Connectivity Analysis of Metagenomic Data

ACH, JP, RCK, RM, JJ, JMT, CTB

January 12, 2012

1 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 [3, 14]. The main bottleneck for these metagenomic studies is the lack of effective strategies to annotate and predict gene functions from the enormous sequencing datasets that are now being generated [4, 6, 10, 16]. De novo metagenomic assembly has been used as a solution to reduce dataset size by collapsing numerous short reads into fewer contigs and providing longer sequences containing multiple genes and operons [8, 13]. Furthermore, because it does not rely on the availability of reference genomes, assembly also produces novel contigs allowing for comparisons of sequences within and between metagenomes [7, 15] or annotations of unknown genomes [3]. The success of de novo metagenomic assembly relies on the ability to store information about the connectivity of sequencing reads within an assembly graph. Thus, its application to large and complex metagenomic datasets is limited by both the amount of sequencing and the availability of computational memory.

To deal with these challenges, new metagenome-specific de novo assemblers use various "divide and conquer" approaches to break apart components of the assembly graph [11] (cite metaVelvet). These assemblers take advantage of the fact that environmental populations contain multiple genomes which have been sampled at varying depths corresponding to their natural abundance. Read coverage (the extent to which sequencing reads contribute to assembled contigs) and/or graph connectivity are used to break apart and simplify metagenomic assembly graphs.

The ability to resolve components of an assembly graph depends on accurately distinguishing variable-coverage genome sequences from sequencing errors and bias. The presence of sequencing biases and errors have been demonstrated in Illumina sequences [2, 5, 9] but very little is known about their effects on assembly graph properties and the resulting assemblies. With large amounts of sequencing (as is needed for complex metagenomes), increases in the number of real biological sequences are accompanied by increases in sequencing errors and biases. In this study, we analyzed the ability to resolve disconnected components of several metagenomic assembly graphs. In doing so, we identified highly-connecting sequences in several metagenomes which we demonstrate originate from sequencing artifacts. We evaluated the effects of removing these sequences on metagenomic assembly and discuss how this approach ultimately enables the assembly of large, complex metagenomes.

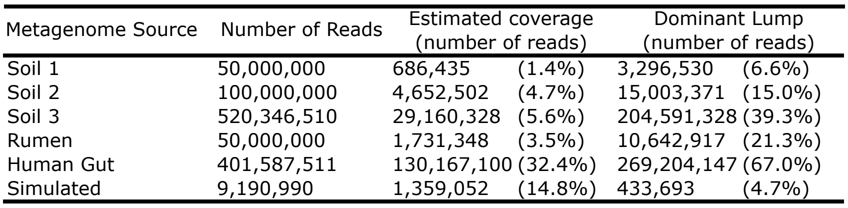
2 Results

2.1 Connectivity analysis of metagenome datasets

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

We selected datasets from three diverse, medium to high complexity metagenomes from the human gut [14], cow rumen [3], and agricultural soil (unpublished) (Table 1). For comparison, we also included one simulated metagenome (error-free) of a high complexity, high coverage (˜10x) microbial community [12]. To study the effects of increased sequencing, we also included two additional subsets of the agricultural soil metagenome containing 50 million and 100 million reads each. We estimate the read coverage of each metagenome dataset by aligning sequencing reads to their corresponding assembled contigs. For the human gut and cow rumen metagenomes, we estimate 32.4% and 3.5%, respectively. For the small, medium, and large soil metagenomes, the coverage was estimated at 1.4%, 4.7%, and 5.60%, respectively.

The connectivity of reads within the assembly graph of each dataset was evaluated within a de Bruijn graph representation (see Methods). Reads contributing to disconnected portions of the assembly graph were separated. In each dataset, we identified a dominant, highly-connected set of sequencing reads which we referred to as the ”lump” (Figure 1, Table 1). In the simulated dataset, this lump consisted of 5% of the reads. In the studied metagenomes, the size of the lump ranged from 7% (in the smallest soil metagenome) to 67% (in the human gut metagenome) of the total reads. For the three soil datasets of increasing size, the size of the lump was disproportionately larger than the increase in sequencing (Figure 1, Table 1). As the number of reads increased by 2-fold and 5-fold, the size of the lump increased by 5-fold and



Metagenome Source	Number of Reads	Estimated coverage (number of reads)	Dominant Lump (number of reads)
Soil 1	50,000,000	686,435 (1.4%)	3,296,530 (6.6%)
Soil 2	100,000,000	4,652,502 (4.7%)	15,003,371 (15.0%)
Soil 3	520,346,510	29,160,328 (5.6%)	204,591,328 (39.3%)
Rumen	50,000,000	1,731,348 (3.5%)	10,642,917 (21.3%)
Human Gut	401,587,511	130,167,100 (32.4%)	269,204,147 (67.0%)
Simulated	9,190,990	1,359,052 (14.8%)	433,693 (4.7%)

Figure 1: The connectivity of sequencing reads from medium to high complexity metagenomes from the soil, rumen, and human-gut were analyzed. Read coverage was estimated by aligning sequencing reads to Velvet-assembled contigs (K=33). A dominant lump, or largest disconnected component of each metagenome assembly graph, was identified in each metagenome.

