Approaches for Enabling Deep, Large Scale Metagenome Assembly

ACH, JJ, ST, JMT, CTB

November 13, 2012

1 Introduction

Complex microbial communities operate at the heart of many crucial environmental, ecological, and biomedical processes, providing critical ecosystem functionality that underpins much 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 community 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 [8]. Both the read lengths and volume of sequencing data pose new challenges to traditional analysis approaches of sequencing data. Short read lengths and their associated sequencing errors and biases contain little biological signal and are noisy, limiting direct annotation approaches against known reference genomes. Further complicating analysis is that 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 for studying sequencing datasets. It both reduces the presence of sequencing errors and the total number of sequences for analysis by identifying consensus sequences from overlapping reads. These resulting assembled contigs are longer than sequencing reads and provide gene order. Importantly, *de novo* assembly does not rely on the existence of reference genomes, thus allowing for 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 evidenced by assembled metagenomes from the rumen, human gut, and permafrost soils

requiring 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, and although many new metagenome-specific assemblers have been developed to address characteristics of mixed population assembly, these are limited in capacity of sample diversity and volume.

Here, we present a novel set of approaches which enable large-scale metagenomic de novo assembly. This is enabled by reducing the dataset volume by normalizing sequencing coverage and removing sequencing errors and biases. Subsequently, reads are binned based on biological connectivity, resulting in partitions which can be assembled with significantly reduced computational requirements. We evaluate these approaches using the assembly of a human gut mock community dataset and find that our methods result in nearly identical assemblies to the unprocessed assembly. Consequently, we apply our approaches towards the assembly of two of the largest published soil metagenomes which have previously been impossible to assemble.

2 Results

2.1 Assembly of the HMP mock metagenome

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

The recovery of reference genomes from de novo metagenomic assembly was evaluated, comparing unfiltered traditional assembly to the the described filtered assembly (See Methods and Supp Info). Initially, the abundance of genomes within the mock dataset was estimated based on the reference genome coverage of sequencing reads in the unfiltered dataset. Coverage (excluding genomes with less than 3-fold coverage) ranged from 6-fold to 2,000-fold coverages (Supp. Table **??** and Supp Fig. **??** and **??**). Overall, the unfiltered dataset reads covered a total of 93% of the reference genomes. During filtering, a total of 5.9 million reads (40% of total reads) contributing to dataset redundancy and sequencing errors and biases (Supp. Info digital normalization and high abundance filtering) were removed. The remaining reads covered a total of 91% of the available reference genomes (Table 1 and Supp. Figures coverage1 and coverage2).

Additionally, 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 (Velvet assembler), 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 unfiltered and filtered assemblies were were similar, sharing 95% genomic content. For the highest abundance genomes (e.g., ref|NC_005008.1, ref|NC_005007.1, and ref|NC_005003.1), the unfiltered assembly recovered significantly more of the original genomes; however, for the large majority of genomes the filtered assembly recovered similar (and sometimes greater) amounts of the reference genomes Supp Fig X and X). The distribution of contig lengths in unfiltered and filtered assemblies were also comparable (Supp Figure X).

The abundance of assembled contigs and reference genomes could be recovered using the coverage of sequencing reads (Supp Fig **??**). In general, a minimum depth of sequencing was observed when contigs were assembled. Above a sequencing coverage of five, the majority of reads which could be mapped to reference genomes were likely to be included in an assembled contig (Supp Fig. **??**). Below this threshold, reads could be mapped to reference genomes but were less likely to be associated with assembled contigs. Using the reference genomes, assembly-based abundance estimations of references could be evaluated, and the estimations based on unfiltered and filtered assemblies compared. The abundance estimations from the filtered assembly were significantly closer to predicted abundances from reference genomes (p-value of 0.032, see Supp Info).

2.1.2 Evaluation of partitioning reads based on connectivity

To divide the the remaining dataset, the filtered dataset was partitioned based on connectivity within a de Bruijn graph representation (previously described in Pell et al. and Howe et al.), and each partition was assembled independently. The resulting assemblies of filtered, unpartitioned and filtered, partitioned datasets were compared and found to be greater than 99% identical. For the mock dataset, a total of 85,818 disconnected partitions (a total of 9 million reads) containing a minimum of five reads were identified (Fig. 1). Among these, only 2,359 partitions contained reads originating from more than one genome. In general, reference genomes with high sequencing coverage were associated with fewer partitions (Supp Table **??**), a total of 112 partitions contained reads from high abundance reference genomes (coverage above 25) compared to 2,771 partitions associated with low abundance genomes (coverage below 25).

