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~~ computationally 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 and any assembler; 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 systemsare~~ DNA sequencing has begun to reveal the enormous diversity and heterogeneity associated within these systems, making them difficult to study *in situ* ~~, and we are only beginning to understand their diversity and functional potential. Advances in DNA sequencingnow provide~~ [2,4,5]. With ultradeep sequencing, we now have unprecedented access to ~~the genomic content of these communities via shotgun sequencing, which produces millions to billions of short-read sequences . Because shotgun sequencing samples communities randomly, ultradeep sequencing is needed to detect~~ even the rare species in ~~environmental samples, with an estimated~~ these environment (i.e., 50 Tbp ~~needed for an individual~~ required to adequately sample a gram of soil [8]~~. Both short read lengths and the large volume of sequencing data pose~~).

Alongside sequencing breakthroughs, new challenges to sequence analysis approaches ~~.~~ have emerged. Sequencing dataset sizes are growing at an exponential rate, requiring significant computational resources for data storage and analysis. 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].

Consequently, investigators of metagenomic datasets are confounded by overwhelming volumes of data which they do not have the computational resources to efficiently analyze or which are unsuitable for the current suite of bioinformatic tools (because of short read lengths or a lack of reference genomes). *De novo* assembly of ~~raw~~ sequence data offers several advantages ~~over analyzing the sequences directly. Assembly removes~~ for analyzing metagenomic datasets. It provides improved accuracy of sequences by removing most random sequencing errors and ~~decreases the total amount of data to be analyzed, yielding assembled~~ results in contigs longer and more specific than unassembled sequencing reads. Further, assembly significantly reduces the total volume of data required for downstream analysis (e.g., gene annotation). Importantly, *de novo* assembly also 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]. Although many metagenome-specific assemblers have recently been developed for community assembly, they cannot work with the volume of reads necessary to achieve high coverage for extremely diverse environmental

metagenomes [9].

Here, we present two pre-assembly read filtering strategies, digital normalization and partitioning, that provide a general strategy for scaling and improving 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 (HGMC) dataset, and find that these filtering methods result in assemblies nearly identical to assemblies from the unprocessed dataset. Moreover, we show that partitioning separates most reads into species-level bins, providing an alternative to abundance-based and k-mer approaches to species clustering.

We next apply these approaches to the assembly of previously intractable 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 of the HGMC dataset by comparing unfiltered traditional assembly to the the described filtered assembly (Fig. 6; see Methods and Supplementary Information). ~~Initially, the~~ The abundance of genomes within the ~~mock~~ HGMC 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~~ 2and 3). Overall, the unfiltered dataset reads covered a total of 93% of the reference genomes. ~~Filtering removed~~ Reads were removed based on their coverage within the dataset (See Methods); we removed a total of 5.9 million reads ~~,~~ (40% of the total reads) from the original HGMC dataset (Table 1)~~; the remaining reads~~ . The remaining reads in the filtered HGMC dataset covered 91% of the reference genomes (Table 1 and Supplementary Fig. 2 and 3).

We next compared the recovery of reference genomes ~~in~~ from contigs assembled from the original and filtered HGMC datasets. Using the Velvet assembler [13],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 [14] and Meta-IDBA [15] , Table 2). Overall, the unfiltered and filtered assemblies were very similar, sharing 95% of genomic content. ~~For the~~

highest abundance In the most abundant 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 readscompared the estimated abundance of genomes in the HGMC dataset using reads aligned to reference genomes and the unfiltered and filtered assemblies. Genome abundance was estimated through the alignment of unassembled reads to either the reference genome or assembled contigs. Sequencing coverage was determined as the median base pair coverage of all aligned reads. (Supplementary Fig. 5). Above a sequencing coverage of fiveFor assembled contigs with coverage greater than 5, the majority of reads which could be aligned contigs were also mapped to reference genomes were included in the assembled contigs (Supplementary Fig. 3 and 4). Below this threshold, reads could be were 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 Comparing the unfiltered and filtered assemblies. The abundance estimations , the estimated abundance of the HGMC genomes 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 99identical. In the mock dataset , we identified 9 million reads in~~ HGMC dataset was partitioned into 85,818 disconnected partitions containing a total of 9 million reads (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. The resulting assemblies of the unpartitioned and partitioned dataset were very similar, sharing 99% identical genomic content.

