Counting k-mers

A novel method for contig abundance estimation in metagenomic sample

Andrew Consroe; Mihai Pop, PhD

Abstract

K-mer count analysis has seen little application in metagenomics and we seek to utilize its simplicity for contig abundance estimation. Counting k-mers provide a foundation for a broad class of analyses and are especially desirable in metagenomics because of the large number of genomes and samples involved. Our results show that despite their simplicity, k-mers retain a trove of information which can be utilized in a interesting applications with the additional benefit of being much faster than existing methods for contig abundance estimation.

Background

Metagenomics studies DNA sequences obtained from samples which contain a heterogeneous mixture of genomes. A popular area of interest in metagenomics is in the human gut microbiome, where complex communities of bacteria interact with human biology. By studying these communities, we can further our understanding of human health and disease. The diverse mixture of genomes increases the difficulty in applying existing methods, like genome assembly, as well as introducing a need for new analyses like abundance estimation. We would like to estimate the relative abundance at which a particular contig – a sequence of DNA such as a gene or a partially assembled piece of a genome for example – is present in a sample. This information can be used to study the community population structure if the contig is from a known genome. Also, contig abundance estimates can be used in downstream analysis to aid in the process of de novo assembly.

Currently, contig abundance estimates are obtained by mapping each read in a sample to a set of contigs of interest. Abundance estimates are then obtained by a measure over the number of reads mapped to each contig. Read mapping is a complex process with many parameters which can affect the estimates. Furthermore, ambiguous read mappings are typically decided with a random choice. A final issue with read mapping is that it is computationally expensive and requires the creation of a new index for each set of contigs to be examined. Many metagenomic studies produce hundreds of samples and we seek a method which can perform these estimates quickly. Also, if each sample produces a set of contigs, we would like to have a method which allows us to be flexible in choosing which combination of contigs to use. Generating indices for each set of contigs is possible, but not ideal.

Introduction

A k-mer is a substring of length k and there are exactly $l - k + 1$ k-mers in a string of length l. The first step in k-mer count analysis is to count the frequency with which each k-mer appears in a given set of reads. This yields a database, call it D, which maps k-mers to counts. Then, given a particular contig, we can iterate over its component k-mers and obtain a list of counts by looking each one up in D.

Constructing a database of k-mers for a single sample is a mature task in software engineering and there are tools which do so – some of which trade memory usage for disk space and vice versa. In general though, counting k-mers for a single sample is very fast. We defer giving exact measurements comparing read-mapping to k-mer counting because read-mapping is dependent on the set of references, whereas k-mer counting is not.

K-mer count analysis has only two parameters at the basic level: the length of the k-mer, k, and whether the orientation of the DNA strand is known. k is typically chosen to be at least 20 and commonly ranges from 20 – 31. Choosing k too small will yield non-specific count information because the likelihood of a k-mer being unique to a genome shrinks as k decreases. However, as k increases, so does the likelihood of a k-mer containing an error. In addition, the performance benefits in time and space usage are only present for some reasonable choice of k. The second parameter determines whether a k-mer and its reverse complement are considered to be the same k-mer or not. If the strandedness of the sample is not known, we collapse the 4^k possible k-mers into 2^{2k-1}.

Motivation

Analyzing k-mer counts alone discards locality information contained in each read. However dire that seems, the goal of this approach is to explore how much information we can retain by counting the frequency of k-mers and obtain new methods for estimating contig abundance which are much faster.

Speed and storage

The explosion in sequence data is a joy for biologists and a challenge for computer scientists. Exponential growth in sequence data cannot be feasibly matched by an exponential growth in computational resources. While there is research in efficiently compressing sequence data due to its inherent redundancy, analysis on this compressed data remains equivalent to analysis on the raw sequence data – so this addresses storage costs but not time to analyze. Storing the database of k-mer counts for a sample provides an alternative lossy compression scheme which allows for analysis over the compressed data directly.

Once k-mers are counted for a sample, its database can be reused in analyzing new contigs as needed. Compared to read-mapping, which requires re-indexing the set of contigs and going through every read again, k-mer counting can provide a huge speedup to repeated analysis with new contigs.

Current methods are sensitive to the reference

There has been active research in applying k-mer count analysis to RNA-seq data for computing transcript abundance. While it may be alluring to use these tools in a new context, like metagenomics, it is not clear whether these tools are suitable for a new domain. Recently, a tool which was originally created for transcript quantification in RNA-seq data, Kallisto, has been applied to metagenomic samples with reported success. However, by removing a small subset of the references and re-running the tools on the same sample, we observe a very high variance as shown below.

