Illumina Sequencing Artifacts Revealed by Connectivity Analysis of Metagenomic Datasets

Adina Chuang Howe[1,2], Jason Pell[1], Rosangela Canino-Koning[1] Rachel Mackelprang[2] Susannah Tringe[2] Janet Jansson[2] James M. Tiedje[1,2] C. Titus Brown[1,*]

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

2 Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI, USA

3 DOE Joint Genome Institute, Walnut Creek, CA, USA

∗ **E-mail: ctb@msu.edu**

Abstract

blah blah blah

Introduction

Given the rapid decrease in the costs of sequencing, we can now achieve the sequencing depth necessary to study even the most complex environments [1,2]. High throughput, deep metagenomic sequencing efforts in permafrost soil, human gut, cow rumen, and surface water have provided insights into the genetic and biochemical diversity of environmental microbial populations [1–3] and the extent to which they are involved in responding to environmental changes [4]. These metagenomic studies have all leveraged *de novo* metagenomic assembly of short reads to assign sequences to microbial taxa and function. *De novo* assembly is an advantageous approach to sequence analysis as it reduces the dataset size by collapsing numerous short reads into fewer contigs and provides longer sequences containing multiple genes and operons [5,6] making annotation-based approaches more practical. Furthermore, it does not rely on the availability of reference genomes to enable identification of novel genetic features and draft genomes [1,3].

Although *de novo* metagenomic assembly is a promising approach for deep sequencing of metagenomes, it is complicated by the variable coverage of sequencing reads from mixed populations in the environment and their associated sequencing errors and biases [7,8]. Several metagenomic-specific assemblers have been developed to deal with variable coverage communities, including Meta-IDBA [9], MetaVelvet, and SOAPdenovo. These assemblers rely on local models of sequencing coverage to help build assemblies and thus are sensitive to the effects of sequencing errors and biases on coverage estimations of the underlying dataset. The effects of sequencing errors on *de novo* assembly has been demonstrated in simulated metagenomes [7, 8, 10], but these datasets do not incorporate models that are representative of real metagenomic data. Specifically, these models exclude the presence of of known non-biological sequencing biases ([11–13]) which would hinder coverage-based assembly approaches.

In this study, we examine real metagenomic datasets for the presence of these artificial sequencing biases, extending previous work to large and complex datasets produced from the Illumina platform. Since these sequencing biases would erroneously connect numerous reads together, they can be characterized by their connectivity within an assembly graph. Here, we take advantage of a de Bruijn assembly graph representation to identify and evaluate highly connective sequences within various metagenomic datasets. We demonstrate that there exist highly connective sequences which originate, at least partially, from sequencing artifacts. In metagenomic datasets, we find that the connectivity of these sequences limit approaches to divide or partition large datasets for further analysis, e.g. de novo assembly. Here, we present approaches to identify and characterize these highly connective sequences and examine the effects of removing these sequences on downstream assemblies.

Results

Connectivity analysis of metagenome datasets

Presence of a single, highly-connected lump in all datasets

We selected datasets from three diverse, medium to high diversity metagenomes from the human gut [2], cow rumen [1], and agricultural soil, representing metagenomes sequenced to various depths (Table 1, column 2). To evaluate the effects of sequencing coverage, we included lower-coverage subsets of the soil metagenome (520 million reads) containing 50 and 100 million reads. We also included a previously published error-free simulated, metagenome based on a mixture of 112 reference genomes [8].

Initially, we evaluated the amount of connectivity between all sequences in each metagenome using an approach similar to the initial step of short read assemblers to identify overlaps of short sequences of length 'k', or k-mers [9, 14, 15]. For complex metagenomes, extremely large diversity and numerous sequencing errors require large amounts of memory to store resulting assembly graphs [1, 2, 4]. To overcome this limitation, we constructed a probabilistic representation of the assembly graph using a bloom filter de Bruijn graph representation within fixed memory as previously described (Pell et al, how to cite this?).

Using this assembly graph representation, we separated reads contributing to disconnected portions of the metagenome assembly graph (e.g., representatives from separate populations in the source environment). For each metagenome, regardless of origin, we found a single dominant, highly-connected set of sequencing reads which we henceforth refer to as the "lump" of the dataset (Table 1, column 3). This lump contained the largest subset of connected sequencing reads and varied in size among the datasets, ranging from 5% of total reads in the simulated metagenome to 75% of total reads in the human gut metagenome. For the soil datasets, as sequencing coverage (e.g., the fraction of reads mapped to an assembly) increased from 1.4 to 4.7 to 5.6%, the lump size increased more dramatically from 7 to 15 to 35% of all reads, indicating increasingly larger connectivity between sequences with more sequencing.

Characterizing the connectivity in the dominant lump

Given the large number of reads connected within metagenomic lumps (up to 182 and 262 million reads in the soil and human gut datasets, respectively), we quantified the degree of connectivity of sequences within the lump by estimating the average local graph density from nodes in the assembly graph (See Methods). We observed that sequences in the identified metagenomic lumps had very high local graph densities, between 22 to 50% of the total nodes in metagenomic lump assembly graphs had average graph densities greater than 20 (Table 1, column 8). In comparison, 17% of the total nodes in the simulated lump had an average local graph density greater than 20, and a mixture of the 112 source genomes for the simulated dataset had fewer than 2% of its nodes with an average graph density greater than 20.

We next assessed the extent to which graph density varied by position along the sequencing reads. The degree of position-specific bias of graph densities was estimated by calculating the average local graph density within ten steps of every k-mer by position in each read. In all environmental metagenomic reads, we observed biases in graph density at the 3'-end region of reads (Figure 1). In soil metagenomes, we observed the most dramatic biases with estimated local graph density increasing at the 3'-end of the reads. Notably, this bias was not present in the simulated dataset. Next, we identified specific sequences within dense regions of the assembly graph which consistently contributed to high connectivity in an exhaustive graph traversal of the reads within each lump. We observed that this subset of sequences were also found to exhibit position-specific biases within sequencing reads (Figure 1, solid lines). Similar to local density trends, position-specific biases of these sequences also varied between metagenomes. As sequencing coverage increased among metagenomes, the amount of 3'-end bias appeared to decrease (e.g., the soils) or inverse (e.g., rumen and human gut), and in the case of the simulated dataset, no such biases were observed.