14-fold, respectively.

@AC - I chose to just leave this out but FYI... In a smaller subset of the human gut metagenome (50 million reads), the relative size of the lump was similar and comprised over 75% of the sequencing reads (data not shown).

@Note that filtering out highly abundant k-mers effects - CTB, need to analyze...need to ask question about this

2.1.2 Characterizing assembly graphs in metagenomic and simulated lumps

To assess the degree of connectivity of sequences within the lump, we measured the local graph density of reads within the de Bruijn assembly graph. The local graph density is defined here as the number of k-mers, or sequences of length k, found within a distance of N divided by N. Examining the connectivity of all 112 bacterial genomes used for the simulated dataset, we found that fewer than 2% of the nodes within the assembly graph had an average graph density greater than 20. The simulated short reads generated from these genomes resulted in a lump made up of 433,693 reads (4.7% of total reads). Within this lump, we calculated the local graph density and found that 17% of the nodes had an average graph density greater than 20. Within lumps resulting from the metagenomic datasets, average graph density greater than 20 ranged from 21 to 50%.

We next determined the extent to which graph density varied by position along sequenced reads by measuring the average local graph density within ten steps of every k-mer by position in a read (Figure 2). For the simulated dataset, the average local density was stable for all positions along reads. In contrast, metagenome datasets had increased average local graph densities which

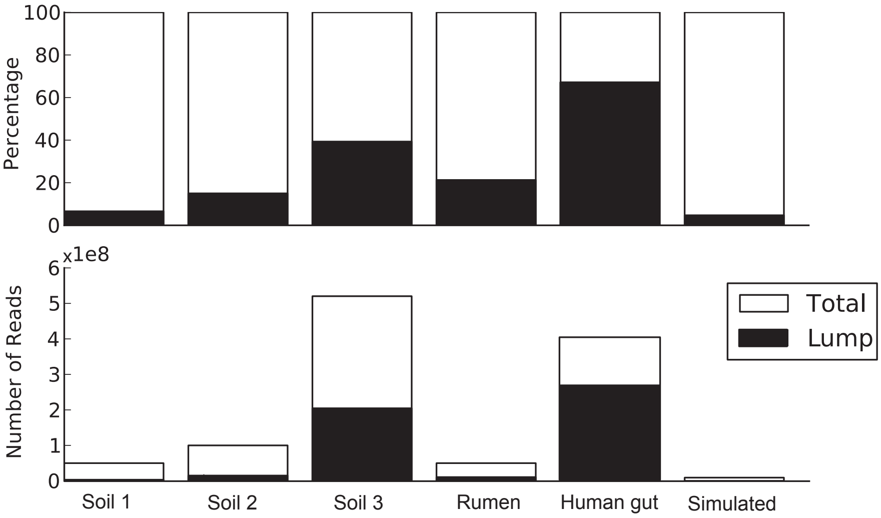
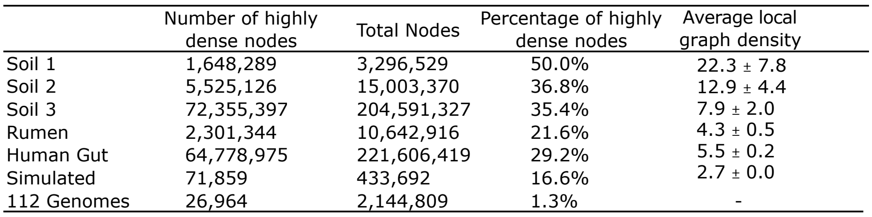


Figure 2: The lump within each metagenome was dominant, regardless of the origin of metagenome.



	Number of highly dense nodes	Total Nodes	Percentage of highly dense nodes	Average local graph density
Soil 1	1,648,289	3,296,529	50.0%	22.3 ± 7.8
Soil 2	5,525,126	15,003,370	36.8%	12.9 ± 4.4
Soil 3	72,355,397	204,591,327	35.4%	7.9 ± 2.0
Rumen	2,301,344	10,642,916	21.6%	4.3 ± 0.5
Human Gut	64,778,975	221,606,419	29.2%	5.5 ± 0.2
Simulated	71,859	433,692	16.6%	2.7 ± 0.0
112 Genomes	26,964	2,144,809	1.3%	-

Figure 3: The number of highly dense k-mer nodes with local assembly graph densities greater than 20 (N=100) for various metagenomes and a mixture of 112 reference genomes [12]. For the human gut metagenome, partial traversal of the graph (82%) was used to calculate the number of highly dense nodes due to computational limitations. The local graph density (N=10) for all nodes in each metagenome was calculated for each k-mer position within reads (see Figure X), and the average density for all k-mer positions is shown.

