Approaches for Large Scale Metagenome Assembly

ACH, JMT, CTB

October 3, 2012

1 Introduction

Complex microbial communities operate at the heart of many crucial environmental, ecological, and biomedical processes, providing critical ecosystem functionality that underpins all of biology ([1, 2, 3, 4, 5, 6, 7]). These systems are difficult to study in situ, and consequently, we lack a fundamental understanding of their diversity and function, much less how they self-assemble, maintain themselves, and evolve through time. Advances in DNA sequencing technologies now provide unprecedented access to these communities in the form of millions to billions of short-read sequences of environmental DNA [2, 4, 5]. Even more sequencing is needed to detect the rare species in environmental samples, e.g., up to 50 Tbp for an individual gram of soil [**?**]. Both the read lengths and scale of sequencing data pose new challenges to traditional analysis approaches of sequencing data. Short read lengths and their associated sequencing errors and biases have little biological signal and are noisy, limiting direct annotation approaches against known reference genomes. Further complicating analysis is the majority of genes sequenced from metagenomes are not similar to known genes [1, 5].

De novo assembly of raw sequencing data offers several advantages to metagenomic sequencing datasets. It significantly reduces the total number of sequences for analysis by identifying consensus sequences from overlapping reads. The assembled contigs are longer than sequencing reads and provide gene order. Importantly, *de novo* assembly does not rely on the existence of reference genomes, allowing the discovery of novel elements. The main challenge for metagenomic applications of *do novo* assembly is that current assembly tools do not scale to the volumes of metagenomic datasets being generated. Assembled metagenomes from the rumen, human gut, and permafrost soils required processing of only the most abundant sequences [2, 4, 5]. Traditional assemblers have been designed for the assembly of single genomes

whose abundance distribution and diversity content are significantly different from the mixed populations of metagenomes. Although many new metagenome-specific assemblers have been developed to address characteristics of mixed population assembly, these are limited to sample diversity and volume.

In this study, we present a set of approaches which enable large-scale metagenomic de novo assembly. We initially reduce the dataset volume by normalizing sequencing coverage and removing sequencing errors and biases. We subsequently partition reads based on biological connectivity, resulting in partitions which can be assembled with significantly reduced computational requirements. Using a human gut mock community dataset, we demonstrate that our approaches at reducing and partitioning datasets are effective at producing nearly identical assemblies and consequently extend our approaches to two of the largest published soil metagenomes which have previously been impossible to assemble.

2 Results

2.1 Case study: Assembly of mock metagenome

2.1.1 Evaluation of data reduction through digital normalization and high abundance filtering

Using the described approaches (see Methods and Supp Info), a total of 5.9 million reads (40% of total reads) were removed from a human gut community mock dataset. We compared the presence of reference genomes known to be present within the mock dataset in the original and filtered sequencing reads. Specifically, the representation of reference genomes by sequencing reads and the ability to recover these reference genomes after assembly were evaluated. The abundance of reference genomes was estimated based on the coverage of sequencing reads in the original dataset, and a total of 30 reference genomes were estimated to have been sequenced to a coverage of greater than 3-fold, ranging from 6-fold to 2,000-fold coverages (Table **??** and Figures 1 and 2). Overall, a total of 91% of the available reference genomes were covered by the reads in the original dataset. After digital normalization and high abundance filtering, the remaining sequencing reads covered a total of 93% of the available reference genomes. After assembly (Velvet), the recovery of reference genomes by the contigs assembled from the original and filtered datasets were compared, resulting in recoveries of 43% and 44% of references, respectively. The assembly of the original dataset contained 29,063 contigs and 38 million

bp compared to the filtered assembly containing 30,082 contigs and 35 million bp (Table 2). Comparable recoveries of references between original and filtered datasets were also obtained for other assemblers (SOAPdenovo and Meta-IDBA). Overall, the genomic content of both assemblies were similar to one another, 94%, and furthermore, the filtered and filtered / partitioned assemblies were greater than 99% identical to each other. For the filtered dataset, the time and memory requirements for de novo assembly were significantly reduced. Specifically, the data reduction took less than four hours, and the time and memory for assembly (Velvet) was reduced from 12 GB and 4 hours for the unfiltered dataset to 3 GB and less than an hour for the filtered dataset.

The abundance of assembled contigs and reference genomes was estimated by the sequencing coverage of reads (Figures 5 and **??**). It was found that above a sequencing coverage of five, the majority of reads which could be mapped to reference genomes were also included in an assembled contig (Figure 4). Below this threshold, reads could be mapped to reference genomes but were less likely to be associated with assembled contigs in either the unfiltered or filtered assemblies. Assembled contigs from the original dataset and filtered datasets were used to estimate the abundance of genomes present within the dataset. Using a one-directional, paired t-test of squared deviations, the difference between the estimated abundances of the original and filtered assembled contigs were compared against those of the associated reference genomes. As we expected the filtered assembly to have increased accuracy due to a reduction of errors, we used a one-sided t-test and found that abundance estimations from the filtered assembly were significantly closer to predicted abundances from reference genomes (p-value of 0.032).

2.1.2 Evaluation of partitioning reads based on connectivity

For the 9 million reads in the filtered dataset, we identified a total of 85,818 partitions containing a minimum of five reads and among these, only 2,359 partitions contained reads originating from more than one genome. With the exception of one partition containing reads from 36 reference chromosomes and/or plasmids, all other partitions contained reads from less than nine genomes. The number of partitions associated with reads which were associated with reference genomes were examined and, in general, reference genomes with high coverage had fewer partitions (Table **??**). There were 112 partitions associated with high abundance reference genomes (coverage above 25) compared to 2,771 partitions associated with low abundance genomes (coverage below 25). In the case of multiple partitions associated with a low abundance genome, reads aligning

to similar regions of a reference genome were associated with a the same partition (SI Figure **??**).

To further evaluate our approaches, spiked reads (containing errors) from known reference genomes were introduced to the mock community dataset and the recovery of these references were evaluated. Initially, a single spiked reference genome (E. coli strain E24377A, NC_009801.1) was added to the a "spiked" mock community dataset and processed identically to the original mock dataset. This resulted in similar amounts of data reduction after digital normalization and partitioning (Table **??**). Among the 81,154 partitioned sets of reads, we identified only 2,580 partitions containing reads from multiple genomes. A total of 424 partitions contained reads from the spiked E. coli genome, among these 201 partitions contained *only* spiked reads (Figure 6). The assembly of these 424 partitions resulted in contigs which when aligned against the E. coli strain E24377A genome overlapped 99.5% (4957067 of 4979619 bp) of the original reference. Next, a similar analysis introducing five closely-related spiked E. coli strains into the mock community dataset was performed. Partitioning this "mix-spiked" mock community dataset resulted in 81,425 partitions, of which only 1,154 partitions contained reads associated with multiple genomes. Among the partitions which contained reads associated with a single genome, 658 partitions contained reads originating from one of the spiked E. coli strains. In partitions containing greater than one genome, 224 partitions contained reads from a spiked E. coli strain and either another spiked strain or from the mock community dataset (Figure 6). The partitions containing any reads originating from the spiked E. coli strains were identified and assembled independently. Among the resulting 6,076 contigs, all but three contigs could be identified as originating from a spiked E. coli genome (e.g., top blast hit). The remaining three contigs were greater than 99% similar to HMP mock reference genomes (NC_000915.1, NC_003112.2, and NC_009614.1). The contigs associated with E. Coli were aligned against the spiked reference genomes, recovering greater than 98% of each of the five genomes. Many of these contigs were associated with reads originating from multiple genomes (Figure 7), 3,075 contigs (51%) contained reads which were mapped from more than one spiked genome. This result is comparable to the fraction of contigs which are associated with multiple genomes when the original (unfiltered) dataset is assembled, resulting in 4,702 contigs associated with the "spiked reads", of which 66% contained reads originating from more than one spiked genome.

