Connectivity Analysis of Metagenomic Data

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

March 1, 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, 19]. 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 produces novel contigs which can be compared within and between metagenomes [7, 16]. With sufficient sequencing, novel draft genomes can also be obtained from these assemblies [3]. The application of de novo assembly to large and complex metagenomic datasets relies on the ability to store information about the connectivity of sequencing reads within an assembly graph. Thus, metagenomic de novo assembly is limited by both the diversity of the metagenome and the availability of computational memory.

Recently developed metagenome-specific de novo assemblers use various "divide and conquer" approaches to break apart components of the assembly graph [11] (cite metaVelvet), taking advantage of the fact that environmental populations contain multiple genomes which have been sampled at varying depths corresponding to their natural abundance. Read coverage and/or graph connectivity are used to break apart and simplify metagenomic assembly graph into subcomponents which are subsequently assembled separately. In order to resolve these components of an assembly graph, variable-coverage sequences must accurately be distinguished from 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 effects of errors and sequencing biases on assembly graphs, and our ability to resolve disconnected components of several complex metagenomic assembly graphs. We identified highly-connecting sequences in several metagenomes which we demonstrate originate, at least partially, 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) . We also included a simulated, error-free metagenome of a high complexity, high coverage (˜10x) microbial community [12] (Table 1). To study the effects of increased sequencing, we included two additional subsets of the agricultural soil metagenome containing 50 million and 100 million reads each. The coverage of each metagenome was estimated by aligning sequencing reads to corresponding assembled contigs (see Methods). The human gut had the highest coverage, estimated at 32.4%. The coverages of the cow rumen, small soil, medium soil, and large soil metagenomes were significantly less, 3.5%, 1.4%, 4.7%, and 5.60%, respectively (Table 1).

The connectivity of reads within the assembly graph of each dataset was evaluated within a de Bruijn graph representation (see Methods). As an initial step, reads contributing to disconnected portions of the assembly graph were separated. For each metagenome, regardless of origin, 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 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 increases in sequencing. As the number of reads increased by 2-fold and 5-fold, the size of the lump

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%)

Table 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.

increased by 5-fold and 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 people always ask if these are highly repetitive - make some comments some where @Note that filtering out highly abundant k-mers effects - CTB, need to analyze...need to ask question about this

2.1.2 Characterizing the dominant lump within the assembly graph

We next assessed the degree of connectivity of sequences within the identified lump of each metagenome. The local graph density of reads within the de Bruijn assembly graph of each metagenome's lump was calculated (Table 2). The local graph density is defined as the number of k-mers, or sequences of length k, found within a distance of N divided by N. For a mixture of 112 complete bacterial genomes, 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 4.7% of total reads. Within this lump, 17% of the nodes had an average graph density greater than 20. Similarly, the average graph density in lumps from the environmental metagenome datasets 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

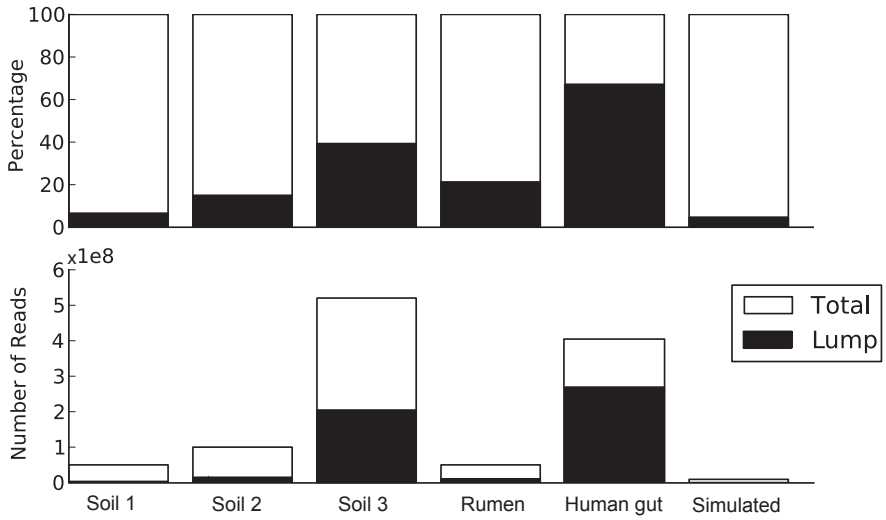

Figure 1: The presence of a dominant lump was identified for all metagenomes studied, regardless of origin of the metagenome.