To further evaluate the effects of partitioning, spiked reads from *E. coli* genomes were in-

troduced into the mock community dataset. A single spiked genome (*E. coli* strain E24377A, NC_009801.1 with 2% error)) was added to the mock community dataset and processed identically to the unfiltered mock dataset. Similar amounts of data reduction after digital normalization and partitioning (Table 1) were observed. 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 (201 partitions contained *only* spiked reads) and when assembled aligned with 99.5% of *E. coli* strain E24377A genome (4,957,067 of 4,979,619 bp) (Fig. 1). Next, the same analysis was performed on the mock dataset after introducing five closely-related *E. coli* strains into the mock community dataset. Partitioning this "mix-spiked" mock community dataset resulted in 81,425 partitions, of which 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 one other reference genome (either another spiked strain or from the mock community dataset) (Fig. 2). The partitions containing 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 (Supp Fig. **??**), 3,075 contigs (51%) could be aligned to reads which originated from more than one spiked genome. This result is comparable to the fraction of contigs which are associated with multiple genomes when the unfiltered dataset is assembled, where in 4,702 contigs associated with "spiked reads", 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 to assemble with normally available memory, e.g., 500 GB. A 75 million reads subset of the Iowa corn dataset alone required 110 GB of memory (Supp Fig.

??). Applying the same filtering approaches as described above, 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. The large majority of k-mers in the soil metagenomes are relatively low-coverage (Fig. 4), 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 six, the large majority of sequences could be assembled (Supp. Fig. **??**). 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 of the corn and prairie soil metagenomes resulted in a total of 1.9 million and 3.1 million contigs (Velvet), 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 where 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 4). Based on these mappings, the read coverage of assembled contigs within the soil metagenomes were estimated (Fig. 5). 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.

Among contigs, the presence of polymorphisms was examined by identifying the amount of consensus obtained by reads mapped (Supp. Info methods). For both the Iowa corn and

prairie metagenomes, nearly all assembled sequences (greater than 99.9%) contained base calls which were supported by 95% consensus from mapped reads over 90% over its length (Supp. Fig Polymorphisms corn and prairie).

2.2.2 Content of soil metagenome assembly

Assembled contigs with their respective bp-coverage (with a median bp coverage greater than 1) were annotated through the MG-RAST pipeline resulting in 2,089,779 and 3,460,496 predicted protein coding regions in the corn and prairie metagenomes, respectively. The large majority of protein coding regions did not share similarity with any gene in the M5NR database, 61.8% in corn and 70.0% in prairie. In total, 613,213 and 777,454 protein coding regions were assigned to functional categories. The functional profile of these annotated features against SEED subsystems were compared (Fig. 7). For both the corn and prairie metagenomes, the most abundant functions were associated with the carbohydrate (e.g., central carbohydrate metabolism and sugar utilization), amino acids (e.g., biosynthesis and degradation), and protein metabolism (e.g., biosynthesis, processing, and modification) subsystems. The subsystem profile of both metagenomes were very similar while the taxonomic profile of the metagenomes based on the originating taxonomy (phyla) were different (Fig. 6, Supp Methods). Within both metagenomes, Proteobacteria were the most abundant taxa. However, in the Iowa corn, Actinobacteria followed by Bacterioidetes and Firmucutes were the next most abundant while in the Iowa prairie, Acidobacteria were the second most abundant phyla, followed by Bacterioidetes and Actinobacteria. The Iowa prairie also had nearly double the fraction of Verruomicrobia compared to the Iowa corn.

3 Discussion

3.1 Filtering approaches effectively reduce datasets

The diversity and sequencing depth represented by the mock community is extremely low compared to that of most environmental metagenomes; however, it represents a simplified, unevenly sampled model for a metagenomic dataset which enables the evaluation of analyses through the availability of source genomes. For this dataset, the filtering approaches described above were effective at reducing the dataset size without significant loss of assembly. This strategic filtering takes advantage of the observed coverage "sweet spot" at which point sufficient sequences are

present for assembly and beyond which further sequencing is not productive (and increases the number of sources of errors) (Supp Fig X). The normalization of sequences also results in a more even distribution of coverage (Fig. 4), minimizing assembly problems caused by highly variable coverage. Additional reduction of the dataset was achieved by the removal of high abundance sequences, previously correlated as Illumina sequencing artifacts (Howe et al).