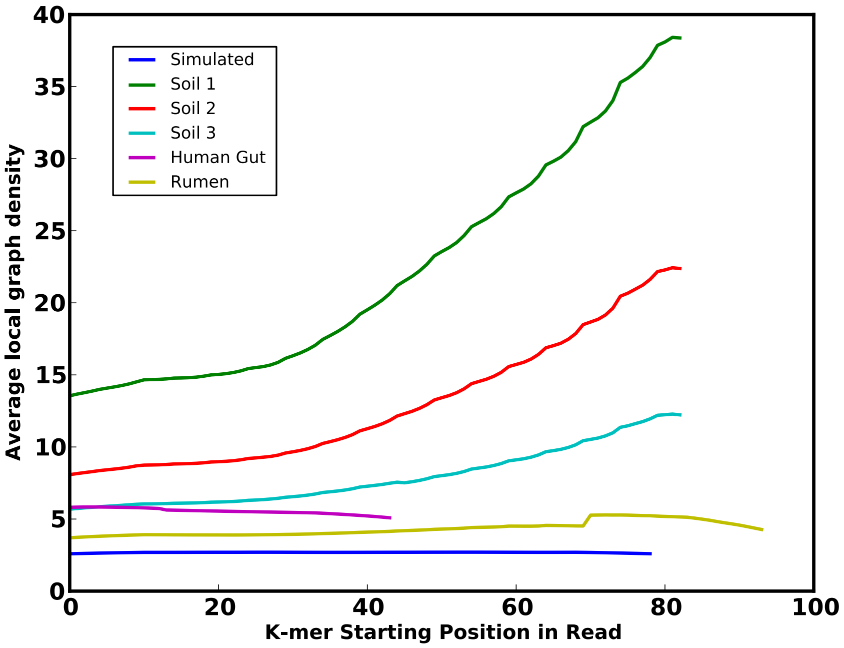


Figure 4: The extent to which average local graph density varies by read position is shown for the lump of various datasets. Unlike the graph density of the simulated metagenome lump, average graph density of sequences in metagenomic lumps varied significantly by position along the read.

varied depending on read position. The reads in the lumps in the soil metagenomes had the most variability of local graph densities by position, with standard deviations ranging from 25% - 35% of the average density, and increases in density occurred at the 3' end end of the read.

2.2 Identification of highly-connecting k-mers and their effects on assembly

2.2.1 Characteristics of highly-connecting sequences in simulated and metagenomic lump reads

We next identified sequences within each lump which were causes of high-connectivity using a systematic assembly graph traveral algorithm (see Methods). For sequences within the lumps of metagenomic datasets, we found that 6 to 8% of the unique k-mers were highly-connective, with the exception of the smallest soil metagenome lump which contained less than 1% of highly-connective k-mers. In contrast, the lump of the simulated metagenome contained fewer of these highly connective k-mers, 3% of total unique k-mers (Table X).

Having identified these highly-connective sequences, we assessed the extent to which these k-mers were found at specific positions along a sequencing read. In the simulated metagenome

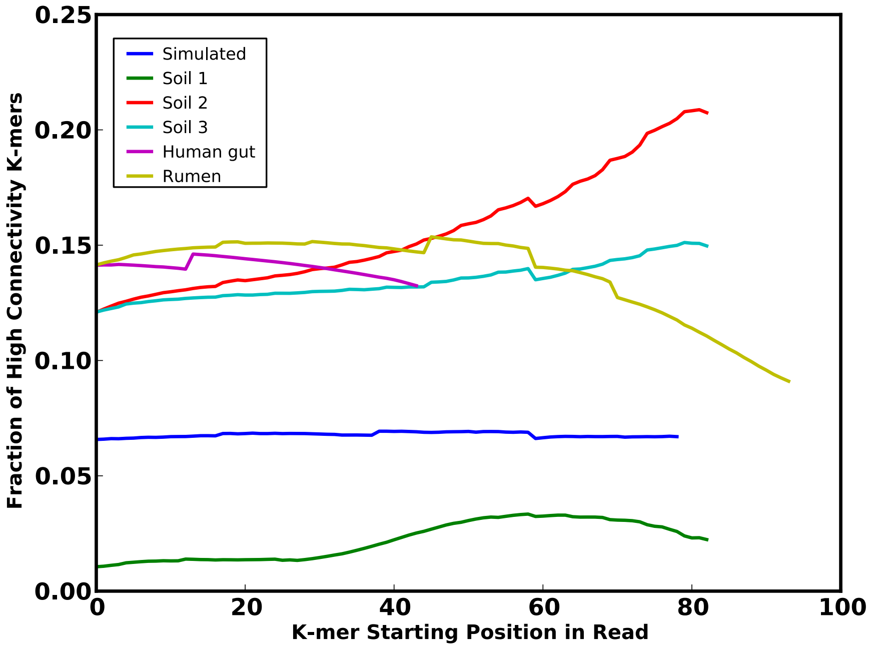
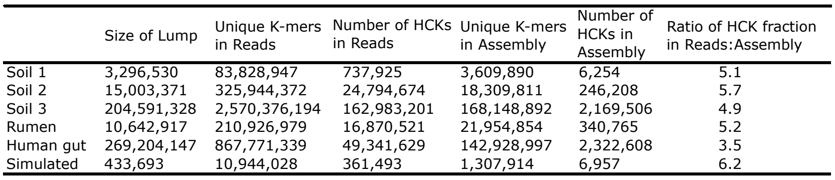


