains uncharacterized, there are now thousands of "reference quality" genomes suitable for the application of reference-based methods.

One of the fundamental metagenomics problems that is amenable to reference-based analysis is that of "sequence classification" or "read assignment". This is the problem of assigning sequenced reads to taxa. The MEGAN program (Huson et al., 2007) was one of the first reference-based read assignment programs and was published shortly after sequencing-by-synthesis methods started to become mainstream. It provided a phylogenetic context to mapped reads by assigning reads to the lowest taxonomic level at which they could be uniquely aligned, and became popular in part because of a powerful accompanying visualization toolkit. One of the drawbacks of MEGAN was that its approach to assigning ambiguously mapped reads limited its application to quantification of individual strains, an issue which was addressed in a number of subsequent programs, for example GRAMMy (Xia et al., 2011) and GASiC (Lindner et al., 2013), which were the first to statistically assign ambiguously mapped reads to individual strains. Unfortunately, these approaches all relied on read alignment, a computational problem that is particularly difficult in the metagenomic setting where reference genome databases are large and read sets gigantic.

In a breakthrough publication in 2014 (Wood & Salzberg) it was shown that it is possible to greatly accelerate read assignment utilizing fast k-mer hashing to circumvent the need for read alignment. An implementation called Kraken was used to show that analyses that previously took hours were tractable in minutes, and the removal of the read alignment step greatly simplified workflows and storage requirements. However the Kraken speed came at a cost. An examination of the Kraken algorithm and output reveals that the method takes a step back from GRAMMy and GASiC by discarding statistical assignment of reads at the strain level in favor of direct

[1] This chapter is joint work with Harold Pimentel, Nicolas Bray, Páll Melsted and Lior Pachter, and this material has been included with their permission.

taxonomic assignment as in MEGAN. The net effect is that while Kraken is more accurate than MEGAN (Lindgreen & Renard, 2015), it is unsuitable for quantification. This is because, unlike GASiC, Kraken is strictly designed to be a read assigner: its only output is a file listing the taxonomic assignment for each read. A natural question to ask is whether the strengths of Kraken and GASiC can be combined, i.e. whether it is possible to leverage fast k-mer based hashing to map reads not at the taxonomic but at the strain level, while assigning the resulting ambiguously mapped reads using a statistical framework that allows for probabilistic assignment of reads.

To answer this question we turned to RNA-Seq (Cloonan et al., 2008; Lister et al., 2008; Nagalakshmi et al., 2008, Mortazavi et al., 2008), an experiment for which there has been extensive methods development that we hypothesized could be adapted and applied to metagenomics. Many of the challenges of metagenomic quantification translate to problems in RNA-Seq via a dictionary that replaces genome targets with transcript targets. For example, ambiguously mapped genomic reads that are difficult to resolve at the strain level in the metagenomics setting are analogous to reads that are difficult to assign to specific isoforms in RNA-Seq. Statistical questions at the heart of "comparative metagenomics" (Huson et al., 2009; Rodriguez-Brito et al., 2006; Tringe et al., 2005) are analogous to the statistical problems in differential expression analysis. In fact, the only significant differences between metagenomics and RNA-Seq are that genome sequences are much larger than transcripts and reference databases are less complete. These differences have engineering implications, but statistically and computationally, metagenomics and transcriptomics are very much the same.

In this chapter we show that technology transfer from RNA-Seq to metagenomics makes it possible to perform read assignment both rapidly *and* accurately. Specifically, we show that it is possible to accurately assign reads at the strain level using a fast k-mer based approach that goes beyond the hashing of Kraken and takes advantage of the principle of pseudoalignment (Bray et al., 2015). The idea of pseudoalignment originates with RNA-Seq, where it was developed to take advantage of the fact that the sufficient statistics for RNA-Seq quantification are assignments of reads to transcripts rather than their alignments. The same applies in the metagenomics setting, and we show that, just as in RNA-Seq, application of the EM algorithm to "equivalence classes" (Nicolae et al., 2011) allows for accurate statistical resolution of mapping ambiguities. Using a published simulated dataset, a biological dataset from the human microbiome project, and an implementation of pseudoalignment coupled to the EM algorithm in kallisto, we demonstrate significant accuracy and performance improvements in comparison to state-of-the-art programs.

2.2 Results

To test the hypothesis that RNA-Seq quantification methods can be applied in the metagenomics setting we began by examining the performance of eXpress, a program that

implements a streaming EM algorithm for RNA-Seq read assignment from alignments, on simulated data (Roberts & Pachter, 2013). We chose eXpress because it utilizes traditional read alignments directly to a transcriptome but is more memory efficient than other approaches (e.g. RSEM (Li & Dewey, 2011)) and therefore more suitable in the metagenomics setting. Other RNA-Seq quantification tools such as Cufflinks (Trapnell et al., 2010) were not suitable for our needs because of their dependence on read alignments to genomes and not transcriptomes, a requirement that does not translate easily to the metagenomics setting.

To test eXpress we aligned a simulated dataset of Illumina-like reads from 100 microbial genomes to a reference database containing only those genomes, allowing us to compare results to a ground truth (the Illumina100 data) (Mende et al., 2012). We began by comparing eXpress to GASiC, which also utilizes read alignments for read assignment. The results are shown in Table 2.1. We found that eXpress outperforms GASiC at the exact genome, species, genus, and phylum levels, which we believe is because the statistical model of eXpress takes into account data-dependent read error profiles in assigning reads.

A major problem with GASiC and eXpress is that the alignments they require are slow to generate. The alignments, made with Bowtie2 (Langmead et al., 2012), took days. As reported in Kraken and the follow-up, Bracken, which has been specialized for quantification (Lu et al., 2016), significant speed-ups are possible using hashing methods. The programs require only 35 minutes 39s to assign reads and then estimate abundances at the species level. We also tested CLARK (Ounit et al., 2015), another recently published k-mer based assignment tool and, in agreement with the benchmarks in (Lindgreen et al., 2015), we found it to be slightly faster, taking 20 minutes 30s to estimate abundance. Kallisto was the fastest of all programs tested, with a run time of 5 minutes 55s. As seen in Table 2.1, both Bracken and CLARK have noticeably worse performance than both eXpress and kallisto.

We next turned to a comparison of kallisto with Bracken and CLARK using the Illumina100 simulated data (i100) but using a full, more realistic reference database of 29,698 bacterial genomes from Ensembl (Kersey et al., 2016). In order to handle such a large database, which is significantly larger than the maximum index size for all three programs, we first performed a pre-filtering step using recently-published metagenome distance estimator Mash (Ondov et al., 2016) (see methods for details). Mash filtered the 29,698 genomes down to 1027 genomes which were judged closest to the i100 reads being quantified; those 1027 genomes contained 83 out of the 100 "true" strains present in the i100 dataset.

The results of estimating reads from all 100 genomes against the Ensembl-based index, listed in Table 2.1 (where the database is called "Ensembl") and Figures 2.1 and 2.2, show that kallisto is significantly more accurate than CLARK at all taxonomic levels, and is only outmatched by Bracken at the genus level. The dramatic decrease in error from the exact genome

to species level (from 17.59% to 1.26%) indicates that kallisto is correctly assigning the reads from the missing strains to closely related strains from the same species.

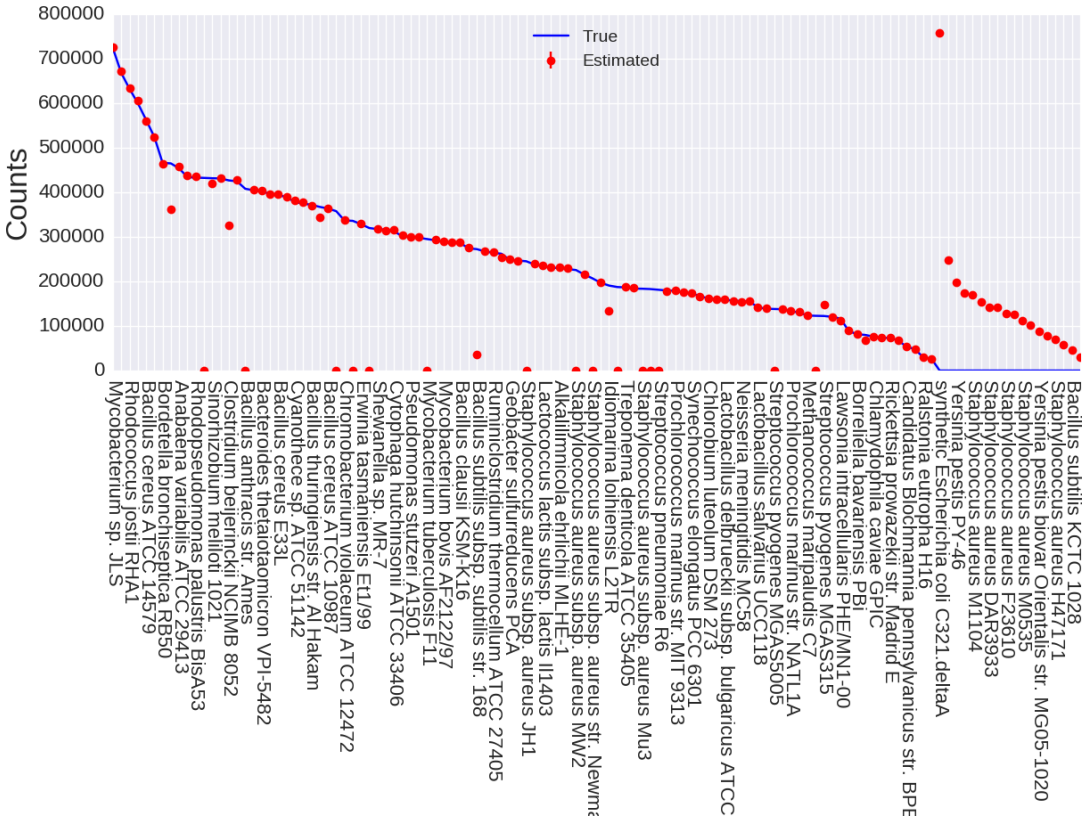

Figure 2.1. Results of kallisto on simulated reads from the Ensembl dataset at the exact genome level.

Even at the exact genome level (where neither Bracken nor CLARK offer estimates), kallisto performs well, given the restriction of missing 17% of the actual genomes present in the reads. To check the effect of the missing genomes on accuracy, we ran kallisto on the i100 reads only from the present 83 genomes and achieved an impressive AVGRE of 2.59% at the exact genome level. Even more promisingly, the species-level error of this 83-genome dataset is 0.77%, which is quite close to the 1.26% species-level error of the full 100-genome dataset. This further supports kallisto's accuracy in assigning reads from missing genomes to closely related genomes.

Mash took 362 minutes on a single core to index the full 30k Ensembl genomes, and another 130 minutes to compare the i100 reads against those genomes; these steps are easily parallelized to multiple cores. On a single core, kallisto was slower than CLARK but faster than Bracken -- kallisto took 111 minutes to index and 60 minutes 40s to quantify, while CLARK took 131 minutes 35s to both index and quantify, and Bracken took 235 minutes 35s to index and 169 minutes 29s to quantify.

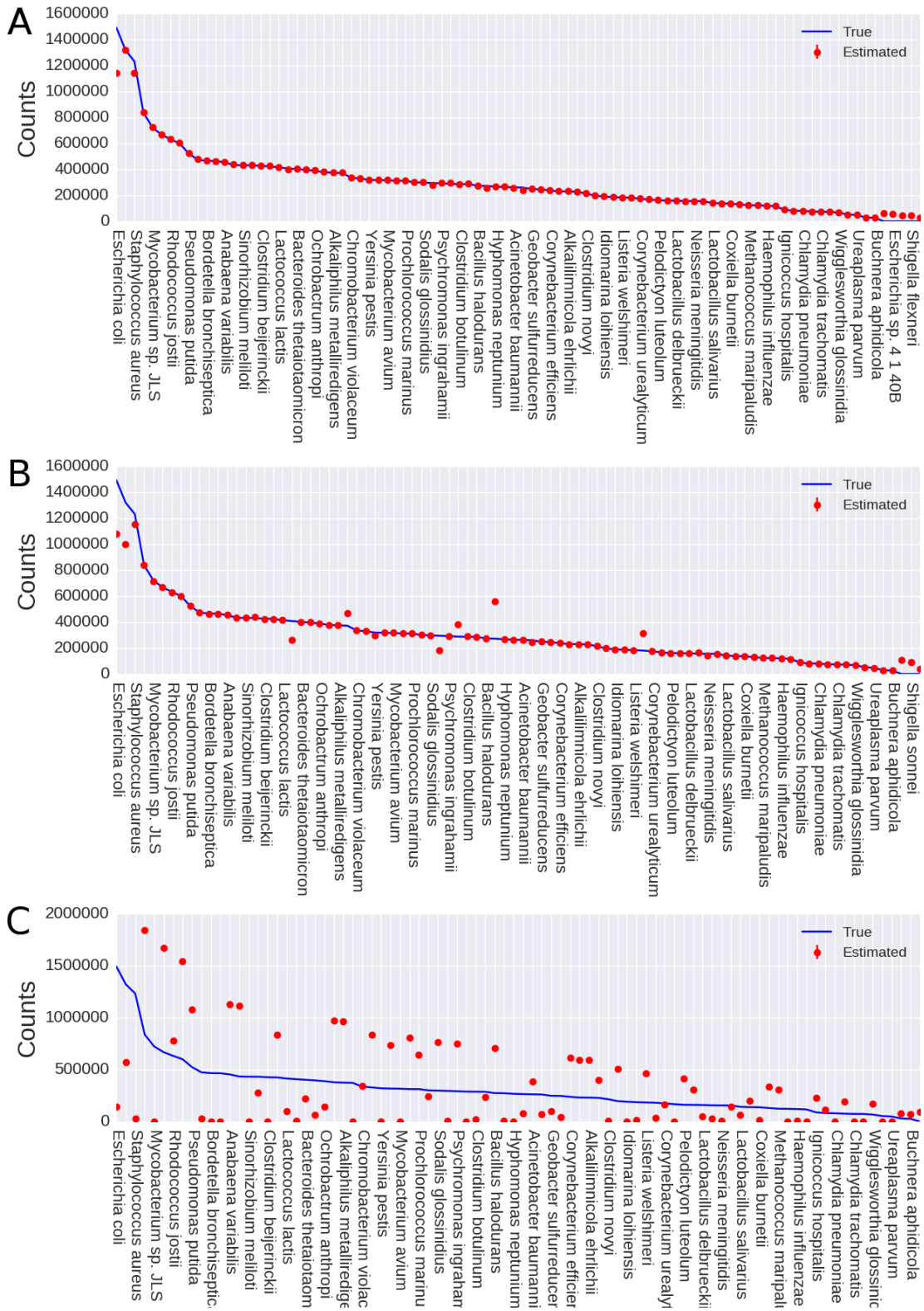

Figure 2.2. Comparison of species-level abundance estimation between metagenomic programs. Results of kallisto (a), Bracken (b) and CLARK (c) on simulated reads from the Ensembl dataset at the species level.