The specific effects of filtering varied depending on differences of reference genomes. Sequencing coverage and conserved regions among references had an impact on filtered assembly recovery. The filtered assemblies of the three plasmids of the *Staphylococcus* epidermidis genome (NC_005008.1, NC_005007.1, and NC_005003.1) were highly abundant (Supp Tab X) and shared several conserved regions (90% identity over 290 bp). During normalization, repetitive elements in these genomes would appear as high coverage elements and be removed, evidenced by a large difference in the number of reads associated with NC_005008.1 in the unfiltered and filtered datasets (Supp Fig X). Consequently, the unfiltered dataset contained more reads spanning these repetitive regions. This most likely enabled assembler heuristics to extend the assembly of these sequences and resulted in the observed increased recovery of these genomes in the unfiltered assemblies. This result, though rare among the mock reference genomes, identifies a shortcoming of our approach, and indeed for most short-read assembly approaches, related to repetitive regions and/or polymorphisms. For the soil metagenomes, our data reduction may cause some information loss which may have been useful for assembly, but the benefit of being able to assemble previously intractable datasets is obvious. 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, if not improved.

3.2 Partitioning effectively separates genomes for assembly

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, sequencing biases which cause artificial connectivity within metagenomic assembly graphs were removed (high abundance sequence filtering) (Howe). As discussed above, tthe removal of these sequences (combined with normalization) did not significantly alter the recovery of reference genomes through de novo assembly and importantly, it enabled the division of the mock community dataset into thousands of disconnected partitions. The resulting assemblies of unpartitioned and partitioned datasets were nearly identical. The large majority of these partitions contained reads from a single reference genome, supporting our previous hypothesis that most connected subgraphs contain distinct genomes. As expected, high coverage, well-sampled genomes were found to contain fewer partitions (highly connected assembly graph), and low coverage, under-sampled genomes contained more partitions (fragmented assembly graph).

To further examine the recovery of sequences through partitioning, one or more *E. coli* strains were computationally spiked into the mock community dataset. For a spike of a single *E. coli* strain, the fraction of partitions containing *E. Coli* associated reads could be reassembled to recover 99% of the original genome (Fig. 1). When five 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 (Supp Fig. **??**). This result is not unexpected nor 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 mock community dataset, this gain was small, reducing unfiltered assembly at 12 GB and 4 hours to less than 2 GB and 1 hour. However, for the soil metagenomes, previously impossible assemblies could be completed 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 Soil assembly

The final assemblies 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 (equivalent to ? genomes). Without references, these assemblies were evaluated to be acceptable based on paired end concordancy and polymorphism presence in reads mapped. 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, indicative of the low sequencing overage of these metagenomes.

As this study represents the largest published soil metagenomic sequencing effort to date, these assembly results demonstrate the enormous amount of diversity within the soil. Even with this level of sequencing, hundred of thousands of functions were identified for each metagenome. More than half of the assembled contigs are not similar to anything in known databases suggesting that de novo assembly of novel sequences holds great benefit for soil metagenomics. Interestingly, among the protein coding sequences which were annotated, comparisons of the two soil datasets suggests that the functional profiles are more similar to one another than the complementing phylogenetic profiles. This result supports previous hypotheses that despite large diversity with two different soil systems, the microbial community provides similar function (cite Kostas?, etc.).

4 Conclusion

CONCLUSION: THEME: PARTITIONING IS NOT ONLY AWESOME FOR ASSEMBLY BUT... THEME: LOTS MORE TO FIND OUT IN SOIL

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 parameter changes are described in Supp Info.

References

[1] Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–80 (2011).

[2] Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–7 (2011).

[3] Iverson, V. *et al.* Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science* **335**, 587–90 (2012).

[4] Mackelprang, R. *et al.* Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480**, 368–71 (2011).

[5] Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

[6] Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–7 (2005).

[7] Venter, J. C. *et al.* Environmental genome shotgun sequencing of the sargasso sea. *Science* **304**, 66–74 (2004).

[8] Gans, J., Wolinsky, M. & Dunbar, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**, 1387–90 (2005).

5 Tables

Table 1: The total number of reads in unfiltered, filtered (normalized and high abundance (HA) k-mer removal), and partitioned datasets and the computational resources required (memory and time).