Figure 5: Highly-connecting sequences (k-mers) were present with position-specific bias within sequencing reads in metagenomic lumps.

lump, the highly-connecting k-mers did not exhibit any bias with regard to position within the read. In other words, these k-mers had equal probability of being located at any position along the read. In the case of metagenome lumps, however, highly connective k-mers were more prevalent at position-specific locations along the read (Figure X). For the three soil metagenomes, the fraction of total k-mers which were identified as highly-connective increased at the 3'-end of the read. In the human-gut and rumen metagenome lumps, the fraction of these k-mers decreased at the 3'-end of the read.

2.2.2 Characteristics of highly-connecting sequences in simulated lump assemblies

We were interested in the incorporation of the highly-connective k-mers in the final assembly of the lumps. For all lumps, we found that the proportion of unique knot-causing k-mers in sequencing reads were significantly more than that of assembled contigs (longer than 500 bp). In the simulated lump, there was an 6-fold enrichment of these sequences in the reads compared to the final assembly. In metagenomic lumps, there were 3 to 6 times more highly-connecting k-mers in sequencing reads than in assembled contigs (Table X). Examining the position of these highly-connecting k-mers within assembled contigs, we found that these sequences were being disproportionately placed on the ends of contigs (Figure X) in every metagenome assembly



	Size of Lump	Unique K-mers in Reads	Number of HCKs in Reads	Unique K-mers in Assembly	Number of HCKs in Assembly	Ratio of HCK fraction in Reads:Assembly
Soil 1	3,296,530	83,828,947	737,925	3,609,890	6,254	5.1
Soil 2	15,003,371	325,944,372	24,794,674	18,309,811	246,208	5.7
Soil 3	204,591,328	2,570,376,194	162,983,201	168,148,892	2,169,506	4.9
Rumen	10,642,917	210,926,979	16,870,521	21,954,854	340,765	5.2
Human gut	269,204,147	867,771,339	49,341,629	142,928,997	2,322,608	3.5
Simulated	433,693	10,944,028	361,493	1,307,914	6,957	6.2

Figure 6: Highly-connecting k-mers were more highly enriched in sequencing reads compared to assembled contigs. All metagenomes (except for the soil 3 and human gut) were assembled with Velvet (as described in Methods) with K=25, 27, 29, 31, 33 and merged. Enrichment ratios for soil 3 and human gut metagenomes calcuated from assemblies at only at K=33 due to computational limitations.

studied.

@Effects of removing these highly connecting lumps on the the lump, discuss breaking up lump by removing these guys...

2.2.3 Effects of removing highly-connecting k-mers on the simulated lump assembly

To study the effects of the highly-connecting k-mers on assembly, we performed assemblies of the simulated metagenome lump with and without removal of the identified highly-connective, knot-causing k-mers from reads. Overall, the number of contigs (longer than 500 bp) and length of assembly was improved by filtering knot-causing k-mers. The unfiltered assembly contained 1,844 contigs (maximum size = 3,787 bp) with 1,365,436 bp, while the filtered assembly contained 2,301 contigs (maximum size = 5,826 bp) with 1,733,645 bp (Table X).

From the resulting assemblies, we predicted open reading frames (ORFs) from assembled contigs with lengths greater than 500 bp (see Methods). The unfiltered and filtered assembly contained 2,401 and 3,049 ORFs, respectively (Table X). To evaluate the accuracy of each assembly, we compared the predicted ORFs to the 112 genomes from which the simulated dataset orginated. In the unfiltered lump assembly, 2,395 (99.8%) ORFs matched the reference genomes while in the filtered lump assembly, more assembled ORFs, 3,037 (99.6%), matched the reference genomes. We found that the unfiltered and knot-filtered assemblies shared a total of 2,018 ORFs, with the remaining 383 and 1,031 ORFs unique to the unfiltered and knot-filtered assemblies, respectively. Comparing the constitutive k-mers making up the unfiltered and knot-filtered assemblies (see Methods), we estimate that these assemblies are approximately