2.2 Characteristics of soil metagenomes

Our approaches were extended to the de novo assembly of two soil metagenomes. Previously, the assembly of the Iowa corn and prairie datasets containing 1.8 billion and 3.3 billion reads, respectively, were impossible in 500 GB of memory. A 75 million reads subset of the Iowa corn dataset alone required 110 GB of memory (Figure 8). Applying the same filtering approaches as applied to the HMP mock dataset, the Iowa corn and prairie datasets were reduced to 1.4 million and 2.2 million reads, respectively, and after partitioning, a total of 1.0 million and 1.7 million reads remained, respectively. Notably, the Iowa corn and prairie were sampled at significantly lower sequencing coverages than the mock community dataset. Whereas has the mock dataset had only 33% of its k-mers (k=20) present less than ten-times in the dataset, the Iowa corn and prairie datasets contained 53% and 43%, respectively, of all k-mers with less than ten-fold coverage. The large majority of k-mers in the soil metagenomes are relatively low-coverage (Figure 10), and consequently, digital normalization did not remove as many reads in the soil metagenomes.

2.2.1 Assembly of soil metagenomes

Based on the mock community dataset, we estimated that above a sequencing depth of 6, the large majority of sequences could be assembled (Figure 4). Given the greater diversity expected in the soil metagenomes, we normalized these datasets to a coverage threshold of 20. After partitioning the filtered datasets, a total 31,537,798 and 55,993,006 partitions (containing greater than five reads) in the corn and prairie datasets, respectively, were identified. For practical assembly, partitions were grouped together such that groups contained partitions with similar numbers of reads and no group contained larger than 10 million reads. Once partitioned, each group of reads could be assembled in less than 14 GB and 4 hours, enabling evaluation of multiple assemblers and various assembly parameters.

The final assembly (Velvet) of the corn and prairie soil metagenomes resulted in a total of 1.9 million and 3.1 million contigs, respectively, and a total assembly length of 912 million bp and 1.5 billion bp, respectively. To estimate abundance of assembled contigs and evaluate incorporation of reads, all quality-trimmed reads were aligned to assembled contigs (greater than 300 bp). Overall, for the Iowa corn assembly, 8% of single reads and 10% of paired end reads mapped to the assembly. Among the paired end reads, less than 0.5% aligned disconcordantly. Similar results were found for the Iowa prairie assembly with only 0.6% paired ends aligned

disconcordantly and slightly increased numbers of reads mapped with 10% of single reads and 11% paired end reads (Table 3). The read coverages of assembled contigs within the soil metagenomes were estimated (Figure 11). Overall, there is a positively skewed distribution of coverage of all contigs from both soil metagenomes, biased towards a coverage of less than ten-fold. The Iowa corn and prairie assemblies contained 48% and 31% of total contigs with a median basepair coverage less than 10.

2.2.2 Content of soil metagenome assembly

Two options -

1) Summarize recA housekeeping results. Can definitely show phylogeny table/chart – tree is a bit tricker. - THIS IS GOING TO BE WORDIER TO EXPLAIN 2) MG-RAST phylogeny distribution and abundance distribution (need to work with ANL on getting this finessed but easy since Im here now) – ADINA'S PREFERENCE FOR WORD COUNT AND EASE OF EXPLAINATION

3 Discussion

3.1 Community sequencing has variable coverage motivating normalization

The gut mock community dataset used in this study originates from an uneven mixture of DNA extracted from 21 different bacterial strains. Although the diversity represented by this mock community is extremely low compared to that of most environmental metagenomes, it represents a simplified dataset with the key advantage of the ability to evaluate analyses through the availability of source genomes. Like metagenomes, this mock community dataset contains uneven sampling of various genomes, evidenced by reads present at a broad range of sequencing depths, from singletons to reads with abundances of greater than 255 and peaks of highly abundant reads at sequencing depths 8 and 21 (Figure 9). This read coverage is reflective of genomes present at high abundances which are readily sampled resulting in reads present at high sequencing depth and conversely, under sampled genomes which result in reads with low sequencing depths.

For this dataset, the sequencing depth of reads necessary for assembly was estimated to be 6-fold. Below this threshold, reads were more unlikely to be present in assemblies, due to both sequences that were unable to be assembled (low coverage) or assembled with errors.

More sequencing depth past this point did not greatly increase the number of assembled contigs suggesting a sequencing coverage "sweet point" at which sufficient sequences are present with which to complete assembly and beyond which further sequencing is redundant. Sequencing beyond this coverage threshold not only is redundant for the purpose of *de novo* assembly but also increases the number of sources of errors. In previous studies of an Illumina E. Coli dataset, the removal of 90% of normalized reads resulted in nearly an identical assembly of the E. Coli genome while removing over 50% of the sequencing errors (Brown, diginorm paper). Removal of redundant sequences from the metagenomic datasets provides more uniform coverage while reducing the volume of sequences and errors prior to assembly. To further remove errors from our dataset and improve assembly, we also targeted Illumina sequencing artifacts in metagenomes which have previously been shown to be correlated to high abundance sequences (Howe paper). The benefits to this approach are observed within the mock dataset with an increase in recovery of reference genomes (2%) after normalization and high abundance filtering.

For the mock community dataset, assemblies obtained with and without our approaches were similar in content and recovery of reference genomes (Figures 1 and 2). For highly abundant genomes (i.e., three plasmids of *Staphylococcus* epidermidis (NC_005008.1, NC_005007.1, and NC_005003.1), the filtered assemblies did not recover large portions of these genomes despite a comparable presence of reads in both the original and filtered datasets for two of the three plasmid genomes. Conserved regions among these genomes shared 90% identity over 290 bp suggests the presence of repetitive elements. During normalization, such sequences would appear as high coverage elements and would be targeted for removal. Compared to the filtered dataset, the original dataset more likely contained a higher coverage of a larger spans of these repetitive regions, enabling assembler heuristics to extend the assembly of these sequences. The observation of this result, though rare among the mock reference genomes, highlights a current shortcoming of our approach, and indeed for most short-read assembly approaches, related to repetitive regions and/or polymorphisms which should be considered in its application. For the soil metagenomes, the data reduction made possible by our methods may cause information loss which may have been useful for assembly, but without doing so, we were previously unable to complete these assemblies. Evaluation with the mock community dataset suggests that this information loss is minimal overall and that our approaches result in a comparable assembly whose abundance estimations are similar (even slightly more accurate).

3.2 Community sequencing is made up of connected genomes motivating partitioning

A broad range of diversity must be represented in metagenomic assembly graphs. These graphs contain continuous paths of short, overlapping sequences which are used to determine read overlaps. Two or more genomes which are thoroughly sequenced would be expected to be connected in a single assembly graph by conserved elements such as those within 16S rRNA genes. For most metagenomes, however, the majority of constituent genomes are undersampled resulting in only fragments of connectivity. Thus, these assembly graphs are expected to contain multiple, separate connected sets of reads or subgraphs representing sequences from different genomes or genomic fragments. Our partitioning approach targets these subgraphs to divide large metagenomes into subsets which reflect the biological characteristics of the originating dataset.

