Title

Assembling large, complex environmental metagenomes

Authors

Howe, Adina Chuang
Jansson, Janet
Malfatti, Stephanie A
et al.

Publication Date

2012-12-12

Peer reviewed

Assembling large, complex environmental metagenomes

Adina Chuang Howe[1,2], Janet Jansson[3,4], Stephanie A. Malfatti[3], Susannah G. Tringe[3], James M. Tiedje[1,2], and C. Titus Brown[1,5*]

[1]*Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA*

[2]*Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI, USA*

[3]*Department of Energy (DOE) Joint Genome Institute, Walnut Creek, CA, USA*

[4]*Lawrence Berkeley National Laboratory, Earth Sciences Division, Berkeley, CA, USA*

[5]*Computer Science and Engineering, Michigan State University, East Lansing, MI, USA*

The large volumes of sequencing data required to sample complex environments deeply pose new challenges to sequence analysis approaches. *De novo* metagenomic assembly effectively reduces the total amount of data to be analyzed but requires significant computational resources. We apply two pre-assembly filtering approaches, digital normalization and partitioning, to make large metagenome assemblies more computationaly tractable. Using a human gut mock community dataset, we demonstrate that these methods result in assemblies nearly identical to assemblies from unprocessed data. We then assemble two large soil

metagenomes from matched Iowa corn and native prairie soils. The predicted functional content and phylogenetic origin of the assembled contigs indicate significant taxonomic differences despite similar function. The assembly strategies presented are generic and can be extended to any metagenome; full source code is freely available under a BSD license.

Complex microbial communities operate at the heart of many crucial terrestrial, aquatic, and host-associated processes, providing critical ecosystem functionality that underpins much of biology [1–7]. These systems are difficult to study *in situ*, and we are only beginning to understand their diversity and functional potential. Advances in DNA sequencing now provide unprecedented access to the genomic content of these communities via shotgun sequencing, which produces millions to billions of short-read sequences [2,4,5]. Because shotgun sequencing samples communities randomly, ultradeep sequencing is needed to detect rare species in environmental samples, with an estimated 50 Tbp needed for an individual gram of soil [8]. Both short read lengths and the large volume of sequencing data pose new challenges to sequence analysis approaches. A single metagenomic project can readily generate as much or more data than is in global reference databases; for example, a human-gut metagenome sample containing 578 Gbp [5] produced more than double the data in NCBI RefSeq (Release 56). Moreover, short reads contain only minimal signal for homology searches and are error-prone, limiting direct annotation approaches against reference databases. And finally, the majority of genes sequenced from complex metagenomes typically contain little or no similarity to experimentally studied genes, further complicating homology analysis [1,5].

De novo assembly of raw sequence data offers several advantages over analyzing the sequences directly. Assembly removes most random sequencing errors and decreases the total amount of data to be analyzed. 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 genomic elements. The main challenge for metagenomic applications of *de novo* assembly is that current assembly tools do not scale to the high diversity and large volume of metagenomic data: metagenomes from rumen, human gut, and permafrost soil sequencing could only be assembled by discarding low abundance sequences prior to assembly [2,4,5]. Traditional assemblers are designed for single genomes whose abundance distribution and diversity content are typically simpler than community metagenomes. Although many metagenome-specific assemblers have recently been developed for community assembly, most of these assemblers do not scale to extremely large samples [9].

Here, we combine two approaches, digital normalization and partitioning, to tackle the problem of *de novo* metagenome assembly. Digital normalization normalizes sequence coverage and reduces the dataset size by discarding reads from high-coverage regions [10]. Subsequently, partitioning separates reads based on transitive connectivity, resulting in easily assembled subsets of reads [11,12]. We evaluate these approaches by applying them to a human gut mock community dataset, and find that these filtering methods result in assemblies nearly identical to assemblies from the unprocessed dataset. Next, we apply these approaches to the assembly of metagenomes from two matched soils, 100-year cultivated Iowa agricultural soil and native Iowa prairie. We compare the predicted functional capacities and phylogenetic origins of the assembled contigs and

conclude that despite significant phylogenetic differences, the functions encoded in both soil data sets are similar. We also show that virtually no strain-level heterogeneity is detectable within the assembled reads.

Results

Data reduction results in similar assemblies

We evaluated the recovery of reference genomes from *de novo* metagenomic assembly by comparing unfiltered traditional assembly to the the described filtered assembly (Fig. 6; see Methods and Supplementary Information). Initially, the abundance of genomes within the mock dataset was estimated based on the reference genome coverage of sequence reads in the unfiltered dataset. Coverage (excluding genomes with less than 3-fold coverage) ranged from 6-fold to 2,000-fold (Supplementary Table 1 and Supplementary Fig. 2 and 3). Overall, the unfiltered dataset reads covered a total of 93% of the reference genomes. Filtering removed 5.9 million reads, 40% of the total (Table 1); the remaining reads covered 91% of the reference genomes (Table 1 and Supplementary Fig. 2 and 3).

We next compared the recovery of reference genomes in contigs assembled from the original and filtered datasets. Using the Velvet assembler, we recovered 43% and 44% of the reference genomes, respectively. The assembly of the original dataset contained 29,063 contigs and 38 million bp, while the filtered assembly contained 30,082 contigs and 35 million bp (Table 2). Comparable recoveries of references between original and filtered datasets were also obtained with other

assemblers (SOAPdenovo and Meta-IDBA, Table 2). Overall, the unfiltered and filtered assemblies were very similar, sharing 95% of genomic content. For the highest abundance references (the plasmids NC_005008.1, NC_005007.1, and NC_005003.1), the unfiltered assembly recovered significantly more of the original sequence; however, for the large majority of genomes, the filtered assembly recovered similar (and sometimes greater) amounts of the reference genomes (Supplementary Fig. 2 and 3). The distribution of contig lengths in unfiltered and filtered assemblies were also comparable (Supplementary Fig. 4).

We estimated the abundance of assembled contigs and reference genomes using the mapped sequencing reads (Supplementary Fig. 5). Above a sequencing coverage of five, the majority of reads which could be mapped to reference genomes were included in the assembled contigs (Supplementary Fig. 3 and 4). Below this threshold, reads could be mapped to reference genomes but were less likely to be associated with assembled contigs. We next compared the abundances of the reference genomes to the abundances of the contigs in the unfiltered and filtered assemblies. The abundance estimations from the filtered assembly were significantly closer to predicted abundances from reference genomes (n = 28,652; p-value = 0.032, see Supplementary Information).

Partitioning separates most reads by species

We next partitioned the filtered data set based on de Bruijn graph connectivity and assembled each partition independently [11,12]. The resulting assemblies of unpartitioned and partitioned were more than 99% identical. In the mock dataset, we identified 9 million reads in 85,818 disconnected

partitions (Supplementary Fig. 6). Among these, only 2,359 (2.7%) of the partitions contained reads originating from more than one genome, indicating that partitioning separated reads from distinct species.

In general, reference genomes with high sequencing coverage were associated with fewer partitions (Supplementary Table 1): a total of 112 partitions contained reads from high abundance reference genomes (coverage above 25) compared to 2,771 partitions associated with lower abundance genomes. This is consistent with the observation that the main effect of low coverage is to "break" connectivity in the assembly graph [13,14].

To further evaluate the effects of partitioning, we introduced spiked reads from *E. coli* genomes into the mock community dataset. First, simulated reads from a single genome (*E. coli* strain E24377A, NC_009801.1 with 2% substitution error and 10x coverage) were added to the mock community dataset and then processed in the same way as the unfiltered mock dataset. We observed similar amounts of data reduction after digital normalization and partitioning (Table 1). Among the 81,154 partitioned sets of reads, we identified only 2,580 (3.2%) 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 to 99.5% of *E. coli* strain E24377A genome (4,957,067 of 4,979,619 bp) (Supplementary Fig. 6).

Next, we introduced five closely-related *E. coli* strains into the mock community dataset and performed the same analysis. Partitioning this "mix-spiked" mock community dataset resulted in 81,425 partitions, of which 1,154 (1.4%) 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 reads from more 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) (Supplementary Fig. 7). We independently assembled the partitions containing reads originating from the spiked *E. coli* strains. Among the resulting 6,076 contigs, all but three contigs originated from a spiked *E. coli* genome. The remaining three contigs were more 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 contained similarities to reads originating from multiple genomes (Supp Fig. 8), and 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 in the unfiltered data set, where 66% of 4,702 contigs associated with spiked reads contain reads that originate from more than one spiked genome.

Data reduction and partitioning enable the assembly of two soil metagenomes

We next applied these approaches to the *de novo* assembly of two soil metagenomes. Iowa corn and prairie datasets (containing 1.8 billion and 3.3 billion reads, respectively) could not be assembled by Velvet in 500 GB of RAM. A 75 million reads subset of the Iowa corn dataset alone required 110 GB of memory, suggesting that assembly of the 3.3 billion read data set might need as much as 4 TB of RAM (Supplementary Fig. 9). Applying the same filtering approaches as described above,

the Iowa corn and prairie datasets were reduced to 1.4 billion and 2.2 billion reads, respectively, and after partitioning, a total of 1.0 billion and 1.7 billion reads remained, respectively. Prefiltering used 300 GB of RAM or less. The large majority of k-mers in the soil metagenomes are relatively low-abundance (Fig. 2), and consequently digital normalization did not remove as many reads in the soil metagenomes as in the mock data set.