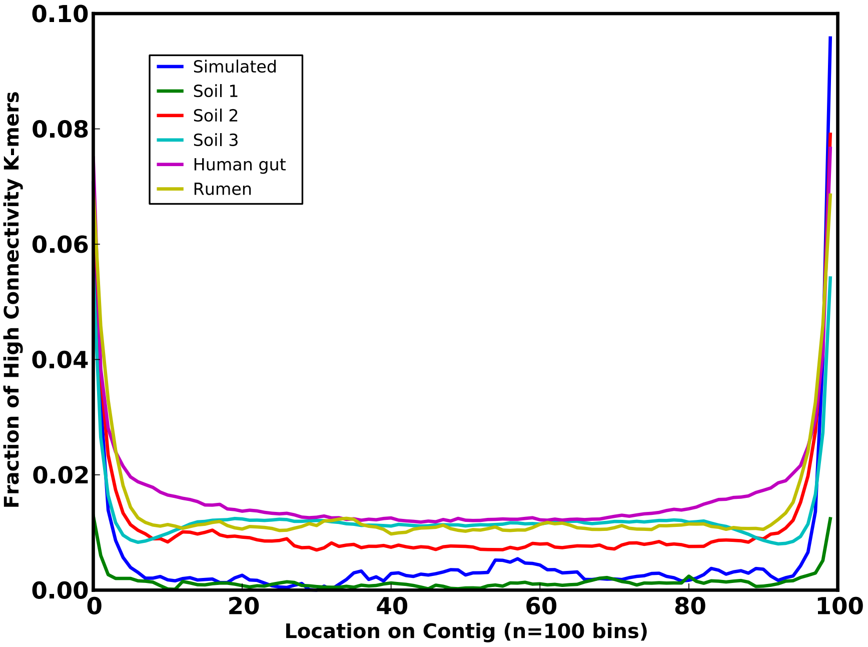
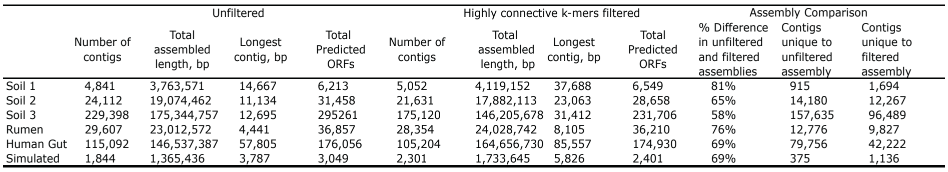


Figure 7: When incorporated into an assembly, highly-connecting sequences (k-mers) were disproportionately present at the ends of contigs.



	Unfiltered				Highly connective k-mers filtered				Assembly Comparison		
	Number of contigs	Total assembled length, bp	Longest contig, bp	Total Predicted ORFs	Number of contigs	Total assembled length, bp	Longest contig, bp	Total Predicted ORFs	% Difference in unfiltered and filtered assemblies	Contigs unique to unfiltered assembly	Contigs unique to filtered assembly
Soil 1	4,841	3,763,571	14,667	6,213	5,052	4,119,152	37,688	6,549	81%	915	1,694
Soil 2	24,112	19,074,462	11,134	31,458	21,631	17,882,113	23,063	28,658	65%	14,180	12,267
Soil 3	229,398	175,344,757	12,695	295261	175,120	146,205,678	31,412	231,706	58%	157,635	96,489
Rumen	29,607	23,012,572	4,441	36,857	28,354	24,028,742	8,105	36,210	76%	12,776	9,827
Human Gut	115,092	146,537,387	57,805	176,056	105,204	164,656,730	85,557	174,930	69%	79,756	42,222
Simulated	1,844	1,365,436	3,787	3,049	2,301	1,733,645	5,826	2,401	69%	375	1,136

Figure 8: Comparison of unfiltered and filtered (removal of highly connecting k-mers) assemblies of various metagenome lumps.

31% different (Table X). Among the 2,018 shared ORFs between the unfiltered and filtered assemblies, the predicted ORFs from the filtered assembly had slightly longer alignment lengths compared to those of the unfiltered assembly (Figure X).

@@are the unique orfs shorter? - small signal, not sure if good enough if strong enough to include here, decided to leave this out @ Note that I left out the edge effects of ORFs...I thought it was just too funky and vague. I think the annotations against the references make for the stronger argument.