	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%	-

Table 2: 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.

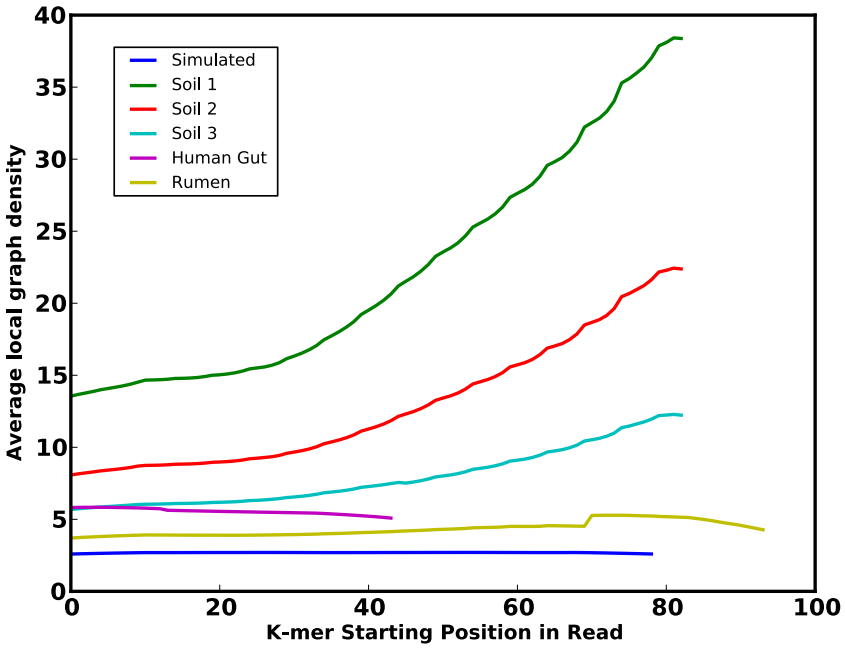

Figure 2: The extent to which average local graph density varies by read position is shown for the lump of various datasets. Unlike the graph density of the simulated metagenome lump, average graph density of sequences in metagenomic lumps varied significantly by position along the read.

densities towards the 3'-end 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.

@is this a real number that people usually report,

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 probable causes of high local graph density using a systematic assembly graph traveral algorithm (see Methods). For environmental metagenomes, we identified 6 to 8% of unique k-mers within each lump were sources of high connectivity, with the exception of the smallest soil metagenome lump which contained less than 1% (Table 3). In contrast, the lump of the simulated metagenome contained fewer of these highly connective k-mers, approximately 3% of total unique k-mers.

We also compared the highly-connecting sequences shared between the soil, rumen, and

Annotation from NCBI-nr database	# of highly-connective sequences with annotation
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

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

human gut metagenomes. In total, 7,586 highly-connecting sequences were shared between the three soil, rumen, and human gut metagenomes. For the smallest metagenome (soil 1), these shared sequences made up approximately a very small fraction of the total identified highly-connecting sequences (7,586/737,925, 1.0%). We identified the closest reference proteins from the NCBI-nr database, requiring complete sequence identity. Among the 7,586 highly-connecting sequences, only 1,018 sequences (13%) matched existing reference proteins. Additionally, many of the annotated sequences matched multiple conserved protein sequences from multiple genomes. In total, the 1,018 sequences matched a total of 118 genomes. The top five annotated proteins conserved in greater than 3 genomes encode for genes involved in protein biosynthesis, DNA metabolism, and biochemical cofactors (Supplementary Table X).

Having identified these highly-connective sequences, we next assessed the extent to which these k-mers were found at specific positions along a sequencing read (Figure 3). In the simulated metagenome 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, highly connective k-mers were more prevalent at position-specific locations along the read. 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 same read region in the human-gut and rumen metagenome lumps, the fraction of these k-mers decreased relative to other regions of the reads.