Based on the mock community dataset, we estimated that above a sequencing depth of five, the large majority of sequences could be assembled into contigs larger than 300 bp (Supplementary Fig. 1). Given the greater diversity expected in the soil metagenomes, we normalized these datasets to a sequencing depth of 20 (i.e., discarding redundant reads within dataset above this coverage). After partitioning the filtered datasets, we identified a total 31,537,798 and 55,993,006 partitions (containing more than five reads) in the corn and prairie datasets, respectively. For assembly, we grouped partitions together into files containing 10 million reads. Data reduction and partitioning were completed in less than 300 GB of RAM; once partitioned, each group of reads could be assembled in less than 14 GB and 4 hours. This readily enabled the usage of multiple assemblers and 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 greater than 300 bp, 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. 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, 95.5% of the reads aligned concordantly. In the Iowa prairie assembly, 10% of the single reads and 11% of the paired end reads aligned to the assembled contigs, and 95.4% of the paired ends aligned concordantly (Table 4). Based on these mappings, we calculated read recovery in assembled contigs within the soil metagenomes (Fig. 3). Overall, there is a positively skewed distribution of read overage of all contigs from both soil metagenomes, biased towards a coverage of less than ten-fold, and 48% and 31% of total contigs in Iowa corn and prairie assemblies respectively had 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 (Supplementary Information Methods). For both the Iowa corn and prairie metagenomes, more than 99.9% of contigs contained base calls which were supported by a 95% consensus from mapped reads over 90% of their lengths, demonstrating an unexpectedly low polymorphism rate (Supplementary Fig. 10).

Annotation of the soil assemblies revealed similar functional profiles but different taxonomy

We annotated assembled contigs through the MG-RAST pipeline, which was modified to account for per-contig abundance. This annotation resulted in 2,089,779 and 3,460,496 predicted protein coding regions in the corn and prairie metagenomes, respectively. The large majority of these 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 (29.3%) and 777,454 (22.5%) protein coding regions were assigned to functional categories. The functional profile of these annotated features against SEED subsystems

were compared (Fig. 5). For both the corn and prairie metagenomes, the most abundant functions in the assembly were associated with the carbohydrate (e.g., central carbohydrate metabolism and sugar utilization), amino acid (e.g., biosynthesis and degradation), and protein (e.g., biosynthesis, processing, and modification) metabolism subsystems. The subsystem profile of both metagenomes were very similar while the taxonomic profile of the metagenomes based on the originating taxonomy (phyla) was different (Fig. 4, Supp Methods). Within both metagenomes, Proteobacteria were most abundant. In Iowa corn, Actinobacteria, Bacteroidetes, and Firmicutes were the next most abundant, while in the Iowa prairie, Acidobacteria, Bacteroidetes, and Actinobacteria were the next most abundant. The Iowa prairie also had nearly double the fraction of Verrucomicrobia than did Iowa corn.

Discussion

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 robust assembly (Supp Fig. 1). The normalization of sequences also resulted in more even coverage (Fig. 2), minimizing assembly problems caused by variable coverage. Additional reduction of the dataset was achieved by the removal of high abundance sequences [11].

The specific effects of filtering varied depending on differences between reference genomes. Variable abundance and conserved regions in 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 (Supplementary Table 1) and shared several conserved regions (90% identity over more than 290 bp). During normalization, repetitive elements in these genomes would appear as high coverage elements and be removed, as evidenced by a large difference in the number of reads associated with NC_005008.1 in the unfiltered and filtered datasets (supplementary Fig. 2). Consequently, the unfiltered dataset contained more reads spanning these repetitive regions. This most likely enabled assemblers 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 have caused some information loss in exchange for the ability to assemble previously intractable data sets. Evaluation of 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 slightly improved.

Metagenomes contain many distinct genomes, which are largely disconnected from each other but which sometimes share sequences due to conservation or lateral transfer. Our prefiltering approach removes both common multi-genome elements as well as artificial connectivity stemming from the sequencing process. As shown above on the mock data set, the removal of these

sequences does not significantly alter the recovery of reference genomes through *de novo* assembly: the resulting assemblies of unpartitioned and partitioned datasets were nearly identical for the mock data. The large majority of these partitions contained reads from a single reference genome, supporting our previous hypothesis that most connected subgraphs contain reads from distinct genomes [12]. As expected, high abundance, well-sampled genomes were found to contain fewer partitions and low abundance, under-sampled genomes contained more partitions, due to fragmentation of the assembly graph.

We further examined the recovery of sequences through partitioning by computationally spiking in one or more *E. coli* strains before applying filtering and partitioning. When we spiked in a single *E. coli* strain, we could reassemble 99% of the original genome (Supplementary Fig. 6). When we spiked in five closely related strains, we could recover the large majority of the genomic content of these strains, albeit largely in chimeric contigs (Supplementary Fig. 8). This result is not unexpected, 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 challenge assemblers, and our approaches are not specifically designed to target such regions. However, the partitions resulting from our approach could provide a much-reduced subset of sequences to be targeted for more sensitive assembly approaches for highly variable regions (i.e. overlap-layout-consensus approaches or abundance binning approaches [15]).

One valuable result of partitioning is that it subdivides our datasets into sets of reads which

can be assembled with minimal 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 MetaIDBA).

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 $\approx$ 500 *E. coli* genomes worth of DNA. We evaluated these assemblies based on paired-end concordance, which showed that the majority of the assembled contigs agreed with the raw sequencing data. Overall, there is a positively skewed distribution of abundance of all contigs from both soil metagenomes, biased towards an abundance of less than ten, indicative of the low sequencing coverage of these metagenomes.

This study represents the largest published soil metagenomic sequencing effort to date, and these assembly results demonstrate the enormous amount of diversity within the soil. Even with this level of sequencing, millions of putative genes were defined for each metagenome, with hundreds of thousands of functions. More than half of the assembled contigs are not similar to anything in known databases, suggesting that soil holds considerable unexplored taxonomic and functional novelty. 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 overall function [16–19].

We present two strategies that readily enable the assembly of very large environmental metagenomes by discarding redundancy and subdividing the data prior to assembly. These strategies are generic and should be applicable to any metagenome. We demonstrate their effectiveness by first evaluating them on the assembly of a mock community metagenome, and then applying them to two previously intractable soil metagenomes.

Partitioning is an especially valuable approach because it enables the extraction of read subsets that should assemble together. These read partitions are small enough that a variety of assembly, abundance analysis, and polymorphism analysis techniques can be easily applied to them individually.

The two soil assemblies also provide a deeper glimpse of the opportunities and challenges of large scale environmental metagenomics in high-diversity systems such as soil: we identified millions of novel putative proteins, most of unknown function.

Acknowledgements This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2010-65205-20361 from the United States Department of Agriculture, National Institute of Food and Agriculture and National Science Foundation IOS-0923812, both to C.T.B. A.H. was supported by NSF Postdoctoral Fellowship Award #0905961 and the Great Lakes Bioenergy Research Center (Department of Energy BER DE-FC02-07ER64494). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-

AC02-05CH11231. We acknowledge the support of Krystle Chavarria and Regina Lamendella for extraction of DNA from Great Prairie soil samples and the technical support of Eddy Rubin and Tijana Glavina del Rio at the DOE JGI and John Johnson and Eric McDonald at MSU HPC.

Author Contributions A.H. and C.T.B. designed experiments and wrote paper. A.H. performed experiments and analyzed the data. J.J., S.T., and J.T. discussed results and commented on manuscript. S.M. managed raw sequencing datasets.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence should be addressed to C. Titus Brown (ctb@msu.edu).