2.2.4 Characteristics of highly-connecting sequences in metagenomic lump assemblies

@Having trouble getting this sized so its not teenie, troubleshoot later

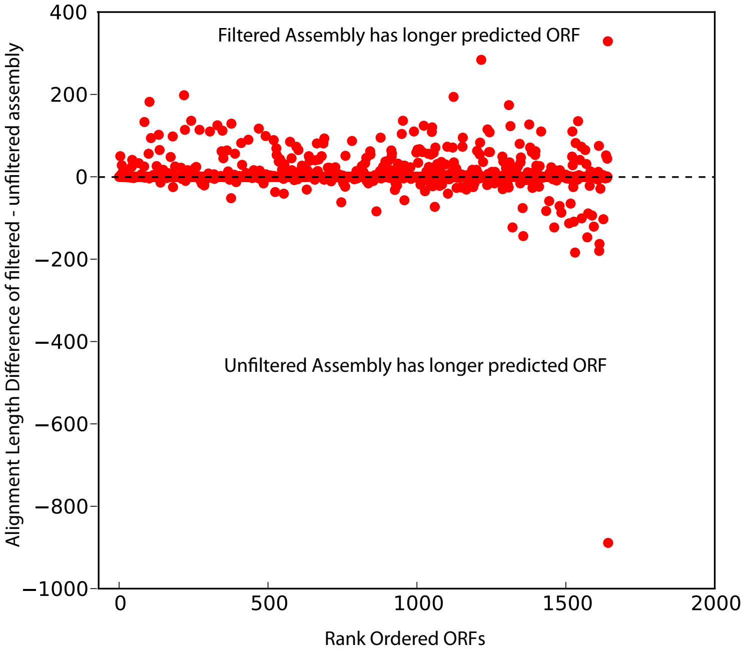


Figure 9: Alignment lengths of ORFs shared between unfiltered and filtered assemblies (removal of highly connective k-mers) of the simulated metagenome are compared. Difference in alignment length of predicted ORFs to contigs is shown (filtered assembly - unfiltered assembly alignment length), average=3.7 bp, stdev=35.8 bp, median=0 bp.

2.2.5 Effects of removing highly-connecting k-mers on metagenome lumps

@@ put in other data, comparing assemblies, rumen, etc

3 Discussion

3.1 Characteristics of the lump did not appear to be biological in origin.

We were able to separate several metagenomic assembly graphs into millions of disconnected assembly subgraphs representing sequences originating from different genomes. The largest of these subgraphs, called the lump, contained a disproportionate number of reads relative to other subgraphs. Initially, we considered that this lump consisted mainly of connecting sequences which were conserved across multiple genomes (i.e. 16S rRNA, ITS regions). However, efforts to remove conserved genes from datasets did not significantly break apart the lump (data not shown). Furthermore, the large size of the lump within metagenomic data compared to that of a simulated metagenomes (i.e., 75% of reads in the human gut metagenome vs. 5% in simulated) suggested that some connectivity within this lump was not biological. Comparing soil metagenomes of increasingly larger sizes, we observed that with increasing sequencing, there was a supra-linear increase in the size of the lump and thus increase in graph connectivity (Figure X). These results combined indicated that possible spurrious connectivity was present in metagenomic lumps, and we proceeded to further analyze the connectivity these sequences.

3.2 Position-specific biases indicate that sequencing artifacts are present in the lump.

We first assessed the degree of connectivity within the lump by measuring the local graph density of metagenomic lumps. For a mixture of genomes, the local graph density was measured to be low, less than 2% of nodes with graph density greater than 20. For the lump of connected reads of a simulated Illumina sequencing dataset of these genomes, the graph density was unsurprisingly larger with 17% of nodes having a graph density greater than 20. For the metagenomic lumps, however, the degree of connectivity was consistently larger than the simulated dataset, with an average of 35% of metagenomic nodes having a graph density of greater than 20. In addition to increased connectivity, we observed varying local graph densities with respect to the position within a read in all metagenomes (Figure X). This position-specific bias, not observed in the simulated dataset, clearly suggested spurrious connectivity among metagenomic sequences.

To further explore the sources of this spurrious connectivity, we identified the highly-connecting sequences within each lump which were likely causes of the lump itself. Similarly to the observed position-specific bias of local graph density along the read, the presence of highly-connecting sequences also had a biased presence in locations of sequencing reads (Figure X). Given that shotgun sequencing is randomly generated, these observed position-specific biases (local graph density and presence of highly connecting k-mers) strongly suggest that some highly-connecting sequences are not biological and are the result of sequencing artifacts.

3.3 Comparing position-specific trends in various metagenomes indicates preferential attachment.

Similar position-specific trends were observed for the local graph density and the fraction of highly-connecting sequences for all metagenomes studied (Figures X and X). The 3'-end biases of the two largest soil metagenomes was the most pronounced. Moreover, for these soil metagenome lumps, the size of the lumps and associated position-specific trends increased at a greater rate than the amount of sequencing, suggesting the presence of spurrious connectivity. We suspect that this connectivity is the result of an effect referred to as ”preferential attachment” [1]. In this case, highly connecting ”X” sequences in a lump would recruit a number of connecting ”Y” reads into the lump. As more sequences are added, these ”Y” reads, which do not necessarily have to be highly-connective, recruit more ”Z” reads into the lump resulting in increasingly larger lump size. For soil metagenomes, where sequencing coverage is relatively low and diversity is high (5.6% coverage for the largest soil metagenome), increased sequencing would cause preferential attachment of ”X”, ”Y”, and ”Z” reads resulting in increasingly larger lump sizes. For metagenomes with less complexity, like that of the human gut and cow rumen (32.4% and 3.5% coverage), the number of the ”Y” and ”Z” reads which are lump-associated but not highly-connective would increase (because of the increased sequencing coverage) at a greater rate than than ”X” reads. This would effectively introduce a greater proportion of sequences in the lump which would not be identified as highly-connective and result in an overall decrease in the total fraction of these sequences. This trend was observed in our metagenome lumps where the total number of unique highly-connecting sequences (6-8/