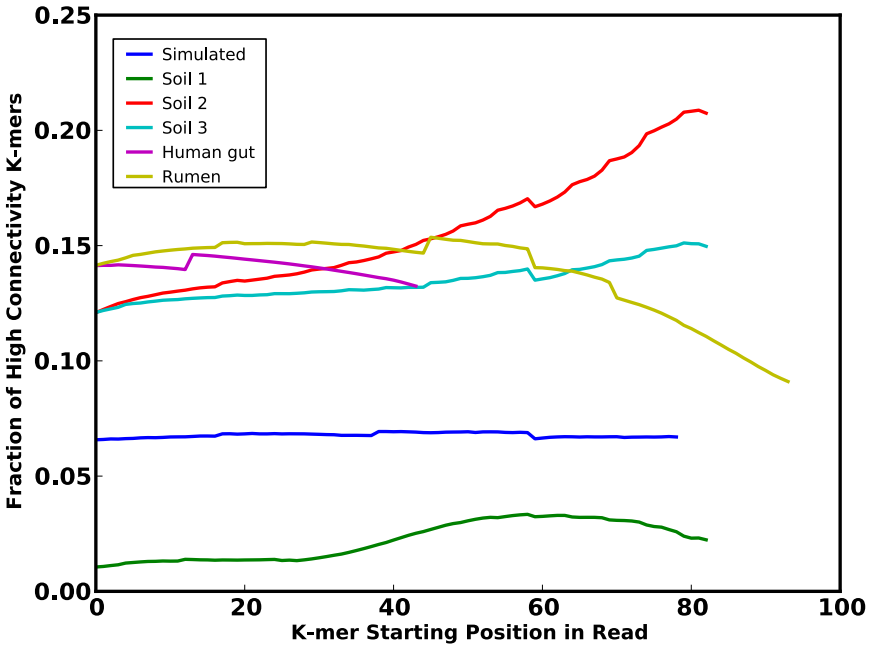

Figure 4: The extent to which highly-connecting k-mers are present at specific positions within sequencing reads for various metagenomes.

2.2.2 Characteristics of highly-connecting sequences in simulated and metagenomic lump assemblies

We were interested in the incorporation of the highly-connective k-mers into the final assembly of the sequences within the lumps. For all lumps, we found that the there was an enriched presence of these k-mers within reads compared to the final assembly (Table 3). In the simulated lump, there was an 6-fold enrichment of these sequences in the reads compared to the assembled contigs. Likewise, in metagenomic lumps, there was a 3 to 6 time enrichment of highly-connective k-mers in reads over the assemblies. 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 in every metagenome assembly studied, including the simulated metagenome (Figure 4) .

	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 (0.9%)	3,609,890	6,254	5.1
Soil 2	15,003,371	325,944,372	24,794,674 (8.0%)	18,309,811	246,208	5.7
Soil 3	204,591,328	2,570,376,194	162,983,201 (6.3%)	168,148,892	2,169,506	4.9
Rumen	10,642,917	210,926,979	16,870,521 (8.0%)	21,954,854	340,765	5.2
Human gut	269,204,147	867,771,339	49,341,629 (6.0%)	142,928,997	2,322,608	3.5
Simulated	433,693	10,944,028	361,493 (3.3%)	1,307,914	6,957	6.2

Table 3: 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. Enrichment ratios for soil 3 and human gut metagenomes calcuated from assemblies at only at K=33 due to computational limitations.