Table 2.1. Normalized count-based classification accuracy at four taxonomic ranks. CLARK and Bracken results are missing at the strain level because they do not output strain level counts. Calculated errors are Average Relative Error and Relative Root Mean Square Error.

	Exact Genome		Species		Genus		Phylum	
	AVGRE	RRMSE	AVGRE	RRMSE	AVGRE	RRMSE	AVGRE	RRMSE
i100								
kallisto	0.97	5.42	0.14	0.36	0.13	0.38	0.09	0.10
Bracken	–	–	1.94	9.51	2.21	10.78	0.91	0.92
CLARK	–	–	12.28	22.73	10.32	18.22	7.52	7.88
GASiC	7.21	19.31	3.80	10.46	3.72	11.43	2.52	3.10
eXpress	2.57	11.92	0.40	0.61	0.34	0.57	0.13	0.18
Ensembl								
kallisto	17.15	39.32	1.26	3.01	0.98	2.17	0.72	0.76
Bracken	–	–	4.94	16.22	1.10	3.97	0.35	0.38
CLARK	–	–	59.15	72.40	52.68	67.04	45.44	56.76

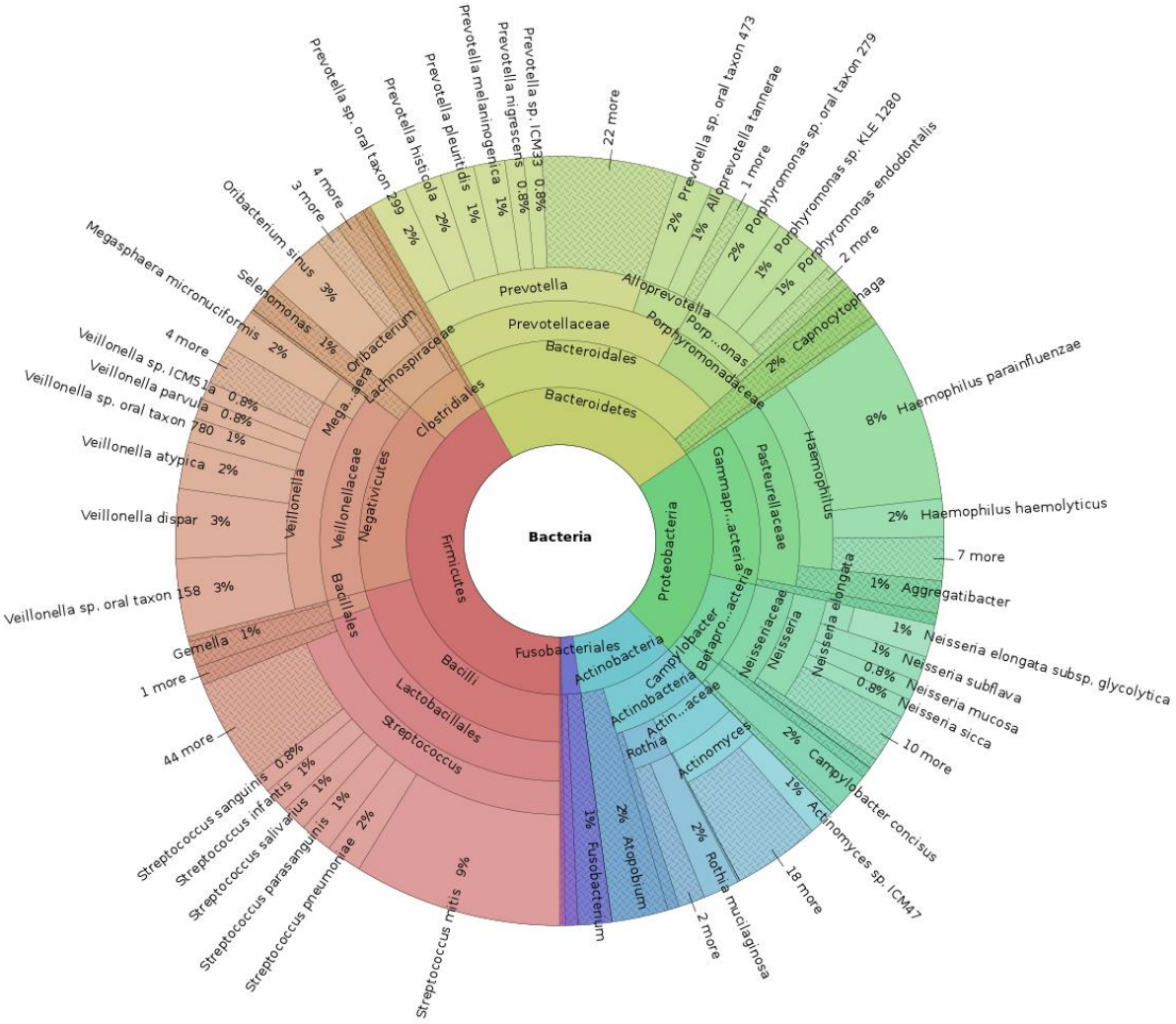

Figure 2.3. Results of kallisto on bacterial reads in human saliva samples at all taxonomic levels.

To test the performance of kallisto on biological data, we analyzed a set of saliva samples from the Human Microbiome Project. These three samples -- SRS014468, SRS015055, and SRS019120 -- consist of a total of 9.3 million 60-100bp paired-end reads, collected from three separate individuals. We pooled them together to analyze the microbes present in the general saliva microbiome. Running the same Mash-based pipeline on 30k Ensembl genomes identified 744 likely genomes, and using kallisto to quantify the saliva reads against those genomes found primarily bacteria of the genera *Streptococcus* (17.5%), *Prevotella* (17.1%), *Veillonella* (11.2%), and *Haemophilus* (9.9%) as well as a number of less abundant genera (shown in Figure 3). The most abundant species are those known to be abundant in the oral microbiome: *Streptococcus mitis*, *Haemophilus parainfluenzae*, *Veillonella sp. oral taxon 158*, and *Prevotella histicola*.

2.3 Methods

Illumina100 dataset

We tested kallisto and alternate programs on a set of simulated reads published in (Mende et al., 2012). The Illumina100 dataset consists of 53.33 million 75bp reads, simulated by the iMESSi metagenomic simulator using an Illumina error model. The reads were simulated from a set of 100 unique bacterial genomes. The set is of genomes from 85 different species and 63 different genera, over a range of abundances from 0.86% to 2.2%.

Reads were trimmed with the program Trimmomatic (version 0.32) (Bolger et al., 2014) to a minimum length of 40bp, using its adaptive trimming algorithm MAXINFO with a target length of 40 and default strictness. Trimming was very permissive, and only 40 reads were dropped due to quality issues.

Taxonomic identification

We analyzed each program's output at four taxonomic ranks: phylum, genus, species, and "exact genome" level. The latter tests the abundance estimation of the actual Illumina100 genomes, which are a combination of strains and substrains and thus aren't taxonomically well defined. The other three ranks are as assigned by NCBI's Taxonomy Database, as of August, 2016.

Count estimation accuracy calculation

Using a simulated dataset with known abundances allowed us to benchmark programs by comparing program outputs with true values for each genome. While kallisto is able to output length-corrected individual genome abundances, most of the programs we compared with only counts, so for consistency we analyzed the accuracy of assigned or estimated counts for each program. We normalized the estimated counts by the percent of assigned reads in order to be able to compare relative count estimates between programs.

We primarily used the error measures AVGRE (Average Relative Error), which computes the mean of the difference between truth and estimate, and RRMSE (Relative Root Mean Square Error), which computes the root mean square average of the difference between truth and estimate, to judge the accuracy of our estimates. Formally, with *n* true genomes/species/genera/phyla, true counts $\tau_i \ (1 \leq i \leq n)$ and estimated counts t_i at the rank, and *A* aligned reads out of *T* total reads we computed

$$AVGRE = \frac{1}{n}\sum_{i}^{n}\frac{|t_i \cdot \frac{T}{A} - \tau_i|}{\tau_i} \quad \text{and} \quad RRMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}\left(\frac{t_i \cdot \frac{T}{A} - \tau_i}{\tau_i}\right)^2}$$

The scripts used to compile the results are available at *http://github.com/pachterlab/metakallisto*

Reference genome database

In addition to aligning the Illumina100 reads against their originating genomes, we tested the more realistic case of aligning against a large bacterial database -- Ensembl's bacterial genomes as of version 30. All 29,698 bacterial genomes were downloaded, combined with the i100 genomes, and used as-is with Mash (see below). For abundance estimation with Bracken, CLARK, and kallisto, constituent contigs, chromosomes, and plasmids were concatenated together with a series of 10 ambiguous bases represented as N, and NCBI's taxonomic ID was manually added to the headers for Kraken's use.

Mash genome pre-filtering

To lower the number of genomes to index to a reasonable level, we ran the Illumina100 dataset against all 30,000 Ensembl genomes using Mash, a genome distance calculator. We used only the top 10 genomes from each species that were judged closest to the reads in subsequent abundance estimation, to get a reasonable number of genomes for indexing.

The scripts used to filter the genomes based on Mash results are available at *http://github.com/pachterlab/metakallisto*

2.4 Conclusions

The idea of translating RNA-Seq methodology to and from metagenomics was, to our knowledge, first proposed in (Paulson et al., 2013) where statistical methods for identifying differential abundances in microbial marker genes were developed. In that paper, there were comparisons between the proposed metagenomics method and RNA-Seq differential analysis methods implemented in DESeq (Anders & Huber, 2010) and edgeR (Robinson et al., 2010).

Notably, the central idea of the paper, the specific consideration of zero inflated distributions to account for undersampling, is also used in single cell expression analysis (McDavid et al., 2013).

Our results show that RNA-Seq methods for quantification are also applicable in the metagenomics setting, and our results with kallisto demonstrate that it is possible to accurately and rapidly quantify the abundance of individual **strains**. With a few exceptions, e.g. (Bradley et al., 2015), most metagenomic analyses have focused on higher taxonomy, a point highlighted in the recent benchmarking paper (Lindgreen et al., 2015) which compares predictions at the phylum level because "[comparisons at that level are] less prone to differences". The phylum level is four levels removed from genus, let alone species or strain. Our results suggest that the door is now open to metagenome analyses at the highest possible resolution.

While our benchmarks are primarily based on simulated data, our experiments are much more realistic than previous analyses. For example, the Kraken and CLARK papers report results on simulations with ten genomes, whereas we have simulated from 100 genomes and mapped against nearly 30,000. One of the difficulties we faced in our analyses was the technical issue of taxonomic naming and annotation in collating results. This seemingly trivial matter is complicated by the lack of attention paid to low taxonomic level analysis in previous studies.

As reference databases grow in size, there will be continued challenges in quantification and downstream analysis. While the two-step Mash-kallisto workflow we have described here can scale for the time being, novel algorithmic ideas are needed to that can leverage large databases for individual genome analysis, yet efficiently discard irrelevant information.

Chapter 3: K-mer based metatranscriptome analysis

3.1 Introduction

While metagenomic sequencing and analysis is useful for determining which microbes are present in a microbiome, it does not tell us the functional activity of these communities. The best approach to answering that question is metatranscriptomic sequencing: sequencing the coding mRNA of a microbial community. This area of microbial ecology is younger and less well-researched than metagenomics, in large part because of the additional challenges of enriching for prokaryotic mRNA: the inability to fully remove rRNA sequences from microbial total RNA means that deeper sequencing is required to get a useful number of mRNA reads, because up to 91% of reads can still be rRNA (He et al., 2010).