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 (coverage below 25). This is consistent with ~~the observation that the main effect of low coverage is to "break" connectivity in~~ previous observations where low coverage in sequences cause "breaks" in connectivity within the assembly graph [16,17].

To further evaluate the effects of partitioning, we introduced spiked, simulated reads from *E. coli* genomes into the ~~mock community~~ HGMC 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~~ HGMC dataset and the resulting dataset, HGMC.Ecoli1, was filtered,

partitioned, and assembled as described below. Similar amounts of data reduction after digital normalization and partitioning (Table 1) were observed. Among the 81,154 partitioned sets of reads ~~,~~ ~~we identified~~ in the HGMC.Ecoli1 dataset, only 2,580 (3.2%) partitions ~~containing~~ contained reads from multiple genomes. ~~A totalof~~ In total, 424 partitions contained reads from the spiked *E. coli* genome (201 partitions contained *only* spiked reads) and when assembled, these reads 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 datasetand performed the same analysis. Partitioning this "mix-spiked" mock community datasetresulted~~ original HGMC dataset. This dataset, referred to as HGMC.EColi5, was filtered, partitioned, and assembled, resulting in 81,425 partitions~~, of which~~ . Among these, 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~~ Independently assembling the partitions containing reads originating from the spiked *E. coli* strains ~~. Among the resulting~~ resulted in 6,076 contigs, all but three ~~contigs originated~~ originating from a spiked *E. coli* genome. The remaining three contigs were more than 99% similar to ~~HMP mock~~ available 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 found in the HGMC (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~~

The assembly of the HGMC.Ecoli5 dataset was also performed without removal of any reads (e.g., no digital normalization or partitioning). Comparing the HGMC.Ecoli5 unfiltered and filtered assemblies, we found that the fraction of contigs which are associated with multiple genomes ~~in~~ were similar. In the unfiltered data set, ~~where~~ 66% of 4,702 contigs were associated with spiked reads ~~contain reads that originate~~ which could have originated from more than one ~~spiked~~ genome.

Data reduction and partitioning enable the assembly of two soil metagenomes

We next applied ~~these~~ digital normalization and partitioning approaches to the *de novo* assembly of two soil metagenomes. Unfiltered Iowa corn and prairie datasets (containing 1.8 billion and 3.3 billion reads, respectively) could not be assembled by Velvet in ~~500 GB~~ 500GB 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~~used for the HGMC dataset, the Iowa corn and prairie datasets were reduced to 1.4 billion and 2.2billion reads, respectively, and after partitioning, a total of 1.0 billion and 1.7 billion reads remained, respectively. ~~Prefiltering~~ Pre-filtering used 300 GB of RAM or less. ~~The~~ Notably, 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~~ HGMC 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 a minimum of 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 contigswithin the soil metagenomes~~ the alignment of sequencing reads to assembled contigs, we estimated the distribution of sequencing coverage in resulting assemblies (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~~ As the resulting assemblies are consensus sequences representative of the unassembled dataset, we investigated the degree of variation (i.e., polymorphism) present among aligned reads to assembled contigs. (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~~ We acknowledge that the diversity and sequencing depth represented by the ~~mock community~~ HGMC dataset is extremely low compared to that of most environmental metagenomes; however, it represents a simplified, unevenly sampled model for a metagenomic dataset which ~~enables~~ allows for the evaluation of ~~analyses~~ novel approaches 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~~ normalizes the abundance of reads in a dataset to a specific sequencing coverage. As there exists an observed coverage "sweet spot" at which point sufficient sequences are present for robust assembly (Supp Fig. 1)~~. The normalization of sequences~~ , this effectively reduces the volume of the dataset for assembly while removing errors introduced by extraneous reads. Further, this normalization also resulted in more even coverage (Fig. 2), minimizing assembly problems caused by variable cover-

age. 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~~ characteristics of genomes. We observed that variable abundance and conserved regions in references had an impact on ~~filtered assembly recovery~~ the recovery of sequences in filtered HGMC datasets. 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 consequently would be removed, as evidenced by a large difference in the number of reads associated with NC_005008.1 in the unfiltered and filtered (normalized) datasets (supplementary Fig. 2). ~~Consequently, the~~ The unfiltered dataset contained comparably more reads spanning these repetitive regions ~~. This most~~ which likely enabled assemblers to ~~extend the~~ more effectively extend the unfiltered assembly of these sequences~~and resulted in the observed~~ , ultimately observed as an increased recovery of these genomes in ~~the unfiltered assemblies.~~ these assemblies.