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

Table 2: Assembly summary statistics (total contigs, total million bp assembly length, maximum contig size bp) of unfiltered (UF) and filtered (F) or filtered/partitioned (FP) datasets with Velvet (V) assembler. Assembly for UF and FP datasets also shown for MetaIDBA (M) and SOAPdenovo(S) assemblers. Iowa corn and prairie metagenomes could not be completed on unfiltered datasets.

	UF	F	FP	Assembler
HMP Mock	29,063 / 38 / 146,795	30,082 / 35 / 90,497	30,115 / 35 / 90,497	V
HMP Mock	24,300 / 36 / 86,445	-	27,475 / 36 / 96,041	M
HMP Mock	36,689 / 37 / 32,736	-	29,295 / 37 / 58,598	S
Iowa corn	N/A	N/A	1,862,962 / 912/ 20,234	V
Iowa corn	N/A	N/A	1,334,841 / 623 / 15,013	M
Iowa corn	N/A	N/A	1,542,436 / 675 / 15,075	S
Iowa prairie	N/A	N/A	3,120,263 / 1,510 / 9,397	V
Iowa prairie	N/A	N/A	2,102,163 / 998 / 7,206	M
Iowa prairie	N/A	N/A	2,599,767 / 1,145 / 5,423	S

Table 3: Assembly comparisons of unfiltered (UF) and filtered (F) or filtered/partitioned (FP) HMP mock datasets using different assemblers (Velvet (V), MetaIDBA (M) and SOAPdenovo (S)). Assembly content similarity is based on the fraction of alignment of assemblies and similarly, the coverage of reference genomes is based on the alignment of assembled contigs to reference genomes (RG).

Assembly Comparison	Percent Similarity	RG Coverage	Assembler
UF vs. F	95%	43.3% / 44.5%	V
UF vs. FP	95%	43.3% / 44.4%	V
UF vs. FP	93%	46.5% / 45.4%	M
**UF vs. FP	98%	46.2% / 46.4%	S

6 Figures

Table 4: Fraction of single-end (SE) and paired-end (PE) reads mapped to Iowa corn and prairie Velvet assemblies.

	Iowa Corn Assembly	Iowa Prairie 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%

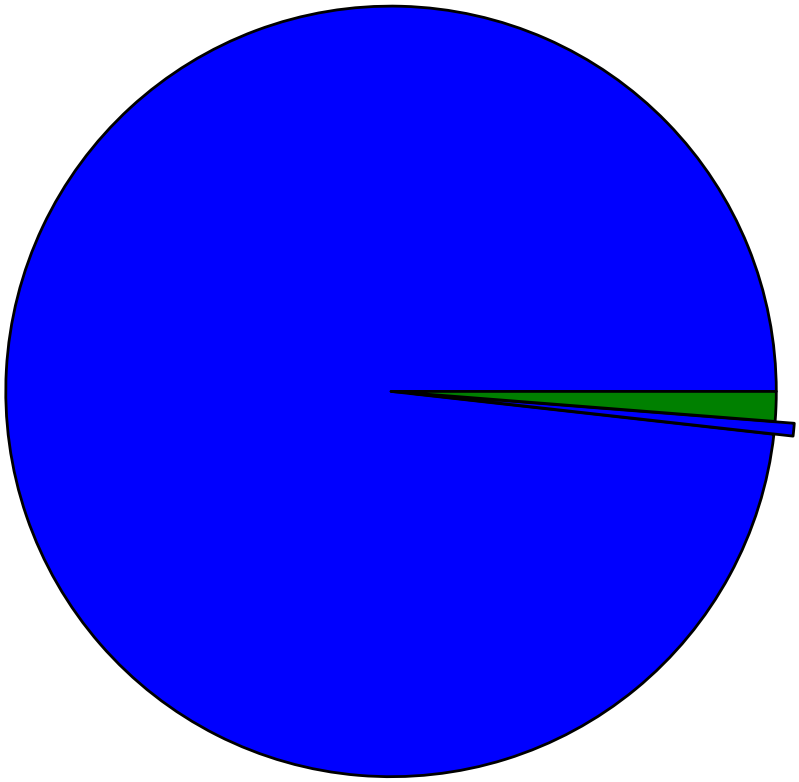

Figure 1: The fraction of partitions in spiked HMP dataset (single *E. Coli*) which contain single genomes (blue) and multiple genomes (green). The exploded pie chart section indicates partitions which contain spiked *E. coli* reads which were subsequently assembled independently.