In addition to the technical difficulties with metatranscriptomics, the pipelines available for analysis remain limited and based almost entirely on BLAST and standard alignment algorithms, with no significant handling of ambiguous reads. Because there is a significant amount of shared sequence between bacterial genes, this is a particularly challenging problem. As a result, metatranscriptomic pipelines are slow and often awkward to run, with an inability to make use of the full range of bacterial reference sequence. Several recent pipelines are all or primarily available online, to deal with the issues of required computational resources and complex install procedures. While this is effective in some respects, it has its drawbacks: users can only utilize the database provided, and are unable to control when their data is analyzed.

Following kallisto's success at metagenomic read assignment (Schaeffer et al., 2015), in this chapter, we apply kallisto to the related problem of metatranscriptome read assignment. At first glance, metatranscriptomics is even more similar to the RNA-Seq analysis kallisto was designed for than metagenomics; however, in addition to the size problem common to all meta'omics analysis, there is also the difficulty of high transcript similarity between transcriptomes. In general, coding sequences are more strongly conserved than non-coding sequences, so metatranscriptomes focus on sequence regions that often show little variation between strains. As kallisto has no way of knowing which transcripts are grouped into strains, it does not preferentially identify transcripts from a smaller number of strains, and so distinguishing between similar transcripts from multiple genomes becomes very difficult. To address this problem, we use a two-step process: an identification stage and a quantification stage. The identification stage (pseudoaligning against a wide variety of targets) attempts to narrow down the possible strains present in the sample, and the quantification stage (pseudoaligning against the strains the previous stage identified as present) uses kallisto's pseudoalignment and EM algorithm to judge how much of each transcript is in the sample.

This two-step process is also intended to address kallisto's biggest flaw: large indexes for pseudoalignment take an enormous amount of memory. Indexing several thousand genomes

takes hundreds of gigabytes of RAM, prohibitive even for large computing resources. While this is not a problem for traditional RNA-Seq, it is clearly a big issue for meta'omics, which often has tens of thousands of reference genomes or transcriptomes. Ideally, the identification stage can accommodate more index flexibility (and thus smaller indexes), because it is only needed for a present/absent determination, while the quantification stage will need a smaller index because it's only comparing between a small number of pre-identified strains.

3.2 Results

In order to test whether kallisto could successfully calculate transcript abundance at all, we pseudoaligned a simulated 7.5-million read single end metatranscriptome dataset directly to its 109 source transcriptomes. As not all of the transcriptomes were available from Ensembl, we removed the simulated reads that were derived from unavailable transcriptomes; with this modification, 72% of the simulated reads were assigned by kallisto. After summing the transcripts into their source transcriptomes, overall accuracy at the strain level was nearly

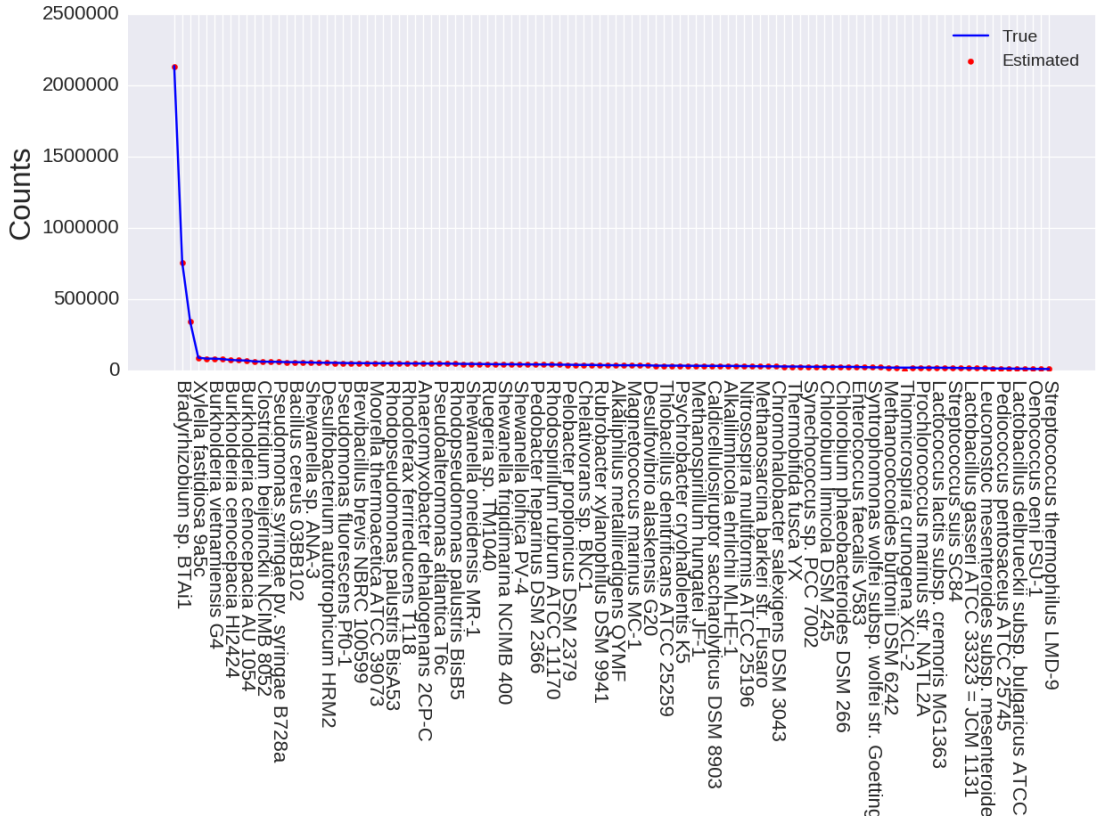

Figure 3.1. Estimated counts at strain level aligned against present transcriptomes. True (blue line) and estimated (red dots) strain-level counts of simulated metatranscriptome reads pseudoaligned against only transcriptomes present in dataset.

perfect, with an average relative error of 0.86%; this is comparable to the accuracy seen at a metagenomic level.

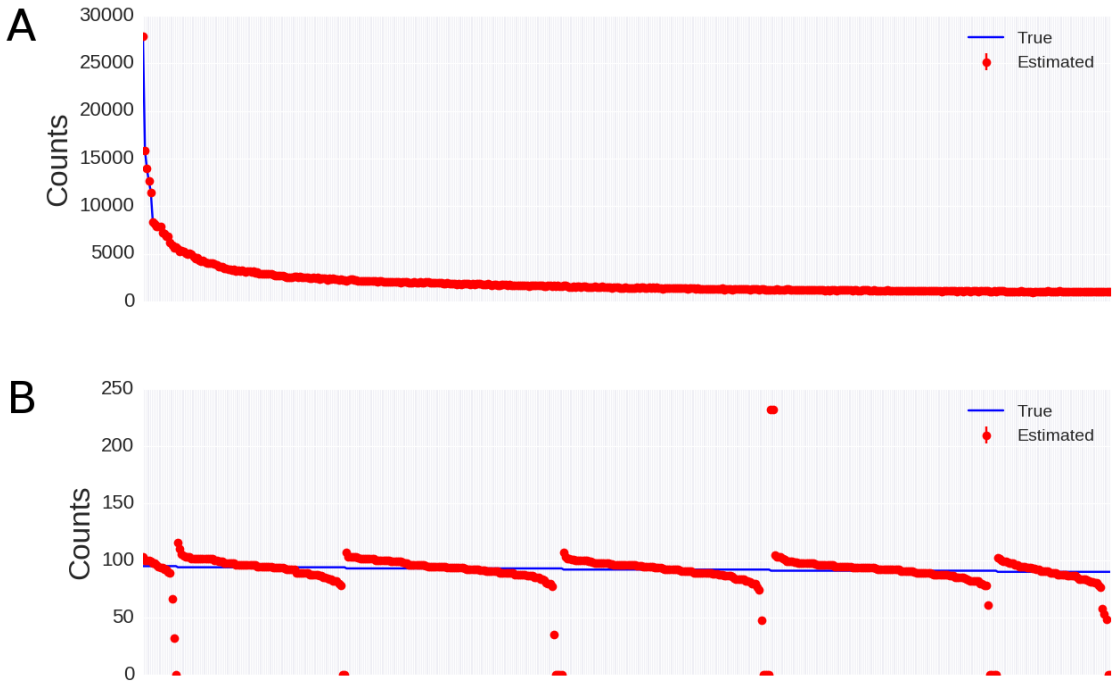

Figure 3.2. Estimated counts of transcripts in simulated data. True (blue line) and estimated (red dots) transcript-level counts of simulated metatranscriptome reads pseudoaligned against only transcripts present in dataset. ***(a)*** *High abundance transcripts with thousands of reads present in dataset;* ***(b)*** *Low abundance transcripts with approximately 100 reads present. Kallisto shows high accuracy estimating the abundance of with highly expressed transcripts, but produces significant errors with lower abundance transcripts. Note that the apparent structure of the errors is an artifact of sorting by first "true" counts, then by estimated counts.*

At the transcript level, calculated errors were much worse, with an average relative error of 34%. Some of the error is due to recent changes in the transcriptomes, causing them to no longer fully match the simulated data being used; this is likely responsible for some of the 28% of the reads failing to align at all. As seen in Figure 3.2, actual alignment was reasonable among the most abundant transcripts, but varied much more in the lower abundance transcripts. It is clear that the errors within a transcriptome balance out, since the above results summed by genome are extremely accurate; this suggests that the abundance estimation errors are essentially randomly distributed, rather than being systemic to certain strains or transcripts.

Despite the lackluster transcript abundance estimation, the highly accurate strain-level abundance estimation was encouraging. In order to test the ability of kallisto to identify the source strains from a large set of possible strains, we created a representative transcriptome index, containing the transcriptome of a single strain from every available species. Pseudoalignment of the simulated dataset to this representative transcriptome index, however,

gave very poor species-level identification results, as seen in the below graph. Any read that did not have an exact-match strain was likely to be misassigned, resulting in a large amount of false positives and an average error of 66%.

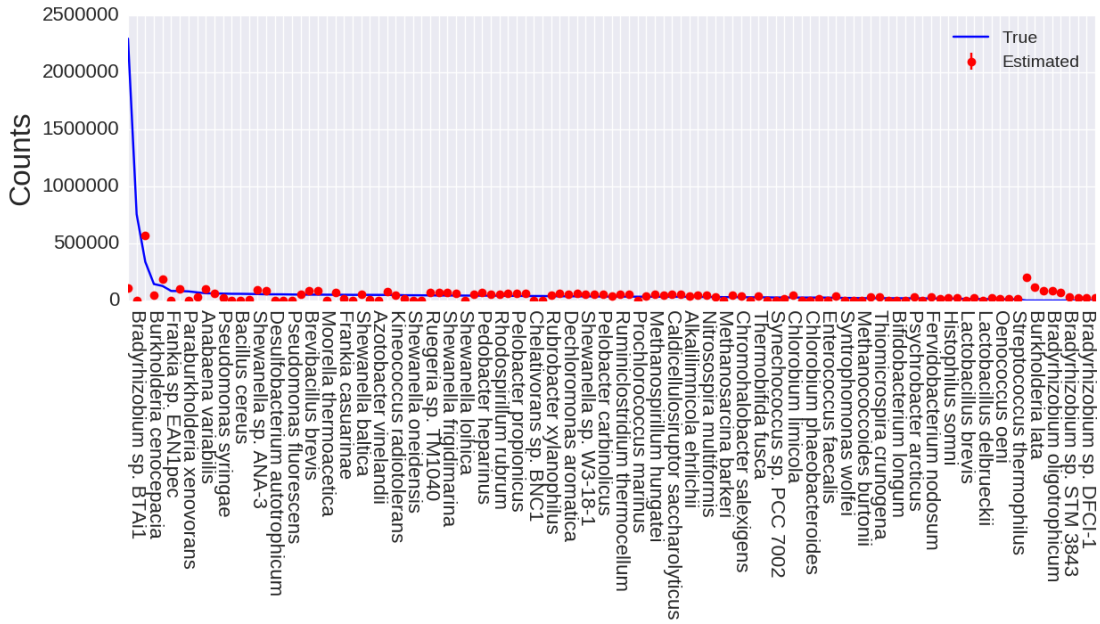

Figure 3.3. Estimated counts at species level aligned against representative transcriptomes. True (blue line) and estimated (red dots) strain-level counts of simulated metatranscriptome reads pseudoaligned against a representative transcriptome index containing only a single strain from 4,412 species.

The above issues are most likely a result of kallisto failing to take into account that transcripts come from single strains. From kallisto's point of view, if a read could come from two transcripts, there's no reason to prefer pseudoaligning to one versus the other, and most likely, kallisto will end up concluding that two identical transcripts are present in equal abundances in the metatranscriptome. However, from an external viewpoint, it would be better if it gave extra weight to transcripts that come from strains with multiple transcripts present in the sample. Future plans include adapting kallisto's model to regularize abundance across genomes, by including the ability to pass along information to kallisto of the form "these transcripts are linked, and should be present or absent as a group". This could be implemented as a penalty for transcripts with highly unbalanced coverage within a single strain during the expectation-maximization step. In the short term, we tested identifying strains using genomic pseudoalignment to improve accuracy of metatranscriptome abundance estimation.