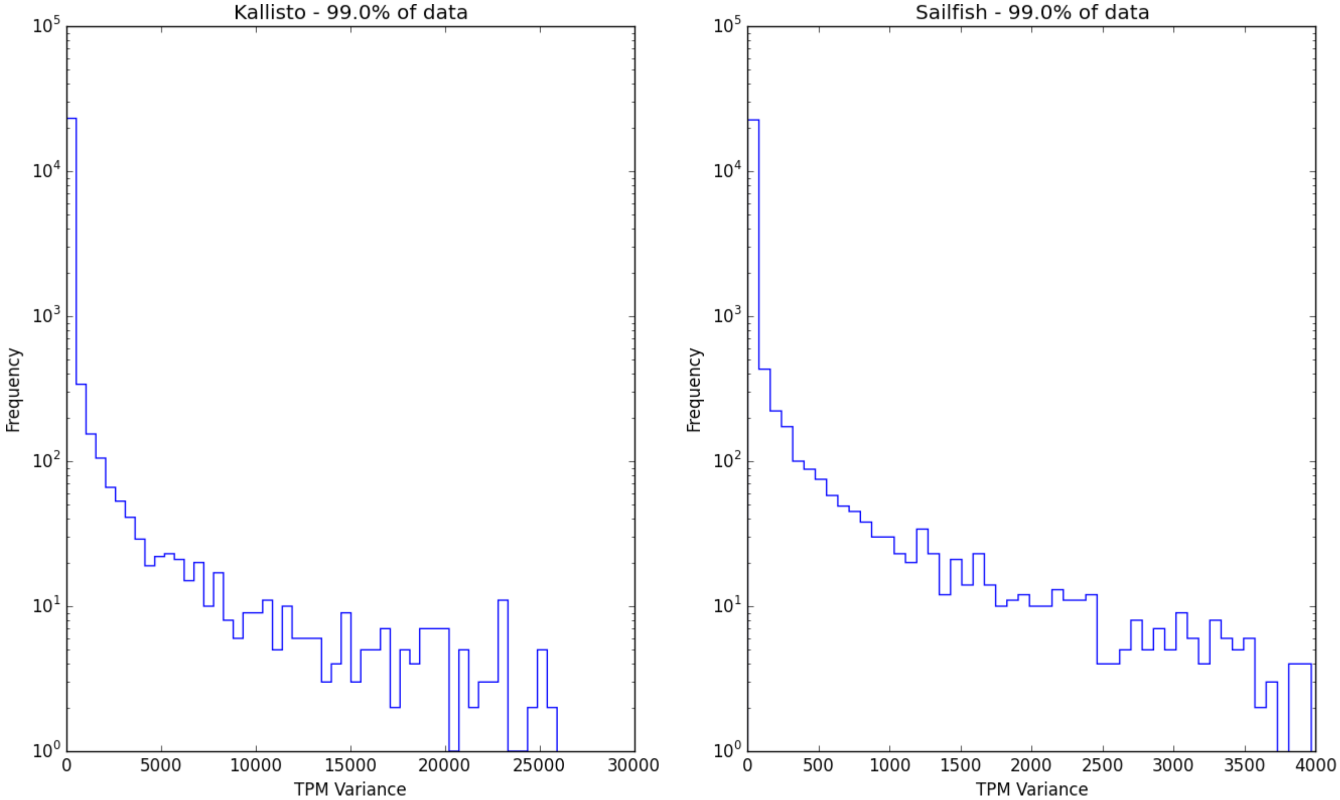

Figure 1: We take the variance of the TPM reported for each contig across 10 samples and plot the distribution of variance. (log-log scale)

Introducing a dependence on the set of contigs used is undesirable in a metagenomic context. We suspect this is less of an issue in RNA-seq experiments because there is more knowledge in what is contained in the sample. But in many metagenomic applications, we have little prior knowledge and want to avoid being penalized for this. While it may be argued this is becoming less of a problem with the growing number of reference genomes, there will always be a need for exploratory methods.

Simulated Community

In order to test our novel methods, we have created three simulated metagenomic communities of increasing complexity. Each sample contains bacterial genomes at 20-fold copy variation and paired-end reads were simulated at 2X coverage.