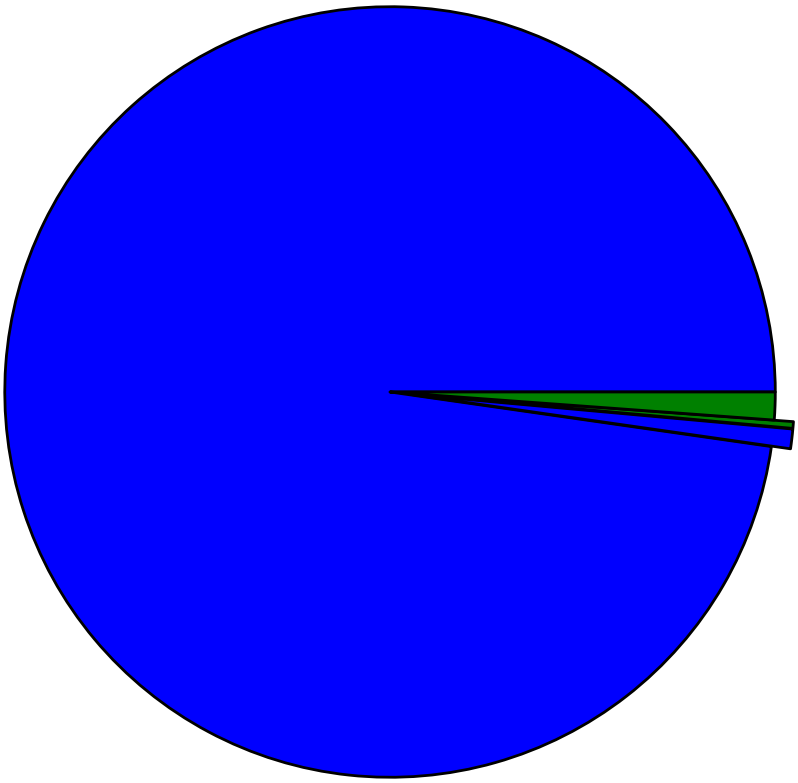

Figure 2: The fraction of partitions in spiked HMP dataset (five *E. Colis*) which contain single genomes (blue) and multiple genomes (green). The exploded pie chart section indicates partitions which contain spiked *E. coli* reads which were subsequently assembled independently.