Effects of removing of highly connective sequences on assembly

In all datasets, we found that our approaches for removal of highly connective k-mers was effective at breaking apart metagenomic lumps and the resulting size of the largest partition of connected reads in each metagenome was reduced to less than 7% of the total reads in the lump. The partitioned sets of sequences could be assembled very efficiently in parallel, greatly reducing the memory and time required for assembly from greater than 100 GB and 100 hours (e.g. largest soil metagenome lump) to less than 2 GB memory and less than 1 hour for all metagenomes.

To explore the extent to which the identified highly-connective sequences impacted assembly, we first evaluated the effects of the removing these sequences from reads in the simulated lump and its resulting assemblies. The assembly of the reads in the original, unfiltered simulated lump and that of the reads remaining after removing highly connective sequences (the filtered assembly) were compared for three assemblers (Velvet, Meta-IDBA, and SOAPdenovo). Based on total assembly length of contigs greater than 300 bp, filtered assemblies of the simulated metagenome resulted in a loss of between 4 - 16% of total assembly length (Table 2). Filtered assemblies contained fewer total contigs than unfiltered assemblies, and the maximum contig size increased in the case of Velvet assembly but decreased in the case of the Meta-IDBA and SOAPdenovo assemblies. Direct comparisons of the two assemblies found that the filtered assemblies comprised on average 88% of the unfiltered assemblies, and the unfiltered assemblies contained nearly all (96%) of the filtered assembled sequences. Overall, despite the removal of over 3% of the total unique 32-mers (Table 1) in the simulated metagenome, the resulting filtered assemblies resulted in only a loss of 0.1 - 0.6% of annotated original reference genes (Table 2).

We next evaluated the effects of using similar approaches on metagenomic datasets. Similar to the simulated assemblies, the removal of highly connective sequences for all metagenomes and assemblers resulted in a loss of total assembly length and number of contigs (Table 2). In general, filtered assemblies were largely contained within unfiltered assemblies and comprised 51-88% of unfiltered assembly. The observed changes in metagenomic assemblies were difficult to evaluate as the source genomes to these datasets are unknown, and a loss in assembly length may actually be beneficial due to the elimination of artifactual contigs. To aid in this evaluation, we used the previously published set of rumen draft genomes which were constructed from *de novo* assembly efforts of the rumen metagenome [1]. Overall, we found that removal of highly connective sequences from the rumen dataset resulted in 1-3% loss of sequences which matched to draft reference genomes.

To further study the effects of highly connective sequences, we examined their incorporation into unfiltered assemblies. Overall, less than 1% of highly connective sequences on average were incorporated by any assembler, the maximum was 3-4% in the Velvet and Meta-IDBA assemblies of the human gut dataset (Table 1 and 3). Each assembled contig was divided into equal length bins (the size of bins was dependent on the total length of the contig) and examined for the presence of the previously identified highly connective sequences. We found that contigs, especially in assemblies from Velvet and Meta-IDBA, incorporated a larger fraction of these sequences at its ends relative to other binned positions (Figure 3). The SOAPdenovo assembler incorporated fewer of the highly connective sequences into its assembled contigs, none of these sequences in the simulated dataset were assembled and only 41 in the small soil dataset. For the human gut metagenome assemblies, millions of the highly connective sequences were incorporated into assembled contigs, comprising nearly 4% of all assembled sequences on Velvet contig ends (Supp. Fig. 1).

Identifying highly connective sequences

For the simulated metagenome, we could identify the original source of highly connective k-mers using available reference genomes. Many of these sequences originated from well-conserved housekeeping genes involved in protein synthesis, cell transport, and signaling. The top reference genes with perfect matches to highly connective k-mers which were present in the dataset a minimum of 50 times were identified (Table

4). To determine possible biological sources of highly connective sequences within real metagenomes, we compared the sequences shared between the soil, rumen, and human gut metagenomes. In total, 7,586 highly-connecting sequences (32-mers) were shared between the three soil, rumen, and human gut metagenomes. We identified the closest reference protein from the NCBI-nr database requiring complete sequence identity. Only 1,018 sequences (13%) matched existing reference proteins, and many of the annotated sequences matched multiple conserved protein sequences from multiple genomes. The top five proteins conserved in greater than 3 genomes are shown in Table 4, and largely encode for genes involved in protein biosynthesis, DNA metabolism, and biochemical cofactors (Table 5).

A potential cause of artificial high connectivity within metagenomes is the presence of high abundance sequences. Thus, we identified the subset of highly connective k-mers which were also present with an abundance of greater than 50 within each metagenome and their location in sequencing reads (Figure 2, dotted lines). These high abundance k-mers comprised a very small proportion of the identified highly connective sequences, less than 1% in the soils, 1.5% in the rumen, and 6.4% in the human gut metagenomes, but the position-specific biases of these sequences were very similar to the biases of the larger set of highly connective k-mers.

To identify consistent patterns within sequences causing position-specific biases, we examined the abundance of distribution unique 5-, 6-, 7-, 8-mers contained within the high abundance subset of each dataset's highly connective k-mers. STILL FINISHING

Discussion

Sequencing artifacts are present in highly connected sequences

Through assessing the connectivity of reads in several metagenomes, we identified a disproportionately large subset of reads which were connected together within an assembly graph, hereafter referred to as the "lump" in each metagenome. The simulated metagenome's lump comprised 5% of its total reads (Table 1). As this dataset contains no errors, this observed connectivity represents conserved sequences within a single genome or between multiple genomes (Table 3). In contrast, the highly connective lump within real metagenomic lumps comprised a significantly larger proportion of reads, ranging from 7% to 75% of the total reads (Table 1), suggesting that anomalous, non-biological connectivity may be present within these lumps. Interestingly, in the soil metagenomes, we observed that the amount of connectivity nearly doubled with less than a 5% increase of sequencing coverage. When sequencing coverage increased slightly from 4.7 to 5.6% in the medium and large soil metagenomes, the number of reads connected in the lump grew significantly from 15 million to 182 million. Given the very high diversity and very low coverage of these soils, the magnitude of the observed increases in connectivity seemed unlikely from biological sources, further supporting the presence of sequencing biases within these datasets.