A Naive Approach

One of the simplest approaches for estimating the abundance of a contig in isolation is to take the mean of it's corresponding k-mer counts in the sample. However, when these estimates are compared to the known copy number of the genome from which they originate, we can see accuracy degrade as complexity increases.

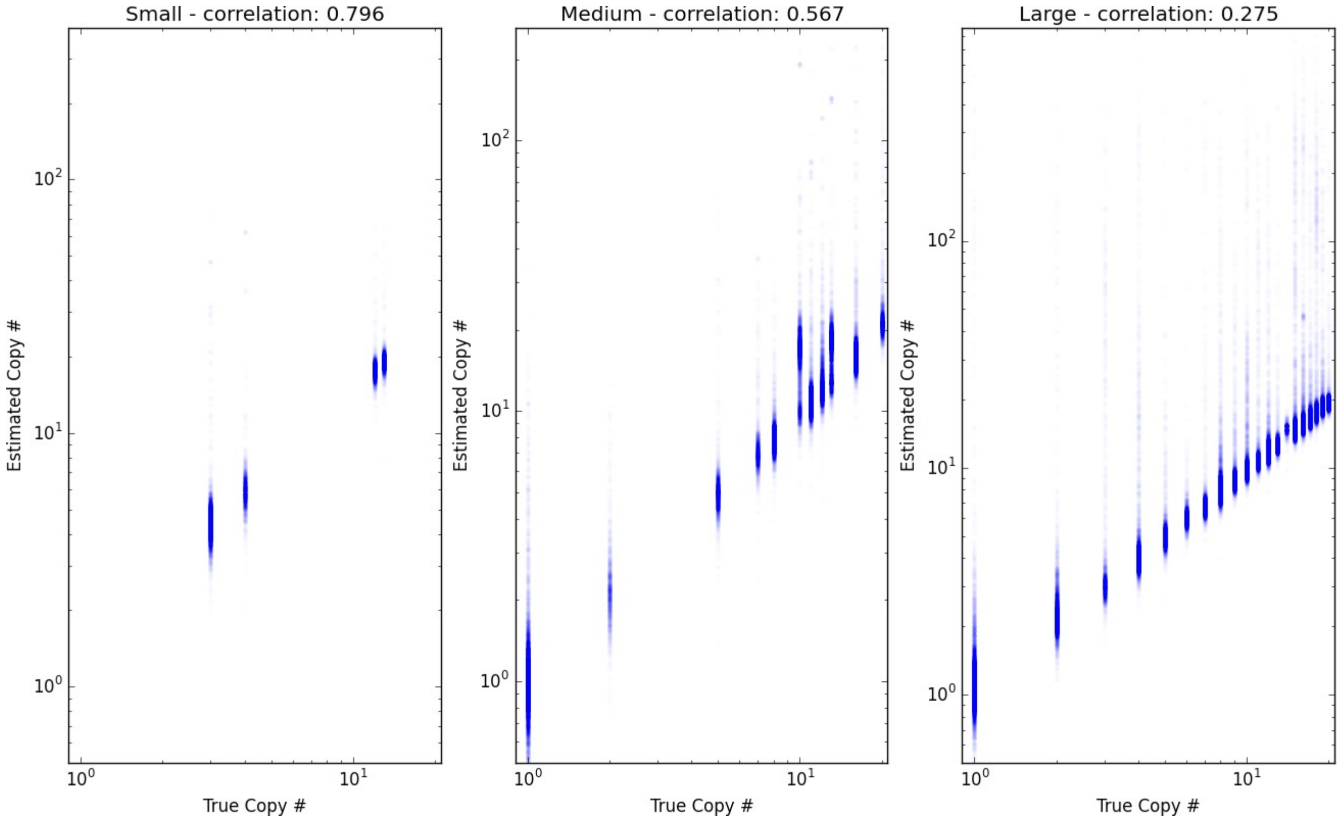

Figure 2: Reference-free contig abundance estimation using the mean of k-mer counts.

Blame

As the complexity of the samples increase, the number of k-mers which are shared by genomes increases, as seen in Figure 3. This causes our simple method to incorrectly assign the full count of each k-mer. We refer to this concept as *blame*, because each contig is only responsible for some portion of the observed counts, not all of them. In order to correct for this, previous methods like Sailfish have done an iterative procedure to distribute counts with proportion to the current abundance estimate. However, since a goal of this work is to explore methods which do not depend on the set of contigs in question, we have explored other options. The approaches we explored seek some transformation of a contig's observed count distribution to compensate for shared counts.

