Connectivity Analysis of Metagenomic Data

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

November 17, 2011

1 Introduction

@ Titus suggests rewrite - condense it down, too broad

Given the rapid decrease in the costs of sequencing, we can now achieve the sequencing depth necessary to study even the most complex environments [2, 13]. Currently, the major challenges in metagenomic studies are the continuously increasing number of sequencing reads and the lack of effective strategies to annotate and predict gene functions from these short reads [3, 5, 9, 16]. De novo metagenomic assembly of these reads offers several solutions. Firstly, assembly significantly reduces the data size by collapsing numerous short reads into relatively fewer contigs, providing longer sequences containing multiple genes and operons [7, 12]. Because it does not rely on the availability of reference genomes, de novo metagenomic assembly also produces novel contigs and create opportunities for further analysis, such as comparisons of novel sequences within and between metagenomes [6, 14] or annotations of novel genomes [2].

The general strategy for metagenomic assembly has been to use de novo *genome* assemblers [2, 13], particularly those targeting the assembly of short read sequences such as de Bruijn graph assemblers (reviewed in [7, 12]). These single genome assemblies are complicated by repetitive sequences, sequencing errors, and sequencing biases, and within the de Bruijn assembly graph, these elements increase graph complexity. For single genome assembly, relatively low rates of polymorphisms and sequencing errors are assumed to enable read coverage to determine the correct assembly path through the graph (cite some assemblers here). For metagenomic assembly, these assumptions do not hold true, and coverage-based resolution of the metagenomic assembly graph is confounded by the presence of multiple organisms which may both be closely related to each other and sampled at unequal depths [10]. De novo metagenomic assembly is further complicated by the presence of sequencing errors and/or artifacts which current assem-

blers cannot distinguish from variable-coverage genomic sequences [10, 15]. The presence of such sequencing biases is known in Illumina sequences [1, 4, 8], but their effects on assembly graph properties and thus resulting assemblies is largely unstudied.

The significant depth of sequencing which is needed to study complex environments presents specific challenges to metagenomic de novo assembly. Assemblers must not only be able to scale to these larger datasets but also be able to deal with the increasing amounts of sequencing errors and biases that accompany real biological sequences. A better understanding of the assembly graph and the effects of sequencing artifacts for these complex metagenomic datasets is critical to improving de novo metagenomic sequence assembly. In this study, we analyze the graph structures and connectivity of several metagenomic datasets and present our findings of highly-connecting sequences which are observed in all metagenomes we studied. We suggest that a significant portion of these sequences are sequencing artifacts and examine the effects of their removal on metagenomic assembly.

@comment on significant depth - undefined - find coverage of metahit and hess and set stage for coverage

2 Results and Discussion

2.1 Connectivity analysis of metagenome datasets

We selected datasets from three diverse, medium to high complexity metagenomes from the human gut [13], cow rumen [2], and agricultural soil (unpublished). For comparison, we also included one simulated metagenome (error-free) of a high complexity, high coverage (˜10x) microbial community [11]. 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.

@put in estimated coverage - histogram of frequency of coverage of contigs to demonstrate average coverage and spread of that average for rumen, metahit, soil 50, 100, 500, and simulated

As metagenomes are sequenced from natural populations, they are composed of sequences originating from multiple genomes. With appropriate sequencing coverage and minimal sequencing errors, reads from genes originating from the same genome will be connected together. To study the connectivity of reads in the above described metagenomes, we used a de Bruijn assembly graph representation (see Methods). Consistent with the diverse populations within

the studied metagenomes, connected reads from each dataset were subdivided into millions of partitions. Suprisingly, each dataset was dominated by the presence of a large, highly connected partition or single "lump" of connected reads (Figure 1, Table 1). Though present in each metagenome, the size of the lump varied in the human gut, soil, and rumen metagenomes. In the human gut dataset, over 75% of the sequencing reads were associated with this lump. In the rumen and 500 million read agricultural dataset, a total of 21% and 39% of all reads, respectively, were observed to be highly connected. Within the simulated dataset, the lump was significantly smaller in size, encompassing approximately 5% of the reads. As the simulated metagenome does not contain any sequencing errors or biases, we expect that its lump represents real biological connectivity or well-conserved regions across multiple genomes (16S rRNA or ITS regions). The large difference in the size of the lump in the human gut, rumen, and soil metagenomes compared to the simulated metagenome suggested the presence of non-biological sequences causing high connectivity in environmental metagenomes.