If sequencing biases were present within these metagenomes, we would expect to observe that the metagenomic lumps would consist not only of artificial sequences but also sequences from reads which would be "preferentially attached" [16]. Consider that there is an original set of highly connecting "X" sequences in a lump. These sequences would recruit a number of connective "Y" reads into the lump. These recruited "Y" reads would then recruit more "Z" reads into the lump which would not necessarily connect to the original "X" reads. In error-free datasets, we would observe this preferential attachment phenomenon as a linear increase of lump size with increasing sequencing coverage. In the case of the presence of highly-connective sequencing biases, however, we'd observe that preferential attachment would cause dramatic increases in the number of recruited "Y" and "Z" reads, such is the observed case in the soil datasets.

To more rigorously demonstrate the presence of artifacts within our datasets, we considered that the sequencing of metagenomes is a random process and thus any position-specific bias within sequencing reads is unexpected and non-biological. For the metagenomes studied here, we used two approaches to

examine characteristics of connectivity correlated to specific positions within sequencing reads. First, we measured the local graph density (as defined in Methods) at specific positions within sequencing reads. Next, we identified the specific k-mers which were consistently present in highly dense regions of the assembly graph and evaluated their location within sequencing reads. When these approaches were applied to the simulated dataset, we observed no position-specific trends when assessing either local graph density (Figure 1) or highly connective k-mers (Figure 2, solid lines) as is consistent with the lack of sequencing errors and biases in this dataset. In all real metagenomes, however, we identified position-specific trends in measurements of both local graph density and the location of highly connective sequences, clearly indicating the presence of artificial sequences. Although present in all metagenomes, the direction of the bias varied between soil, rumen, and human gut datasets, particularly for the position-specific presence of highly connective sequences. It is possible that in higher coverage datasets, such as the rumen and human gut, there is a larger presence of indirectly preferentially attached reads which are connected to high coverage sequences of biological origins. This preferential attachment of such reads would result in increasing the number of total reads and consequently the decrease the total fraction of highly connective k-mers (Figure 2, y-axis). This trend is observed in the decreasing fractions of highly connective sequences at the 3' end of reads as sequencing coverage increased in the small, medium, to large soil metagenomes and in the soil, rumen, to human gut metagenomes (Figure 2).

Removal of highly connective sequences partitions lump without significant loss of reference genomes

As is apparent from conserved biological sources of high connectivity within the simulated metagenome, not all the observed connectivity within real metagenomes is artificial, and our approaches are limited in that they cannot differentiate between sequencing artifacts and sources of real biological connectivity. Regardless of the origin of highly connective sequences, we suspected that these sequences would challenge assemblers which rely on resolving the complex "lump" in the assembly graph. Indeed, very few highly connective sequences with abundances greater than 50 were incorporated into any assembly (Table 3) and those which were assembled were disproportionately placed at the ends of contigs (Figure 3), suggesting that assembly could not extend beyond these sequences. Although this trend was observed for all assemblers, it was more prevalent in the Velvet and Meta-IDBA assemblers, highlighting differences in assemblers.

Given that these sequences were found to have position-specific biases within reads and challenged multiple assemblers, we removed them from metagenomic lumps. The removal of these highly connective sequences had two key advantages: firstly, it removed artificial sequences which should not be assembled, and secondly, it resulted in the dissolution of the high connectivity within the metagenomic lump and allowed for the partitioning of all metagenomes. Having successfully divided metagenomic lumps, we compared the combined assembly of the partitioned sets of reads to the original lump dataset with several assemblers. For the partitioned reads, we were able to assemble subsets of reads in parallel, resulting in significantly reduced assembly requirements (time and memory) (Table 2). For the largest soil metagenome, containing over 500 million reads, we could not complete the Meta-IDBA assembly in less than 100 GB of memory. After partitioning, the assembly could be completed in less than 2 GB of memory. Additionally, we could use multiple k-mer length assemblies (demonstrated here with Velvet) and subsequently merge resulting assembled contigs, which was previously either impossible (due to memory requirements) or impractical (due to time). In this study, we used consistent parameters for each assembler to compare unfiltered and filtered assemblies. However, because the accuracy of an assembly relies on the characteristics of the underlying dataset, it is often beneficial to optimize assembly parameters and such optimization is much easier performed on assemblies which require less time and memory. As metagenome dataset sizes grow increasingly larger, the ability to efficiently analyze partitioned datasets and/or evaluate multiple assemblies will be increasingly important.

The advantages of removing highly connective sequences must be balanced against consequences to resulting assemblies. We compared several metagenome assemblies before and after the removal of these sequences. Comparing the simulated dataset's assemblies, the removal of highly connective sequences resulted in very little loss of annotated reference genes and a similar assembly compared to the unfiltered data (85% similarity), supporting the removal of these highly connective sequences for more efficient assembly. Unlike the simulated dataset, for the studied metagenomes, we compare filtered metagenomic assemblies against original assemblies of unknown quality. For the rumen metagenome, we performed a partial evaluation of the assemblies using draft reference genomes previously constructed from the assembly of high abundance k-mers. Similar to the simulated assemblies, we observed only a small loss of rumen reference genomes assembled (Table 2). In general, for all metagenomes, we observed 25% loss in assembly after removing highly connective sequences, much more than observed in assemblies of reference genes and genomes in the simulated and rumen datasets. Some of this loss is likely beneficial, resulting in the removal of sequencing artifacts; it is also possible that our approach removes sequences which can accurately be assembled but cannot be distinguished due to lack of reference genomes. However, without the removal of these sequences, many of the assemblies of the larger metagenomes would not be as practical.