This result, though rare among ~~the mock reference genomes~~ genomes in the HGMC dataset, identifies a shortcoming of our approach, and indeed for most short-read assembly approaches,related to repetitive regions and/or polymorphisms. ~~For the soil metagenomesour~~ We acknowledge that in our assembled soil metagenomes, data reduction may ~~have caused some information loss in exchange~~ cause information loss but exchange this disadvantage for the ability to assemble previ-

ously intractable data sets. ~~Evaluation of the mock community~~ Overall, evaluation of the HGMC dataset suggests that this information loss is minimal ~~overall~~ and that our approaches result in a comparable assembly whose abundance estimations are even slightly improved.

Metagenomes contain many distinct genomes, which are largely disconnected from each other but which ~~sometimes~~ often share sequences due to sequence conservation or lateral transfer. Our ~~prefiltering~~ pre-filtering approach removes both common multi-genome elements as well as artificial connectivity stemming from the sequencing process~~.~~ ([11]). As shown above on the ~~mock data set~~HGMC dataset, the removal of these sequences does not significantly alter the recovery of reference genomes through *de novo* assembly: the resulting assemblies of ~~unpartitioned and~~ unfiltered, filtered, and filtered and partitioned datasets were nearly identical~~for the mock data. The~~ . Further, 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~~ To further examine the recovery of sequences through partitioning~~by computationally spiking in~~ , computationally-derived sequences from one or more *E. coli* strains ~~before applying filtering and partitioning~~were amended to the HGMC dataset. 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,~~ unique to our approaches, however, 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~~ 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 [18]).

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~~ HGMC 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 1.9million 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 [19–22].

We present two strategies that readily enable the assembly of very large environmental metagenomes by discarding redundancy and subdividing the data prior to assembly. The strategies are generic and broadly 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 belong to individual species. These read partitions are small enough that a variety of assembly, abundance analysis, and polymorphism analysis techniques can be easily applied to them individually.

These strategies filter and partition reads prior to assembly and can work with any assembler, considerably reducing the computational resources needed to complete an assembly. By acting as

pre-filters, digital normalization and partitioning let downstream assemblers focus on improving their performance on low-coverage or high variability data without a strong consideration for computational resources. This should enable significant improvement of metagenome assembly techniques going forward and provide the critical references which will enable future investigations of complex environments.