To enable partitioning of metagenomic datasets, we must first remove sequencing biases which cause artificial connectivity within metagenomic assembly graphs (Howe). In the mock community dataset, the removal of these sequences (combined with normalization) did not significantly alter the recovery of reference genomes through de novo assembly and enabled the division of the mock community dataset into thousands of disconnected partitions. The resulting assembly of these partitions was nearly identical to the assembly of the reduced dataset. The large majority of these partitions contained reads from a single reference genome, supporting our previous hypothesis that most connected subgraphs contain distinct genomes. In general, high coverage genomes contained fewer partitions because they were well-sampled with sequencing, and most reads could be connected together within the assembly graph. Low coverage genomes contained more partitions as these assembly graphs were fragmented due to undersampling. The same partition was associated with reads from fragmented regions (Figure **??**) since these reads were connected within the assembly graph.

Our approaches also successfully could recover sequences from one or more E. Coli strains computationally spiked into the mock community dataset. For a spike of a single E. Coli strain, we identified the fraction of partitions containing associated reads and from these partitions alone could recover 99% of the original genome (Figure **??**). When closely related strains were spiked into the mock dataset, we could recover the large majority of the genomic content of these strains but largely in chimeric contigs which contained reads from multiple reference genomes

(Figure 7). However, this result is not unique to our approach as assemblies of the unfiltered dataset resulted in a slightly higher fraction of assembled contigs associated with multiple references. Overall, closely related sequences which result from either repetitive or inter-strain polymorphisms are a challenge to assemblers, and our approaches are not specifically designed to target such regions. However, the partitions resulting from our approach (without digital normalization) could provide a subset of sequences which could be targeted for more sensitive assembly approaches for such regions (i.e. overlap-layout-consensus approaches or abundance binning approaches (cite Itai)).

A valuable result of our partitioning approach is that it effectively subdivides our datasets into sets of reads which can be assembled in parallel, and consequently, with less computational resources. For the soil metagenomes, grouped partitions could be assembled in less than a day and in under 14 GB of memory enabling the usage of multiple assembly parameters (e.g., k-length) and multiple assemblers (Velvet, soapdenovo, and meta-idba).

3.3 What we gained from the soil assembly

Discussion of above analysis

4 Conclusion

CONCLUSION: THEME: PARTITIONING IS NOT ONLY AWESOME FOR ASSEMBLY BUT...

4.1 Assembly Pipeline

The entire assembly pipeline for the mock community is described in detail in an iPython notebook available for download at XXX accompanied by a web-based tutorial. Soil assembly was performed with the same pipeline and paramater changes are described in Supp Info.

4.1.1 Estimation of assembly requirements for soil metagenomes

Subsets of the Iowa corn metagenome were assembled with the Velvet assembler (v1.2.07) with the following parameters: velveth K=45, -short and velvetg -exp_cov auto -cov_cutoff auto, -scaffolding no. The time and memory for each assembly was estimated up to a maximum of 150 hours and 100 GB.

4.1.2 Partitioning and *de novo* assembly of disconnected reads

Normalized and filtered datasets were loaded into a probablistic representation of the assembly graph as previously described in Pell et al, and disconnected partitions of the resulting graph were separated. Partitions containing less than 5 reads were discarded. Each partition was subsequently assembled using the Velvet assembler with the same setting as describe as above, with the exception that K=35-59 and shortPaired setting was used for paired end reads. The resulting contigs greater than 300 bp from multiple-K assemblies were dereplicated with CD-HIT (XXXX, 99% similarity) and merged with Minimus2 (XXXX).

4.2 Comparing coverage of reference genomes by reads

Reads in the HMP mock unfiltered and filtered datasets were mapped back to originating genomes using default settings in Bowtie2 (citation). For cases where reads could be mapped back to multiple genomes, a single genome was randomly selected to be identified with each read. Sequencing coverage was estimated for the whole genome as the median base pair coverage for all base pairs in reference genome.

4.3 Comparing assemblies

Resulting assemblies (contigs greater than 300 bp) were compared using the total number of contigs, assembly length, and maximum contig size for each assembly. Assemblies were also aligned to each another using blastn and the resulting coverage of each assembly was calculated. In the case of the mock community, the resulting assemblies were also aligned to sequenced draft genomes of the original isolates and, if applicable, spiked reference genomes. Abundance of assembled contigs and reference genomes were estimated by mapping raw reads with Bowtie (allowing up to 2 mismatches for a match). The median base pair coverage was used to estimate abundances. Associated assembled contigs (greater than 300 bp) from the unfiltered and filtered (digital normalized) assemblies were identified using a blastn alignment (requiring E-value cutoff of 1e-5). Contigs were associated with reference genomes through an identical alignment approach.