3.4 Removing highly-connecting sequences improves assembly overall.

Although some of the highly-connecting sequences in metagenomic lumps are sequencing artifacts (given their position-specific bias), it is apparent that not all of these sequences are non-biological as these sequences are also present in the error-free simulated dataset (3% of the unique k-mers identified as highly-connective). Regardless of the origin of these sequences, we were interested in their incorporation into resulting assemblies. For all datasets studied, we observed that these sequences were under-represented in the final assembly compared to their presence in original sequencing reads (Table X). Moreover, when these sequences were incorporated into assembly, they tended to begin or end contigs (Figure X). These results suggest that, overall, the assemblers are challenged by characteristics of these sequences regardless of their biological or non-biological origins and that the removal of these sequences would have little effect on the final assembly.

We evaluated the effects of removing these sequences with the simulated dataset. This was an ideal case study because it contains no sequencing errors or biases and could be validated by the original reference genomes. We compared the assembly of the simulated dataset before and after filtering out highly-connected sequences to evaluate the biological effects of these sequences on assembly. We found that the unfiltered and filtered assemblies were quite different based on constituent k-mer composition (31

@Next add in rumen validation where we have the genomes from the assembly generated, and stats for assembly for the rest of the metagenomes, need to run metasim...

4 Methods

4.1 Metagenomic datasets

All datasets, with the exception of the agricultural soil metagenome, were 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. Human-gut associated sequences (Illumina) of samples MH0001 through MH0010 were obtained from ftp://public.genomics.org.cn/BG The agricultural soil metagenome was from the sequencing (Illumina) of Iowa corn soil and is currently unpublished. All reads used in this study 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 reads used in this study are available at X. The number of reads after quality-trimming is shown in Table 1 for each metagenome. The simulated high complexity, high coverage dataset was previously published (Pignatelli, 2011). Coverage of each metagenome was estimated by aligning trimmed sequencing reads to assembled contigs with lengths greater than 500 bp. For coverage estimates, the 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. For datasets larger than 50 million reads, metagenomes were partitioned with the methods described in (cite PNAS paper) prior to assembly to overcome computational memory limitations for assembly. Reads were aligned to assembled contigs with Bowtie (v0.12.7), allowing for a maximum of two mismatches.

4.2 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). 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 the size of the dataset and the memory available. For analyzing the graph connectivity of the studied datasets, we used 4 x 48e9 bit bloom filters for soil, rumen, and human gut datasets, and 4 x 1e9 bit bloom filters for the simulated dataset. 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 referre to above as a single, highly-connected lump.

4.3 Local graph density and identifying highly-connected k-mers

We implemented a systematic traversal algorithm to 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 are 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 N divided by the distance N. Thus, the local graph density of a linear sequence would be 2, and additional branches or repeats would increase graph density. To evaluate the extent of connectivity within each metagenomic lump, we calculated the number of nodes (K=32) with a local graph density of greater than 20 when N=100 and determined these nodes to be "highly dense" (Table X). We also evaluated the degree to which graph density varied by position along a read. For each k-mer in a read, the local graph density within N=10 nodes was calculated. This average local graph density by k-mer position for all reads was subsequently calculated (Figure X).

@brain fart - fix highly dense - there's better phrases I'm sure

We examined the position of these highly-connective k-mers in reads contributing to the lump. Each sequence in the lump was broken into its constituent k-mers and evaluated to be highly-connective or not highly-connective. The total fraction of k-mers within each dataset lump which were identified as highly-connective and are shown in Figure X.

4.4 De Novo Metagenomic Assembly

De novo metagenomic assembly reads within each metagenomic lump were completed with Velvet (v1.1.02, cite Zerbino) with the following parameters: velveth -short -shortPaired (if applicable to the dataset) and velvetg -exp_cov auto -cov_cutoff 0 -scaffolding no. For soil 1, soil 2, rumen, and simulated metagenomes, assemblies were performed at K=25, 27, 29, 31, and 33 and merged with Minimus (Amos v3.1.0, cite). For soil 3 and human gut metagenomes, assemblies were performed at only K=33 due to the size of the datasets.

To study the effects of highly connecting sequences on metagenomic assembly, these k-mers were filtered from reads by truncating the reads at the location of these sequences and the remaining reads assembled as described above.