Highly connective sequences are challenging to identify

We attempted to identify any biological characteristics of highly connective sequences. Among these sequences in the simulated dataset and those shared by all metagenomes, we identified a small fraction of these sequences (13% in simulated and ¡ 7% in metagenomes) which matched reference genes, mostly associated with housekeeping functions (Tables 4 and 5). Speculating that the remaining highly connective sequences originated from high abundance reads (both from biological sources of high connectivity or sequencing biases), we attempted to identify characteristics of the most abundant sequences. We found that this subset (sequences which were present greater than 50x) displayed similar trends for position-specific biases compared to their respective sets of highly connective sequences (Figure 2), indicating that they are comprised of sequencing biases. We attempted to identify signatures in the these sequences but found that they are largely random, making them difficult to efficiently identify and evaluate. Currently, we are evaluating a promising approach to improve the identification and removal of probable sequencing artifacts based on targeting high abundance sequencing.

Conclusion

As datasets from NGS technologies continue to increase in size, our ability to analyze this sequencing data must reevaluated. Here, we demonstrate the presence of sequencing artifacts within several metagenomic datasets that are a cause of non-biological connectivity within assembly graphs. We show that these abundant, highly connective sequences are sources of sequencing artifacts in metagenomes and are difficult to assemble, regardless of their origin. These sequences also add erroneous diversity and high coverage into metagenomes, significantly increasing memory requirements for de novo assembly. Our previous efforts to resolve components of complex metagenome assembly graphs have been bottlenecked by the presence of these highly-connective sequences. Here, we have developed approaches to identify these sequences and demonstrate that their removal results in comparable assemblies. The removal of such sequences results not only in eliminating artificial sequences but also allows for the partitioning of disconnected subgraphs, significantly reducing assembly requirements. Our analysis provides an understanding of the nature of these sequences and how their removal is an important first step for scalable *de novo* assembly. As datasets from NGS technologies continue to increase in size, this study highlights the importance of re-evaluating the nature of new sequencing data for both accurate and efficient downstream analysis approaches.

Methods

Metagenomic datasets

All datasets, with the exception of the agricultural soil metagenome, originate from previously published datasets. Rumen-associated sequences (Illumina) were randomly selected from the rumen metagenome available at ftp://ftp.jgi-psf.org/pub/rnd2/Cow_Rumen [1]. Human-gut associated sequences (Illumina) of samples MH0001 through MH0010 were obtained from ftp://public.genomics.org.cn/BGI/gutmeta/Raw_Reads [2]. The simulated high complexity, high coverage dataset was previously published (Pignatelli, 2011). All reads used in this study, with the exception of those in simulated metagenome, were quality-trimmed for Illumina's read segment quality control indicator, where a quality score of 2 indicates that all subsequent regions of the sequence should not be used. After quality-trimming, only reads with lengths greater than 30 bp were retained. All quality trimmed datasets, including the previously unpublished agricultural soil metagenome, are available on a public Amazon EC2 snapshot, XXX. (temporarily on scratch hpc:/mnt/scratch/howead/to-transfer-to-amazon/. The number of reads after quality-trimming is shown in Table 1 for each metagenome. The sequencing coverage of each metagenome was estimated as the fraction of reads which could be aligned to assembled contigs with lengths greater than 500 bp. For the coverage estimates, an assembly of each metagenome was performed using Velvet (v1.1.05) with the following parameters: K=33, exp cov=auto, cov cutoff=0, no scaffolding. Reads were aligned to assembled contigs with Bowtie (v0.12.7), allowing for a maximum of two mismatches.

Lightweight, compressible de Bruijn graph representation

We used a lightweight probabilistic de Bruijn graph representation to explore k-mer connectivity of the assembly graph (cite PNAS paper, software available at https://github.com/ctb/khmer). The de Bruijn graph stores k-mer nodes in Bloom filters and keeps edges between nodes implicitly, i.e. if two k-mer nodes exist with a k-1 overlap, then there is an edge between them. Bloom filters are a probabilistic set storage data structure with false positives but no false negatives, thus the size of the bloom filters were selected to be appropriate for each dataset and the memory available. For analyzing the graph connectivity of the studied datasets, we used 4 x 48e9 bit bloom filters. As metagenomic sequencing contains a mixture of multiple organisms, we could exploit the biological structure of the sequencing by partitioning the assembly graph into disconnected subgraphs that represent the original DNA sequence components. The set of the largest number of reads which were connected in the assembly graph is referred to above as a single, highly-connected lump. Data and examples of scripts used for this analysis are available on the Amazon EC2 public snapshot: data-in-paper/lumps and method-examples/0.partitioning-into-lump.

Local graph density and identifying highly-connected k-mers

We implemented a systematic traversal algorithm to identify highly connected components of the assembly graph. Waypoints were labeled to cover the graph such that they are a minimum distance of L apart. Originating from a waypoint, all k-mers (throughout the study k=32) were systematically and exhaustively traversed within a region that is the distance N. The local graph density was calculated as the number of X k-mers reachable within a distance of N nodes (k-mers) divided by the distance N. In this study, N was equal to 10 nodes within the assembly graph. For the largest metagenomes, the metahit and large soil datasets, local graph density was calculated on a representative subset of reads due to computational limitations. To identify specific highly-connective sequences within the lump assembly graphs, graph traversal to a distance of 40 nodes was attempted from marked waypoints. If more than 200 k-mers were found within this traversal were identified, all k-mers within this traversal were identified as candidates for highly connective sequences. If the same k-mers were consistently identified in other graph traversals, up to five times, the k-mer was flagged as a highly connective sequence.

Aligning theses k-mers to original sequencing reads, we identified the position-specific location of these k-mers. Data and examples of scripts used for this analysis are available on the Amazon EC2 public snapshot: data-in-paper/density-bias, data-in-paper/hc-kmer-bias, method-examples/1.density-analysis, method-examples/2.identifying-hc-kmers, and method-examples/3.hc-kmer-analysis.