4.4 Figures and Tables

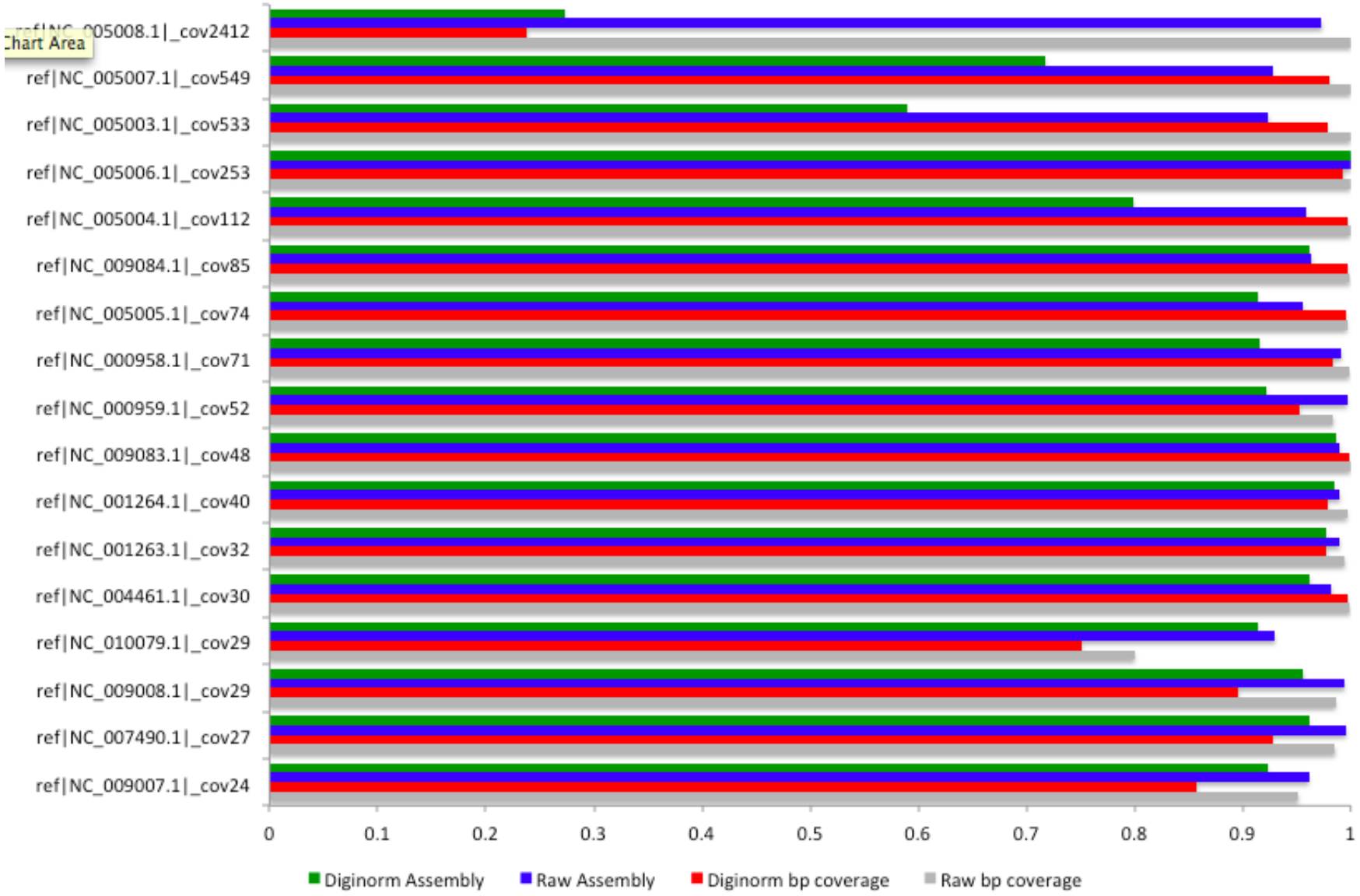

Figure 1: Coverage of reference genomes by unfiltered and filtered assembled contigs and unfiltered and filtered reads.

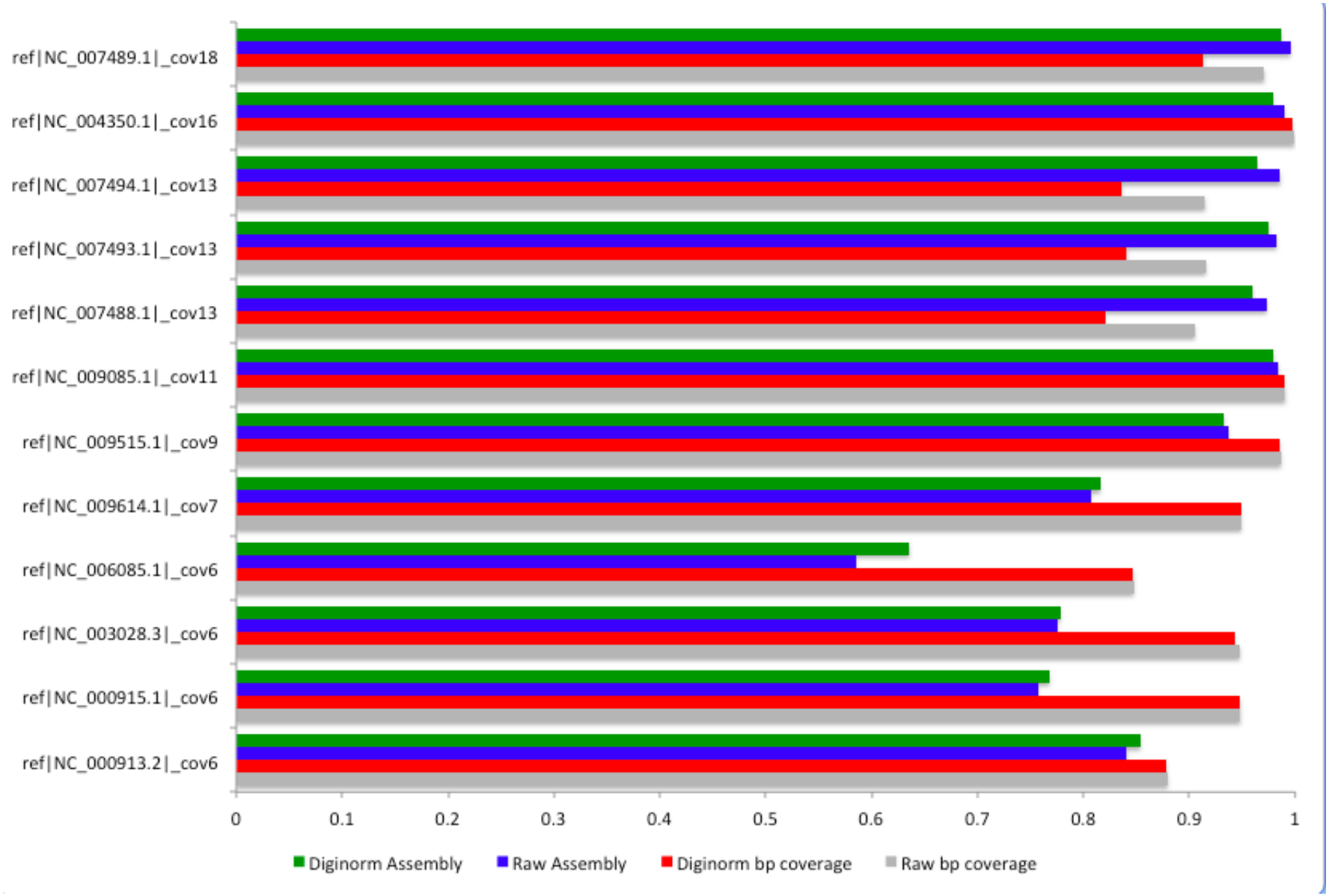

Figure 2: Coverage of reference genomes by unfiltered and filtered assembled contigs and unfiltered and filtered reads.

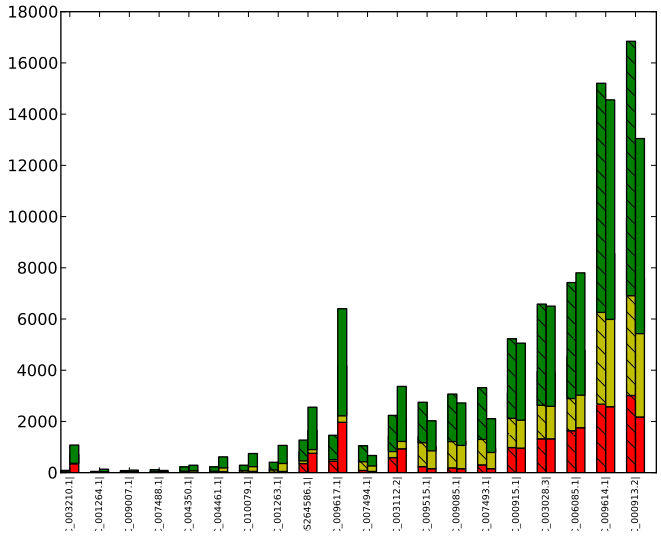

Figure 3: Total number of contigs for unfiltered (bars with hashed lines) and filtered (solid bars) for top twenty references with most assembled contigs. Red indicates contig lengths less than 500 bp, yellow indicates contig lengths between 500 bp and 3000 bp, and green indicates contig lengths greater than 5000 bp.