Pseudoaligning metatranscriptome reads to bacterial genomes should lead to sufficiently accurate identification for pre-filtering potential source genomes, since kallisto doesn't consider coverage when determining abundance, thus the uneven coverage of a transcriptome is not a problem. Additionally, as these are prokaryotes, there are no splicing issues to confuse genome-level alignment. For identification purposes, we created a species-level representative index,

picking one strain from each species present in Ensembl. Due to the fact that Ensembl has fewer genomes than transcriptomes, some of the source transcriptomes of my simulated dataset were not available as genomes (even at the species level), so the dataset used for this test had 30 additional strains removed, leaving a total of 80 strains remaining.

Kallisto's performance on the simulated dataset, when aligning to the representative genome index, showed an impressive degree of strain- and species-level accuracy in its abundance calculations, given that many of the true source strains of the simulated reads were not present. At the species level, average relative error was 27%, with only 9% of the reads being attributed to species not actually present in the sample. As can be seen in the species-level graph below, the overall trend of abundance estimation is reasonably accurate, given the limitations.

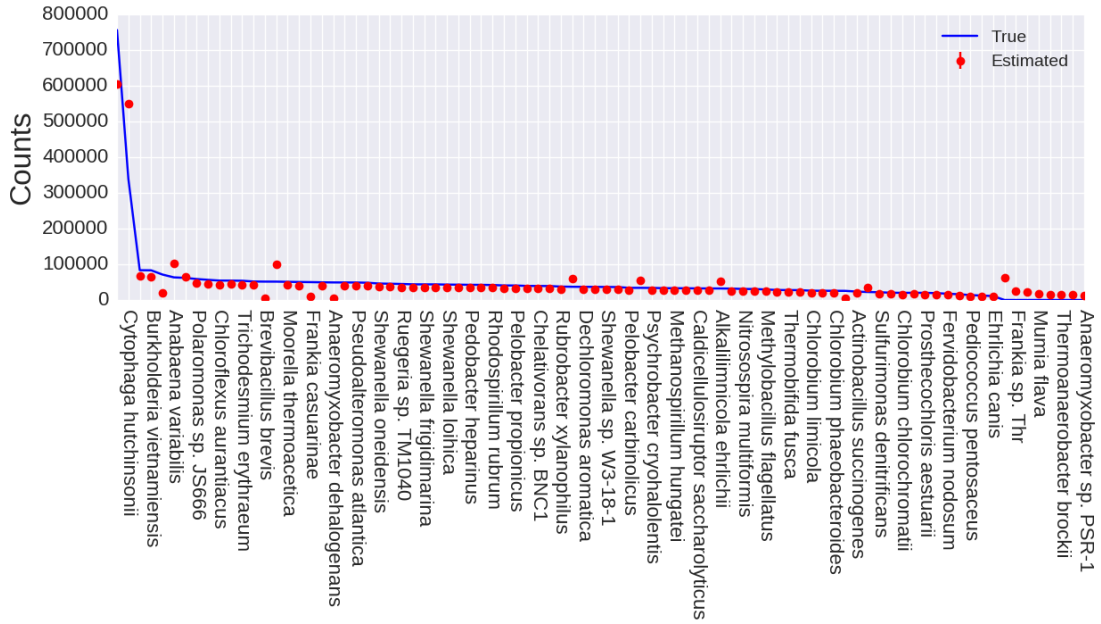

Figure 3.4. Estimated counts at species level aligned against representative genomes. True (blue line) and estimated (red dots) species-level counts of simulated metatranscriptome reads pseudoaligned against a representative genome index containing only a single strain from 4,412 species.

We used the successful genome-level taxonomic identification to determine which transcriptomes were possible read sources. At the predefined cutoff of 1000 estimated counts, 109 species were considered as sources of the metatranscriptomic data, containing all 80 source transcriptomes of the simulated reads. Ensembl contains 500 transcriptomes from those species, which were used to generate the index for the abundance estimation stage. The resulting quantification was quite accurate, with an average strain-level relative error of 9.22%, and only 5.79% of the reads misassigned to absent taxa.

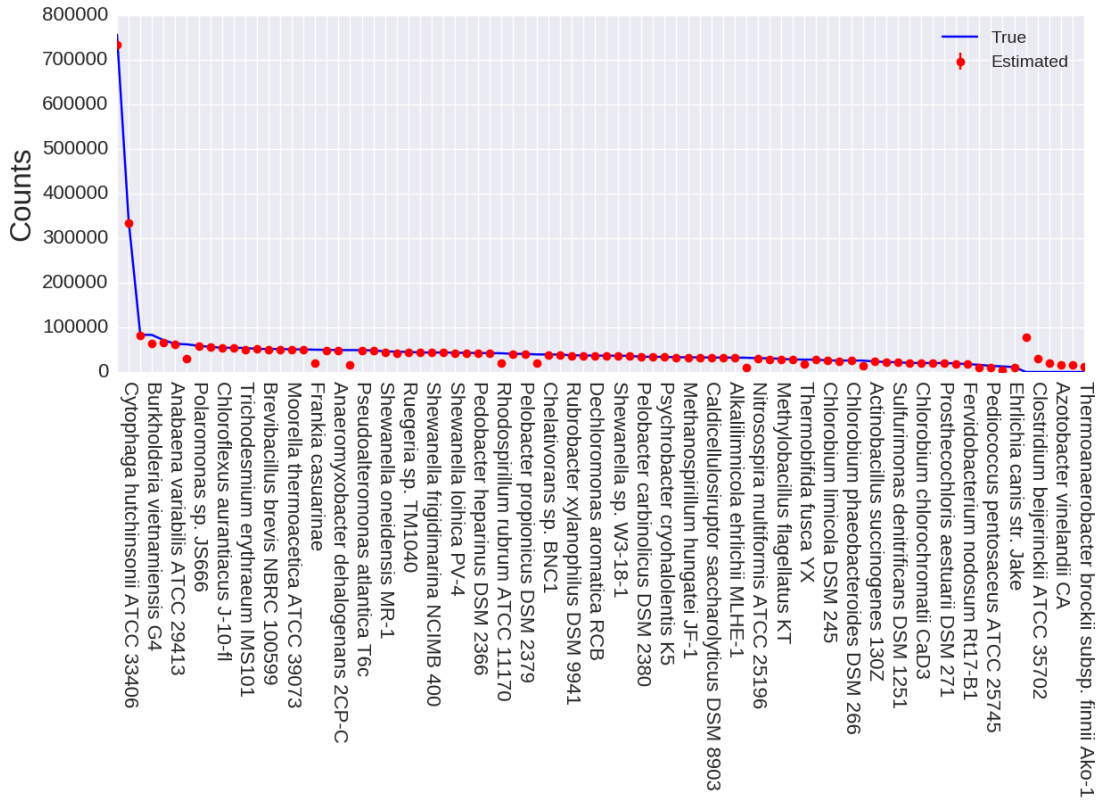

Figure 3.5. Estimated counts at strain level aligned against pre-filtered transcriptomes. True (blue line) and estimated (red dots) strain-level counts of simulated metatranscriptome reads pseudoaligned against an index containing only transcriptomes that had significant read assignment from a representative genome index.

On the strength of the simulated results, we then applied this pipeline to sequenced human gut microbiome reads collected by Franzosa et al. (2014). As with the simulated data, we first pseudoaligned the reads to a representative species genome index, to identify potential strains in the sample. 255 species passed the cutoff of 1000 counts, for a total of 5072 transcriptomes. As this is beyond the workable limits of kallisto's indexing, we ran a second level of genome-based identification, using the 5692 genome strains associated with the 255 species found in the first pass.

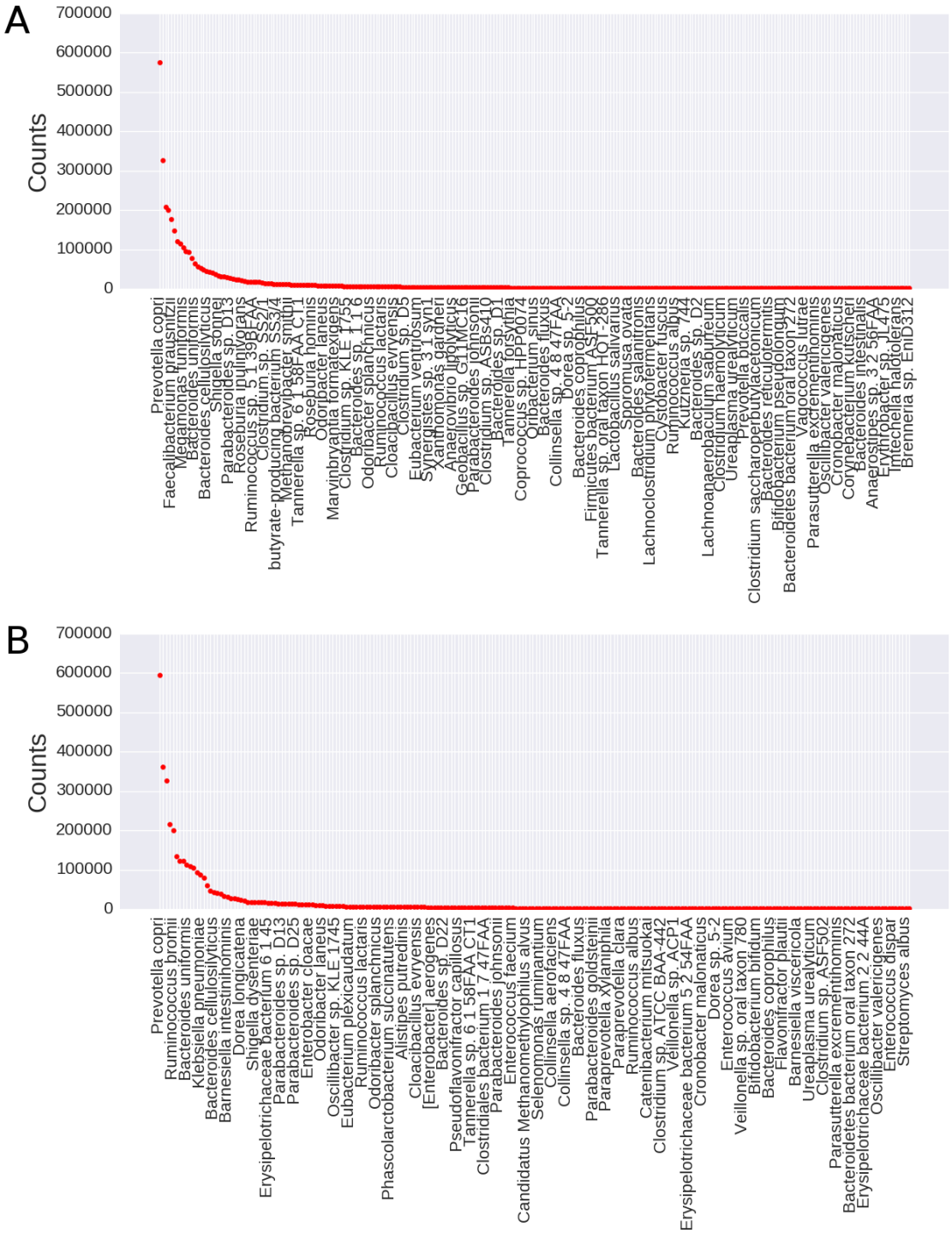

Figure 3.6. Estimated counts of human gut metatranscriptome at species level aligned against representative genomes. (a) First pre-filtering pass. Representative index contained a single strain from 4,412 species, and resulted in 255 species considered 'present'. ***(b)*** *Second pre-filtering pass. As the 255 species had too many strain-level transcriptomes for a single index, another set of indexes was built with the 5,692 strains associated with 255 species, and sequenced reads were pseudoaligned to them. This resulted in only 71 species considered 'present', but they are still associated with 4,641 individual strains, too many for a single index.*

Our hope was that this further refinement would result in a smaller list of species from which to select transcriptomes. Even with the second-pass list of potential genomes passing a higher threshold of 10,000 counts, although the number of species was lowered to 71, the number of actual strains only dropped to 4641, because *E. coli* remained among the top 3 species in overall abundance, and it has 2597 sequenced strains by itself. This is beyond kallisto's indexing ability as a single index, and more than two thousand *E. coli* strains alone meant that partitioning to 1000 transcriptomes per index would not work for quantification purposes.

Instead, we used the strain-level results directly, which gave 556 strains with more than 1000 counts assigned, 477 of which had Ensembl transcriptomes. We created a new index with only these transcriptomes, and pseudoaligned the gut microbiome data to it. The resulting estimated abundances at the genus level are seen in the figure below. Most notably, the top 3 genera in this metatranscriptomic sample are Prevotella, Bacteroides, and Megamonas, with significantly more counts than other genera. All three are known human gut inhabitants, and both Bacteroides and Prevotella are considered two of the most common "core" human gut genera (Xiao et al., 2015).