To identify the sources of highly connective k-mers identified in the simulated metagenome, these sequences were aligned against the reference genes originating from the 112 source genomes using Bowtie (v0.12.7) requiring exact matches. Highly connective k-mers shared between all the metagenomes were also aligned against the NCBI non-redundant genome database (ftp://ftp.ncbi.nih.gov/blast/db, March, 1, 2011) using blastn [17] and requiring an exact match over the entire k-mer.

We also identified the subset of highly connective k-mers which were present at greater than 50 times within lumps. Data used for this analysis are available on the Amazon EC2 public snapshot: data-in-paper/lumps/HC-kmers/HA-HC-kmers and method-examples/4.abundant-hc-kmers. These high abundance, highly connective sequences were also aligned to sequencing reads to demonstrate position specific biases. We evaluated the existence of short k-mer (k=5-8) characteristics within high abundance, highly connective k-mers which did not have an exact match to the NCBI non-redundant database. Each identified 32-mer was broken up into shorter k-mers, and the abundance of various k-mers was calculated.

De novo Metagenomic Assembly

To evaluate the effects of these k-mers on assembly, sequencing reads were trimmed at the location where a highly connective k-mer could be aligned and the resulting assemblies are referred to as "filtered" assemblies. Untrimmed assemblies are referred to as "unfiltered" assemblies. *De novo* metagenomic assembly of reads within each unfiltered metagenomic lump was completed with Velvet (v1.1.02) with the following parameters: velveth -short -shortPaired (if applicable to the dataset) and velvetg -exp_cov auto -cov_cutoff 0 -scaffolding no [15]. For the small and medium soil, rumen, and simulated datasets, Velvet assemblies were performed at K=25-49, resulting contigs were dereplicated to remove contigs with 99% similarity using CD-HIT (v 4.5.6), and final contigs were merged with Minimus (Amos v3.1.0, [18]). For the largest soil and human gut metagenomes, assemblies were performed at only K=33 due to the size of the datasets and memory limitations. Additional assemblies were performed with meta-IDBA (v0.18) [9] : –mink 25 –maxk 50 –minCount 0 and with SOAPdenovo: -K 31 -p 8 max_rd_len=200 asm_flags=1 reverse_seq=0. After removal of highly connective k-mers in metagenomic lumps, each filtered lump was loaded into a new lightweight probabilistic de Bruijn graph representation to separate disconnected subgraphs. Sequences in multiple subgraphs were grouped together such that assembly could be performed in parallel on each group of sequences. Identical assembly parameters and methods as described above were performed on these partitioned sequences. Unfiltered and filtered assemblies were compared using the total number of contigs, total assembly length, and maximum contig size. Additional, the coverage of each assembly was calculated through estimating the average base pair coverage of the BLAST alignment of each assembly to one another (E-value ¿ 10-5) or, in the case of the simulated and rumen assemblies, to reference genomes. The simulated and rumen reference genomes were previously published in [1] and [8], respectively. Data used for this analysis are available on the Amazon EC2 public snapshot: /data-in-paper/assembly*.

We examined the location of the identified high abundant, highly connecting k-mers within assembled contigs. The location of these k-mers within assembled unfiltered contigs was examined by dividing each contig into 100 equally-sized regions. The fraction of highly-connecting k-mers which aligned exactly to each region was calculated for each metagenome. Data and examples of scripts used for this analysis are available on the Amazon EC2 public snapshot: method-examples/5.hc-kmer-contigs/.

References

1. Hess M, Sczyrba A, Egan R, Kim... T (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science .

2. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65.

3. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, et al. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. Science 335: 587–90.

4. Mackelprang R, Waldrop M, DeAngelis... K (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature .

5. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95: 315–27.

6. Pop M (2009) Genome assembly reborn: recent computational challenges. Briefings in bioinformatics 10: 354.

7. Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, et al. (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS ONE 7: e31386.

8. Pignatelli... M (2011) Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PLoS ONE .

9. Peng Y, Leung H, Yiu S, Chin F (2011) Meta-idba: a de novo assembler for metagenomic data. Bioinformatics 27: i94.

10. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, et al. (2006) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nature Methods 4: 495–500.

11. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. The ISME Journal 3: 1314–7.

12. Keegan KP, Trimble WL, Wilkening J, Wilke A, Harrison T, et al. (2012) A platform-independent method for detecting errors in metagenomic sequencing data: Drisee. PLoS Comput Biol 8: e1002541.

13. Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 11: 187.

14. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, et al. (2009) Abyss: a parallel assembler for short read sequence data. Genome Res 19: 1117–23.

15. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res 18: 821–9.

16. Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286: 509.

17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–10.

18. Sommer DD, Delcher AL, Salzberg SL, Pop M (2007) Minimus: a fast, lightweight genome assembler. Bmc Bioinformatics 8: 64.

	Total Reads (millions)	% Reads Mapped to Assembly	Largest Unfiltered Partition "Lump" (millions of reads)	Total Highly Connective 32-mers	Total 32-mers in Unfiltered Lump	% Highly Connective	Nodes with Density > 20
Small Soil	50.0	1.4	3 .0 (7%)	6,429,673	84,906,521	8%	50%
Medium Soil	100.0	4.7	15.0 (15%)	33,266,397	326,454,473	10%	37%
Large Soil	520.3	5.6	182.2 (35%)	230,353,299	2,198,140,432	10%	40%
Rumen	50.0	32.0	10.3 (21%)	25,400,121	201,532,081	13%	22%
Human Gut	350.0	3.5	263 (75%)	136,594,783	860,627,857	16%	28%
Simulated	9.2	14.8	0.5 (5%)	364,816	11,592,284	3%	17%

Table 1. The original size and proportion of highly connective 32-mers in the largest subset of partitioned reads ("lump") in several medium to high complexity metagenomes. Read coverage was estimated with the number of aligned sequencing reads to Velvet-assembled contigs (K=33). The dominant lump, or largest disconnected component of each metagenome assembly graph, was found to contain highly connecting k-mers responsible for high local graph density.