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	Filtered	Partitioned
	Reads (Mbp)	Reads (Mbp)	Reads (Mbp)
HMP Mock	14,494,884 (1,136)	8,656,520 (636)	8,560,124 (631)
HMP Mock Spike	14,992,845 (1,137)	8,189,928 (612)	8,094,475 (607)
HMP Mock Multispike	17,010,607 (1,339)	9,037,142 (702)	8,930,840 (697)
Iowa Corn	1,810,630,781 (140,750)	1,406,361,241 (91,043)	1,040,396,940 (77,603)
Iowa Prairie	3,303,375,485 (256,610)	2,241,951,533 (144,962)	1,696,187,797 (125,105)

	Unfiltered (GB / h)	Filtered and Partitioned (GB / h)
HMP Mock	4 / <2	4 / <2
HMP Mock Spike	4 / <2	4 / <2
HMP Mock Multispike	4 / <2	4 / <2
Iowa Corn	188 / 83	234 / 120
Iowa Prairie	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

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%

Figures

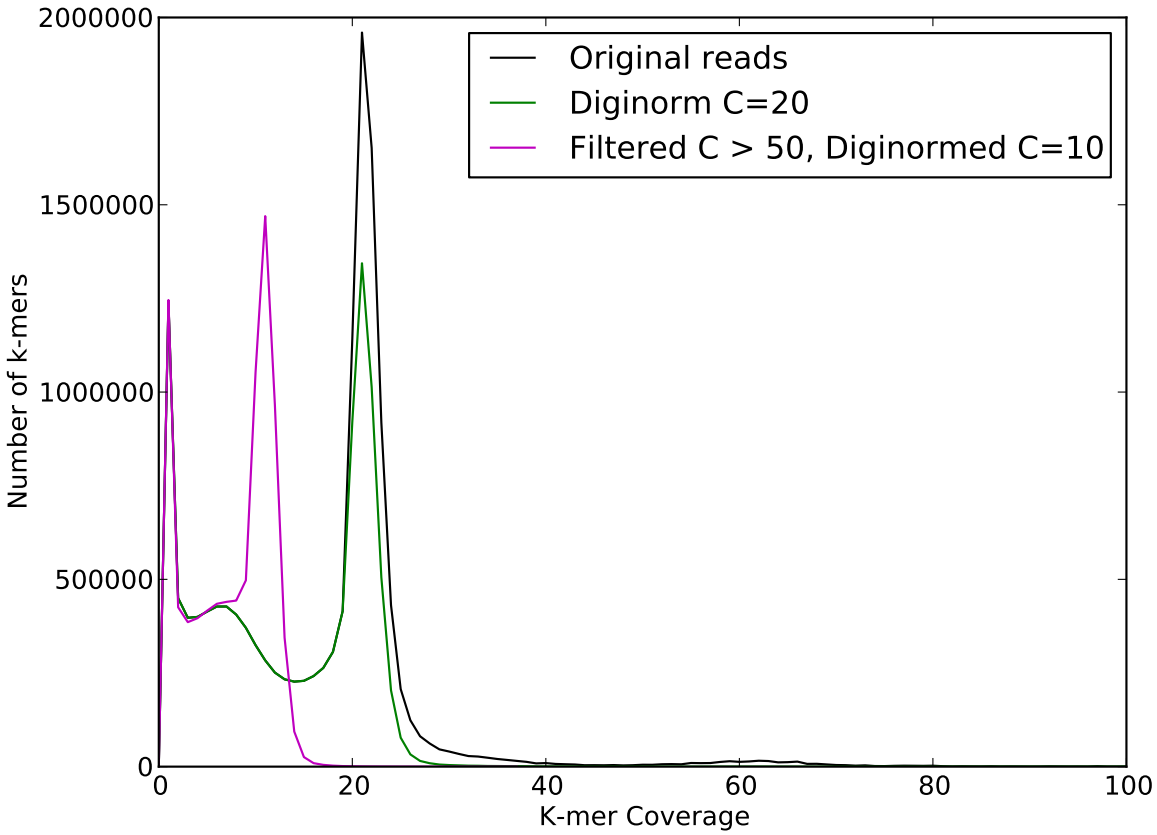


Figure 1: K-mer coverage of HMP mock community dataset before and after filtering approaches.

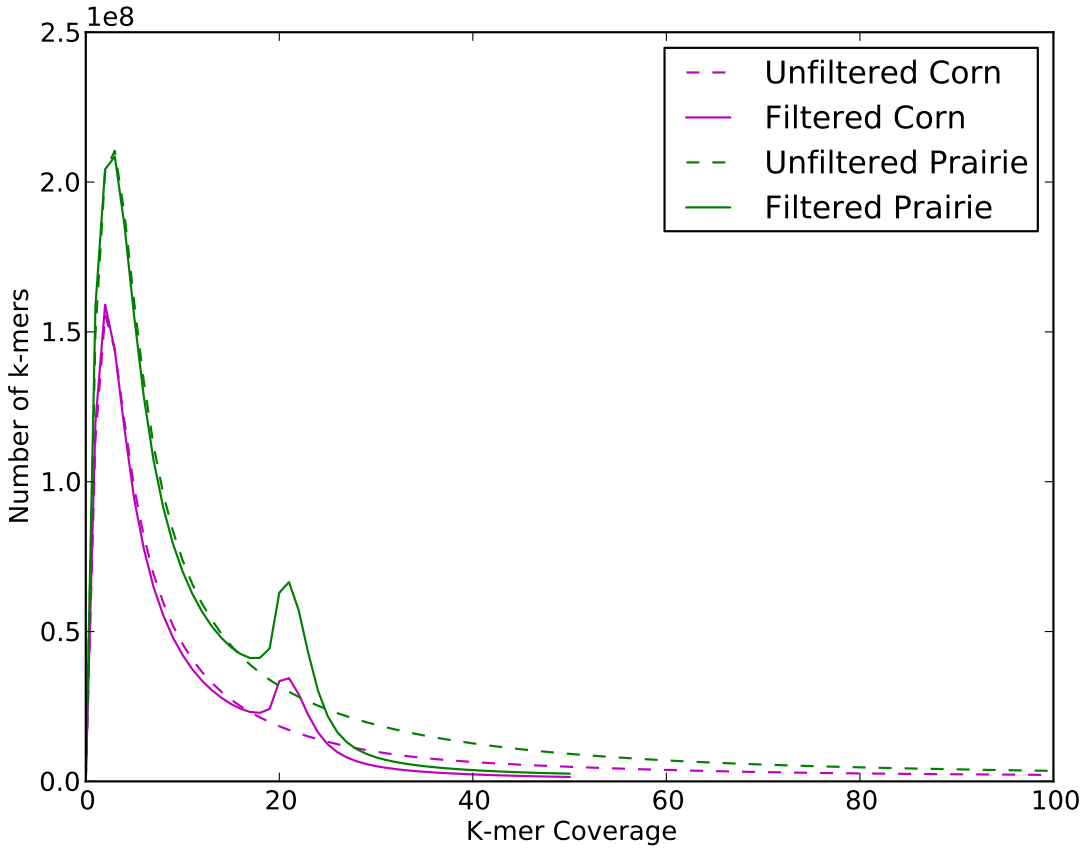


Figure 2: K-mer coverage of Iowa corn and prairie metagenomes before and after filtering approaches.

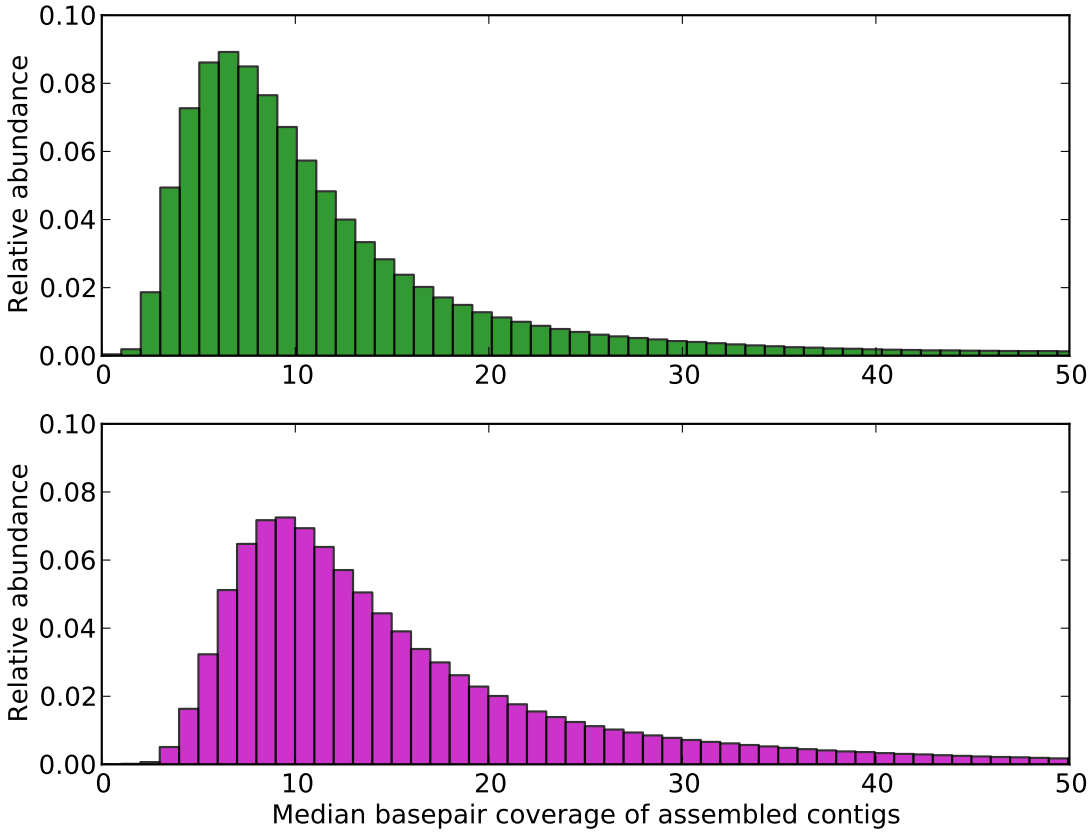


Figure 3: Coverage (median basepair recovered) distribution of assembled contigs from Iowa corn soil (top) and Iowa prairie soil (bottom) metagenomes.