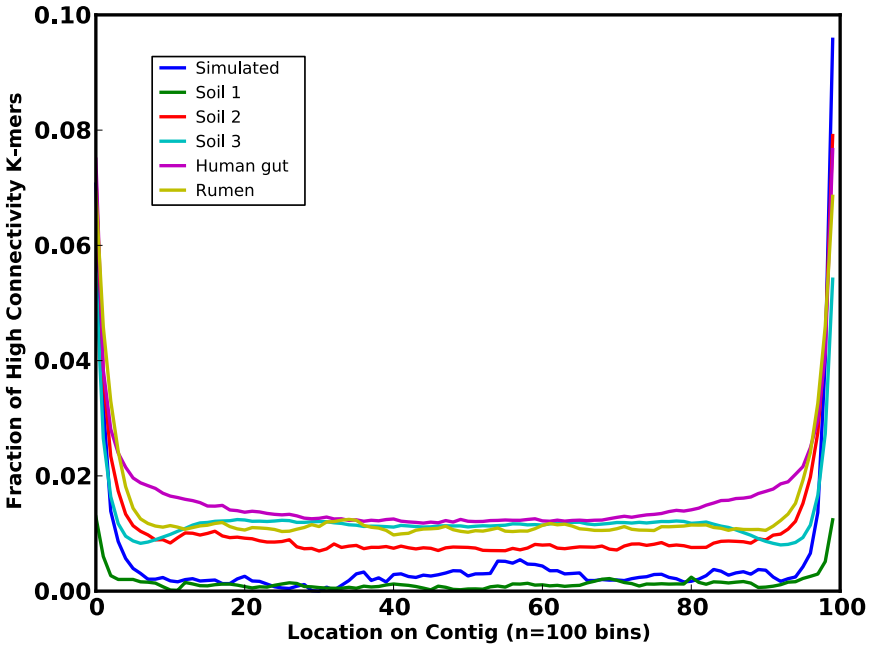

Figure 5: When incorporated into an assembly, highly-connecting sequences (k-mers) were disproportionately present at the ends of contigs.

2.3 Effects of removal of highly-connecting k-mers on assembly

2.3.1 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 k-mers from reads. As a result of removing these sequences, several assembly statistics improved for the simulated data. Whereas the unfiltered assembly contained 1,844 contigs (maximum size = 3,787 bp) with 1,365,436 bp, the filtered assembly contained 2,301 contigs (maximum size = 5,826 bp) with 1,733,645 bp (Table 4). Overall, we found the unfiltered and filtered assemblies to be quite different. Based on comparing constituent k-mer composition, we estimated that the assemblies are approximately 69% similar. Furthermore, of the 1,844 contigs assembled from the unfiltered assembly, we estimate that 375 (20%) are unique. In the filtered assembly, we estimate that 1,136 of the 2,301 contigs (49%) are unique.

We next identified protein coding regions from the simulated assemblies by predicting open reading frames (ORFs) from assembled contigs with lengths greater than 500 bp. The filtered assembly contained more ORFs compared to the unfiltered assembly, 3,049 filtered ORFs and 2,401 unfiltered ORFs (Table 4). To evaluate the accuracy of each assembly, we compared the predicted ORFs to the proteins within the original 112 genomes which were used to generate the simulated metagenome. In the unfiltered 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. Overall, the unfiltered and filtered assemblies shared shared a total of 2,018 ORFs which varied in lengths between unfiltered and filtered assemblies. Comparing the alignment length of predicted ORFs to reference proteins, we found that the filtered assembly had slightly longer ORF alignment lengths compared to the unfiltered assembly (460 ORFs), and few ORFs had longer alignments in the unfiltered assembly (402 ORFs). The large majority of shared ORFs had no difference in alignment lengths (1,156 ORFs) (Figure 5).

@@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. We can put it back in though if you like it.

	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	% similarity 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 (19%)	1,694 (34%)
Soil 2	24,112	19,074,462	11,134	31,458	21,631	17,882,113	23,063	28,658	65%	14,180 (59%)	12,267 (57%)
Soil 3	229,398	175,344,757	12,695	295,261	175,120	146,205,678	31,412	231,706	58%	157,635 (69%)	96,489 (55%)
Rumen	29,607	23,012,572	4,441	36,857	28,354	24,028,742	8,105	36,210	76%	12,776 (43%)	9,827 (35%)
Human Gut	115,092	146,537,387	57,805	176,056	105,204	164,656,730	85,557	174,930	69%	79,756 (69%)	42,222 (40%)
Simulated	1,844	1,365,436	3,787	3,049	2,301	1,733,645	5,826	2,401	69%	375 (20%)	1,136 (49%)

Table 4: Comparison of unfiltered and filtered (removal of highly connecting k-mers) assemblies of various metagenome lumps.