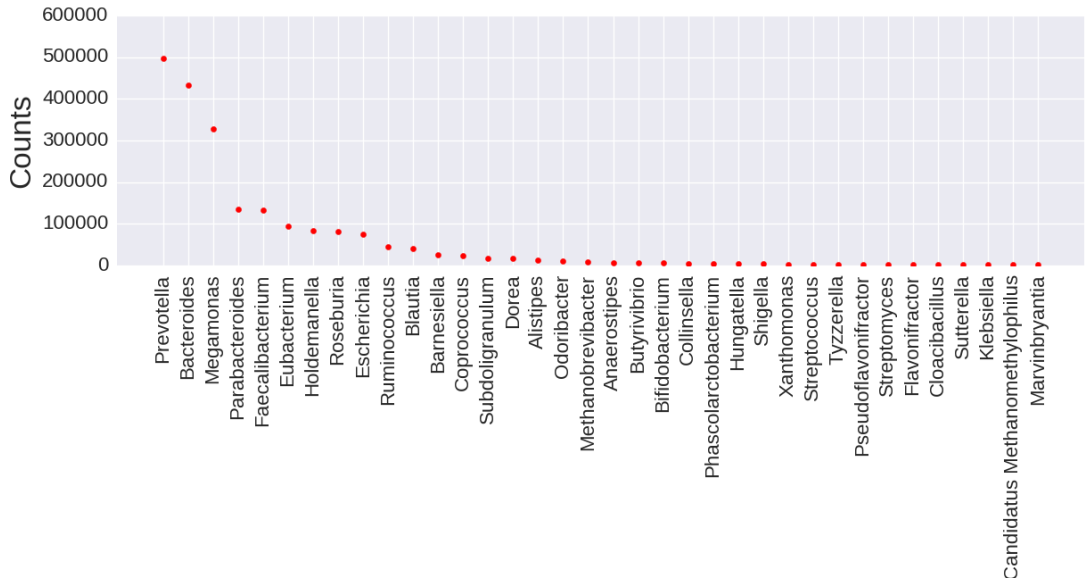

Figure 3.7. Estimated counts of human gut metatranscriptome at genus level aligned against pre-filtered transcriptomes. RNA-Seq reads were pseudoaligned against 477 microbial transcriptomes, derived from a representative transcriptome index containing a single strain from each of 4,412 species.

As this metatranscriptome dataset has a paired metagenome dataset -- DNA and RNA collected from the same fecal sample -- we ran the paired DNA sample through kallisto as well. We used the same two-step filtering process to filter then pseudoalign the metagenomic dataset. We pre-filtered with the same representative-strain genome indexes, which resulted in 142 strains being indicated as present in the sample. Much as with the metatranscriptome data, this

included species with thousands of strains present in the reference genome database, so we couldn't index all indicated species. Instead we used only these strains to make a sample-specific index; 6 million of the 18 million metagenomic reads aligned to this index. The estimated counts at the genus level are listed in Figure 3.8.

The top three most abundant genera in the revised, DNA-Seq-based estimate are *Bacteroides*, *Prevotella*, and *Faecalibacterium*, with *Megamonas* no longer in the top 10. All three of these genera are among the most common human gut microbes, as in fact are 9 of the top 10, and the remaining genus *Barnesiella* is a known gut inhabitant that was found to be associated with the reduction of vancomycin-resistant *Enterococcus* (VRE) colonization (Ubeda et al., 2013).

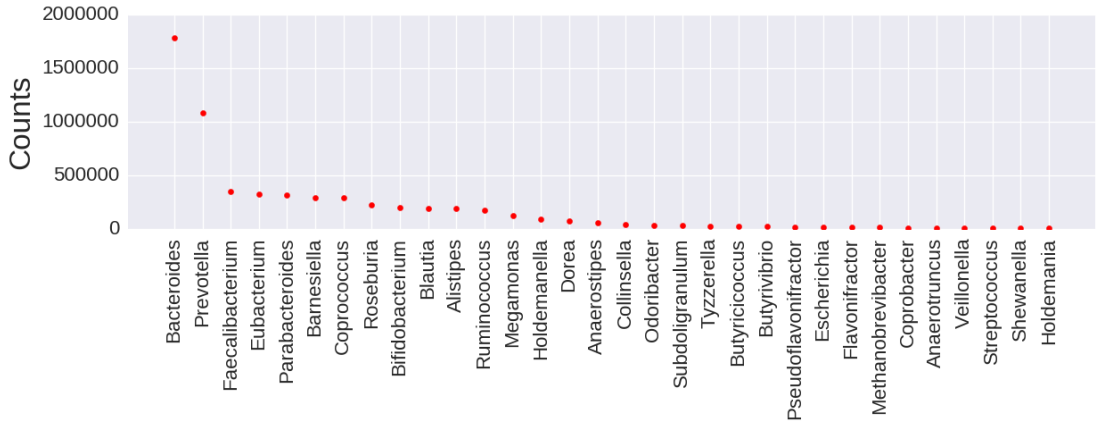

Figure 3.8. Estimated counts of human gut metagenome at genus level aligned against pre-filtered transcriptomes. DNA-Seq reads pseudoaligned against 142 microbial genomes, derived from a representative genome index containing a single strain from 4,412 species.

While obviously we cannot calculate the true accuracy of this quantification, we can compare it to the results from (Franzosa et al., 2014) as seen below, which lists the top 10 genera found in all samples (unfortunately not listing which sample was associated with which column). Their top genera are all present in both the metatranscriptome and metagenome abundance estimations performed by kallisto, albeit in slightly different abundances and with additional genera interspersed. Most notable is *Prevotella*, a genus present in high abundance in both the metatranscriptome and metagenome, but not reported at all in Franzosa et al.'s abundance summary. The specific species that was assigned the most reads was *Prevotella copri*, which according to Franzosa et al.'s supplementary dataset was found only in subject X316192082's stool samples, and had several hundred thousand reads aligned to their custom reference pan-genome of *Prevotella copri*. Thus, both kallisto and Franzosa et al. agree on the genera present in this gut microbiome sample.

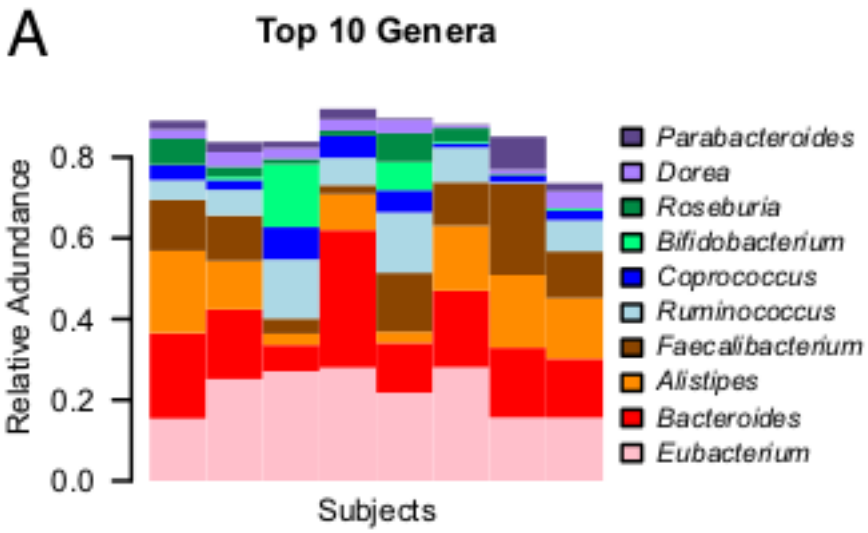

Figure 3.9. Estimated relative abundance of top genera in human gut metagenomes. Only genera that make up at least 1% of 2/8 samples are included. Taxa identification was determined based on marker genes. Figure from (Franzosa et al., 2014).

The lack of complete consistency between Franzosa et al.'s results and kallisto's is not surprising, as the former was primarily based on results from MetaPhlAn, which estimates taxonomic identity via clade-specific marker genes rather than full genome alignment. Franzosa et al. performed more detailed alignment to the clades identified by MetaPhlAn via Bowtie2 and a custom database of concatenated clustered genes from each species, although these results were not indicated in their abundance summary figure. Their final mapping rate was 31%, approximately the same as kallisto's 33%, suggesting that the unmapped reads are most likely from unsequenced microbes. Unfortunately for further comparison, MetaPhlAn does not list which species are contained in their core gene catalog, but it is between 1,221-3000 species, depending on which version was used. This is of course significantly fewer than the nearly 30,000 genomes that kallisto was able to effectively use for pseudoalignment, which would be expected to change the resulting distribution of read assignments.

Of course, metatranscriptomic data is not usually used for strain-level abundance estimation, but rather for functional analysis, in hopes of understanding what the community is doing. While the source paper analyzed the transcriptome only for comparative purposes between samples, we looked at the abundance of functional categories for the single sample used above. We used the KEGG MGENES functionally-annotated microbiome gene database to construct a series of kallisto pre-filtering indexes (see methods), then selected the highly abundant genes to create a final index. Summing the counts associated with each top-level

KEGG pathway showed a high prevalence of genes associated with metabolism pathways, translation, membrane transport, and replication and repair.

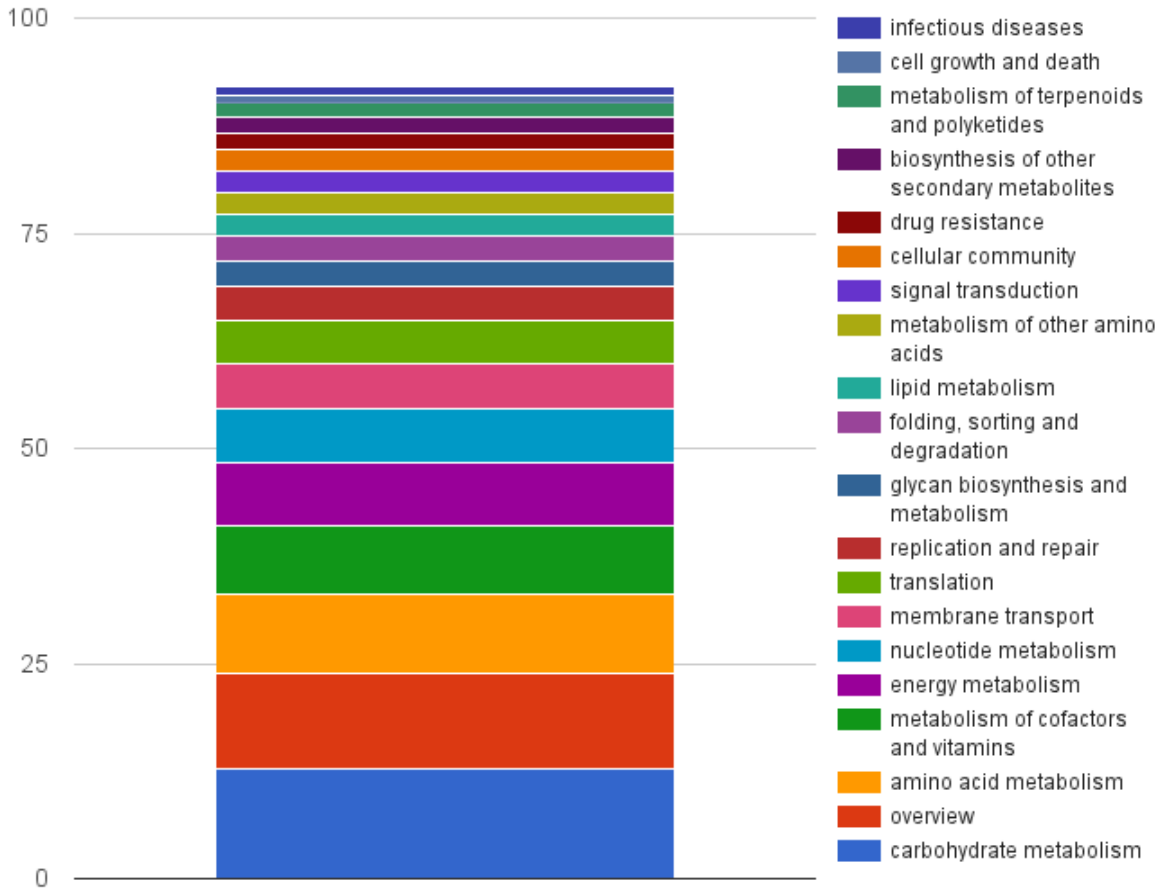

Figure 3.10. Percentage of kallisto-estimated human gut microbiome transcripts assigned to listed KEGG functional pathways. Remainder of transcripts were assigned to functional pathways with less than 1% overall abundance.

To compare these results to a standard functional analysis pipeline, we used COMAN, an online metatranscriptomic analysis pipeline (Ni, Li, & Panagiotou, 2016) which uses the aligner DIAMOND to align the sequenced transcript reads to 2700 complete microbial genomes from RefSeq. Subsequently, it assigns functional annotations from KEGG to genes with a 1e-5 cutoff using KOBAS 2.0. The overall method is very similar to BLAST, but significantly faster.

The COMAN functional analysis shows an extremely similar pattern to kallisto's results. Specifically, the top most abundant functional categories that kallisto identified, with close to or more than 200,000 counts assigned, are all the same as the top categories that COMAN found: carbohydrate metabolism, overview genes, amino acid metabolism, cofactor and vitamin metabolism, energy metabolism, membrane transport, and translation. The minor differences -- such as kallisto indicating a higher expression of amino acid metabolism genes than COMAN -- may be the result of using different gene databases, or could be an effect of kallisto more

effectively handling ambiguously-aligned reads. Distinguishing between these would require indexing COMAN's gene database in kallisto (difficult due to the lack of descriptive information about it), or ideally creating a simulated metatranscriptome with known functional content -- the simulated metatranscriptome used here is insufficient, because COMAN accepts only inputs in the FASTQ format.