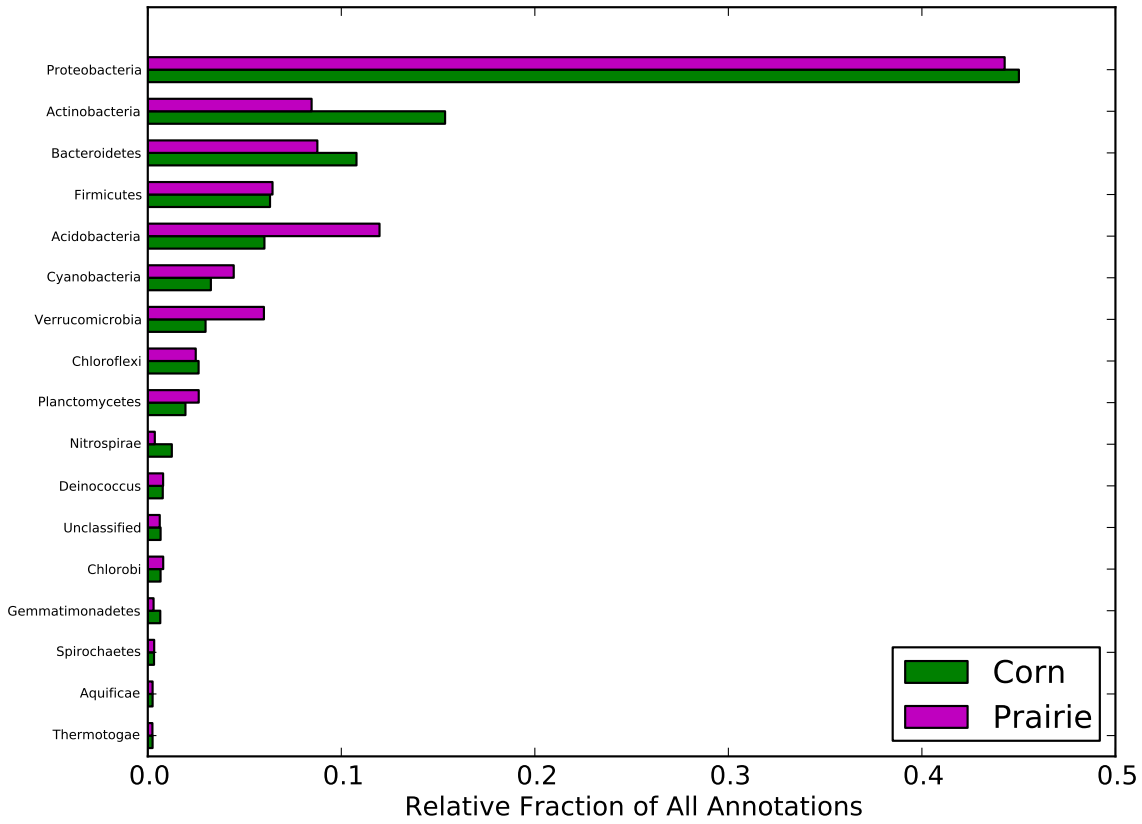


Figure 4: Phylogenetic distribution from SEED subsystem annotations for Iowa corn and prairie metagenomes.

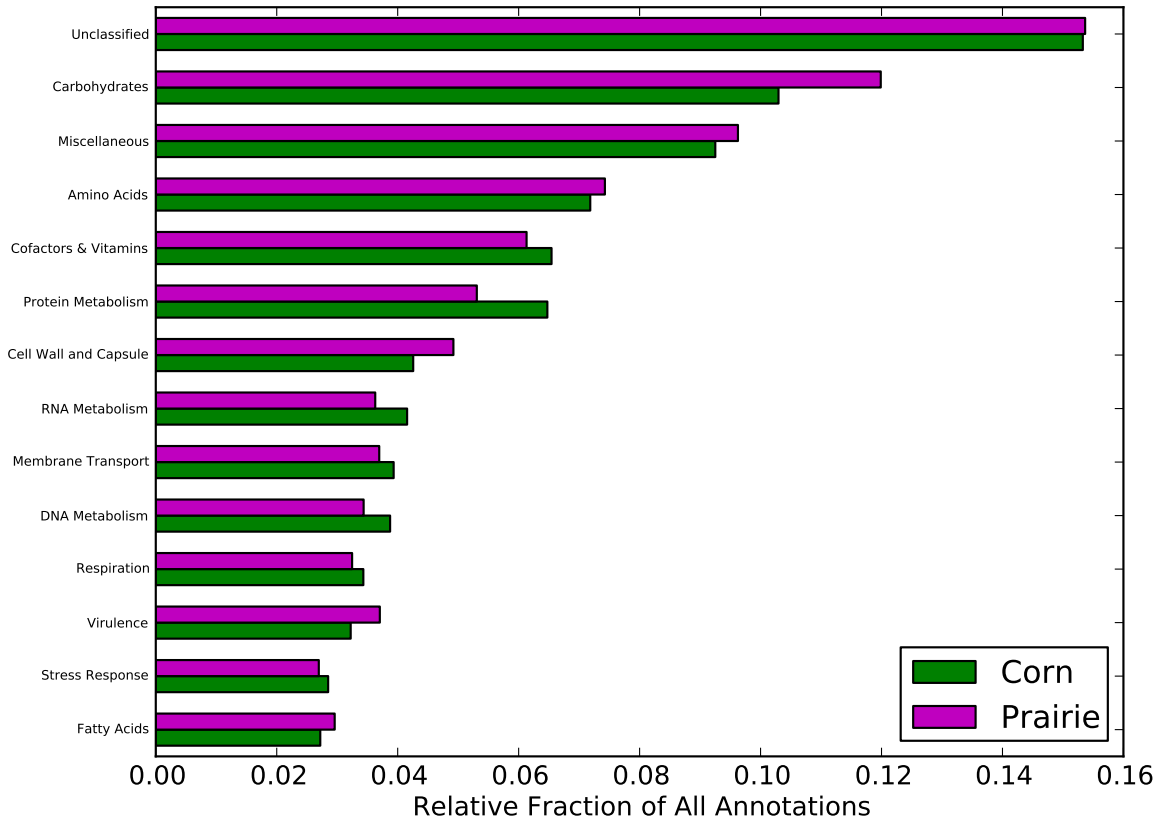


Figure 5: Functional distribution from SEED subsystem annotations for Iowa corn and prairie metagenomes.

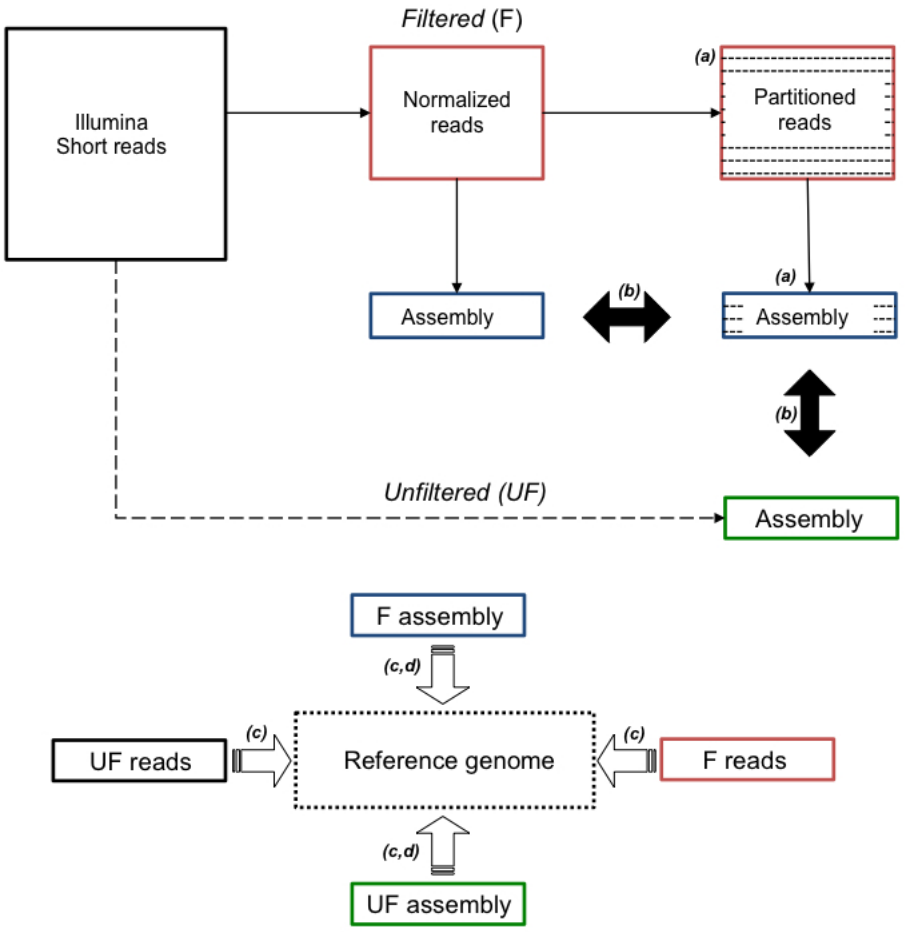


Figure 6: Flowchart describing methods for *de novo* metagenomic assembly. Using the HMP mock community dataset, alternative approaches for data reduction and assembly were compared. (a) Disconnected subgraphs of the assembly graph were partitioned. Most connected subgraphs contained reads and contigs aligning to distinct genomes (Supplementary Fig. 6). (b) The genomic content of all assemblies were found to be comparable in genomic content. (c) Reads and assembled contigs could be aligned to reference genomes to determine effectiveness of recovery. (d) The abundance of contigs (based on read mapping) could be compared to estimated abundances of corresponding reference genomes.