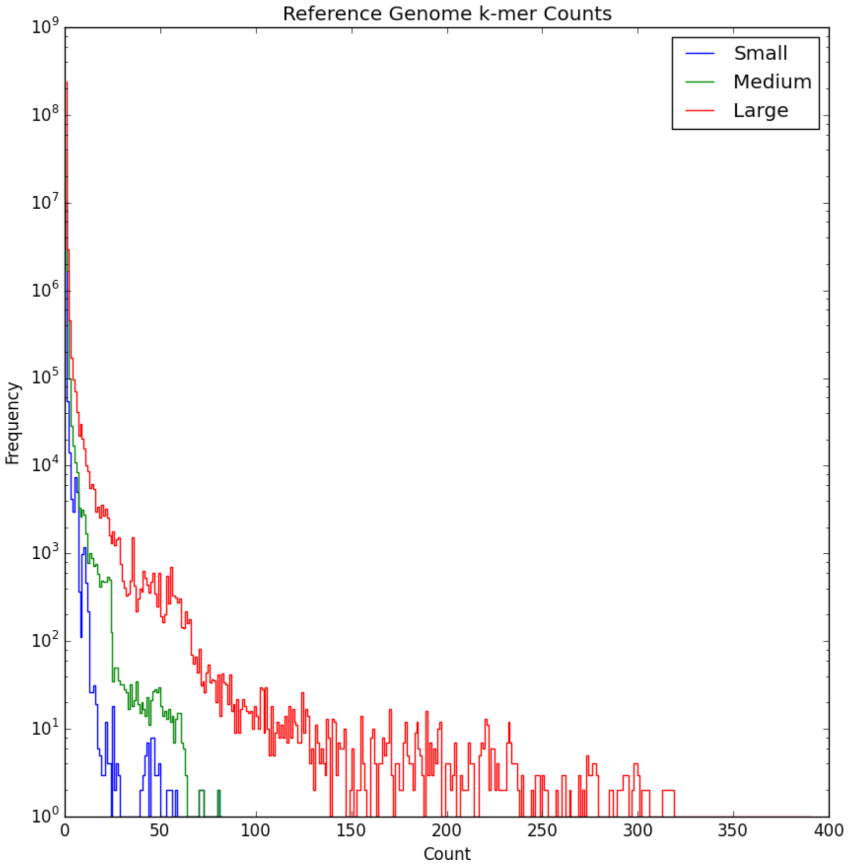

Figure 3: K-mer count distribution for each set of full reference genomes. (log scale)

Transforming The Count Distribution

In Figure 4, we observe that contigs which are among the highest 1% of estimated abundance – and thus incorrectly quantified – have multiple peaks at very high counts. If we eliminate counts above a certain threshold, we will still incorrectly classify those which lack a defined peak at some other count. A promising modification on our previous method is to take the mean of the log counts because this lessens the impact of high counts, while still allowing them to contribute to the estimation. These results can be seen in Figure 5.