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

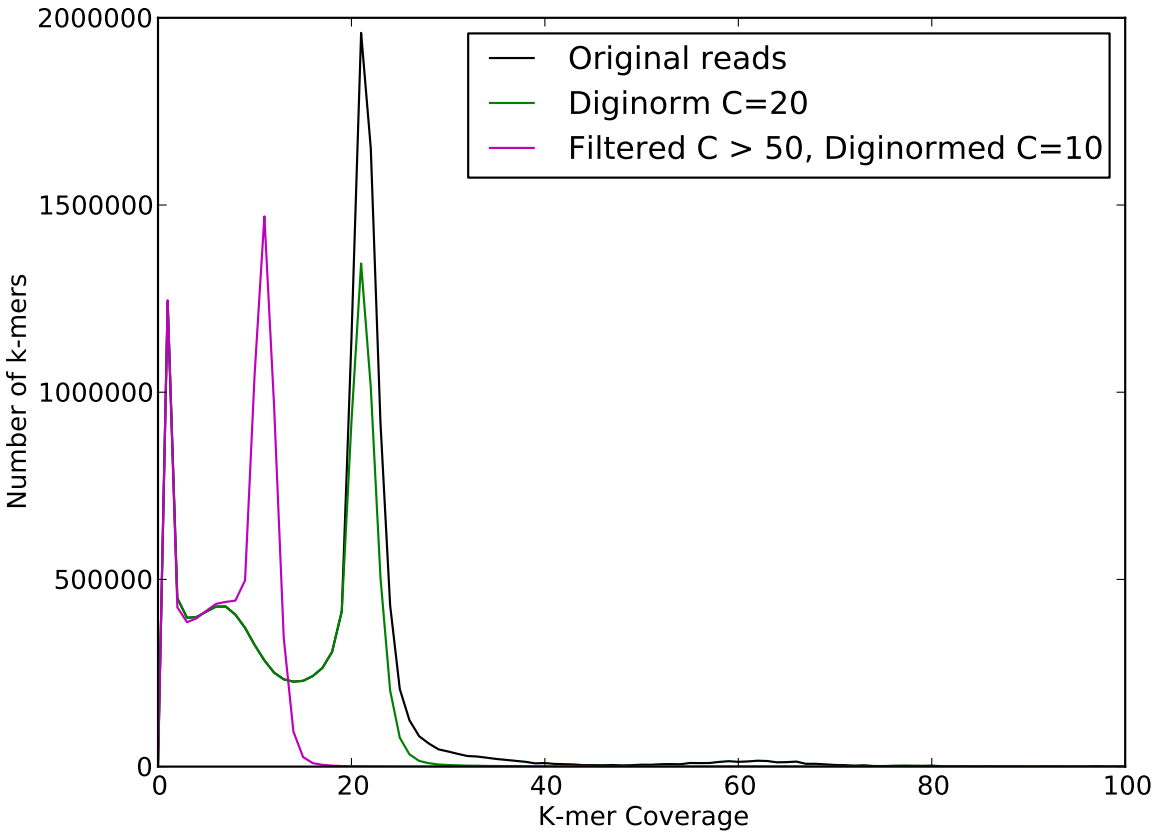

Figure 1: K-mer coverage of HMP mock community dataset before and after filtering approaches.

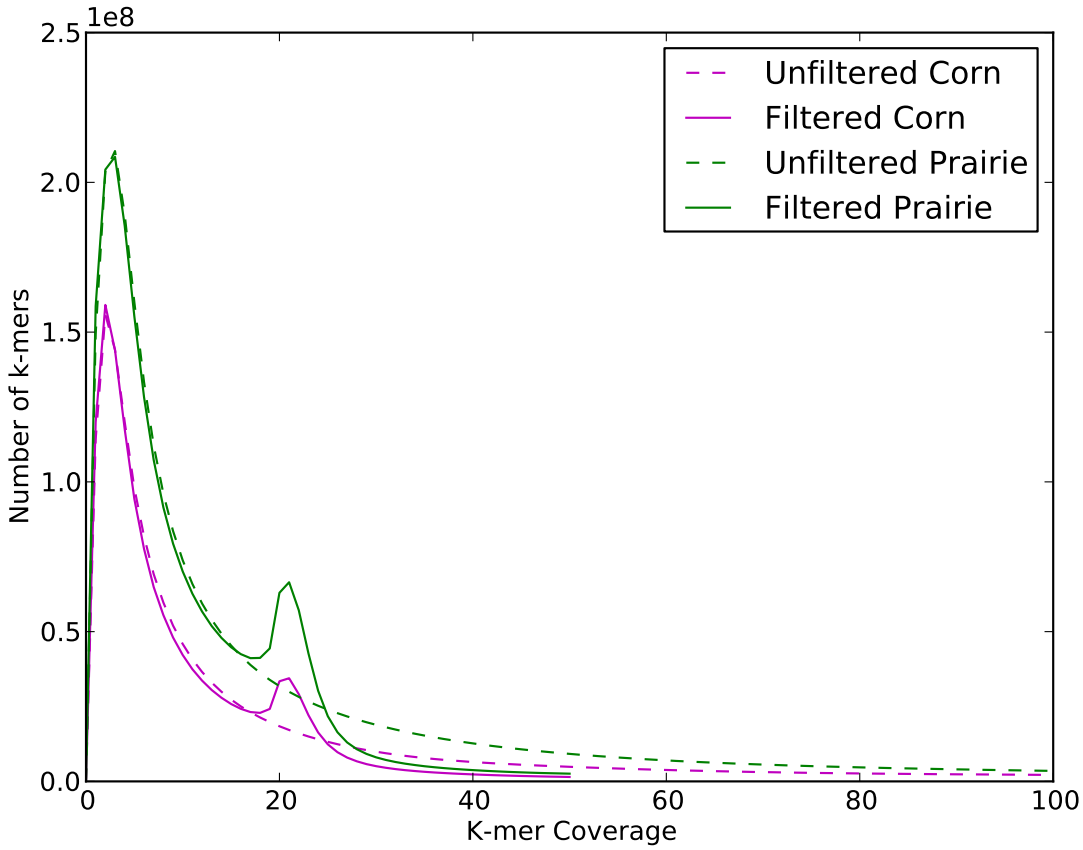

Figure 2: K-mer coverage of Iowa corn and prairie metagenomes before and after filtering approaches.