Online Methods

Assembly Pipeline

The entire assembly pipeline for the mock community is described in detail in an IPython notebook available for download at *http://nbviewer.ipython.org/urls/raw.github.com/ngs-docs/ngs-notebooks/master/ngs-70-hmp-diginorm.ipynb* and *http://nbviewer.ipython.org/urls/raw.github.com/ngs-docs/ngs-notebooks/ master/ngs-71-hmp-diginorm.ipynb*. Soil assembly was performed with the same pipeline and parameter changes as described in Supplementary Information. The annotated metagenome for Iowa corn can be found at *http://metagenomics.anl.gov/linkin.cgi?metagenome=4504797.3* and Iowa prairie at *http://metagenomics.anl.gov/linkin.cgi?metagenome=4504798.3*.

Statistical Methods

The reference-based abundance (from reads mapped to reference genomes) and assembly-based abundance (from reads mapped to contigs) of genomes were compared. Using a one-directional, paired t-test of squared deviations, the abundance estimates of the unfiltered and filtered assemblies were compared. The mean and standard deviation of the abundances of unfiltered contigs, filtered contigs, and reference genes were 6.8 +/- 7.1, 8.1 +/- 7.7, and 7.8 +/- 5.2, respectively. We expected the filtered assembly to have increased accuracy due to a reduction of errors (e.g. normalization and high abundance filtering) and used a one-sided t-test which indicated that abundance estimations from the filtered assembly were significantly closer to predicted abundances from reference genomes (n=28,652, p-value of 0.032).

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).

9. Scholz, M. B., Lo, C.-C. & Chain, P. S. G. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology* **23**, 9–15 (2012).

10. Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv:1203.4802* (2012).

11. Howe, A. C. *et al.* Illumina sequencing artifacts revealed by connectivity analysis of illumina sequencing artifacts revealed by connectivity analysis of metagenomic datasets. *arXiv:1212.0159* (2012).

12. Pell, J. *et al.* Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 13272–13277 (2012).

13. Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome Research* **18**, 324–30 (2008).

14. Pevzner, P. A., Tang, H. & Waterman, M. S. An eulerian path approach to dna fragment assembly. *Proc Natl Acad Sci USA* **98**, 9748–53 (2001).

15. Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research* (2012).

16. Girvan, M. S., Campbell, C. D., Killham, K., Prosser, J. I. & Glover, L. A. Bacterial diversity promotes community stability and functional resilience after perturbation. *Environmental Microbiology* **7**, 301–313 (2005).

17. McGrady-Steed, J., Harris, P. M. & Morin, P. J. Biodiversity regulates ecosystem predictability. *Nature* **390**, 162–165 (1997).

18. Müller, A. K., Westergaard, K., Christensen, S. & Sørensen, S. J. The diversity and function of soil microbial communities exposed to different disturbances. *Microbial ecology* **44**, 49–58 (2002).

19. Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 3160–3165 (2004).

Supplementary Information: Approaches for Large Scale Metagenome Assembly

January 1, 2013

Summary of approaches used on mock community dataset

The HMP mock community dataset and its available draft reference genomes were used to evaluate our approaches towards data reduction and partitioning for *de novo* metagenomic assembly. Reads of the mock community dataset were initially digitally normalized to a coverage threshold of 20 (as previously described in [1]), reducing the total number of reads from 14 to 11 million. Additionally, to remove possible sequencing artifacts associated with high coverage sequences (previously described in Howe et al., in preparation), highly-abundant sequences (20-mers present at coverage greater than 50-fold) were filtered and the dataset was further normalized to a coverage of 10, resulting in a total of 9 million reads (Fig. 3). Finally, the remaining reads were divided into disconnected sets of reads resulting in a total of 85,818 partitions containing greater than five reads (summarized in Table 1).

Supplementary Methods

Datasets

In this study, we examined two large soil metagenomes generated from soils collected from Iowa corn and native prairie soils. Sequencing was performed at the DOE Joint Genome Institute (Walnut Creek, CA). Reads were quality trimmed at where Phred scores indicated a score of '2'. The total quality-trimmed reads in the Iowa corn and prairie datasets were 1.8 million and 3.3 million, respectively. We also include a human gut mock community dataset (combined from SRA SRX055381 and SRX055380). For this mock community dataset, DNA from bacterial isolates originally recovered from within or on the human body was mixed together at

staggered concentrations (over 5 orders of magnitude based on genomic DNA concentrations) and sequenced. The mock community dataset originally contained 14.5 million reads.

To evaluate our approaches, we added simulated reads from either a single E. Coli (str. K-12 substr DH10B) or five E. coli strains (K-12 substr DH10B, E24377, O147:H7 str. EC4115, UMN026, SE15) into select metagenomes. We computationally generated 100 bp reads from each reference genome to a coverage of 10x and with a 2% error rate and subsequently randomly shuffled these reads with select datasets.

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.

Digital normalization

Digital normalization was previously describe in [1]. For the mock community dataset, digital normalization was performed with the following parameters: K=20, coverage=20, and Bloom filter size = 1 GB x 4. For Iowa corn metagenome, digital normalization parameters were as follows: K=20, coverage=20, and Bloom filter size = 48 GB x 4. Similar parameters were used for the Iowa prairie metagenome, with the exception that the Bloom filter size was 60 GB x 4.

Removal of high abundance sequences

To eliminate known sequencing artifacts in Illumina metagenomes (previously described in Howe et al., in preparation), high abundance sequences (coverage greater than 50) were removed using the count-min-sketch datastructure used for digital normalization. For the relatively high coverage mock community dataset, filtered reads were subsequently normalized to a coverage of 10 (K=20, bloom filter size = 1 GB x 4).

Partitioning and *de novo* assembly of disconnected reads

Normalized and filtered datasets were loaded into a probabilistic representation of the assembly graph as described in [2], and disconnected partitions of the resulting graph were separated. Partitions containing less than five reads were discarded. Each partition was subsequently

assembled using the Velvet assembler with the same setting as described 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 ([3], 99% similarity) and merged with Minimus2 [4].

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 [5]. 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 the reference genome.

Read coverage by assemblies

All quality trimmed reads for Iowa corn and prairie were aligned with assembled contigs (length greater than 300 bp) using default parameters in Bowtie2 [5]. Paired end reads were evaluated according to concordance with paired end library preparation (i.e. paired end reads on opposite DNA strands) and the alignment of both pairs of reads to an assembled contig. The base pair coverage of each contig was estimated with the median base pair coverage of all reads across the length of the contig. Additionally, for each position in a contig (with the exception of the external 100 bp on each end), the percentage of the mapped consensus base pairs was calculated. The fraction of positions with greater than 95% base consensus was calculated to estimate the presence of polymorphisms within the assembled contig.

Annotation of assemblies

Assembled contigs and their corresponding median bp coverage for the Iowa corn and prairie metagenomes were upload into MG-RAST annotation pipeline [6] and are available on MG-RAST as 4504979.3 (Iowa corn) and 4504798.3 (Iowa Prairie). The resulting MG-RAST blat annotations were compared to the M5NR database using a maximum e-value of 1e-5, a minimum identity of 60%, and a minimum alignment length of 15 aa . Both the phylogenetic distribution of bacteria (phyla) and functional distibution of subsystems were compared between the Iowa and corn metagenomes.

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 other 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.

The reference-based abundance (from reads mapped to reference genomes) and assembly-based abundance (from reads mapped to contigs) of genomes were compared. Using a one-directional, paired t-test of squared deviations, the abundance estimates of the unfiltered and filtered assemblies were compared. We expected the filtered assembly to have increased accuracy due to a reduction of errors (e.g. normalization and high abundance filtering) and used a one-sided t-test which indicated that abundance estimations from the filtered assembly were significantly closer to predicted abundances from reference genomes (p-value of 0.032).