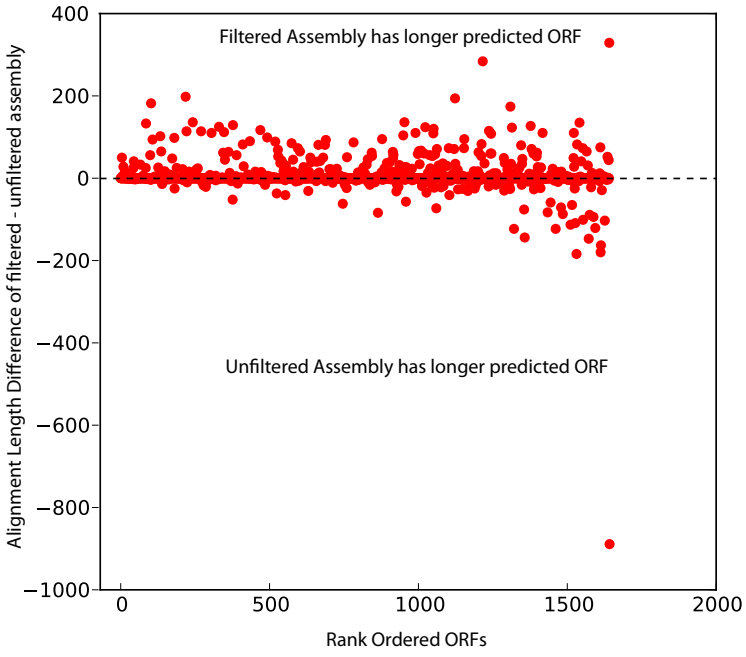

Figure 6: Alignment lengths of ORFs shared between unfiltered and filtered assemblies (removal of highly connective k-mers) of the simulated metagenome are compared. Difference in alignment length of predicted ORFs to contigs is shown (filtered assembly - unfiltered assembly alignment length).

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

Comparing the unfiltered and filtered assemblies of the rumen, human gut, and soil metagenomes, we observed that removing highly-connecting k-mers consistently resulted in increased maximum contig lengths (ranging from 48% to 157%). The effects of filtering highly-connecting k-mers on the total length of assembly and number of predicted ORFs varied (Table 4). For the smallest soil, human gut, and rumen metagenomes, filtered assemblies had a longer total length. For the soil metagenomes containing 100 million and 500 million reads, the total length of assembly decreased 6 and 17%, respectively. The total number of predicted ORFs decreased for filtered assemblies (ranging from 1 to 22% fewer ORFs), with the exception of the smallest soil metagenome which had an increase in number of predicted ORFs. Comparing the constituent k-mers between unfiltered and filtered assemblies, we observed that the assemblies of the environmental metagenomes, like the simulated assemblies, were also quite different. Similarity, based on constituent k-mer comparisons, ranged from 58% to 81% for unfiltered and filtered assemblies. A large proportion of contigs were also unique to either the unfiltered or filtered assembly.

We compared the assemblies of the ruman metagenomic lump to 15 draft genomes predicted from the published assembly of the full rumen metagenome [3]. We compared the number of predicted ORFs from unfiltered and filtered assemblies which matched these genomes. For the unfiltered assembly, 4,607/36,857 (12%) predicted ORFs had a best match to the draft genomes while in the filtered assembly, more matches 4,618/36,210 (13%) ORFs matched the genomes. A total of 117 ORFs were identified in both assemblies. Among these ORFs, 51 ORFs when aligned to draft genomes were longer in the filtered assembly, 50 ORFs were longer in the unfiltered assembly, and 16 ORFs had identical alignment lengths to draft genomes.