Further evidence for these non-biological sequences came from the soil datasets where the increase in the size of the lump was not proportional to the increase in the amount of sequencing (Figure 1, Table 1). In general, increased sequencing would result in a proportional increase in coverage. For metagenomes with low sequencing coverage (i.e., soil) in which well-conserved genes of the population are not thoroughly sampled, an increase in sequencing coverage would also result in a proportional increase in sequence connectivity. In the soil metagenomes, as the number of reads increased by 2-fold and 5-fold, the size of the lump increased by 5-fold and 14-fold, respectively. This observed supra-linear increase of the lump size suggest that with more sequencing, increasingly larger numbers of reads connected to this lump are being generated. This bias towards reads in the lump combined with the low coverage of the soil metagenomes suggested the probable presence of sequencing biases within metagenomes. With sequencing biases (or high coverage of conserved genes), we would expect a rapid accumulation of sequences into the lump as more sequencing is added. Highly connecting "X" sequences in a lump would recruit a number of connecting "Y" reads into the lump. Subsequently, as more sequences are added, these "Y" reads, which do not necessarily have to be highly-connective, would recruit more "Z" reads into the lump resulting in increasingly larger lump size. This phenomenon, called "preferential attachment" [cite cite], was observed in the soil metagenomic lumps.

To further explore this hypothesis and better understand the connectivity of metagenomic reads, we measured the local graph density of reads within the de Bruijn assembly graph. The

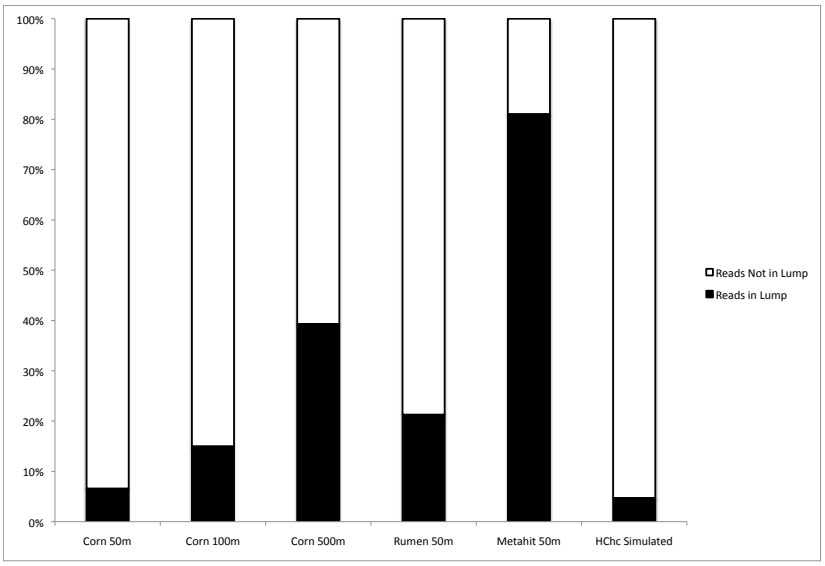

Figure 1: The largest proportion of total reads which are connected to each other within a "lump".

local graph density is defined here as the number of k-mers found within a distance of N divided by N. In the case of a linear sequence, the local graph density would be 2, and increased graph complexity in the form of additional branches or repeats would increase this value. For each of the studied datasets, we compared the local graph densities of the k-mers (nodes) within the de Bruijn assembly graph. For the human gut, rumen, and corn metagenome (500 million reads), over 90% of the nodes had an average graph density greater than 20 (Need to rerun on metahit). In comparison, the density of the assembly graphs of a mixture of microbial genomes (112 genomes) and the complementary simulated dataset both had fewer than 2% of the nodes with an average graph density greater than 20. The combination of the presence of a large, highly-connected lump with large local graph densities in environmental metagenome assembly graphs indicated substantial spurious connectivity within these metagenomic sequences. Systematic biases in base calling would create such connectivity and can often be observed as position-specific within sequencing reads. We proceeded to look for non-uniform properties of highly-connecting sequences within sequencing reads.