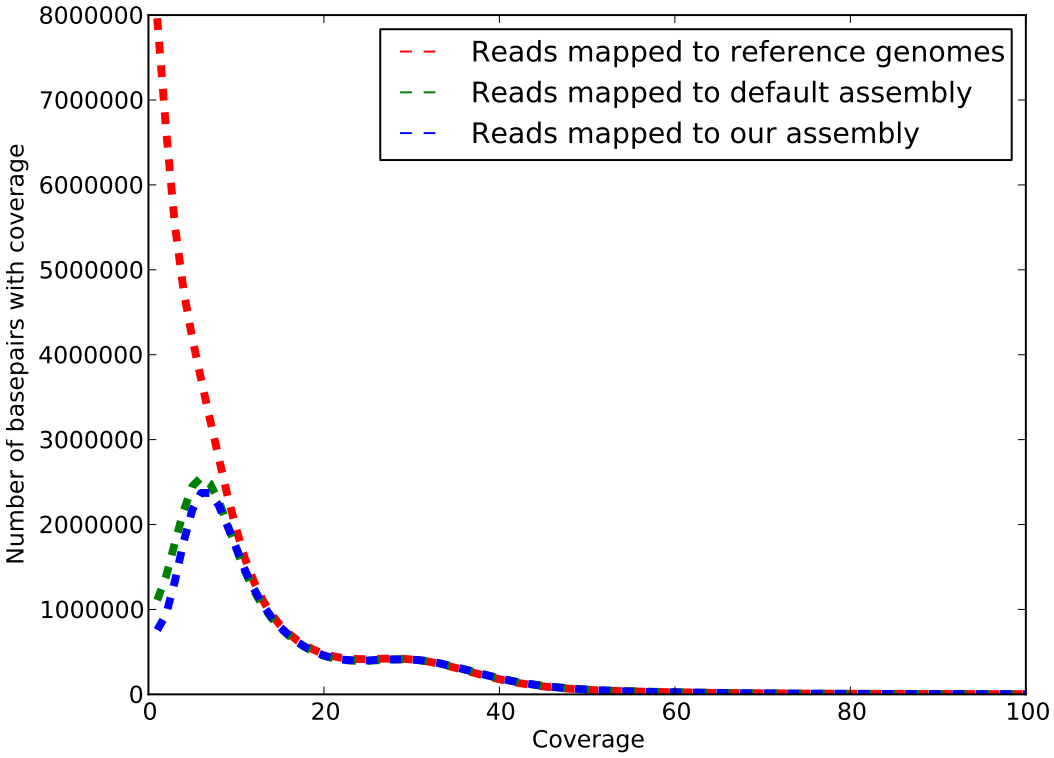
Annotations against the M5NR database were obtained through the MG-RAST annotation pipeline. The phylogenetic and functional distribution of SEED subsystems between the Iowa corn and prairie metagenome were compared (Fig. 6 and 7). For each subsystem, the relative abundance of each subsystem was calculated and the ratio of the fraction present in the Iowa corn and prairie was determined (e.g., the relative abundance of a subsystem which was equally represented in both corn and prairie metagenomes would equal 1). To estimate similarity among all subsystems and phyla, the following was calculated: $((1 - \text{ratio})^2)^{0.5}$ where a value closer to 0 indicates higher similarity. Overall, for the phylogenetic and functional distribution of SEED annotations, this value was 0.35 +/- 0.57 and 0.10 +/- 0.08.

Supplementary Tables

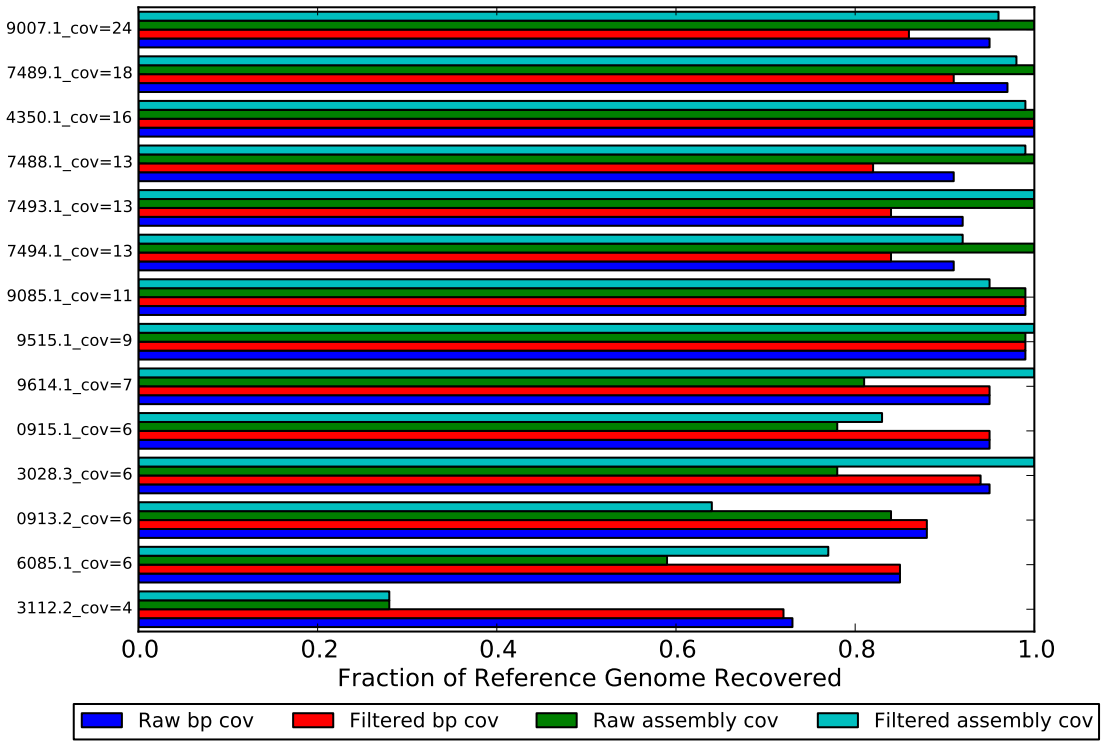
Supplementary Table 1: HMP mock dataset reference genomes estimated sequencing depth (median bp coverage of reads), number of partitions, total length (bp), coverage of reference genomes by unfiltered reads (UF Cov), coverage of reference genomes by filtered reads (F Cov), coverage of reference genomes by unfiltered assembled contigs (UFA Cov), and coverage of reference genomes by filtered assembled contigs (FA Cov).

Reference Genome	Coverage	No. Partitions	Length (bp)	UF Cov (bp)	F Cov (bp)	UFA Cov	FA Cov
gi|32470588|ref|NC_005008.1|	2,412	9	4,439	4,439	1,058	100 %	28 %
gi|32470581|ref|NC_005007.1|	549	16	4,679	4,679	4,585	100 %	77 %
gi|32470520|ref|NC_005003.1|	533	21	6,585	6,585	6,441	100 %	64 %
gi|32470572|ref|NC_005006.1|	253	2	8,007	8,004	7,953	100 %	100 %
gi|32470532|ref|NC_005004.1|	112	52	24,365	24,358	24,291	100 %	83 %
gi|126640109|ref|NC_009084.1|	85	3	11,302	11,295	11,270	100 %	100 %
gi|32470555|ref|NC_005005.1|	74	12	17,261	17,202	17,180	100 %	100 %
gi|10957398|ref|NC_000958.1|	71	73	177,466	177,261	174,614	100 %	95 %
gi|10957530|ref|NC_000959.1|	52	37	45,704	44,974	43,557	100 %	92 %
gi|126640097|ref|NC_009083.1|	48	2	13,408	13,405	13,383	100 %	100 %
gi|15807672|ref|NC_001264.1|	40	63	412,348	410,970	403,553	100 %	99 %
gi|15805042|ref|NC_001263.1|	32	546	2,648,638	2,634,512	2,589,566	100 %	99 %
gi|27466918|ref|NC_004461.1|	30	476	2,499,279	2,498,081	2,492,248	100 %	98 %
gi|125654693|ref|NC_009008.1|	29	14	37,100	36,585	33,250	94 %	96 %
gi|161508266|ref|NC_010079.1|	29	442	2,872,915	2,298,758	2,157,196	100 %	92 %
gi|77404776|ref|NC_007490.1|	27	27	100,828	99,385	93,550	100 %	96 %
gi|125654605|ref|NC_009007.1|	24	92	114,045	108,526	97,860	100 %	96 %
gi|77404693|ref|NC_007489.1|	18	12	105,284	102,212	96,169	100 %	99 %
gi|24378532|ref|NC_004350.1|	16	131	2,030,921	2,029,376	2,025,544	100 %	99 %
gi|77404592|ref|NC_007488.1|	13	30	114,178	103,351	93,637	100 %	99 %
gi|77461965|ref|NC_007493.1|	13	628	3,188,609	2,919,441	2,681,855	100 %	99 %
gi|77464988|ref|NC_007494.1|	13	262	943,016	862,781	788,626	100 %	98 %
gi|126640115|ref|NC_009085.1|	11	683	3,976,747	3,939,190	3,936,208	99 %	99 %
gi|148642060|ref|NC_009515.1|	9	552	1,853,160	1,828,231	1,826,639	99 %	98 %
gi|150002608|ref|NC_009614.1|	7	7,751	5,163,189	4,899,622	4,896,808	81 %	82 %
gi|15644634|ref|NC_000915.1|	6	2,888	1,667,867	1,581,502	1,581,024	78 %	79 %
gi|194172857|ref|NC_003028.3|	6	4,123	2,160,842	2,047,832	2,037,347	78 %	78 %
gi|49175990|ref|NC_000913.2|	6	5,913	4,639,675	4,080,605	4,074,119	84 %	85 %
gi|50841496|ref|NC_006085.1|	6	6,459	2,560,265	2,169,547	2,169,056	59 %	64 %
gi|77358697|ref|NC_003112.2|	4	9,269	2,272,360	1,655,023	1,626,301	28 %	33 %