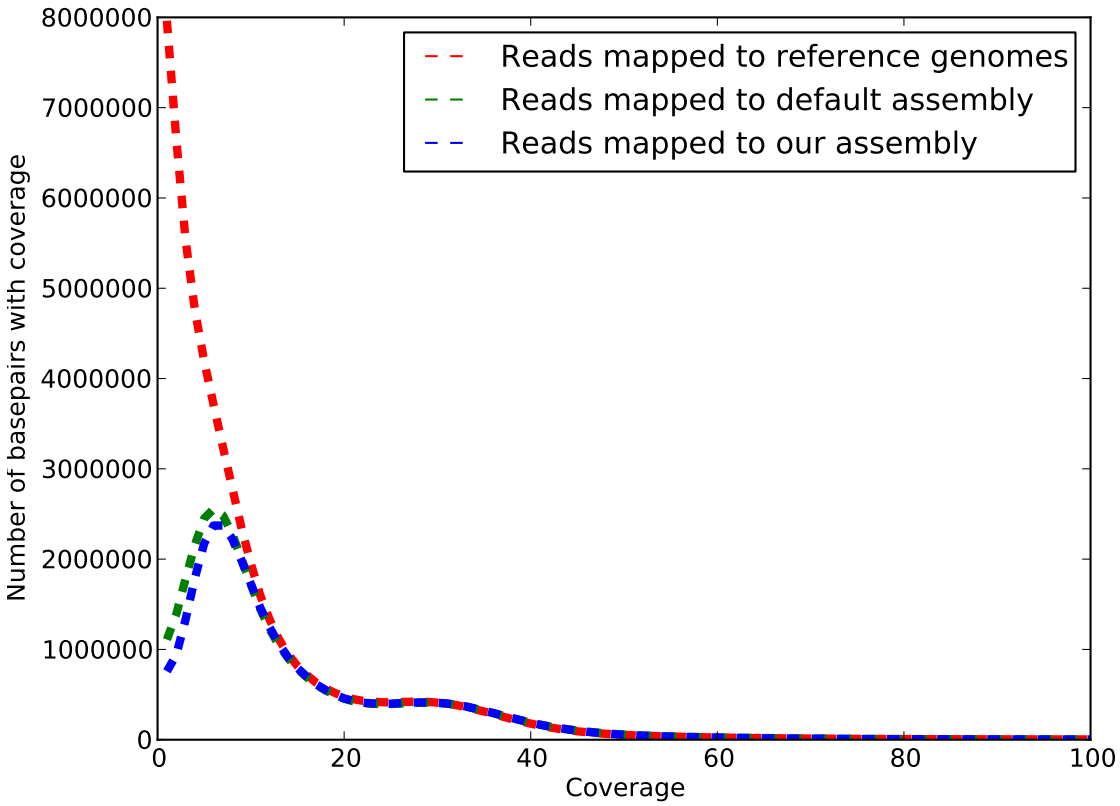

Figure 4: Number of basepairs with specified coverage for reads which map to reference genomes and unfiltered and filtered assembled contigs greater than 300 bp.

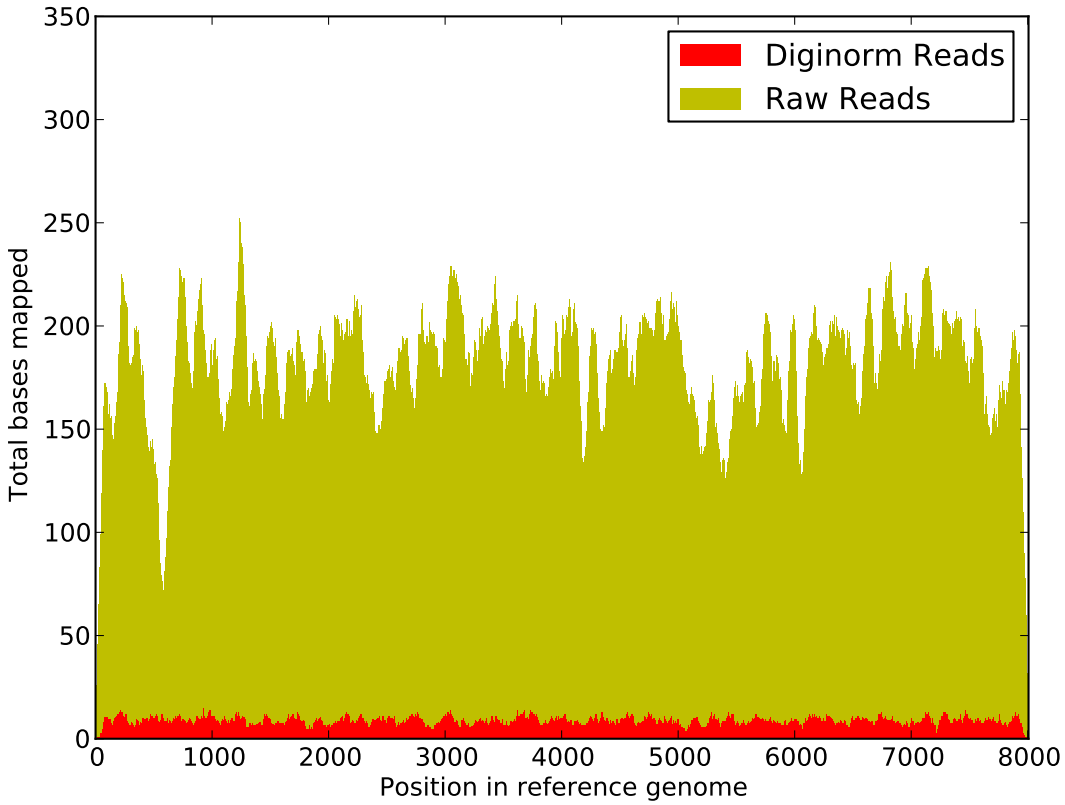

Figure 5: Alignment of reads (colored by originating partition) to reference genome NC_00745901