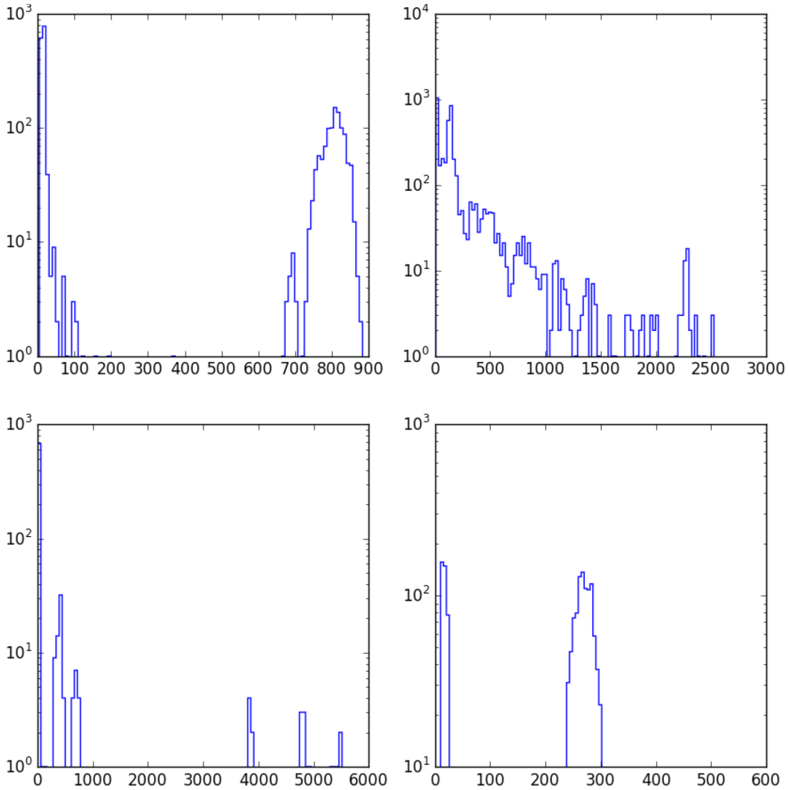

Figure 4: Contigs from the large simulated community with abundance estimates in the highest 1% show multiple peaks in their count distribution as well as very high counts.

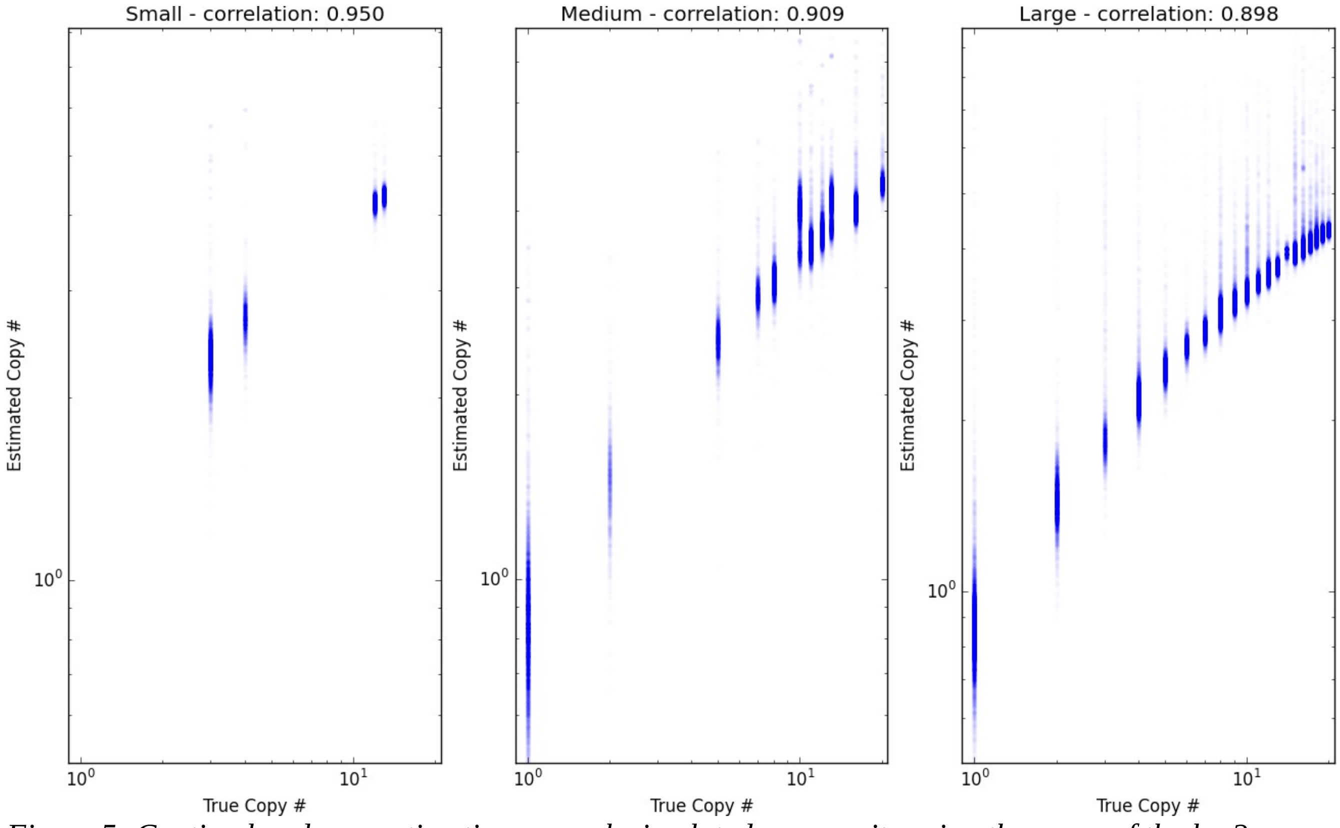

Figure 5: Contig abundance estimation on each simulated community using the mean of the log2 counts.