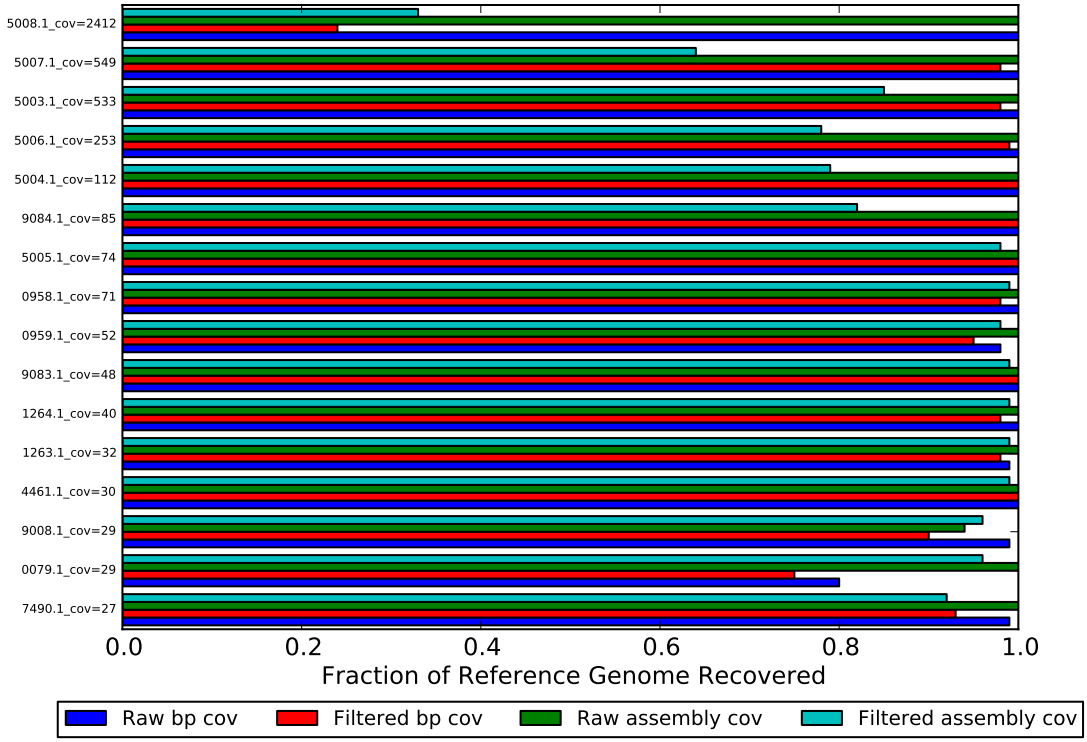
Supplementary Figures



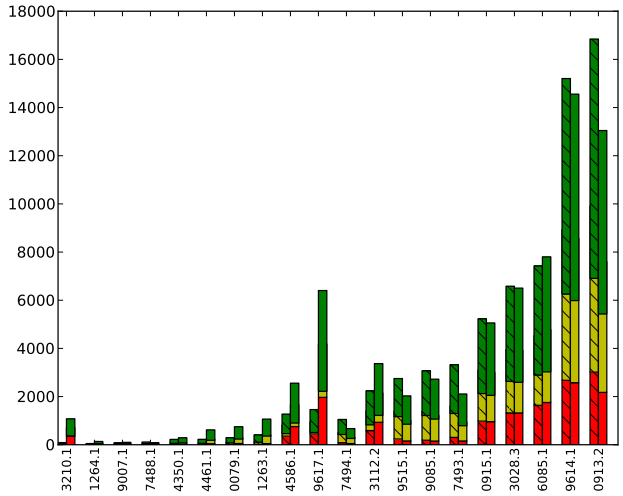
Supplementary Figure 1: Number of basepairs with specified coverage for reads which map to reference genomes and unfiltered and filtered assembled contigs greater than 300 bp.



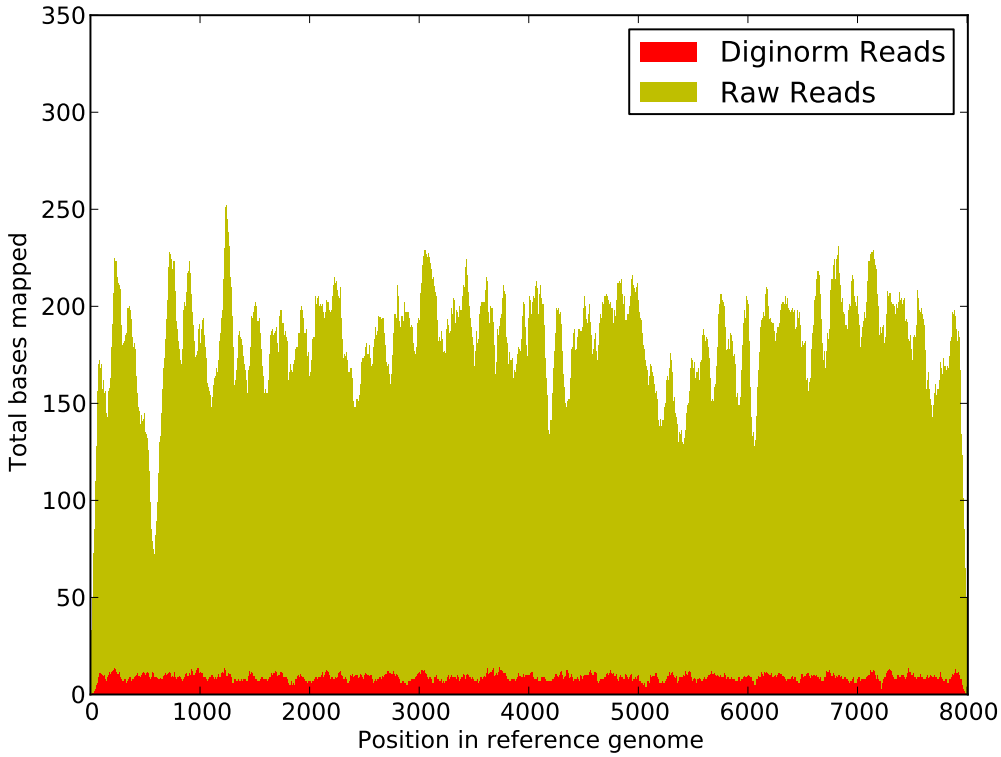
Supplementary Figure 2: Coverage of reference genomes by unfiltered and filtered assembled contigs and unfiltered and filtered reads.



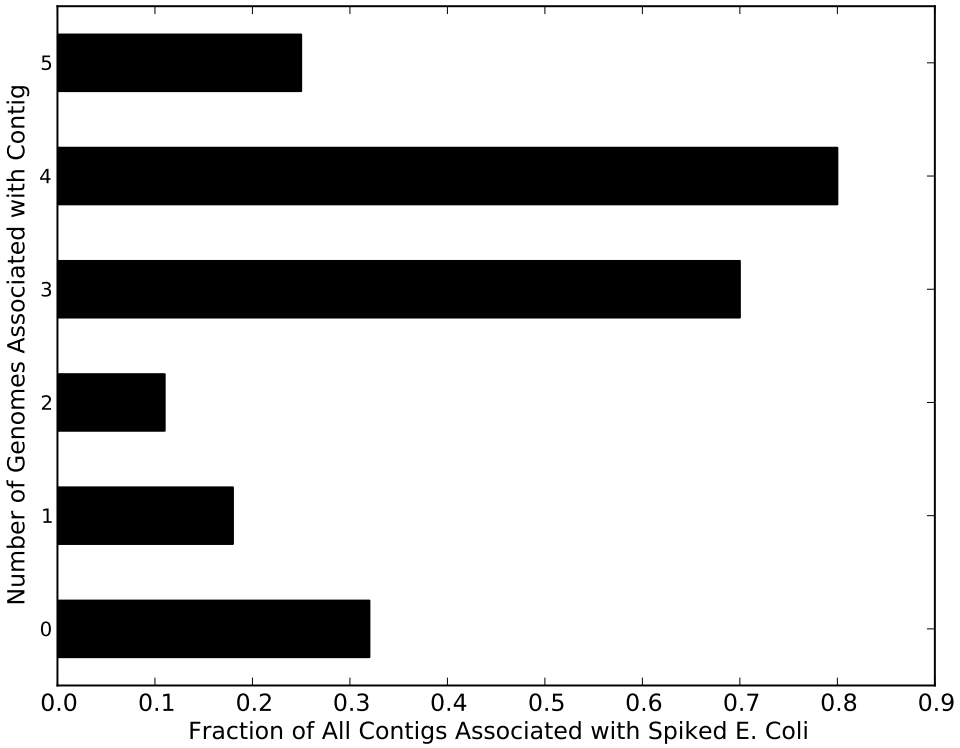
Supplementary Figure 3: Coverage of reference genomes by unfiltered and filtered assembled contigs and unfiltered and filtered reads.