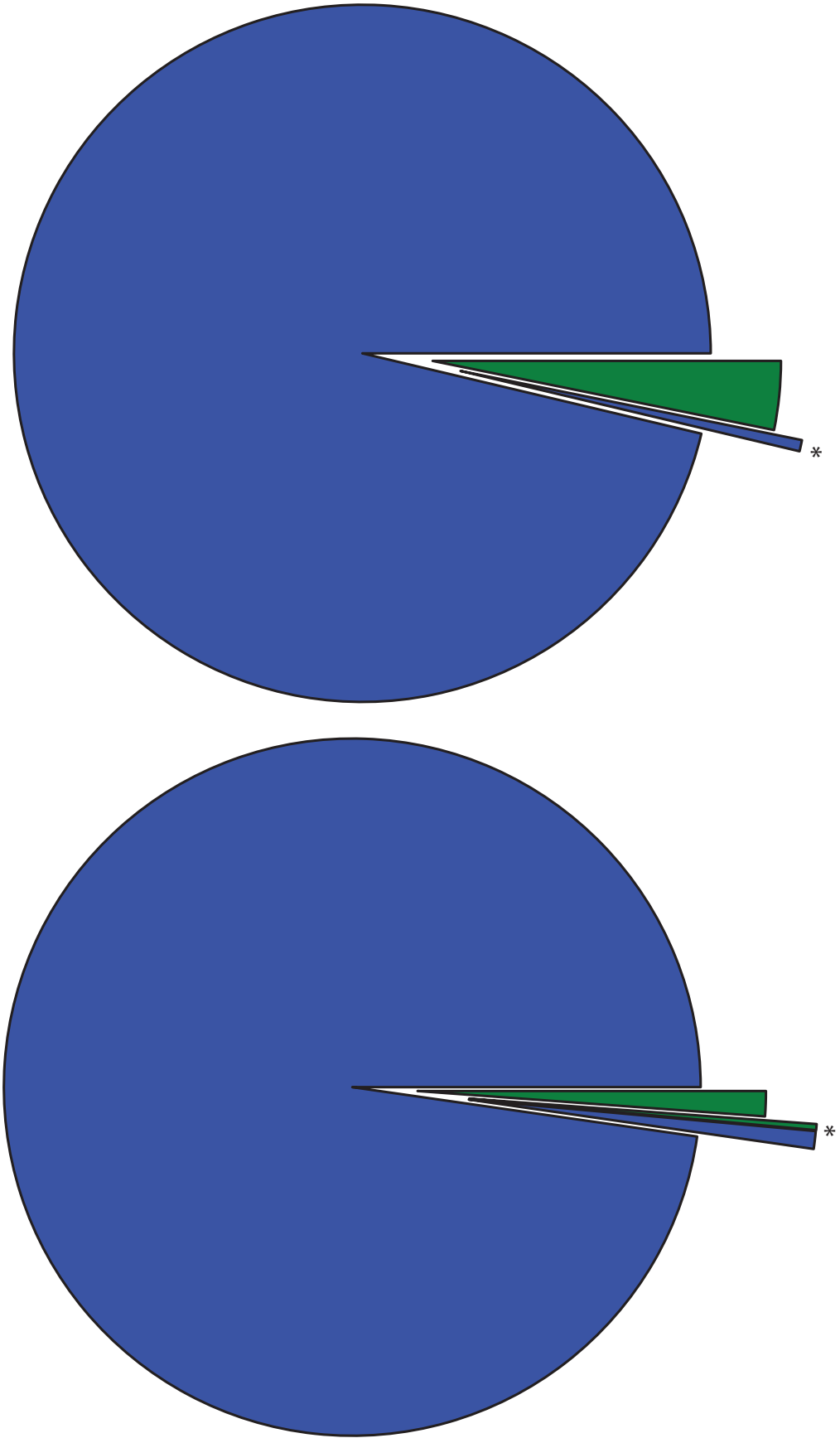

Figure 6: The fraction of partitions in spiked HMP datasets which contain single genomes (blue) and multiple genomes (green). The section marked as * indicates partitions which contain spiked E. Coli reads which were subsequently assembled independently. The top piechart is the single E. Coli spiked dataset and the bottom piechart is the multiple E. Coli spiked dataset.

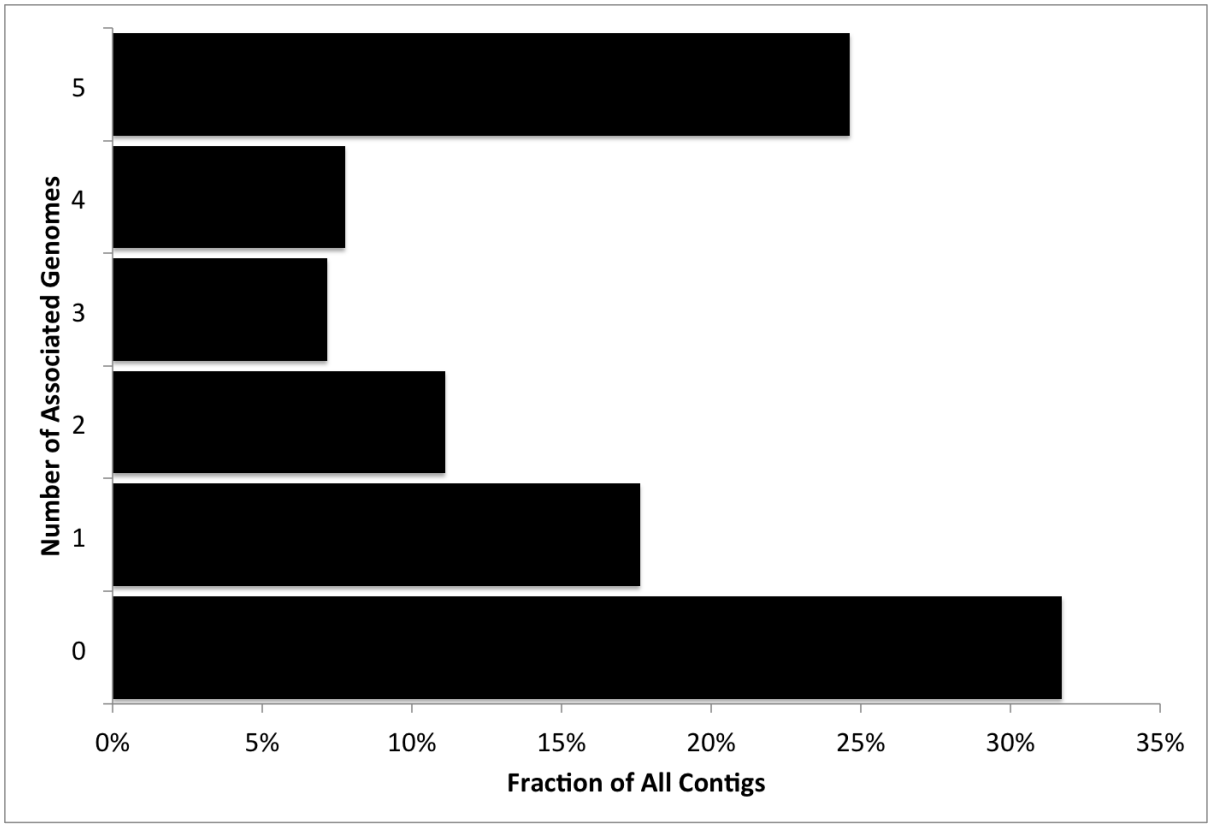

Figure 7: The fraction of assembled contigs assembled from partitions containing spiked E. Coli reads associated with 0 to five of the E. Coli reference genomes. The large majority of contigs contain reads associated with multiple genomes or to no genome.