@AC...Assemblies were also performed with ABYSS (v1.2.0, cite) with the following parameters: ABYSS -k 33 - do we want to anything with this?

4.5 Comparison of shared constituent k-mers

To calculate the number of shared unique k-mers between assembly A and assembly B, constituent k-mers of contigs from assembly A were loaded into bloom filters (4 x 1e9 bits). Subsequently, the constituent k-mers from the assembly B were queried against the assembly A k-mers. The number of shared unique k-mers is dependent on which assembly is initially loaded into the bloom filters. Thus, each comparison was completed twice, once with the unfiltered assembly and once with the filtered assembly initially loaded into the bloom filter. Assembly similarity was determined by the lowest fraction of shared unique k-mers between these two comparisons (Figure X).

4.6 Identifying properties of highly-connecting k-mers

The enrichment of knot-causing k-mers in unfiltered reads was studied by identifying the fraction of unique k-mers in unfiltered sequencing reads and in their resulting assembled contigs. The ratio of these k-mer fractions (unfiltered reads/assembled contigs) estimates the enrichment of knot-causing k-mers in the reads.

To further understand the contribution of the knot-containing contigs to unfiltered and filtered assembly differences, we calculated the difference in constituent unique k-mers between knot-containing contigs and the filtered contigs (resulting from assembly of knot-filtered reads) using Bloom filters as described above. The fraction of total knot-causing k-mers between the two assemblies was calculated by dividing the number of different k-mers in knot-containing contigs by the total number of different k-mers in unfiltered and filtered contigs.

The location of knots in unfiltered contigs was also studied. Contigs containing knot-causing k-mers were divided into 100 equally-sized regions. For each contig, the total number of knot-causing k-mers and total number of k-mers was calculated. For each dataset, the total fraction of knot-causing k-mers in each region for all contigs was calculated and is shown in Figure X.

The presence of knot-causing k-mers in ORFs was examined. Fraggenescan (v1.1.15, cite) with the following parameters: -complete=0 -training=454_10 was used to identify ORFs in unfiltered contigs. We defined the "edge" of an ORF within a contig to be between 32 bp (k-mer size used in our de Bruijn graph representation) outside of of an ORF to within 16 bp (k/2) inside the ORF. The remaining internal ORF bases were defined as inside the ORF, and external bases were defined as outside the ORF. For each base within a contig, we determined if it was the initial base of a knot causing k-mer and if it was located inside, outside, or at the edge of an ORF. The distribution of knot-causing bases (k-mers) between the inside, outside, and edge were then compared to the total distribution of all bases.

To evaluate assemblies of the simulated dataset, ORFs and contigs were aligned to the original 112 genomes used to generate the simulated metagenome using BLAST (v2.2.25). For specific contig regions (the ORFs which were edge-enriched for knot-containing sequences), mismatches between assembled sequences and reference genomes were identified to evaluate the accuracy of assembly. To determine the effects of removing knot-causing sequences, the alignment lengths of contigs from unfiltered and knot-filtered assemblies which shared identical top blast alignments were compared.

5 Conclusion

Removing these lumps is computationally very useful, enabling MetaIDBA as well as scaling approaches

References

[1] A.L Barabási and R Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[2] Olivier Harismendy, Pauline C Ng, Robert L Strausberg, Xiaoyun Wang, Timothy B Stockwell, Karen Y Beeson, Nicholas J Schork, Sarah S Murray, Eric J Topol, Samuel Levy, and Kelly A Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 10(3):R32, Jan 2009.

[3] M Hess, A Sczyrba, R Egan, and T Kim.... Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, Jan 2011.

[4] K Hoff, T Lingner, and P Meinicke.... Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Research*, Jan 2009.

[5] S Hoffmann, C Otto, S Kurtz, and C Sharma.... Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Computational ...*, Jan 2009.

[6] V Kunin, A Copeland, A Lapidus, K Mavromatis, and P Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*, 72(4):557–578, Dec 2008.

[7] W Li. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *Bmc Bioinformatics*, 10(1):359, 2009.

[8] Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–27, Jun 2010.

[9] K Nakamura, T Oshima, and T Morimoto.... Sequence-specific error profile of illumina sequencers. *Nucleic Acids ...*, Jan 2011.

[10] H Noguchi and J Park.... Metagene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research*, Jan 2006.

[11] Y Peng, H Leung, SM Yiu, and F.Y.L Chin. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94, 2011.

[12] M Pignatelli.... Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS ONE*, Jan 2011.

[13] M Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354, 2009.

[14] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage,

Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, MetaHIT Consortium, Peer Bork, S Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, Mar 2010.

[15] PD Schloss and J Handelsman. A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *Bmc Bioinformatics*, 9(1):34, 2008.

[16] Yuan Zhang and Yanni Sun. Metadomain: A profile hmm-based protein domain classification tool for short sequence. *Pacific Symposium on Biocomputing*, 2012.