Figures and Tables

Velvet Assembler

	Coverage of Unfiltered by Filtered Assembly	Coverage of Normalized Filtered by Unfiltered Assembly	Coverage of Reference Genes by Unfiltered	Coverage of Reference Genes by Filtered	Unfiltered Assembly Statitics (No. of Contigs/Assembly Length(bp)/Max Contig Size (bp)	Normalized Assembly Statitics (No. of Contigs/Assembly Length(bp)/Max Contig Size (bp)	Unfiltered Assembly Requirements (Memory, GB / Time (hours)
Small Soil	74.5%	98.6%	-	-	25,470 / 16,269,879 / 118,753	17,636 / 10,578,908 / 13,246	5 / 4
Medium Soil	75.4%	98.1%	-	-	113,613 / 81,660,678 / 57,856	79,654 / 54,424,264 / 23,663	18 / 21
Large Soil	50.8%	86.3%	-	-	554,825 / 306,899,884 / 41,217	290,018 / 159,960,062 / 41,423	33 / 12*
Rumen	75.1%	98.3%	17.2%	14.6%	92,044 / 74,813,072 / 182,003	72,705 / 49,518,627 / 34,683	11 / 14
Human Gut	79.5%	88.5%	-	-	543,331 / 234,686,983 / 85,596	203,299 / 181,934,800 / 145,740	76 / 8*
Simulated	84.6%	98.3%	4.5%	3.9%	11,204 / 6,506,248 / 5,151	9,859 / 5,463,067 / 6,605	< 1 / < 1

Meta-IDBA Assembler

	Coverage of Unfiltered by Filtered Assembly	Coverage of Normalized Filtered by Unfiltered Assembly	Coverage of Reference Genes by Unfiltered	Coverage of Reference Genes by Filtered	Unfiltered Assembly Statitics (No. of Contigs/Assembly Length(bp)/Max Contig Size (bp)	Normalized Assembly Statitics (No. of Contigs/Assembly Length(bp)/Max Contig Size (bp)	Unfiltered Assembly Requirements (Memory, GB / Time (hours)
Small Soil	75.6%	93.9%	-	-	15,739 / 9,133,564 / 37,738	12,513 / 7,012,036 / 17,048	< 1 / < 1
Medium Soil	67.5%	94.5%	-	-	76,269 / 45,844,975 / 37,738	52,978 / 30,040,031 / 18,882	2 / 2
Large Soil	N/A	N/A	-	-	N/A	395,122 / 228,857,098 / 37,738	> 116 / incomplete
Rumen	70.4%	94.4%	15.5%	13.0%	60,330 / 47,984,619 / 54,407	48,940 / 33,276,502 / 22,083	12 / 3
Human Gut	74.0%	96.5%	-	-	173,432 / 211,067,996 / 106,503	132,614 / 142,139,101 / 85,539	58 / 15
Simulated	86.5%	93.4%	3.8%	3.5%	8,707 / 4,698,575 / 5,113	7,726 / 4,078,947 / 3,845	< 1 / <1

SOAPdenovo Assembler

	Coverage of Unfiltered by Filtered Assembly	Coverage of Normalized Filtered by Unfiltered Assembly	Coverage of Reference Genes by Unfiltered	Coverage of Reference Genes by Filtered	Unfiltered Assembly Statitics (No. of Contigs/Assembly Length(bp)/Max Contig Size (bp)	Normalized Assembly Statitics (No. of Contigs/Assembly Length(bp)/Max Contig Size (bp)	Unfiltered Assembly Requirements (Memory, GB / Time (hours)
Small Soil	86.6%	95.8%	-	-	14,275 / 7,100,052 / 37,720	12,801 / 6,343,110 / 13,246	3 / < 1
Medium Soil	82.2%	95.7%	-	-	66,640 / 33,321,411 / 28,695	56,023 / 27,880,293 / 15,721	10 / < 1
Large Soil	78.7%	94.2%	-	-	412,059 / 215,614,765 / 32,514	334,319 / 171,718,154 / 41,423	48 / 11
Rumen	84.7%	97.3%	14.7%	13.4%	62,896 / 40,792,029 / 22,875	55,975 / 34,540,861 / 19,044	5 / < 1
Human Gut	84.9%	98.5%	-	-	190,963 / 171,502,574 / 57,803	161,795 / 139,686,630 / 56,034	35 / 5
Simulated	93.2%	96.1%	2.5%	2.4%	6,322 / 2,940,509 / 3,786	6,029 / 2,821,631 / 3,764	< 1 / < 1

Table 2. Comparison of unfiltered and filtered assemblies of various metagenome lumps using Velvet, SOAPdenovo, and Meta-IDBA assemblers. Assemblies were aligned to each other, and coverage was estimated (columns 1-2). Simulated and rumen assemblies were aligned to available reference genes/genomes (columns 3-4). Total number of contigs, assembly length, and maximum contig size was estimated for each assembly, as well as memory and time requirements of unfiltered assembly (columnes 5-7). Filtered assemblies required less than 2 GB of memory. Velvet assemblies of the unfiltered human gut and large soil datasets (marked as *) could only be completed with K=33 due to computational limitations. The Meta-IDBA assembly of the large soil metagenome could not be completed in less than 100 GB.

	Velvet	SOAPdenovo	Meta-IDBA
Small Soil	0	41	8,717
Medium Soil	32,328	852	23,881
Large Soil	653,071	279,519	N/A
Rumen	45,721	14,858	33,046
Human Gut	4,661,447	1,749,387	5,528,054
Simulated	5,118	0	5,480

Table 3. Total number of abundant (greater than 50x), highly connective sequences incorporated into unfiltered assemblies (percentage of total highly connective sequences).

	Number of Hits to Unique Genes in 112 Reference Genomes
ABC transporter-like protein	306
Methyl-accepting chemotaxis sensory transducer	210
ABC transporter	173
Elongation factor Tu	94
Chemotaxis sensory transducer	51
ABC transporter ATP-binding protein	44
Diguanylate cyclase/phosphodiesterase	36
ATPase	36
S-adenosyl-L-homocysteine hydrolase	36
Adensylhomocysteine and downstream NAD bindin	36
Ketol-acid reductoisomerase	34
S-adenosylmethionine synthetase	34
Elongation factor G	34
ABC transporter ATPase	33

Table 4. Annotation of highly-connecting sequences from the simulated metagenome with most hits to conserved genes within the 112 reference genomes [8].