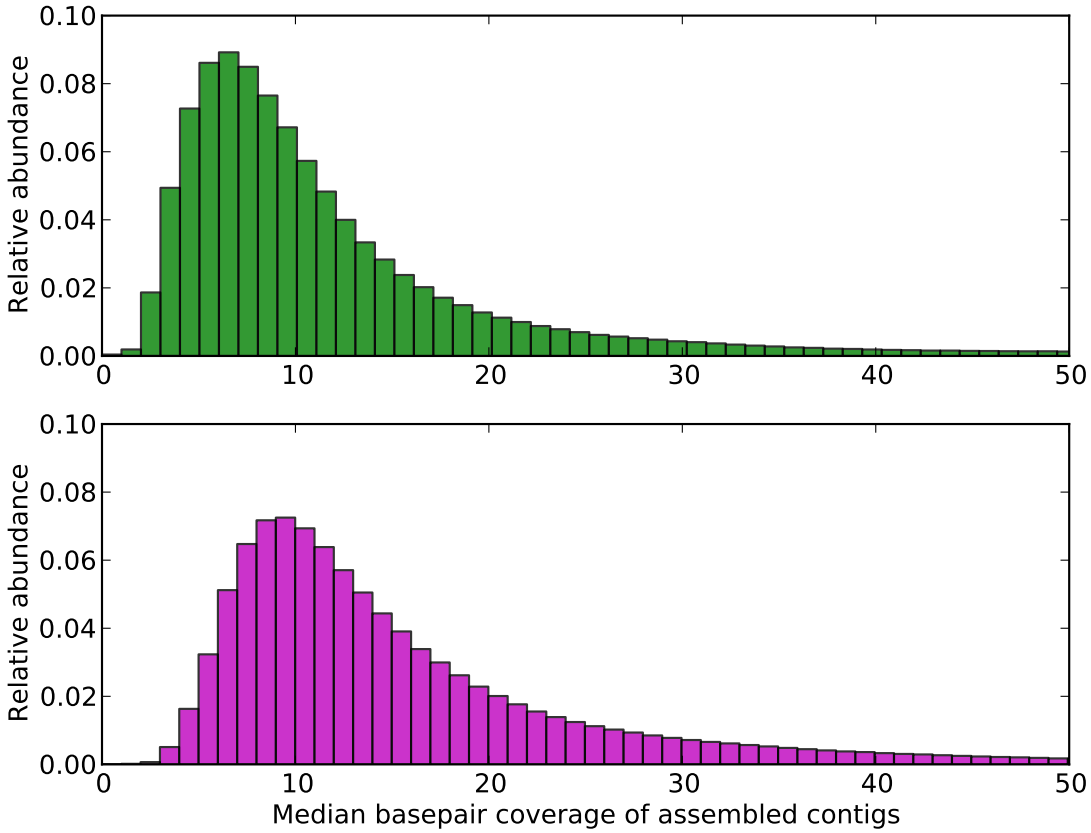

Figure 3: Coverage (median basepair recovered) distribution of assembled contigs from Iowa corn soil (top) and Iowa prairie soil (bottom) metagenomes.