@ Make one of those pretty subfigure graphs - top figure is the one above, bottom figure would show different sizes of dataset, % of lump, and in particular highlight corn subsets (maybe

	Total Reads	Reads in Lump
Corn 50m	50,000,000	3,296,530
Corn 100m	100,000,000	15,003,371
Corn 500m	520,346,510	204,591,328
Rumen 50m	50,000,000	10,642,917
Metahit 50m	35,285,448	28,601,416
HChc Simulated	9,190,990	433,693

Table 1: Total number of reads in each dataset and corresponding largest connected subset lump

as an inset)

2.2 Properties of highly connected sequences in sequencing reads

We used a systematic traversal algorithm to identify the highly-connecting k-mers within the de Bruijn graph of each metagenome's lump (see Methods). These sequences were flagged as probable causes to "knots" in our highly-connected lump and likely causes of the lump itself. To study position-specific effects of these sequences, the location of knot-causing sequences (k-mers) were identified in originating sequencing reads. The presence of position-specific bias of a sequence within metagenomic reads is inconsistent with the random selection of DNA fragments by shotgun sequencing. If no sequencing error or biases were present, there would be no position-specific biases within sequencing reads as is the case of knot-causing sequences within reads of the simulated dataset. Cumulatively, the fraction of these sequences normalized to total sequences in the lump is relatively stable at all positions in the read (Figure 2). In contrast, within metagenomic lumps regardless of origin, we found that the position of knot-causing sequences within the read were biased towards one end of the read. In the soil metagenomes, a large fraction of knot-causing sequences were found at the 3' end of the read. In the human gut

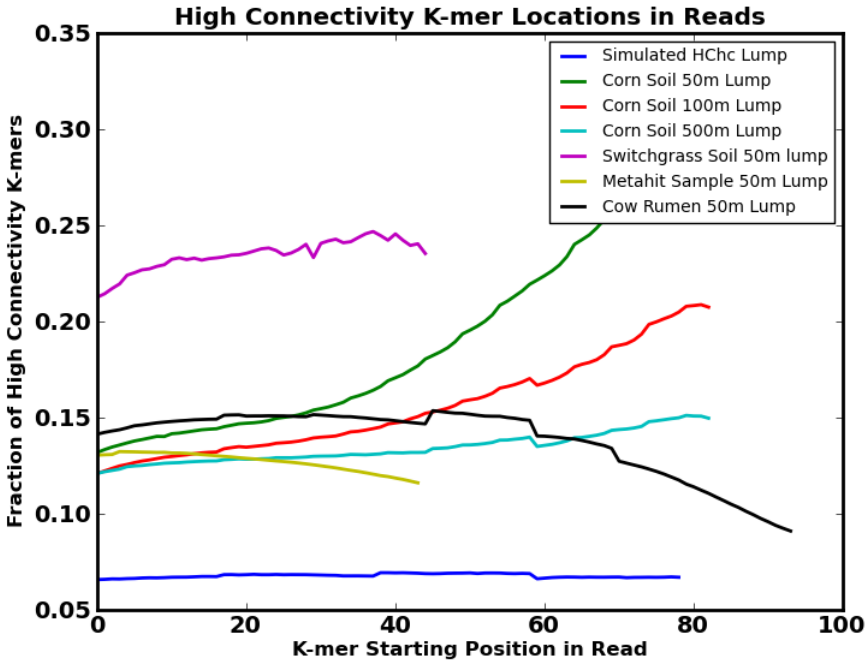

Figure 2: Position specific biases are observed in metagenomes but not in simulated data. For each position, the ratio of total number of knot-causing 32-mers and total number of 32-mers for all reads was determined.

and rumen reads, smaller fractions of these sequences were located at the 3' end of the reads.