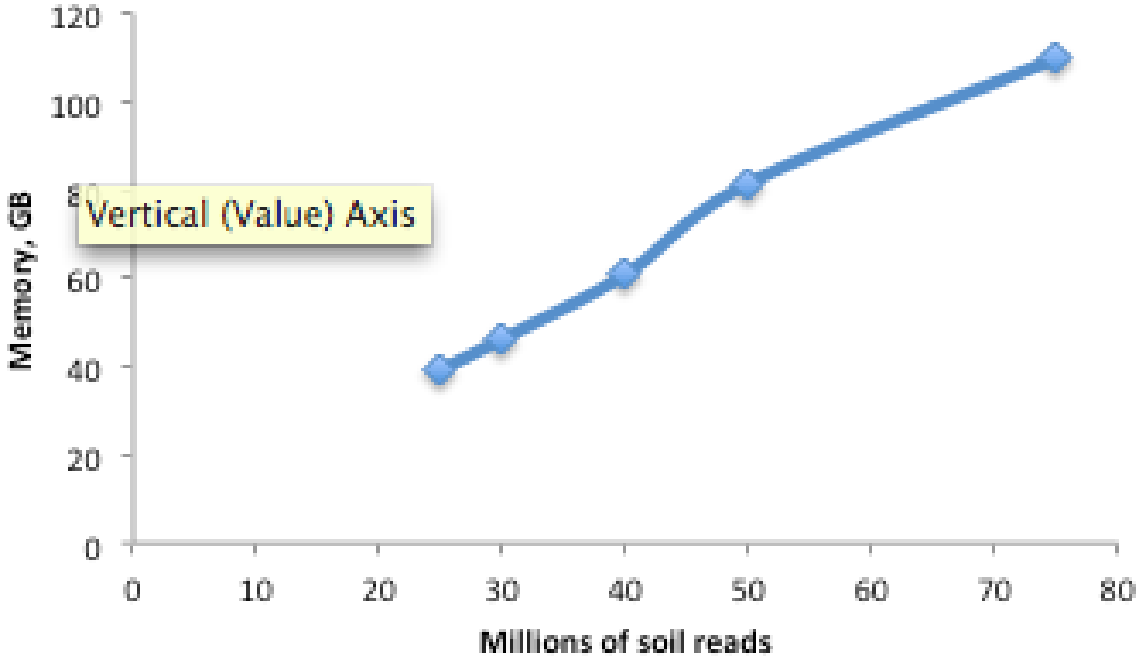

Figure 8: Memory requirements to assemble subsets of Iowa corn soil metagenome

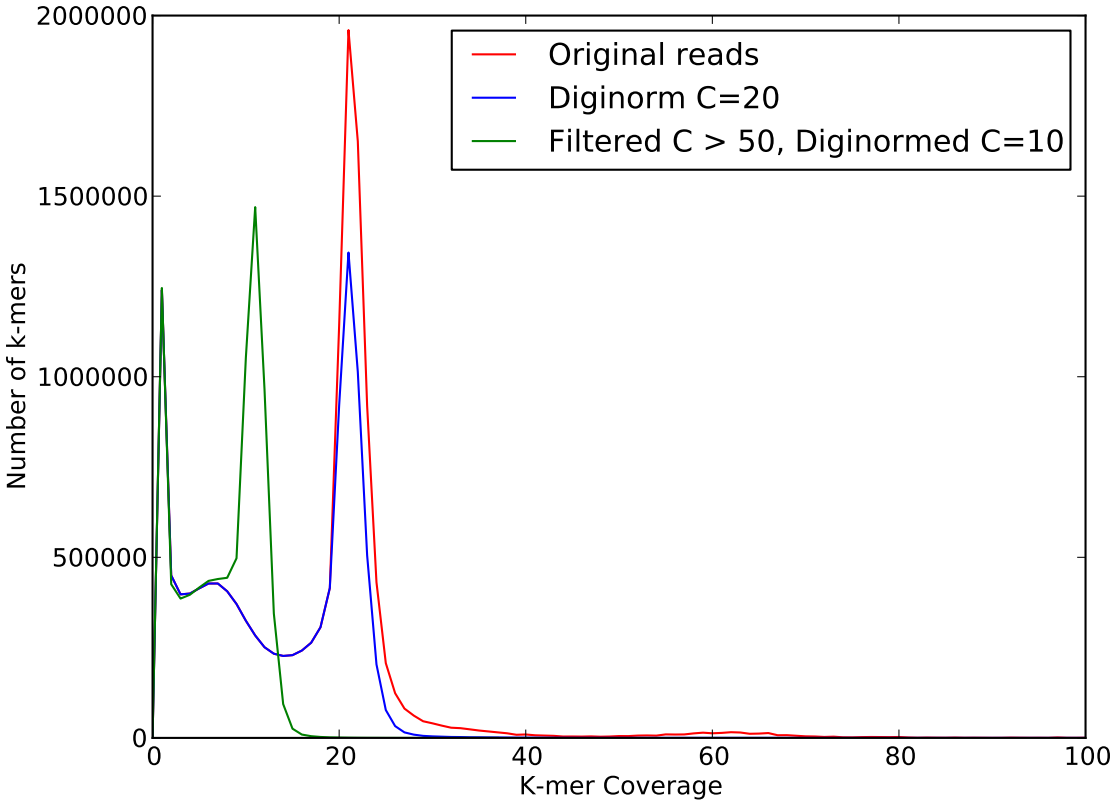

Figure 9: K-mer coverage of HMP mock community dataset before and after filtering approaches.

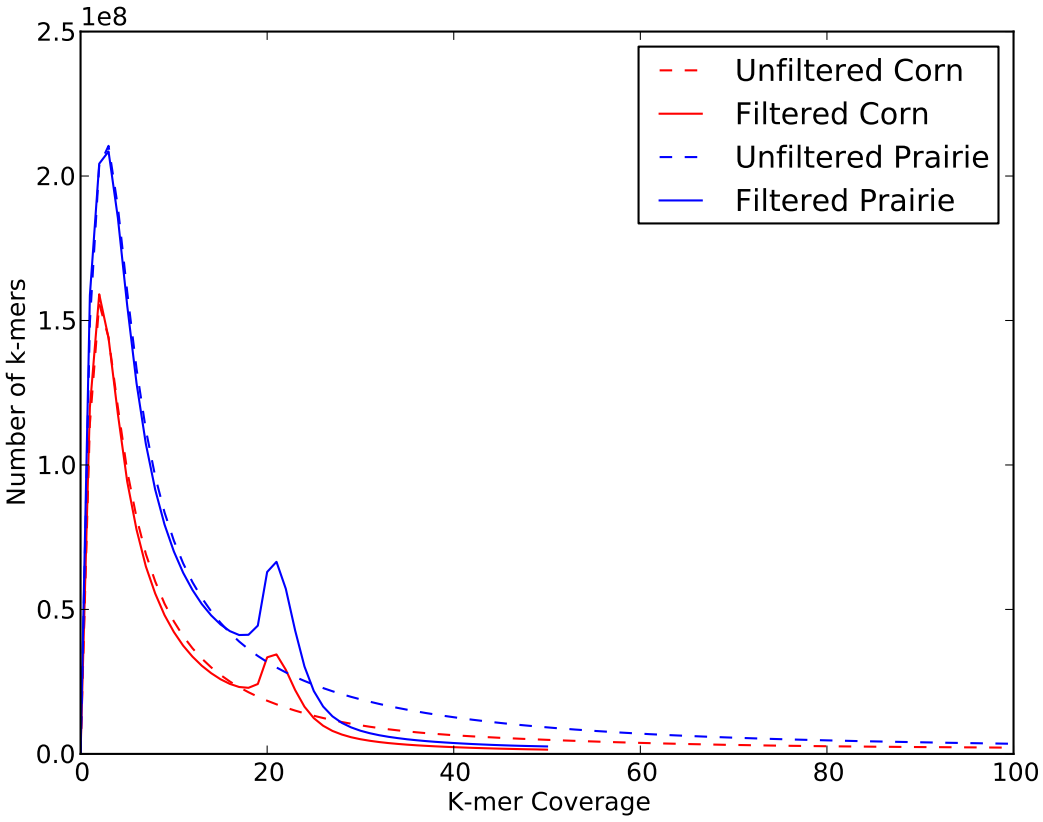

Figure 10: K-mer coverage of Iowa corn and prairie metagenomes before and after filtering approaches.