	Number of NCBI Genomes
Translation elongation factor/GTP-binding protein LepA	11
S-adenosylmethionine synthetase	8
Aspartyl-tRNA synthetase	8
Malate dehydrogenase	7
V-type H(+)-translocating pyrophosphatase	6
Acyl-CoA synthetase	6
NAD synthetase / Glutamine amidotransferase chain of NAD synthetase	5
Ribonucleotide reductase of class II	4
Ribityllumazine synthase	4
Heavy metal translocating P-type ATPase, copA	3
GyrB	3
Glutamine amidotransferase chain of NAD synthetase	3
ChaC family protein	3

Table 5. Annotation of highly-connecting sequences to conserved nucleotide sequences originating from 3 or more reference genomes. Shown are protein annotations whose nucleotide sequences matched 3 or more highly-connecting sequences shared in the three soil, rumen, and human gut metagenomes.

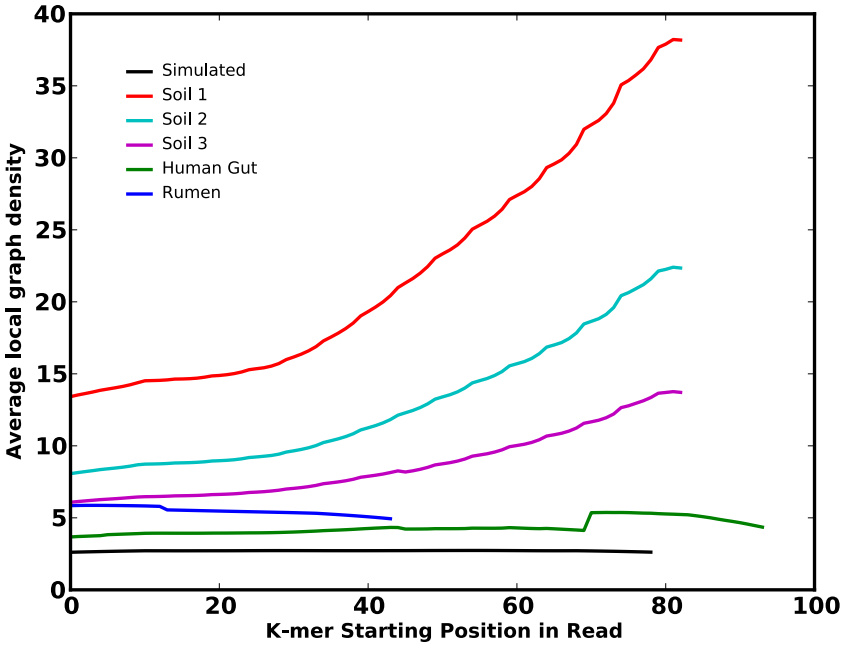


Figure 1. The extent to which average local graph density varies by read position is shown for the lump of various datasets.