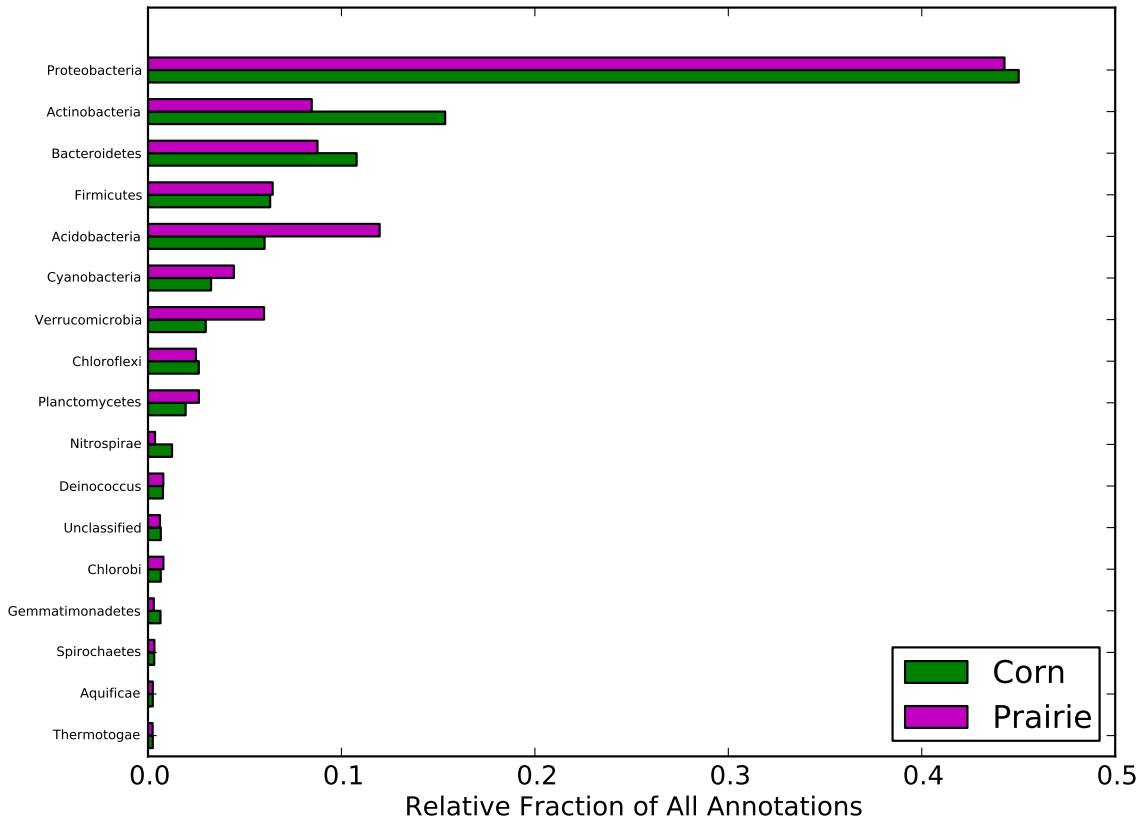

Figure 4: Phylogenetic distribution from SEED subsystem annotations for Iowa corn and prairie metagenomes.