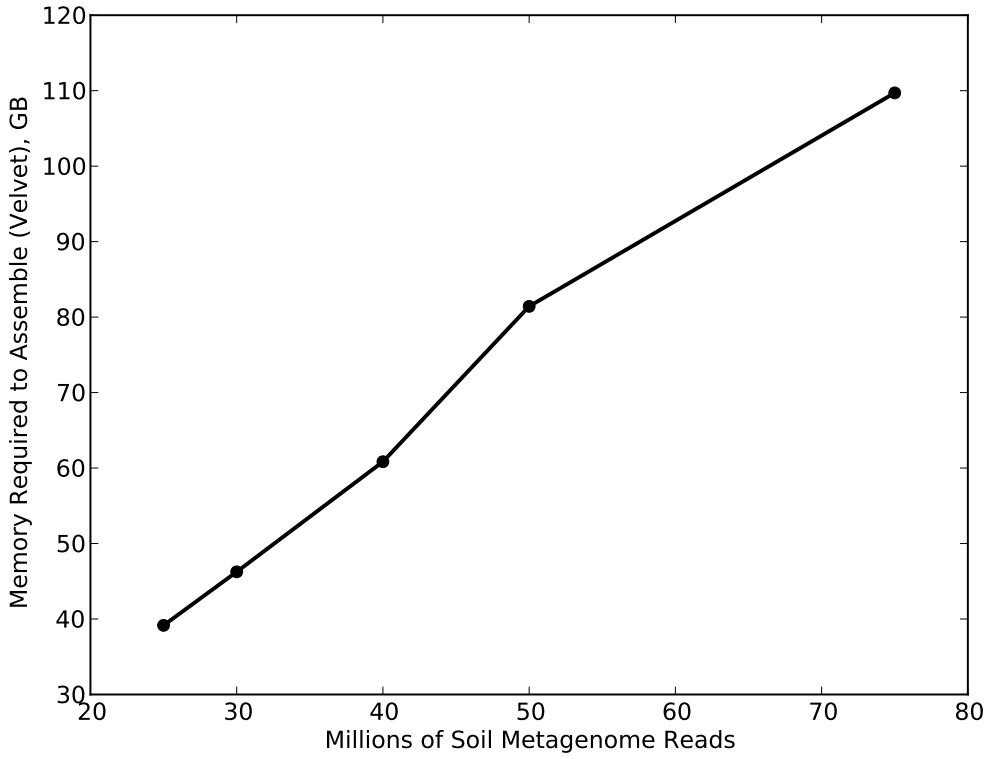
Supplementary Figure 4: Total number of contigs for unfiltered (bars with hashed lines) and filtered (solid bars) for top twenty references with most assembled contigs (ranked by unfiltered assembly). Red indicates contig lengths less than 500 bp, yellow indicates contig lengths between 500 bp and 3000 bp, and green indicates contig lengths greater than 5000 bp. Reference genome IDs shown here are last 5 digits of RefSeq ID.



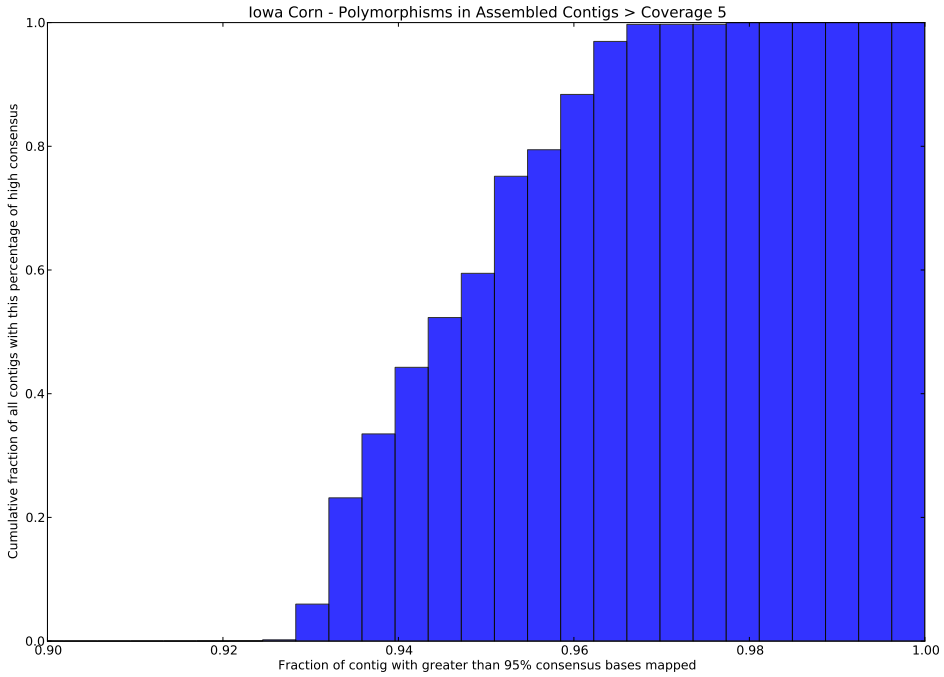
Supplementary Figure 5: Alignment of reads (colored by originating partition) to reference genome NC_00745901



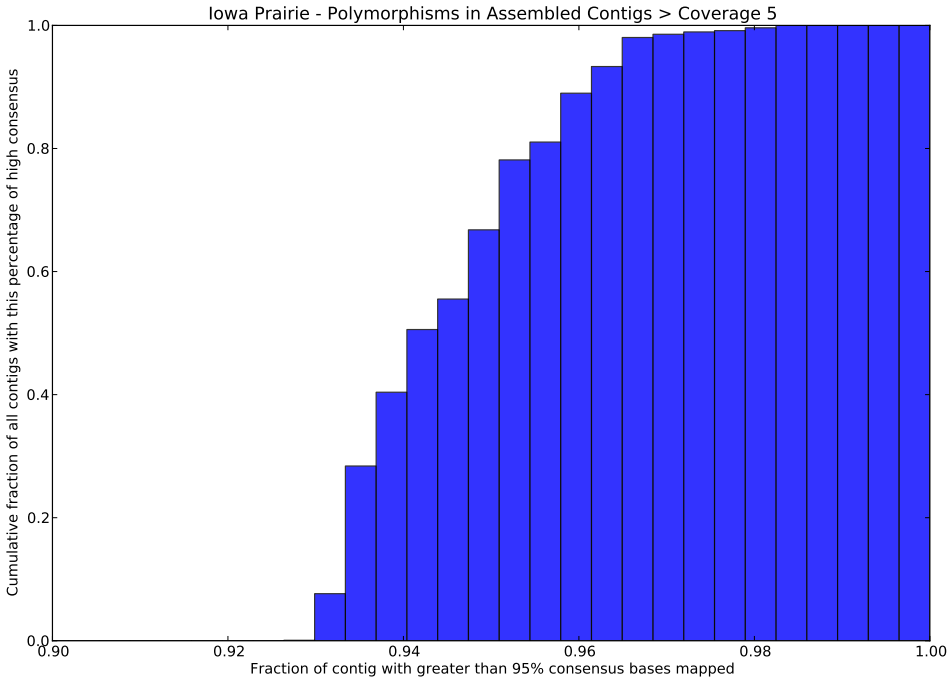
Supplementary Figure 6: The fraction of assembled contigs assembled from partitions containing spiked *E. coli* reads associated with 0 to five of the *E. coli* reference genomes. The large majority of contigs contain reads associated with multiple genomes or to no genome.



Supplementary Figure 7: Memory requirements to assemble subsets of Iowa corn soil metagenome



Supplementary Figure 8: The presence of polymorphic sequences in assembled contigs of Iowa corn metagenome.



Supplementary Figure 9: The presence of polymorphic sequences in assembled contigs of Iowa prairie metagenome.

Supplementary References

References

[1] Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv:1203.4802* (2012).

[2] Pell, J. *et al.* Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 13272–13277 (2012).

[3] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics (Oxford, England)* (2012).

[4] Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. Minimus: a fast, lightweight genome assembler. *Bmc Bioinformatics* **8**, 64 (2007). URL `http://www.biomedcentral.com/1471-2105/8/64`.

[5] Langmead, B. & Salzberg, S. Fast gapped-read alignment with bowtie 2. *Nature Methods* **9**, 357–359 (2012).

[6] Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* **9**, 386 (2008).