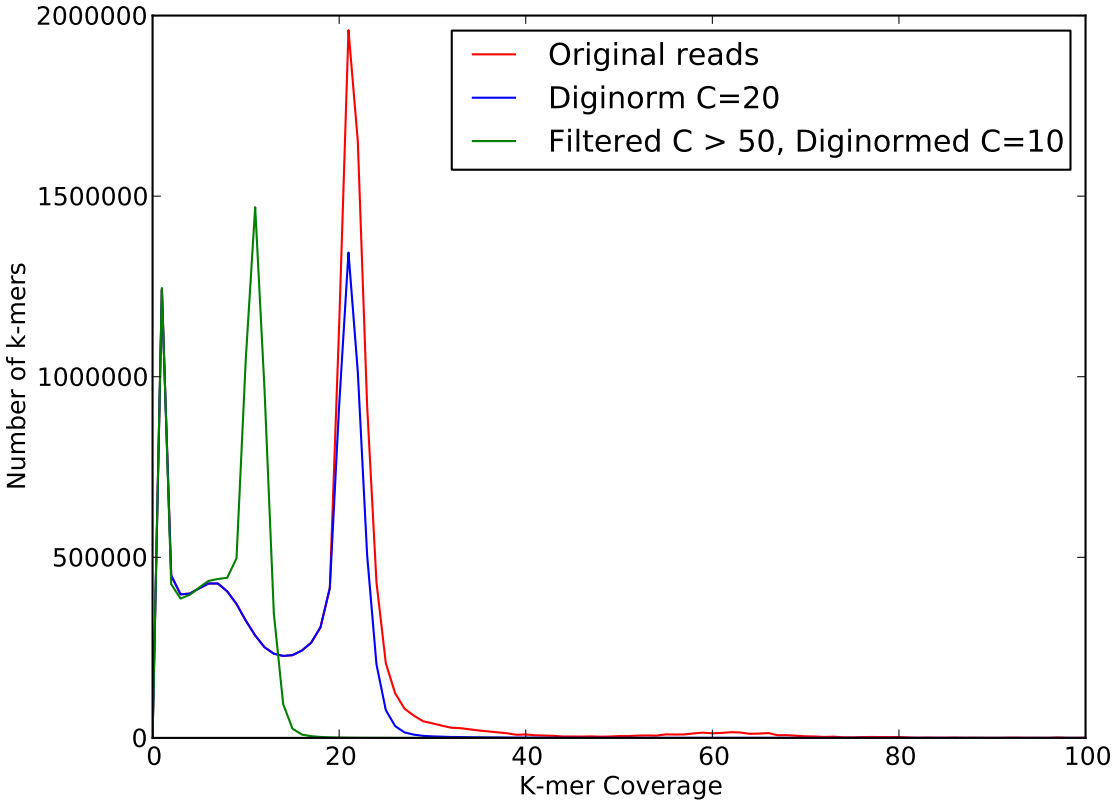

Figure 3: K-mer coverage of HMP mock community dataset before and after filtering approaches.

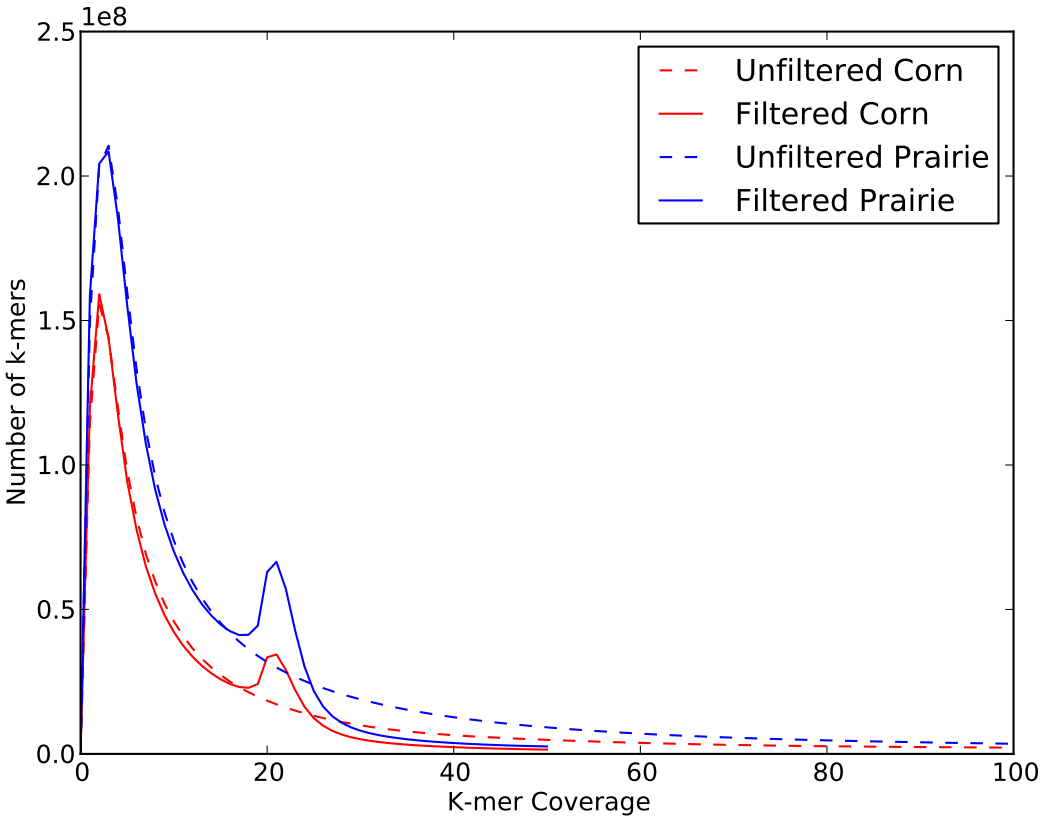

Figure 4: K-mer coverage of Iowa corn and prairie metagenomes before and after filtering approaches.