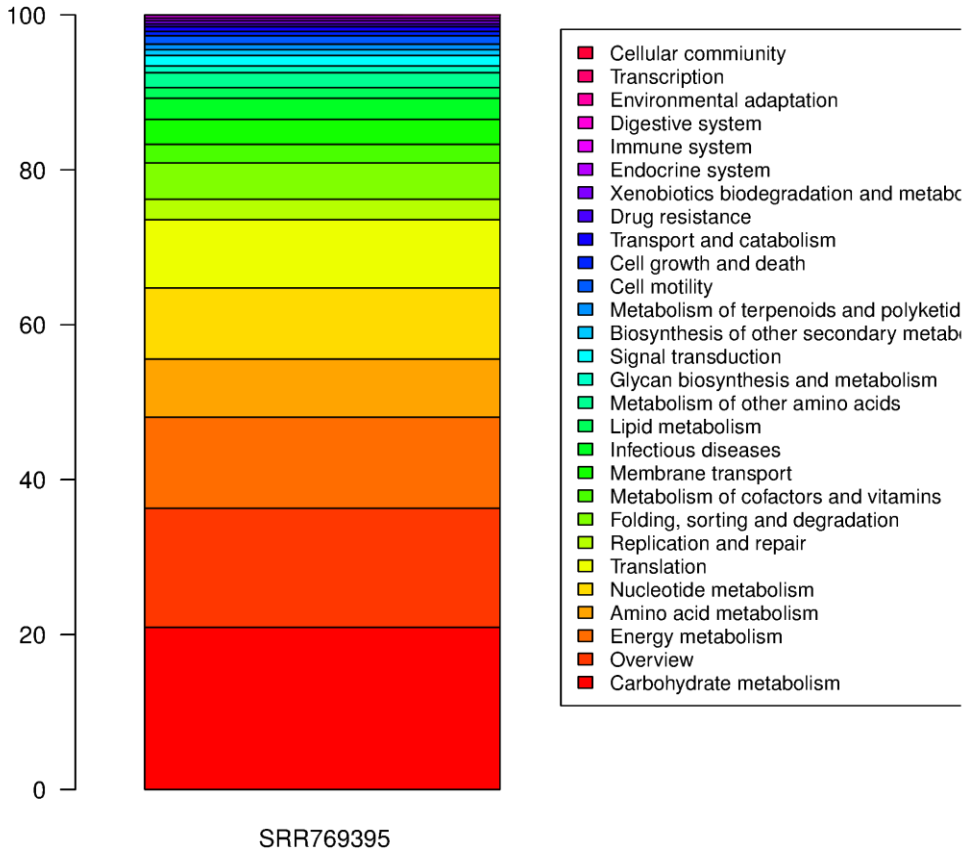

Figure 3.11. Estimated percentage of genes present in KEGG functional pathways, as estimated by COMAN.

3.3 Methods

Transcriptome and Genome Databases

All reference transcriptomes and genomes used are from Ensembl's bacterial database at *ftp://ftp.Ensemblgenomes.org/pub/current/bacteria/fasta/*. 39,586 complete and partial transcriptomes were downloaded in May 2016, containing a total of 137,567,837 transcripts, and 29,698 complete and partial genomes were downloaded in May 2016, matching approximately 4412 bacterial species.

Both transcriptomes and genomes were left as individual FASTA files, but were processed so FASTA headers contained the strain name, and spaces were replaced with underscores so full header information would be retained by kallisto.

Representative Species Indexes

To handle kallisto's memory limitations while indexing, instead of indexing all tens of thousands of available strains, representative indexes were created from a single strain for each species. Pseudoaligning against these representative strains allows for some level of taxa identification, if not quantification. The largest strain (transcriptome or genome) of each species present in the database was chosen for indexing. While this does select for the most sequence, it does not select for the most contiguous sequence, so presumably many of the genomes are incompletely assembled.

Since there are several thousand unique species in Ensembl's databases of both genomes and transcriptomes -- still too many for kallisto to index at once -- the collection of representative strains was broken into chunks of approximately 1000 strains for indexing, grouped by genus where possible. This meant, for most of the indexes in this chapter, 5 to 7 chunks per "representative index." While these separate indexes are definitely less accurate than a single index, in this case we care far more about false negatives than false positives, so the accuracy hit is acceptable.

Simulated Dataset

To test the performance of kallisto with a known ground truth, we used a simulated metatranscriptomic dataset of 100bp single-end reads, modeled in abundance off the popular "simulated low complexity" Sanger metagenomic dataset. The transcript reads were simulated by Toseland et al. (2014), and the reads and descriptions were generously shared by Andrew Toseland via private correspondence. "True" abundance of strains and transcripts was calculated by raw read counts, unadjusted for sequencing bias or error rate.

The simulated dataset consists of 7.5 million single-end reads from 112 strains, with the majority of strains being at approximately equal abundance, and a few strains being at significantly higher abundance; this is a relatively common form for microbiome populations to take, with a few abundant strains, and the remainder being present at low levels. Due to limitations of the Ensembl bacterial transcriptome database, three of the strains present in the simulated dataset were not available: "*Burkholderia cepacia 383*", "*Cronobacter turicensis z3032*", and "*Prochlorococcus marinus MIT 9312*." I removed all reads from those strains from the dataset fasta for the majority of experiments, unless indicated otherwise, leaving 7,351,496 reads.

Biological Dataset

To test the performance of kallisto on a biological sequenced metatranscriptome, we used a human gut microbiome RNA-Seq dataset with 5.5 million 100bp paired-end reads (SRA SRR769395), originally generated and analyzed by Franzosa et al. (2014). The dataset is the

microbiome of a stool sample preserved in RNAlater, from a single healthy individual. This dataset is paired with a DNA-Seq dataset (SRR769516) from the same individual's stool sample.

Taxa-Level and Transcript-Level Analysis

The estimated abundance of the simulated and sequenced datasets was calculated and analyzed using the uploaded python script "compare_metagenomic_results_to_truth.py." In short, the kallisto output file is processed and fasta headers are converted to NCBI taxonomic IDs (in the case of taxa-level analysis) or the format "strain name_bp location" (in the case of transcript-level analysis). Counts are summed in the case of duplicates, and a normalization factor is computed based on the percentage of reads that were assigned. The normalized estimated counts are then subtracted from the true counts for a given taxa/transcript, divided by the true counts, and the mean is given as the average relative error.

Functional Analysis

The estimated abundance of KEGG functional pathways was determined by pseudoaligning to the KEGG MGENES annotated environmental gene catalog (Kanehisa & Goto, 2000). As the full catalog is 83Gb, we divided it into 87 1Gb indexes for gene identification purposes. These were run in parallel, and took approximately 3-5 minutes to quantify against each index. Genes with an estimated TPM less than 10 were then removed from the FASTA, and the remaining genes were reindexed and quantified against. The resulting gene abundances were subsequently labeled with their KEGG pathways, and sums of total transcript count per pathway were computed.

3.4 Conclusions

Our results show that using a two-step process of taxa identification followed by quantification allows kallisto to functionally take advantage of much larger databases than it can actually index. For the purposes of transcriptome analysis, it is clear that the identification stage works much better on genomes rather than transcriptomes, while quantification works best on exact-match transcriptomes. Identification also does not require the target representative sequences to be in a single index; thousands of taxa can be spread across multiple indexes, and the combined abundance outputs will still give a fairly accurate picture of which taxa are present in the sample. Of concern, however, is that accuracy of both identification and quantification goes down when exact-match transcriptomes are not present; this is a significant problem, as most microbes are not sequenced, and thus exact-match sequences are only guaranteed to be available if they are assembled by the user. While this is certainly a doable task, and is commonly performed in metagenomic analysis, it would be nice to remove this obstacle.

As currently implemented, kallisto can effectively use metatranscriptome data to estimate strain-level abundance, but only performs well at estimating transcript-level abundance for high

abundance transcripts. This is likely due to the extremely low count of transcripts that are not highly expressed, in the moderate coverage datasets used in these example analyses; as seen in figure 3.2(b), the true counts are on the order of 100 reads in the simulated metatranscriptomic dataset, which may very well be insufficient for kallisto to determine a reasonable abundance.

Of course, most actual analysis of metatranscriptome datasets is not at the transcript level, but the functional level, grouping together transcripts that have a similar purpose (and thus grouping together those with similar sequences). This solves the transcript similarity problem, while giving information on what interests most researchers in RNA-Seq data: the likely pathways currently active. Kallisto's functional-level results were highly similar to those of standard functional analysis pipelines, indicating it is well-suited for this informative form of analysis.

While the field of metatranscriptomics grows in importance, the pipelines available for analysis remain limited and based almost entirely on BLAST and standard alignment algorithms, with no significant handling of ambiguous reads. This is, to my knowledge, the first use of a k-mer based algorithm on metatranscriptomic data, as well as the most complex handling of ambiguities. These results indicate that these RNA-Seq-based methods are equally applicable to metatranscriptomics, and should be examined further.

Chapter 4: Concluding remarks on low-memory k-mer indexing improvements

As demonstrated in the previous chapters, the primary limitation to kallisto's ability to pseudoalign to multiple genomes or transcriptomes is the high memory requirements of its indexes. The current form of the index is such that more than a couple thousand microbial genomes overwhelms even large servers with extensive RAM, making it impossible for kallisto to be used on meta'omic data for the average individual laptop, as well as making it impossible for kallisto to directly pseudoalign against the current 50,000 microbial genomes contained by Ensembl, or the 83Gb of annotated microbial genes available from KEGG.

In this work, we investigated two separate options for handling this issue: pre-filtering using the genome distance estimator Mash, and pre-filtering by breaking up a collection of genomes or genes into smaller, more feasible indexes, then iteratively pseudoaligning to each one. There are advantages and disadvantages to each: Mash is extremely fast, even on tens of thousands of genomes, but has a hard time distinguishing between similar genomes, which is obviously a significant limitation. This results in a significant level of false negatives at the strain level, which is an area of increasing interest for metagenomic analysis. Conversely, while using kallisto sub-indexes for pre-filtering is extremely accurate -- no false negatives were observed in the simulated metatranscriptomic analysis -- it is also much slower than Mash alone. Pre-filtering to the split 83Gb KEGG annotated gene database took over 7 hours, on 10 cores. This obviates the primary advantage of a k-mer based method, that of speed, and adds a significant level of complexity to the abundance estimation process.

Ideally, we need a method that combines the speed of Mash with the accuracy of kallisto. This is most achievable by discarding k-mers to reduce computational costs. Most k-mers in a genome are not necessary or even useful for discriminating between microbes. As seen in many of the current k-mer based metagenomic identification programs, and as implemented in Mash itself, a subset of k-mers that are distinctive between genomes can be used instead of the full genome. Previous applications of this idea have demonstrated that, while workable, this results in lower accuracy than using all available k-mers from the reference genomes. This makes it entirely suitable for a fast "first-pass" filtering step to determine which genomes are present in the sample, with detailed abundance estimation occurring with a "full" index of just present genomes subsequently.

The most obvious implementation of a sparse k-mer index is to pick only k-mers that are unique to a genome, which is how CLARK selects its indexed k-mers. However, this can fail in the case where two strains are identical except for an indel: in such a case, the only k-mers unique to the smaller genome would be those spanning the deletion junction, and per-strain

metagenomic coverage is not always high enough to be assured of sequencing those k-mers. So for the sake of robustness to low sequencing depth, we need a significant number of k-mers from each genome, which means we cannot limit ourselves to truly unique ones.

Another implementation that is frequently used is selecting an evenly-distributed set of k-mers across each genome. MiniKraken, for instance, throws out 18 of each 19 sequential k-mers in its index in order to remain within 4GB. While this ensures a reasonable coverage of each genome (roughly 5% in this case), it does not enrich for discriminatory k-mers, and thus can make it harder to distinguish between very similar genomes.

Other approaches include that used by LMAT, which groups k-mers by which genomes they are found in, and indexes all such groups that contain at least 1000 different k-mers. While this does reduce the size of the index, it neither ensures the k-mers are the most discriminatory, nor ensures that all genomes contribute a significant number of k-mers.

For alternate approaches that maintain both high coverage and high discrimination, consider the set of all genomes to be indexed, $\mathbf{G} = \{g_1, g_2, ... g_n\}$, and the set of all k-mers contained in those genomes, $\mathbf{K} = \{k_1, k_2, ... k_n\}$. In order to reliably judge the approximate abundance of each genome, we should have a set number of k-mers from each one; specifically, we should have at least enough k-mers to cover $\boldsymbol{\varepsilon}$ % of a given genome g_n. We want the k-mers associated with g_n to be the most discriminatory (that is, to contain the most information regarding g_n) while also minimizing the set **S** of all k-mers selected to index, in order to keep the index as small and memory-efficient as possible.