@Titus notes confusing to be 3' and 5', I'm not sure if I fixed this well enough

The observed differences in position specific biases between the soil, human gut, and rumen metagenomes may have resulted from actual differences in sequencing biases originating from the creation of sequencing libraries (i.e. changes in Illumina chemistry, different sequencing centers, etc). An alternate explanation is that preferential attachment is causing the recruitment of specific reads which are correlated to the coverage of the metagenome. The human gut and rumen environments are much less complex than the soil and have higher sequencing coverage. In the case of the human gut and rumen metagenomes, more overlapping high-coverage reads may be recruited into the lump by preferential attachment, and these reads would increase the total number of non-knot-causing sequences and dilute the signal of knot-causing sequences. This explanation is supported by the trends observed with increasing dataset size of the soil metagenomes as the fraction of knot-causing 3' position-specific bias decreased with increasing sequencing coverage. Although the origins of metagenome-specific biases are unclear at this point in time, the identification of such sequences in all the studied metagenomes is indicative of the general presence of systematic biases.

2.3 Effects of removing highly connected sequences in assemblies

As a consequence of identifying knot-causing sequences as sequencing artifacts, we were interested in the effects of these sequences on the assemblies of each metagenome (Table 2). We compared the set of shared constituent k-mers between original and knot-filtered assemblies as well as several standard assembly metrics (total number of contigs, maximum contig size, and number of assembled base pairs). For the simulated dataset, removal of knot-causing sequences resulted in an assembly which contained 90% of the same constituent 32-mers as the original assembly. The simulated original and knot-filtered assemblies also shared similar totals of assembled contigs (greater than 500 bp), number of assembled base pairs, and maximum contig size. In real metagenomes (human gut, rumen, and soil), original and knot-filtered assemblies had similar numbers of contigs but had significant differences in the number of base pairs and maximum contig sizes within the assemblies. The Velvet-assembled rumen knot-filtered assembly resulted in approximately 1,000 less contigs, over 1 million more assembled base pairs, and a 2-fold increase in maximum contig size with the removal of nearly 8% of the total unique k-mers. In general, most of the unfiltered and knot-filtered assemblies were similar, sharing greater than 70% of constituent k-mers (with the exception of the soil metagenome with 100 million reads).

2.4 Effects of highly connected sequences in unfiltered assemblies

As the removal of knot-causing sequences did not greatly change the assemblies, we expected that the knot-causing sequences were highly present in assembled contigs. We compared the presence of the knot-causing sequences in the original reads to their presence in assembled contigs. We found that in all the datasets, sequencing reads were significantly more enriched for knot-causing sequences compared to the final assembled contigs. In the simulated dataset, nearly 7 times more knot-causing sequences were present in reads than in assembled contigs. In metagenomic reads, these sequences were 1.8 to 6.5 times more prevalent in reads than assembled contigs (Table 3). These results suggest that highly-connecting sequences, regardless of biological (as in the simulated dataset) or artifactual origin, are not effectively incorporated into assembly (observed in both Velvet and Abyss). This result is consistent with our observations of overall similarity between unfiltered and filtered assemblies after removing knot-causing sequences.

	Velvet Assembly (multi velvet, k=33, exp_cov=auto, cov_cutoff=0, scaffolding=no)						
	Unfiltered Lump Assembly			Filtered Lump Assembly			
	# of Contigs	# of bp	Max Contig Size	# of Contigs	# of bp	Max Contig Size	% Similarity
Metahit Lump	33,514	56,291,031	32,807	33,019	43,437,927	32,781	74%
Rumen Lump	22,640	16,075,203	3,355	21,674	17,202,480	7,949	75%
Iowa Corn 50m Lump	3,311	2,437,537	13,261	3,113	2,313,352	37,673	76%
Iowa Corn 100 m Lump	16,900	12,595,019	9,160	20,277	15,668,156	15,722	56%
Iowa Corn 500 m Lump	229,383	175,281,757	12,695	218,176	177,611,822	26,958	72%
Simulated Lump	1,519	1,110,539	3,414	1,485	1,086,280	3,407	90%

Table 2: Comparison of assembly (Velvet) of metagenome reads in lump with and without removal of knot-causing sequences. Percent similarity of assemblies was determined by determining overlap of constituent 32-mers between assemblies.