Discussion

Using the mean of the log counts seems to perform well in these initial tests. However, low abundance contigs have a very large variance in their estimates. Also, extremely high counts in the sample can lead to overestimates even with the log transform. We suspect that an adaptive transform should be used based on the distribution of counts in the sample. At its core, this amounts to finding some function on counts based on the whole sample distribution to maximize the accuracy of contig abundance estimates. A second dimension can be added to this transformation to account for the frequency of observed counts. While machine learning methods could aid in the process of finding insight into functions which work, we seek a method with theoretical background.

Future Directions

Besides direct extensions of this line of work as mentioned in the discussion, we have come across many ideas which would be interesting to explore in their own right.

Differential k-mer analysis

Even though k-mer count analysis has only two hyperparameters at its core, it is still an open question on how to appropriately choose *k*. One complicating factor in this decision process is that the number of possible k-mers is exponential with *k* which leads to non-smooth changes in the resulting analysis. It would be interesting to use a range of *k* values in a single estimation procedure and see if there are features of a contig's count distribution which are persistent.

Multi-sample k-mer database

It is desirable to store k-mer counts across multiple samples compactly and with fast lookup times. This has direct application in our current work, where we would like to estimate the abundance of a set of contigs across hundreds of samples. In addition, k-mers which appear in many samples could lead to higher storage efficiency than storing each database individually. However, many k-mers appear only in a few samples, leading to a sparse table in some parts and dense in others. These challenges present a nice software engineering project.

Approximate counts

If our final procedure for estimating abundance uses log counts, then we can drastically reduce storage in both the counting process and the resultant database. There has been lots of work in approximate counting in other fields and it would be nice to apply this idea to our work.

Topology of k-mer space

Mathematicians study high-dimensional surfaces using topological methods and these methods can also be applied to a discrete sampling of a space. K-mers naturally live in a sparse *k* dimensional space and we would like to explore the structure of k-mers which occur in all known genomes. Equivalently, one could examine properties (cliques, distribution of degree, etc) of the Hamming graph constructed by all observable k-mers. A deeper exploration should also consider the frequency of each k-mer.

Conclusion

We have shown that while k-mers discard locality information, they retain enough information to be applied in novel ways. Our method for contig abundance estimation in metagenomic samples exhibits no dependence upon the set of contigs being examined and is very fast to perform. Further work will improve the accuracy of abundance estimates by accounting for the distribution of counts in each sample. We have also suggested four new areas of exploration because even though k-mers are simple to explain, they are ripe with information.

References

Bray, Nicolas, Harold Pimentel, Páll Melsted, and Lior Pachter. “Near-Optimal RNA-Seq Quantification.” *arXiv:1505.02710 [cs, Q-Bio]*, May 11, 2015. http://arxiv.org/abs/1505.02710.

Carlsson, Gunnar. “Topology and Data.” *Bulletin of the American Mathematical Society* 46, no. 2 (2009): 255–308. doi:10.1090/S0273-0979-09-01249-X.

Marçais, Guillaume, and Carl Kingsford. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers.” *Bioinformatics* 27, no. 6 (March 15, 2011): 764–70. doi:10.1093/bioinformatics/btr011.

Mitchell, Scott A., and David M. Day. “Flexible Approximate Counting.” In *Proceedings of the 15th Symposium on International Database Engineering & Applications*, 233–39. IDEAS ’11. New York, NY, USA: ACM, 2011. doi:10.1145/2076623.2076655.

Nielsen, H. Bjørn, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, et al. “Identification and Assembly of Genomes and Genetic Elements in Complex Metagenomic Samples without Using Reference Genomes.” *Nature Biotechnology* 32, no. 8 (August 2014): 822–28. doi:10.1038/nbt.2939.

Patro, Rob, Stephen M. Mount, and Carl Kingsford. “Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms.” *Nature Biotechnology* 32, no. 5 (May 2014): 462–64. doi:10.1038/nbt.2862.

Schaeffer, Lorian, Harold Pimentel, Nicolas Bray, Páll Melsted, and Lior Pachter. “Pseudoalignment for Metagenomic Read Assignment.” *arXiv:1510.07371 [q-Bio]*, October 26, 2015. http://arxiv.org/abs/1510.07371.