There is another reason to minimize **S**: if two strains g_1 and g_2 are mostly similar with a small number of differences, obviously **S** will contain all the discriminatory k-mers. However, if those k-mers are less than $\boldsymbol{\varepsilon}$ % of g_1 and g_2, more k-mers must be chosen for **S**, from the sequences shared between g_1 and g_2. If the k-mers chosen for g_1 are different than those chosen for g_2, then an additional source of error has been created, and meaningless differences in sequenced coverage over the genomes can cause abundances to be misjudged. Making sure the k-mers that are not discriminatory between g_1 and g_2 are the same will prevent that problem, while keeping the index as small as possible.

The simplest memory-efficient algorithm would be as follows: first, count the number of genomes that each k-mer k_n appears in, across the whole reference genome. Next, for genome g_1, identify the k-mers present in that genome, and select those k-mers that have the lowest total number of genomes they are present in -- these will be the most unique k-mers available in that genome. Most genomes will not have $\boldsymbol{\varepsilon}$ % of their sequence covered only by unique k-mers, of course, and so some k-mers added to **S** will be present in other genomes. This is fine. For genome g_2, do the same, but when you have a choice of k-mers that are present in the same number of genomes, always select those that are already in **S**. Repeat for all genomes in the

reference. While this will not strictly minimize **S**, it will act to keep **S** reasonably compact, and will keep the non-unique k-mers that are shared between genomes as shared as possible.

An algorithm that would be more complex but more certain to give us the results we need would be to use column subset selection (CSS) methods to reduce the dimensions of a matrix containing all k-mers (McCurdy, Ntranos, & Pachter, 2016). In this case, the columns would be genomes and the rows would be k-mers, with each element indicating the count of that k-mer in that genome, and the selection algorithm would be acting on rows. CSS has been shown to preserve clustering structure in single-cell RNA-Seq datasets while reducing the number of features, keeping the data that is most distinctive and representative. While this algorithm usually requires full matrix creation, there is an online streaming CSS algorithm that may be more useful for our low-memory uses (Cohen, Musco, & Musco, 2015).

Once **S** is finalized, each k-mer is linked to the genomes it is present in and hashed to create the minimized index. The k-mers from the sequenced reads are then compared to the index, with the EM algorithm handling reads that could come from multiple genomes, as usual. The output will be a set of estimated abundances for each genome in the reference, with the accuracy of these estimates primarily depending on the size of **ε**. This allows for a direct trade-off between accuracy and memory efficiency, making this ideal for pre-filtering using tens of thousands of genomes.

References

Ames, S.K., Hysom, D.A., Gardner, S.N., Lloyd, G.S., Gokhale, M.B., and Allen, J.E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics *29*, 2253.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology *11*, R106.

Angly, F.E., Willner, D., Prieto-Davó, A., Edwards, R.A., Schmieder, R., Vega-Thurber, R., Antonopoulos, D.A., Barott, K., Cottrell, M.T., Desnues, C., et al. (2009). The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. PLoS Comput Biol *5*.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bradley, P., Gordon, N.C., Walker, T.M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L.J., Anson, L., Cesare, M. de, et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nature Communications *6*, 10063.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotech *34*, 525–527.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat Meth *12*, 59–60.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., and Holmes, S.P. (2015). DADA2: High resolution sample inference from amplicon data. bioRxiv 024034.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat Meth *7*, 335–336.

Carvalhais, L.C., and Schenk, P.M. (2013). Sample Processing and cDNA Preparation for Microbial Metatranscriptomics in Complex Soil Communities. ResearchGate *531*, 251–267.

Chen, K., and Pachter, L. (2005). Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. PLoS Comput Biol *1*, e24.

Chen, L., Hu, M., Huang, L., Hua, Z., Kuang, J., Li, S., and Shu, W. (2015). Comparative metagenomic and metatranscriptomic analyses of microbial communities in acid mine drainage. ISME J *9*, 1579–1592.

Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Meth *5*, 613–619.

Cohen, M.B., Musco, C., and Musco, C. (2015). Input Sparsity Time Low-Rank Approximation via Ridge Leverage Score Sampling. arXiv:1511.07263 [Cs].

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res *42*, D633–D642.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol *72*, 5069–5072.

Dunbar, J., Barns, S.M., Ticknor, L.O., and Kuske, C.R. (2002). Empirical and Theoretical Bacterial Diversity in Four Arizona Soils. Appl Environ Microbiol *68*, 3035–3045.

Elya, C., Zhang, V., Ludington, W., and Eisen, M.B. (2016). Stable host gene expression in the gut of adult Drosophila melanogaster with different bacterial mono-associations. bioRxiv 053512.

Escobar-Zepeda, A., Vera-Ponce de León, A., and Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. Front. Genet. *6*.

Fox, G.E., Magrum, L.J., Balch, W.E., Wolfe, R.S., and Woese, C.R. (1977). Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci U S A 74, 4537–4541.

Franzosa, E.A., Morgan, X.C., Segata, N., Waldron, L., Reyes, J., Earl, A.M., Giannoukos, G., Boylan, M.R., Ciulla, D., Gevers, D., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. PNAS *111*, E2329–E2338.

Freitas, T.A.K., Li, P.-E., Scholz, M.B., and Chain, P.S.G. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucl. Acids Res. *43*, e69–e69.

Friedman, B.A., and Maniatis, T. (2011). ExpressionPlot: a web-based framework for analysis of RNA-Seq and microarray gene expression data. Genome Biol *12*, R69.

Ghurye, J.S., Cepeda-Espinoza, V., and Pop, M. (2016). Metagenomic Assembly: Overview, Challenges and Applications. Yale J Biol Med *89*, 353–362.

Goncalves, A., Tikhonov, A., Brazma, A., and Kapushesky, M. (2011). A pipeline for RNA-seq data processing and quality assessment. Bioinformatics *27*, 867–869.

Grice, E.A., and Segre, J.A. (2011). The skin microbiome. Nat Rev Microbiol *9*, 244–253.

Gupta, A., and Sharma, V.K. (2015). Using the taxon-specific genes for the taxonomic classification of bacterial genomes. BMC Genomics *16*.

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R., et al. (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Meth *7*, 807–812.

Hugenholtz, P., and Pace, N.R. (1996). Identifying microbial diversity in the natural environment: A molecular phylogenetic approach. Trends in Biotechnology *14*, 190–197.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res *17*, 377–386.

Huson, D.H., Richter, D.C., Mitra, S., Auch, A.F., and Schuster, S.C. (2009). Methods for comparative metagenomics. BMC Bioinformatics *10*, S12.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A.L., Madsen, K.L., et al. (2016). Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front Microbiol *7*.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kembel, S.W., Wu, M., Eisen, J.A., and Green, J.L. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. PLOS Computational Biology *8*, e1002743.

Kersey, P.J., Allen, J.E., Armean, I., Boddu, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C., et al. (2016). Ensembl Genomes 2016: more genomes, more complexity. Nucl. Acids Res. *44*, D574–D580.

Kukutla, P., Steritz, M., and Xu, J. (2013). Depletion of Ribosomal RNA for Mosquito Gut Metagenomic RNA-seq. Journal of Visualized Experiments.

Lagier, J.-C., Edouard, S., Pagnier, I., Mediannikov, O., Drancourt, M., and Raoult, D. (2015). Current and Past Strategies for Bacterial Culture in Clinical Microbiology. Clin. Microbiol. Rev. *28*, 208–236.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., et al. (2015). Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics *15*, 141–161.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357–359.

Lee, Z.M.-P., Bussema, C., and Schmidt, T.M. (2009). rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Res *37*, D489–D493.

Leimena, M.M., Ramiro-Garcia, J., Davids, M., van den Bogert, B., Smidt, H., Smid, E.J., Boekhorst, J., Zoetendal, E.G., Schaap, P.J., and Kleerebezem, M. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics *14*, 530.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Lindgreen, S., Adair, K.L., and Gardner, P.P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports *6*, 19233.

Lindner, M.S., and Renard, B.Y. (2013). GASiC: Metagenomic abundance estimation and diagnostic testing on species level. Nucleic Acids Res *41*, e10.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell *133*, 523–536.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2016). Bracken: Estimating species abundance in metagenomics data. bioRxiv 051813.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature *437*, 376–380.

Martinez, X., Pozuelo, M., Pascal, V., Campos, D., Gut, I., Gut, M., Azpiroz, F., Guarner, F., and Manichanh, C. (2016). MetaTrans: an open-source pipeline for metatranscriptomics. Scientific Reports *6*, 26447.

McCurdy, S.R., Ntranos, V., and Pachter, L. (2016). Column Subset Selection for Single Cell RNA-Seq Clustering. NIPS Workshop on Machine Learning in Computational Biology, Barcelona, Spain.

McDavid, A., Finak, G., Chattopadyay, P.K., Dominguez, M., Lamoreaux, L., Ma, S.S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics *29*, 461–467.

McFall-Ngai, M. (2008). Are biologists in "future shock"? Symbiosis integrates biology across domains. Nat Rev Micro *6*, 789–792.

Mende, D.R., Waller, A.S., Sunagawa, S., Järvelin, A.I., Chan, M.M., Arumugam, M., Raes, J., and Bork, P. (2012). Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data. PLoS ONE *7*, e31386.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., 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.

Mondav, R., Schmidt, S., and Tyson, G.W. (2010). Metatranscriptomics of microbial communities in agricultural and forest soils. ResearchGate, p. 85.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth *5*, 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science *320*, 1344–1349.

Navas-Molina, J.A., Peralta-Sánchez, J.M., González, A., McMurdie, P.J., Vázquez-Baeza, Y., Xu, Z., Ursell, L.K., Lauber, C., Zhou, H., Song, S.J., et al. (2013). Advancing our understanding of the human microbiome using QIIME. Methods Enzymol *531*, 371–444.

Nguyen, N., Mirarab, S., Liu, B., Pop, M., and Warnow, T. (2014). TIPP: taxonomic identification and phylogenetic profiling. Bioinformatics *30*, 3548–3555.

Ni, Y., Li, J., and Panagiotou, G. (2016). COMAN: a web server for comprehensive metatranscriptomics analysis. BMC Genomics *17*, 622.

Nicolae, M., Mangul, S., Măndoiu, I.I., and Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms Mol Biol *6*, 9.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., and Stahl, D.A. (1986). Microbial Ecology and Evolution: A Ribosomal RNA Approach. Annual Review of Microbiology 40, 337–365.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology *17*, 132.

Ounit, R., and Lonardi, S. (2016). Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics btw542.

Ounit, R., Wanamaker, S., Close, T.J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics *16*, 236.

Pascault, N., Loux, V., Derozier, S., Martin, V., Debroas, D., Maloufi, S., Humbert, J.-F., and Leloup, J. (2015). Technical challenges in metatranscriptomic studies applied to the bacterial communities of freshwater ecosystems. Genetica *143*, 157–167.

Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. Nat Meth *10*, 1200–1202.

Pepper, I., and Gerba, C. (2016). Culturing and Enumerating Bacteria from Soil Samples. JoVE Science Education Database.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., and Manichanh, C. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. *41*, D590–D596.

Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Meth *10*, 71–73.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Schaeffer, L., Pimentel, H., Bray, N., Melsted, P., and Pachter, L. (2015). Pseudoalignment for metagenomic read assignment. arXiv:1510.07371 [Q-Bio].

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Segata, N., Boernigen, D., Tickle, T.L., Morgan, X.C., Garrett, W.S., and Huttenhower, C. (2013). Computational meta'omics for microbial community studies. Molecular Systems Biology *9*, 666.

Sunagawa, S., Mende, D.R., Zeller, G., Izquierdo-Carrasco, F., Berger, S.A., Kultima, J.R., Coelho, L.P., Arumugam, M., Tap, J., Nielsen, H.B., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. Nat Meth *10*, 1196–1199.

Toseland, A., Moxon, S., Mock, T., and Moulton, V. (2014). Metatranscriptomes from diverse microbial communities: assessment of data reduction techniques for rigorous annotation. BMC Genomics *15*.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat Biotechnol *28*, 511–515.

Triant, D.A., and Whitehead, A. (2009). Simultaneous Extraction of High-Quality RNA and DNA from Small Tissue Samples. J Hered *100*, 246–250.

Ubeda, C., Bucci, V., Caballero, S., Djukovic, A., Toussaint, N.C., Equinda, M., Lipuma, L., Ling, L., Gobourne, A., No, D., et al. (2013). Intestinal Microbiota Containing Barnesiella Species Cures Vancomycin-Resistant Enterococcus faecium Colonization. Infect. Immun. *81*, 965–973.

Větrovský, T., and Baldrian, P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS ONE *8*, e57923.

Westreich, S.T., Korf, I., Mills, D.A., and Lemay, D.G. (2016). SAMSA: a comprehensive metatranscriptome analysis pipeline. BMC Bioinformatics *17*, 399.

Wong, C.N.A., Ng, P., and Douglas, A.E. (2011). Low-diversity bacterial community in the gut of the fruitfly Drosophila melanogaster. Environmental Microbiology *13*, 1889–1900.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology *15*, R46.

Xia, L.C., Cram, J.A., Chen, T., Fuhrman, J.A., and Sun, F. (2011). Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads. PLOS ONE *6*, e27992.

Xiao, L., Feng, Q., Liang, S., Sonne, S.B., Xia, Z., Qiu, X., Li, X., Long, H., Zhang, J., Zhang, D., et al. (2015). A catalog of the mouse gut metagenome. Nat Biotech *33*, 1103–1108.

Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W., and Church, G.M. (2006). Sequencing genomes from single cells by polymerase cloning. Nat Biotech *24*, 680–686.

Wetterstrand, K.A. (Accessed November 2016). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata.

Wu, Y.-W., and Ye, Y. (2011). A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-tuples. J Comput Biol 18, 523–534.

Appendix A: Notes on collecting microbiome samples from *D. melanogaster* guts

While the gut microbiomes of mammals has been extensively studied, sequencing gut microbes from invertebrates is more difficult. The primary issues are related to the near-impossibility of separating gut contents from host tissue; because of this, host DNA and RNA can easily overwhelm metagenomic and metatranscriptomic sequences. This is exacerbated by the yeast-based diet, which results in significant yeast DNA and RNA in gut samples; up to 60% of a DNA library can consist of non-microbial DNA (Elya, et al., 2016).

Because of this, published analyses of the *Drosophila melanogaster* gut microbiome are exclusively based on 16S rRNA gene sequences, because they can be easily amplified and sequenced while avoiding host and food contamination. However, as explained earlier, this limits the detail in which the microbiome can be examined, and makes it impossible to judge functional activity. In order to get a more complete picture of the D. melanogaster microbiome, I attempted to develop the following protocol to extract paired DNA and RNA samples from the same guts, and build them as libraries.

A.1 Gut dissection

Previous studies have found that third instar D. melanogaster have the most diverse gut microbiome, as based on 16S rRNA gene sequencing (Wong, Ng, & Douglas, 2011). For this reason, I focused on extracting guts from this stage alone. Third instar larva are easy to identify and have the longest stage length of any larval stage, so collecting a significant number is straightforward. Following collection, the larvae are bleached to remove surface microbes, and then the gut is dissected in RNAlater to protect RNA from degradation.

Dissection protocol

Let larva feed on food mixed with food coloring, for easy gut identification, for at least half an hour. Collect third instar larva in mesh egg collecting dish, and rinse food away with water. Transfer larva to fresh 10% bleach solution and surface sterilize for 5 minutes. Transfer larva to PBS solution to rinse. Transfer larva individually to RNAlater under a dissecting scope, and remove cuticle and surrounding fat, leaving only gut tissue. Be careful not to pierce the gut while dissecting, to avoid losing microbial contents.

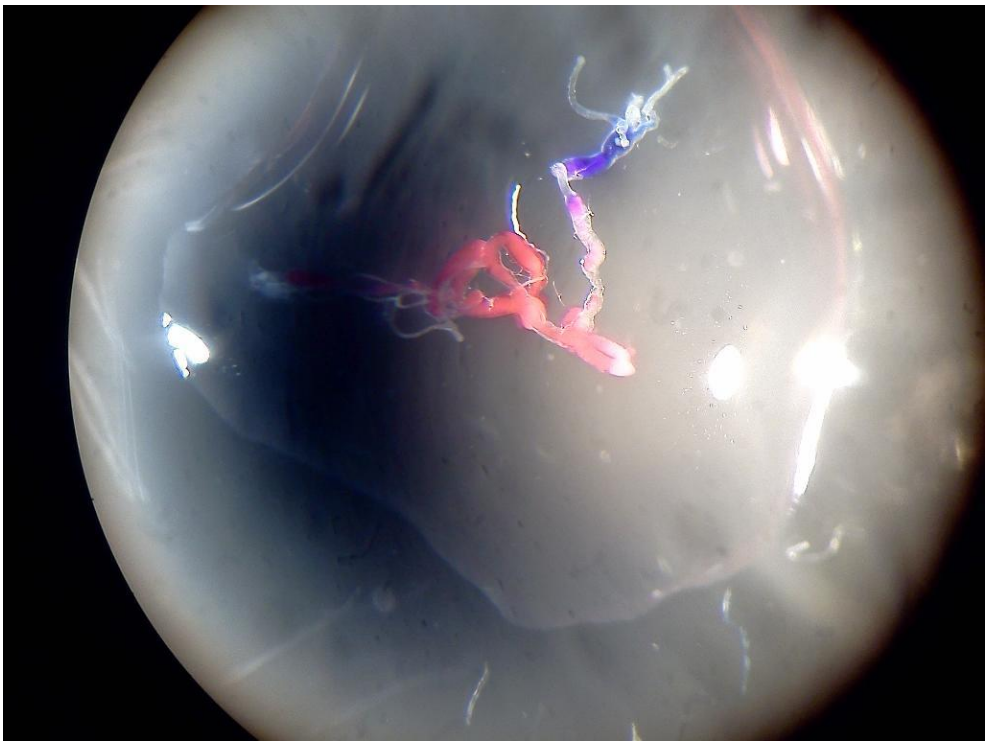

Figure A.1 Dissected third instar larva gut. Pink and purple color is due to food coloring added to food before dissection.

A.2 DNA/RNA extraction

Extracting either DNA or RNA from *D. melanogaster* guts is straightforward, and there are a number of commercial kits and protocols that will work. However, collecting both DNA and RNA simultaneously, at significant concentrations and with acceptable quality, required more trial and error. I was guided in this process by a paper by Triant & Whitehead (2009), which compared a number of protocols for simultaneous DNA/RNA extraction.

Extracting both nucleic acids from prokaryotes, especially gram positive bacteria, requires both enzymatic and mechanical treatments. First, the host gut cells are dissociated using proteinase K, then the bacterial cells are opened up by being vortexed with 0.1 mm zirconium beads. While RNA can be extracted without the proteinase K digestion, DNA requires it, which limits buffers to those that allow the digestion and also protect RNA during the mechanical bead beating process.

TRIzol, while enormously effective at protecting and extracting RNA, does not allow for proteinase K digestion, and the addition of any buffer used during digestion to TRIzol later in the protocol interferes with the phenol/chloroform extraction later. The buffer that finally worked for both purposes is RLT Plus, which is the Qiagen buffer that's used in the RNeasy Plus Micro kit. Because it’s used for homogenization in an RNA extraction kit, it clearly is sufficient for RNA protection, and it does not interfere with proteinase K digestion.

The homogenized supernatant resulting from the bead beating protocol is in a Qiagen buffer that should be compatible with a variety of Qiagen columns. It could also be ethanol precipitated, but I found that resulted in low yields, as well as a higher risk of ethanol contamination in downstream library-building steps. I wasn't able to get any of the protocols to separate the DNA and RNA from the same sample working, so I attempted to split the homogenate and extract DNA from half and RNA from the other half.

Initially, it looked like both protocols were successful: I extracted 1.4ug of RNA and 400ng of DNA, according to Qubit measurements. However, the DNA sample also contained 1.8ug of RNA, according to Qubit measurements, which I successfully removed after treatment with RNaseA for 1 hour at 37 C. Surprisingly, though, the DNA was also nearly eliminated: from 20ng/ul to 1ng/ul, as measured by Qubit. This was a very unexpected outcome, and I have no satisfying explanation for it.

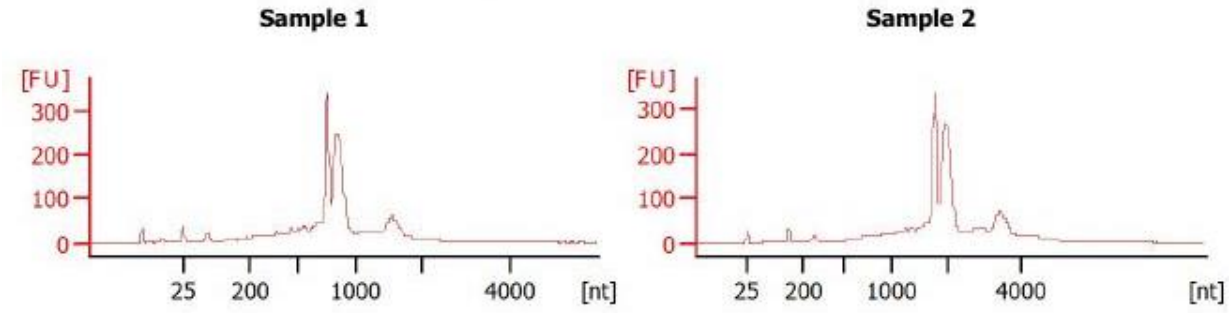

Figure A.2 Bioanalyzer trace of RNA sample extracted from gut.

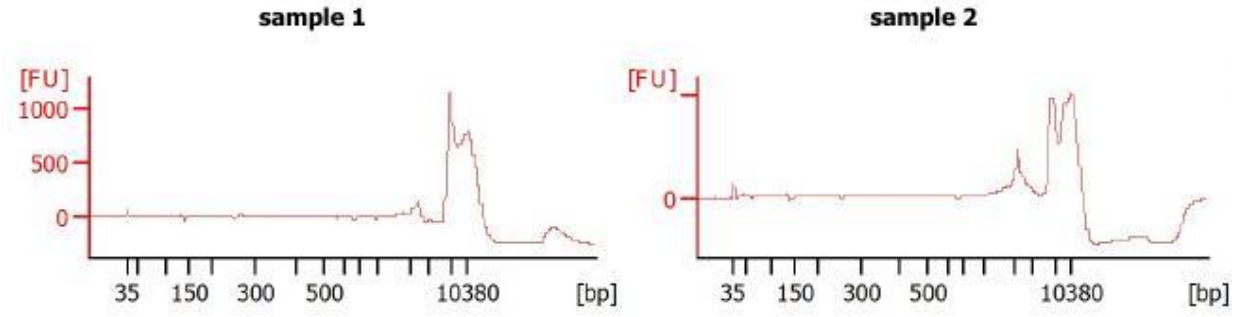

Figure A.3 Bioanalyzer trace of DNA sample extracted from gut.

In the absence of further time to refine this protocol, I did not explore additional avenues of extraction, but a reasonable next step would be to try the Qiagen AllPrep DNA/RNA Micro kit, as recommended by Triant & Whitehead. It should allow for simultaneous extraction of DNA and RNA from the same sample, although in lower concentrations than individual extractions might produce. Given the difficulties of extracting nucleic acids from these samples, this commercial kit seems like a good avenue to explore.

Homogenization protocol

Homogenization protocol is modified from Elya, et al.'s Drosophila gut homogenization (2016). Transfer dissected gut to clean nonstick tube, and add 180ul of buffer RLT Plus and 20ul of proteinase K, then incubate at 56 C for half an hour. After incubation, food coloring should be dispersed throughout the solution, no longer concentrated in the gut tissue (although the gut tissue will still appear coherent). Add 50-100ul of 0.1mm zirconium beads and 200ul of chilled RLT Plus to the tube, and perform subsequent bead beating in 4 C cold room:

1. Vortex at maximum power for 1 minute
2. Allow samples to rest for 30 seconds
3. Vortex at maximum power for 1 minute
4. Allow samples to rest for 30 seconds
5. Centrifuge samples 5 minutes at maximum speed
6. Transfer supernatant to a new tube and save
7. Wash beads with 400 uL cold RLT Plus (mix beads and additional buffer by pipetting up and down)
8. Vortex at maximum power for 1 minute
9. Allow samples to rest for 30 seconds
10. Centrifuge samples 5 minutes at maximum speed
11. Transfer supernatant to tube from step 6
12. Wash beads with 400 uL cold RLT Plus (mix beads and additional buffer by pipetting up and down)
13. Centrifuge samples 5 minutes at maximum speed
14. Transfer supernatant to tube from step 11
15. Centrifuge pooled supernatant from step 14 5 minutes at maximum speed
16. Remove supernatant (leaving beads behind) and transfer to new tube
17. Let sit for 5 minutes at room temperature before proceeding, to allow proteins to dissociate.

A.3 Microbial mRNA enrichment

One of the chief difficulties of metatranscriptomics is enriching the microbial mRNA over the microbial rRNA, in addition to the host and yeast RNA. There are a number of systems and kits designed to preferentially remove rRNA reads or enrich non-rRNA reads, several of which I have tested on *Drosophila* gut RNA.

Initially I attempted an unreleased protocol developed by Dr. Alexandra McCorkindale (2015) to deplete Drosophila rRNA, combined with a protocol to deplete microbial rRNA (Kukutla, Steritz, & Xu, 2013). These protocols use custom biotin-labeled RNA probes to bind to and then remove rRNA sequences. However, they require a significant amount of optimization and modification for use with specific samples. Because of this, I was unable to get these protocols to work effectively.

There are commercial kits that perform the same technique, but each generally only removes the rRNA from a single source, meaning several kits are required to fully deplete rRNA. For instance, the MicrobExpress Bacterial mRNA Enrichment Kit uses capture oligonucleotides and magnetic microbeads to remove up to 95% of the 16S and 23S rRNA from total RNA of some bacterial species. The Ribominus Transcriptome Isolation Kit has biotin-labeled probes that can similarly remove up to 98% of large yeast rRNA molecules, 18S and 25/26S subunits.

Use of both these kits leaves, hopefully, only Drosophila rRNA to be removed. As Drosophila is a less popular model organism for sequencing experiments than mammals like mice or humans, there are fewer rRNA removal kits available. One such kit is the Ovation RNA-Seq System V2, which preferentially primes and transcribes non-rRNA reads, leaving behind Drosophila rRNA. Because this kit transcribes RNA into cDNA, it can only be used as the last step of rRNA removal, after removing all other unwanted RNA.