	% Unique Knots in Reads	% Unique Knots in Assembly	Ratio (Knots in Reads/Knots in Assembly)
Corn 50 m Lump	0.88%	0.18%	4.96
Corn 100 m Lump	7.61%	1.17%	6.48
Corn 500 m Lump	6.34%	1.19%	5.33
Rumen 50 m Lump	8.00%	1.48%	5.39
Metahit 50 m Lump	6.31%	3.37%	1.87
HChc Lump	3.30%	0.49%	6.75

Table 3: Lack of knot-causing sequences in assembled contigs in all metagenomes and simulated data.

2.5 Properties of highly connected sequences in assembled contigs

Given that the knot-causing sequences are not significantly being incorporated into the final assembly, we were interested in the location of these sequences within assembled contigs. We examined the position of the knot-causing sequences within contigs and observed that they were disproportionately being assembled on the ends of the contigs (Figure 3). Notably, the presence of these sequences on the ends of contigs would contribute to observed differences in unfiltered and knot-filtered assemblies. A specific contig cutoff length was used to filter contigs for assembly accuracy. In this study, we included only contigs greater than 500 bp. If knot-causing sequences are located on the ends of contigs, their removal would affect the total length of the contig and consequently its consideration in the final assembly. For example, a 32-bp knot-causing sequence fragment on the end of a 499 bp assembled contig would result in a 531 bp contig in an unfiltered assembly. When comparing constituent k-mers of these two assemblies, the k-mers contributed by this contig (length=499 bp) would be entirely lost in the knot-filtered assembly if the contig cutoff length were 500 bp. The loss of such contigs and their constitutive k-mers would contribute to the differences observed between unfiltered and knot-filtered assemblies. Further differences would also occur from trimming the sequences of the original reads and thus changing their connectivity in the assembly graph.

@ From Titus, did we validate this? I dont think we ever did, did I? Suggestion for validation: Change the amount of contig cutoff for one dataset, and see how the assemblies change. Oh wait, I did this for MSB2 but maybe not all these datasets.....Should be plot showing percent similarity decreasing with contig cutoff increasing for corn 50m, metahit, rumen, and simulated. Data is somewhere...mebbe...its on the list of stuff to do

2.6 Properties of highly connected sequences in assembled contigs

Since knot-causing sequences are preferentially being assembled at the ends of contigs, we suspect that many of these sequences may be erroneously assembled. Evaluating the accuracy of metagenomic assemblies is challenging because we have limited knowledge of the original microbial composition. Ideally, known genomes from the studied system could be used to partially evaluate assemblies. Given the lack of reference genomes for the studied metagenomes, we used an alternative method to evaluate assemblies. The location and length of protein coding genes , or open reading frames (ORFs), was determined in assembled contigs. Within a protein coding region, misassembled sequences would likely cause truncation of ORFs. In this case, these