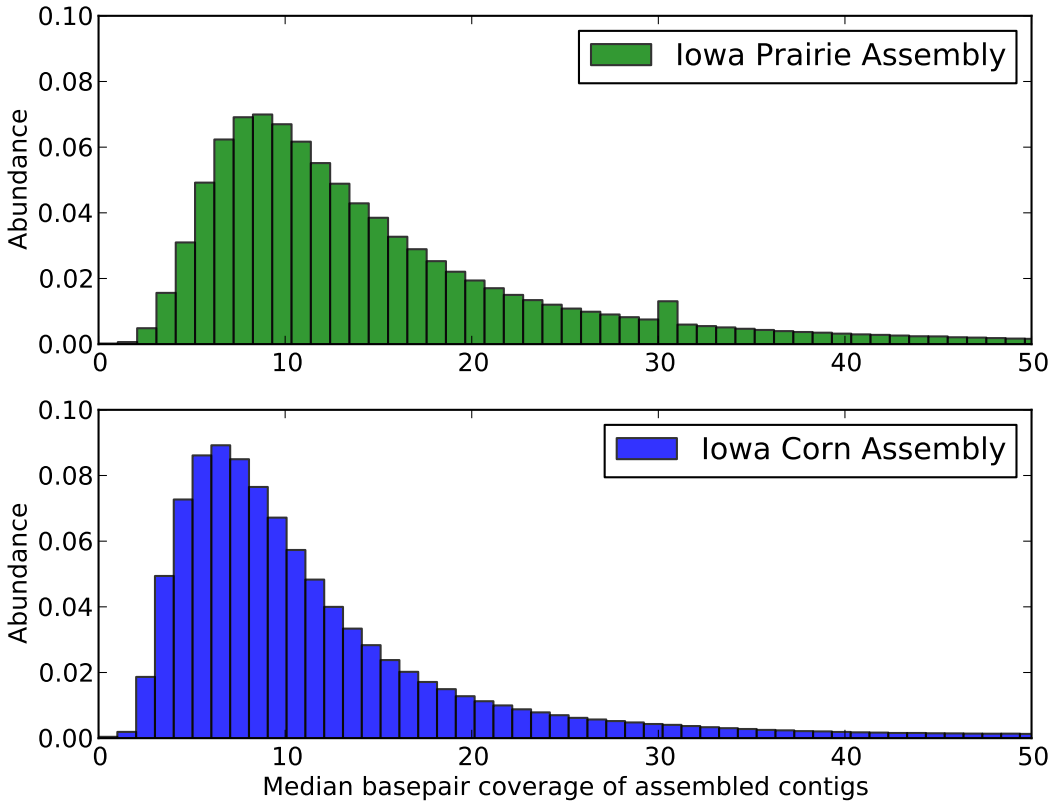

Figure 11: Coverage (median basepair) distribution of assembled contigs from soil metagenomes.

	HMP Mock	HMP Mock Spike	HMP Mock Multispike	Iowa Corn	Iowa Prairie
Raw Reads	14,494,884	14,992,845	17,010,607	1,810,630,781	3,303,375,485
Diginormalization/High Abund. Filter	8,656,536	8,189,928	9,037,142	1,406,361,241	2,241,951,533
Partitioning	8,560,124	8,094,475	8,930,840	1,040,396,940	1,696,187,797
Memory (GB) / Time (h) Filtering	4 GB / < 2	4 GB / < 2	4 GB / < 2	188 / 83	258 / 178
Memory (GB) / Time (h) Partitioning	4 GB / < 2	4 GB / < 2	4 GB / < 2	234 / 120	287 / 310

Dataset	Assembly Comparison	Asssembler	Assembly Content Similarity	Coverage of Reference Genomes	Unfiltered Total No. Contigs / Assembly Length (bp) / Max contig size (bp)	Total No. Co (bp) / M
HMP Mock	Unfiltered/ Filtered	Velvet	94%	42.6 / 43.7	29,063 / 37,776,354 / 146,795	30,082 /
HMP Mock	Unfiltered / Filtered Partitioned	Vevlet	94%	42.6 / 43.7	29,063 / 37,776,354 / 146,795	30,115 /
HMP Mock	Unfiltered / Filtered Partitioned	MetaIDBA	93%	45.6 / 44.6	24,300 / 36,470,912 / 86,445	27,475 /
HMP Mock	Unfiltered / Filtered Partitioned	SOAPdenovo	97%	45.4 / 45.6	36,689 / 36,568,753 / 32,736	29,295 /
Iowa Corn	-	Velvet	-	-	N/A	1,862,962 /
Iowa Corn	-	MetaIDBA	-	-	N/A	1,334,841 /
Iowa Corn	-	SOAPdenovo	-	-	N/A	1,542,436 /
Iowa Prairie	-	Velvet	-	-	N/A	3,120,263 /
Iowa Prairie	-	MetaIDBA	-	-	N/A	2,102,163
Iowa Prairie	-	SOAPdenovo	-	-	N/A	2,599,767 /

	Iowa Corn Velvet Assembly	Iowa Prairie Velvet Assemby
Total Unfiltered Reads	1,810,630,781	3,303,375,485
Total Unfiltered SE Reads	141,517,075	358,817,057
SE aligned 1 time	11,368,837	32,539,726
SE aligned > 1 time	562,637	1,437,284
% SE Aligned	8.43%	9.47%
Total Unfiltered PE Reads	834,556,853	1,472,279,214
PE aligned 1 time	54,731,320	110,353,902
PE aligned > 1 time	1,993,902	3,133,710
% PE Aligned Disconcordantly	0.47%	0.63%
% PE Aligned	9.68%	11.20%

References

[1] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin, Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M de Vos, Søren Brunak, Joel Doré, MetaHIT Consortium, María Antolín, François Artiguenave, Hervé M Blottiere, Mathieu Almeida, Christian Brechot, Carlos Cara, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Konrad U Foerstner, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Wolfgang Huber, Johan van Hylckama-Vlieg, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Jan Knol, Omar Lakhdari, Severine Layec, Karine Le Roux, Emmanuelle Maguin, Alexandre Mérieux, Raquel Melo Minardi, Christine M'rini, Jean Muller, Raish Oozeer, Julian Parkhill, Pierre Renault, Maria Rescigno, Nicolas Sanchez, Shinichi Sunagawa, Antonio Torrejon, Keith Turner, Gaetana Vandemeulebrouck, Encarna Varela, Yohanan Winogradsky, Georg Zeller, Jean Weissenbach, S Dusko Ehrlich, and Peer Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–80, May 2011.

[2] Matthias Hess, Alexander Sczyrba, Rob Egan, Tae-Wan Kim, Harshal Chokhawala, Gary Schroth, Shujun Luo, Douglas S Clark, Feng Chen, Tao Zhang, Roderick I Mackie, Len A Pennacchio, Susannah G Tringe, Axel Visel, Tanja Woyke, Zhong Wang, and Edward M Rubin. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–7, Jan 2011.

[3] Vaughn Iverson, Robert M Morris, Christian D Frazar, Chris T Berthiaume, Rhonda L Morales, and E Virginia Armbrust. Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science*, 335(6068):587–90, Feb 2012.

[4] Rachel Mackelprang, Mark P Waldrop, Kristen M DeAngelis, Maude M David, Krystle L Chavarria, Steven J Blazewicz, Edward M Rubin, and Janet K Jansson. Metagenomic

analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, 480(7377):368–71, Dec 2011.

[5] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, MetaHIT Consortium, Peer Bork, S Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, Mar 2010.

[6] Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, Peer Bork, Philip Hugenholtz, and Edward M Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–7, Apr 2005.

[7] J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, Derrick E Fouts, Samuel Levy, Anthony H Knap, Michael W Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O Smith. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74, Apr 2004.