3 Discussion

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

For each metagenome, the assembly graph was separated into 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. We initially 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 (Table 1, Figure 1). 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. These results combined suggested the presence of spurrious connectivity within 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. We compared local graph densities within metagenomic lumps to those of a mixture of whole genomes and the lump of simulated reads. For a mixture of genomes, the local graph density was measured to be low with less than 2% of nodes having a graph density greater than 20. For the lump of connected reads of a simulated Illumina sequencing dataset of these genomes, the graph density was larger with 17% of nodes having a graph density greater than 20. For the metagenomic lumps, the degree of connectivity was consistently larger than the simulated dataset lump, with an average of 35% of metagenomic nodes having a graph density of greater than 20. In addition to this observed increased connectivity, we also identified position-specific biases of local graph densities within a read for all metagenomes studied (Figure 2). As this bias should not be present in randomly sampled shotgun sequences and is not observed in the simulated dataset, spurrious connectivity clearly exists within the metagenomic sequences within the lump.

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 these highly-connecting sequences also had a biased presence in locations within sequencing reads (Figure 3). This bias supports the presence of non biological sequences within the metagenomic lumps.

To identify the origin of the highly-connecting sequences, we identified the shared highly-connecting sequences between our three soil, rumen, and human gut metagenomes. Overall, a

very small fraction of total identified highly-connecting sequences were found to be shared between environmental metagenomes suggesting minimal shared biases due to technical variations in sequencing, e.g. sequencing center or sequence preparation. To identify possible biological sources of the highly-connecting sequences, we compared the identified sequences to known reference sequences in the NCBI-nr database. The large majority of the highly connecting sequences (¿80/

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

The observation of position-specific trends in the various metagenomes further indicates that the highly-connecting sequences within the metagenomes are from non-biological origin. Similar position-specific trends were observed for both the local graph density and the fraction of highly-connecting sequences within the metagenomic lumps of all samples studied (Figures 2 and 3). The 3'-end biases of the two largest soil metagenomes were the most pronounced. For these soil metagenome lumps, the size of the lumps and associated position-specific trends increased at greater rates than the amount of sequencing, further supporting the presence of sequencing artifacts. 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 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. In the case of the studied metagenomes, 6-8% of unique k-mers were identified to be highly-connnective. In our example, these would be the "X" sequences, and the number of "Y" and "Z" sequences which are connected to these sequences would be affected by the coverage of the metagenome. For soil metagenomes, where sequencing coverage is 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 (32.4% coverage), the number of the "Y" and "Z" reads which are lump-associated but not highly-connective would increase 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 (as seen in the rumen and human gut metagenomes in Figure 3).

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 artifacts as they are also present in the error-free simulated dataset (3% of the unique k-mers identified as highly-connective). However, regardless of the origin of these sequences, we were interested in their incorporation into resulting assemblies. All sequences within the simulated and metagenomic lumps were separately assembled. We observed that the highly-connective sequences which we previously identified were under-represented in the final assembly compared to their presence in original sequencing reads (Table 3). Moreover, when these sequences were incorporated into assembly, they tended to begin or end contigs (Figure 4). These results suggest that, overall, the assemblers are challenged by characteristics of these sequences regardless of their origin and that the removal of these sequences should have little effect on subsequent assembly efforts.

We evaluated the effects of removing these sequences with the simulated dataset. This dataset is ideal for benchmarking because it contains no sequencing errors or biases and can be validated by its original reference genomes. We compared the de novo assembly of the simulated dataset before and after filtering out highly-connected sequences to evaluate the biological effects of these sequences on assembly. Overall, we found that the unfiltered and filtered assemblies were quite different based on constituent k-mer composition (31% different) and the total number of unique ORFs (34% ORFs unique to only the filtered assembly). In general, the removal of the highly-connective sequences resulted in an overall improvement of the assembly. The filtered assembly contained 29% more contigs and 51% longer assembly length (for contigs greater than 500 bp) than the unfiltered assembly (Table 4). The filtered assembly also resulted in 27% more predicted ORFs which had matches to reference genomes. Additionally, when comparing similar ORFs between assemblies, the filtered assembly contained a greater number of equal length or longer ORFs than the unfiltered assembly (Figure 5). In combination, these results indicate that removing highly-connecting sequences, regardless of their origin, improves the overall assembly.