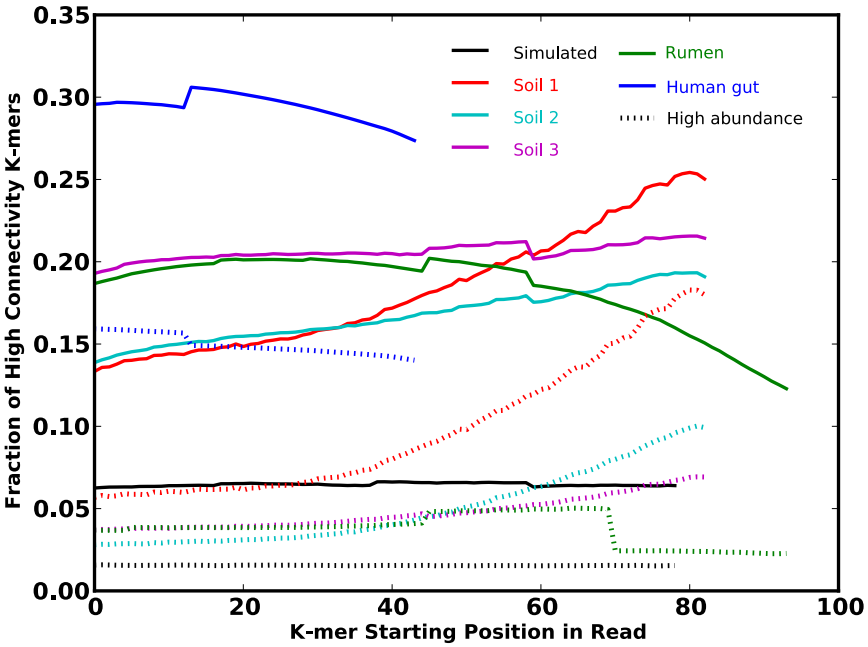
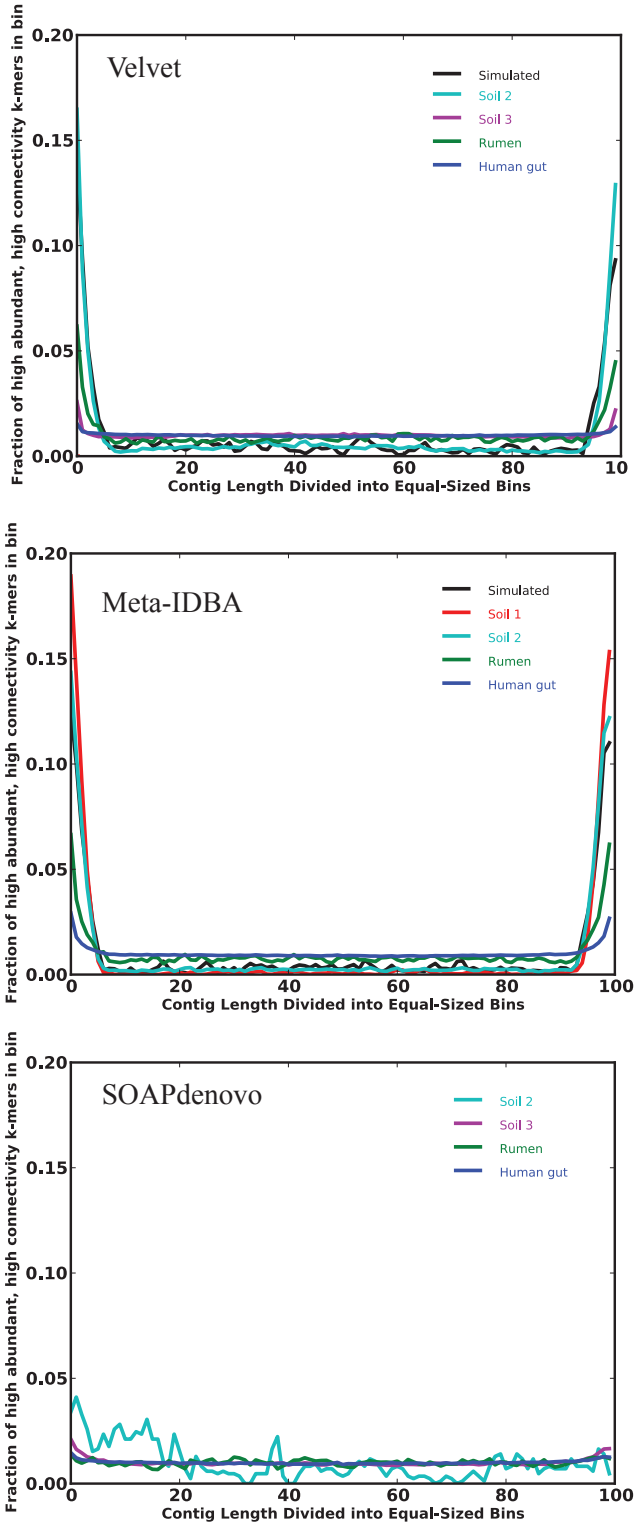
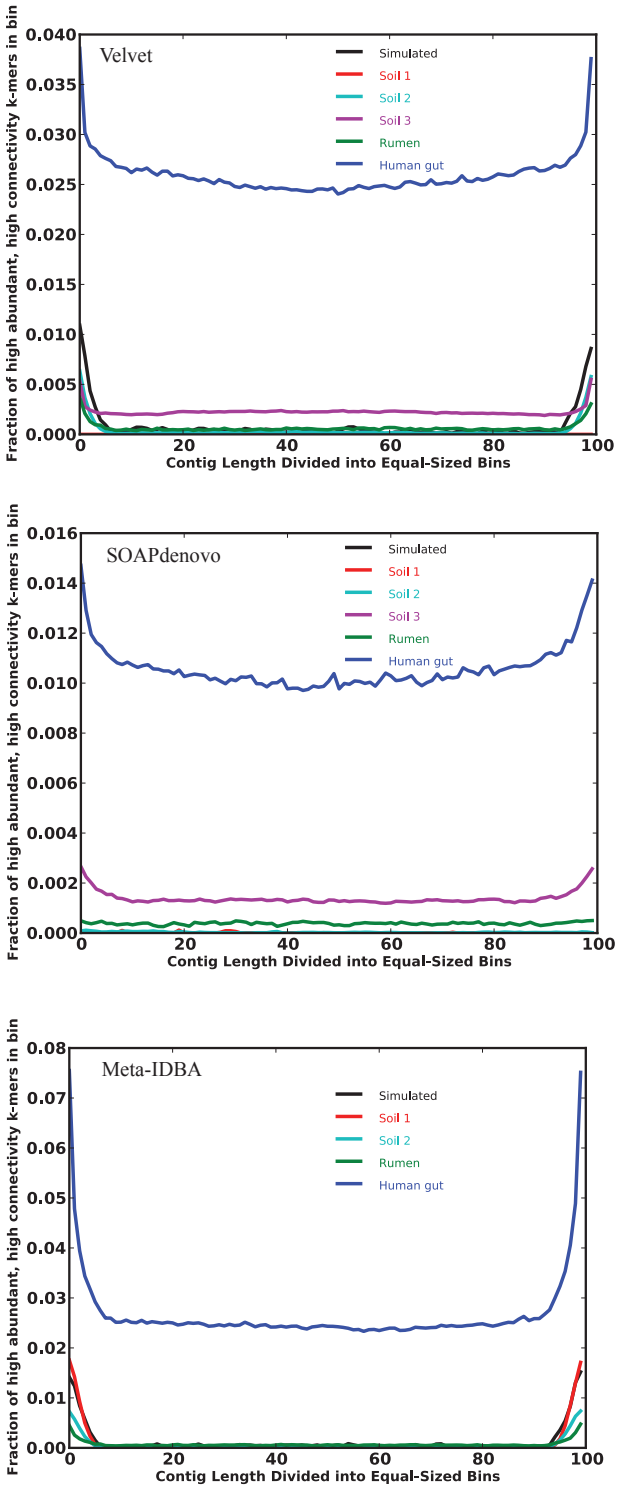


Figure 2. The extent to which highly-connecting k-mers (solid lines) and the subset of highly abundant (greater than 50) k-mers (dashed lines) are present at specific positions within sequencing reads for various metagenomes.



5in

Figure 3. When incorporated into an assembly, abundant (greater than 50 times), highly-connecting sequences (k-mers) were disproportionately present at the ends of contigs. The total fraction of highly-connecting k-mers which are incorporated into each contig binned region.



5in

Figure 4. When incorporated into an assembly, abundant (greater than 50 times), highly-connecting sequences (k-mers) were disproportionately present at the ends of contigs. The total fraction of all k-mers which are identified as high abundant, high connective sequences and incorporated into each contig binned region is shown.