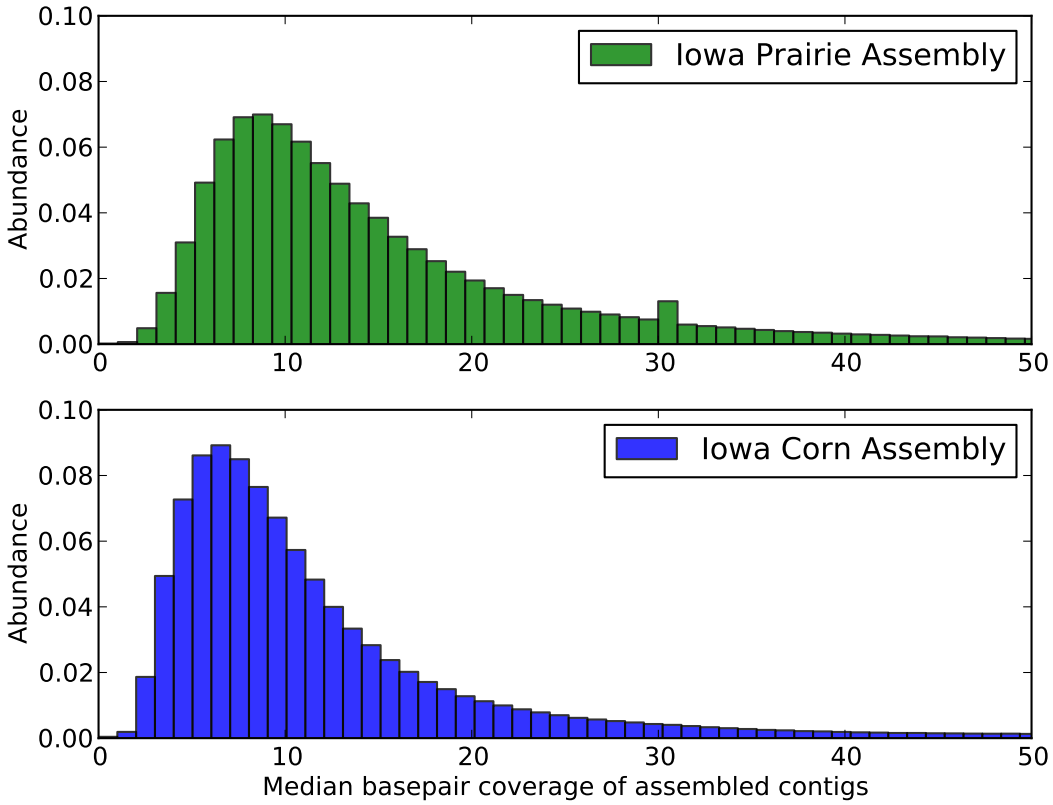

Figure 5: Coverage (median basepair) distribution of assembled contigs from soil metagenomes.