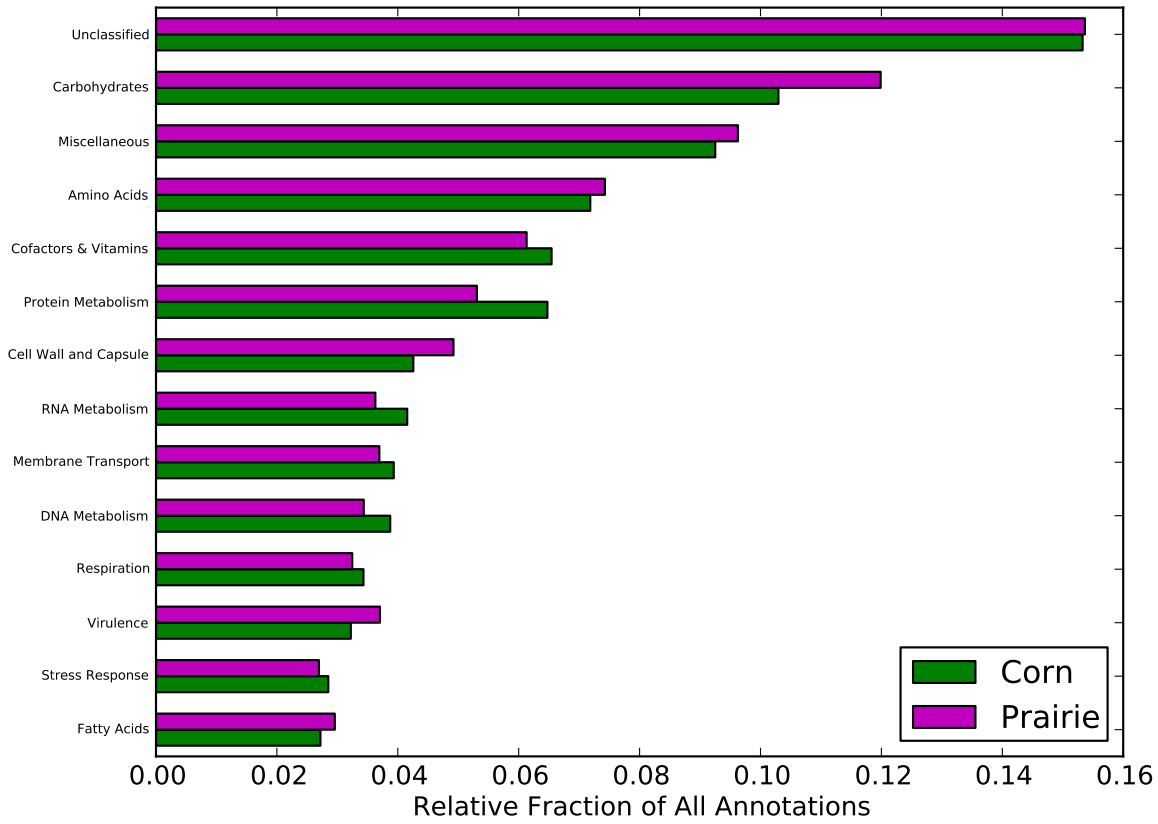

Figure 5: Functional distribution from SEED subsystem annotations for Iowa corn and prairie metagenomes.

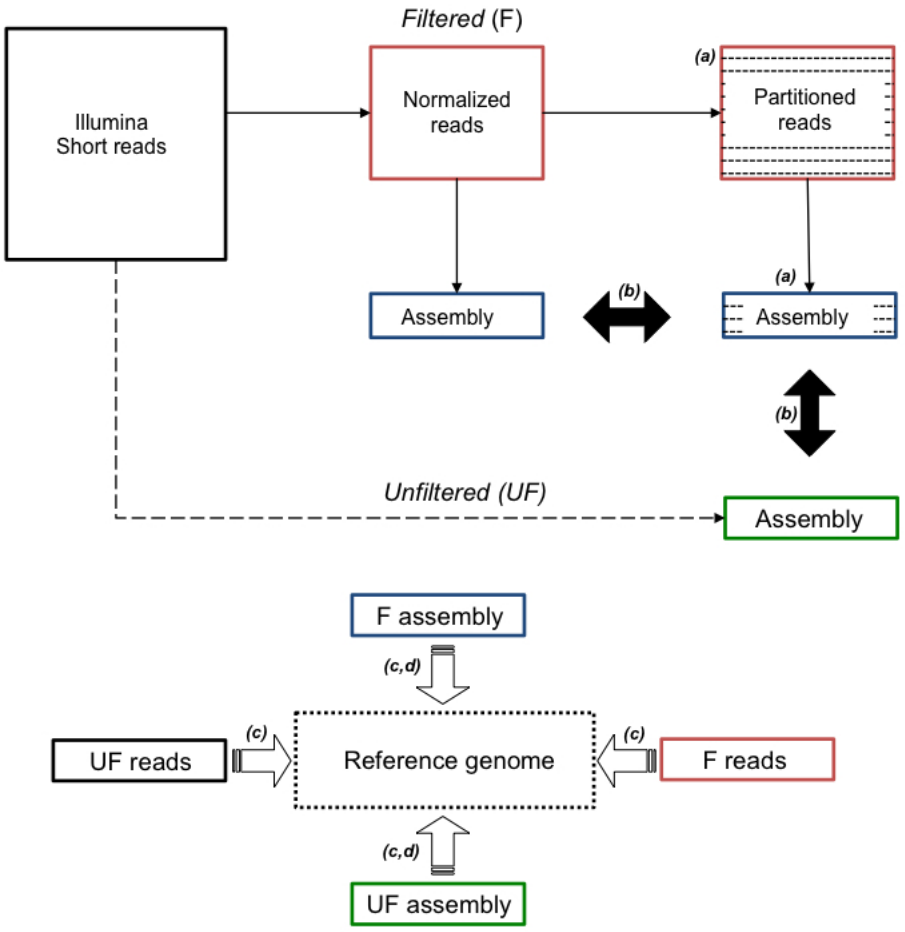

Figure 6: Flowchart describing methods for *de novo* metagenomic assembly. Using the HMP mock community dataset, alternative approaches for data reduction and assembly were compared. (a) Disconnected subgraphs of the assembly graph were partitioned. Most connected subgraphs contained reads and contigs 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. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res* **18**, 821–9 (2008).

14. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**, 265–272 (2010).

15. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics* **27**, i94–101 (2011).

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

17. 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).

18. 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).

19. 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).

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

21. 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).

22. 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).