Comparing soil, human gut, and rumen metagenomes, we also observed large differences in unfiltered and filtered assemblies. The removal of highly-connecting sequences resulted in gaining new contigs and losing unfiltered contigs (Table 4). These metagenomic assemblies are more challenging to evaluate as the original source genomes are unknown. Based on a small

number of draft reference genomes which were previously validated in cite Hess, we evaluated the rumen unfiltered and filtered assemblies. In comparing the predicted ORFs of the rumen assemblies, we found that the filtered assembly predicted slightly more ORFs compared to the unfiltered assembly. Additionally, the rumen filtered assembly contained more ORFs which had improved alignments to the reference genomes.

4 Conclusion

As datasets from NGS technologies continue to increase in size, our ability to analyze this sequencing data must reevaluated. Here, we demonstrate that efforts to resolve components of the complex metagenome assembly graphs are bottlenecked by the presence of highly-connective sequences that are have both biological and artificial origins. We show that in an error-free simulated dataset, the removal of these sequences (though biological in origin) improves the overall assembly. We thus propose the identification and subsequent removal of these sequences from metagenomes. This approach results in not only discarding sequences which we demonstrated to be artifactual but also allow one to break apart the dominant, highly-connected subgraph which contains these sequences. As a consequence of being able to resolve this component of the assembly graph, de novo assembly can effiently be performed on the separate, smaller subgraphs. Overall, our efforts provide a better understanding of the connectivity of metagenomes and gives us the ability to accurately scale de novo assembly which is critical to for successful metagenomic analysis.

5 Methods

5.1 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 [3]. Human-gut associated sequences (Illumina) of samples MH0001 through MH0010 were obtained from ftp://public.genomics.org.cn/BG [14]. The agricultural soil metagenome was from an Iowa corn soil metagenome 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.

5.2 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 each 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.

5.3 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 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 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 greater than 20 when N=100 and determined these nodes to be "highly dense" (Table 2). 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, and the total average local graph density by k-mer position for all reads was subsequently calculated (Figure 2). @brain fart - fix highly dense - there's better phrases I'm sure

5.4 Identifying properties of highly-connecting k-mers

The enrichment of knot-causing k-mers in unfiltered reads was estimated by identifying the fraction of unique k-mers in unfiltered sequencing reads and in their corresponding assembled contigs. The ratio of these k-mer fractions (unfiltered reads/assembled contigs) was used to estimate the enrichment of knot-causing k-mers in the reads.

5.5 De Novo Metagenomic Assembly

De novo metagenomic assembly of reads within each 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 [18]. 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, [17]). For soil 3 and human gut metagenomes, assemblies were performed at only K=33 due to the size of the datasets and memory limitations.

To study the effects of removing highly connecting sequences on assembly, the identified highly connective k-mers were filtered from reads by truncating the reads at the location of these sequences and the remaining reads assembled as described above. To compare unfiltered and filtered assemblies, we compared the total number of contigs, total assembly length, maximum contig size, and total number of predicted ORFs. ORFs were predicted using Fraggenescan(v1.1.15) with the following parameters: -complete=0 -training=454_10 [15]. The k-mers which constituted the unfiltered and filtered assemblies were compared. 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. 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 was 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).

@Need to add in vector assembly comparisons - not sure how this is done well enough to write about it, vector blah blah

@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?

The location of highly-connecting k-mers within assembled unfiltered contigs was examined by dividing each contig into 100 equally-sized regions. The fraction of highly-connecting k-mers within each region was calculated for each metagenome and is shown in Figure 4.

To evaluate assemblies of the simulated and rumen datasets, ORFs and contigs were aligned to the original 112 genomes used to generate the simulated metagenome or 15 draft rumen genomes [3] using BLAST (v2.2.25). Mismatches between assembled sequences and reference genomes were identified to evaluate the accuracy of assembly. To determine the effects of removing highly-connecting sequences, the alignment lengths of contigs from unfiltered and knot-filtered assemblies which shared identical top blast alignments were compared.

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] M Rho, H Tang, and Y Ye. Fraggenescan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, 2010.

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

[17] Daniel D Sommer, Arthur L Delcher, Steven L Salzberg, and Mihai Pop. Minimus: a fast, lightweight genome assembler. *Bmc Bioinformatics*, 8:64, Jan 2007.

[18] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18(5):821–9, May 2008.

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