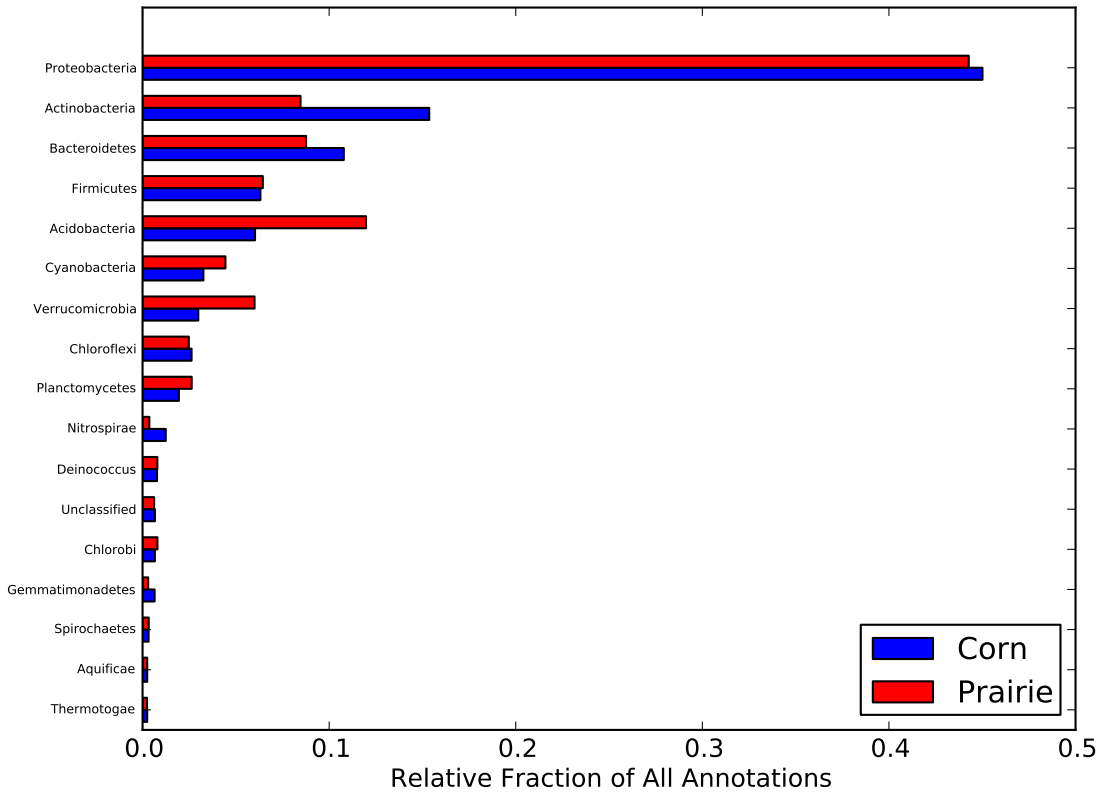

Figure 6: Phylogenetic distribution from SEED subsystem annotations for Iowa corn and prairie metagenomes.

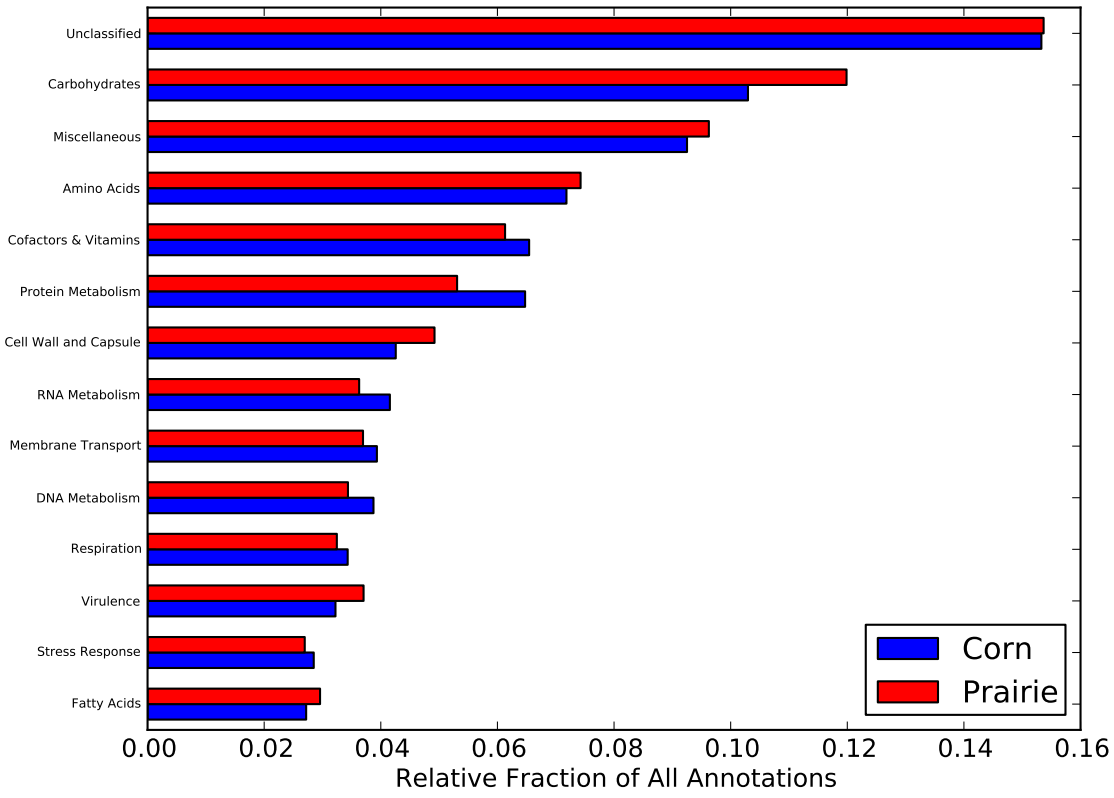

Figure 7: Functional distribution from SEED subsystem annotations for Iowa corn and prairie metagenomes.