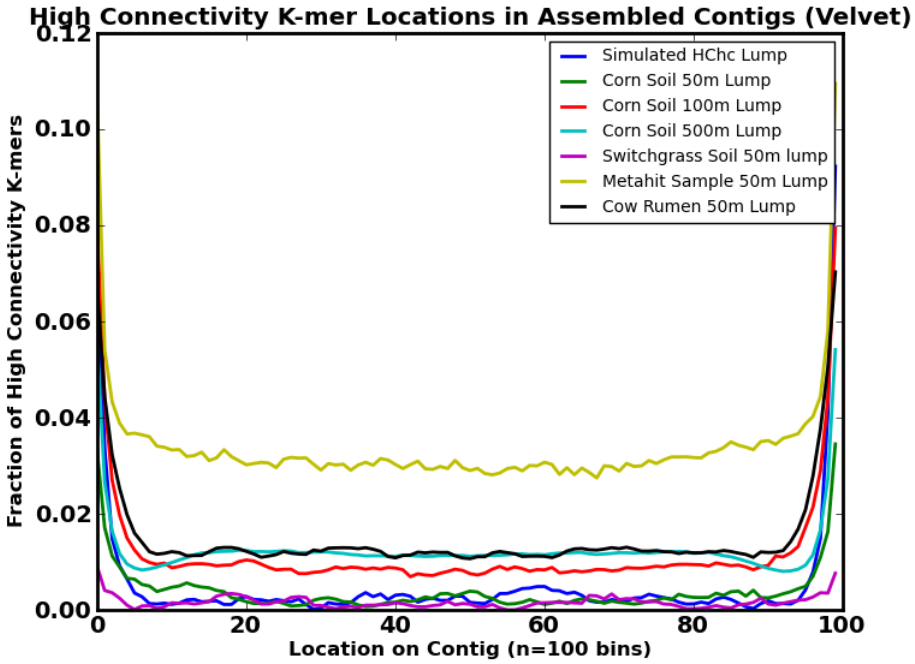

Figure 3: Knot-causing sequences are disproportionately being placed at the ends of assembled contigs. Each contig was into 100 equal-length bins. For all contigs, the ratio of total number of knot-causing 32-mers and total number of total 32-mers within each bin was determined.

misassembled sequences would be located at the "edge" of an ORF (or partially contained inside an identified ORF and partially extended beyond the ORF). Within all datasets, we observed a significant enrichment of knot-causing sequences at the edges of ORFs. In the simulated dataset, there was a 5x enrichment of knot-containing sequences at ORF edges. In the metagenomes, we observed between a 2x and 4x enrichment of these sequences at ORF edges. These results suggest that knot-causing sequences are likely truncating ORFs and their removal, if this is the case, should not change and may improve assemblies. To validate this, we compared unfiltered and knot-filtered assemblies in the simulated dataset. Using the reference genome annotations, we identified genes which were present in both assemblies and compared the alignment length of unfiltered and knot-filtered ORFs. We found that the large majority of alignment lengths did not change (83%, 1229 ORFs), some of the alignment lengths were worsened by filtering (10%, 151 ORFs), and some of the alignment lengths were improved by filtering (7%, 105 ORFs). These results support that, overall, the removal of knot-causing sequences does not largely affect assembly even in the case of the simulated dataset which does not contain sequencing biases. We performed a similar analysis on the assemblies of the rumen metagenome because of the availability of reference genomes. From the larger metagenome from which this dataset

	OUTSIDE	INSIDE	EDGE	OUTSIDE	INSIDE	EDGE	Edge Enrichment
sim	0.09358542	0.87635108	**0.030063**	0.22309011	0.62488093	**0.152029**	5.056928821
50m	0.05521889	0.91564374	**0.029137**	0.04845709	0.84353905	**0.108004**	3.706713693
100m	0.05459073	0.91639772	**0.029012**	0.05942956	0.82805892	**0.112512**	3.878163309
500m	0.05542326	0.91589817	**0.028679**	0.05249973	0.87641024	**0.07109**	2.478855141
rumen	0.03594191	0.93577729	**0.028281**	0.03612522	0.88100373	**0.082871**	2.93029312
metahit	0.05419932	0.92761761	**0.018183**	0.04211528	0.92201423	**0.03587**	1.972740902

Figure 4: I'll make this better later.

was generated (cite Hess), 15 draft genomes were successfully assembled. We considered the possibility that these draft genomes contained sequencing biases which we have identified in the subsampled metagenome. However, the draft genome assemblies were rigorously validated with support from similar tetranucleotide frequencies, high levels and uniform read coverage, and mate-pair correspondance within scaffolds, and we thus proceeded with the usage of these draft genomes for evaluating our assemblies. We identified the genes present in our unfiltered and knot-filtered assemblies which had the best alignments to these references and evaluated the effects of removing knot-causing sequences on their alignment lengths. Among these genes, 1 ORF had the same alignment length, 5 ORFs had shorter alignment lengths, and 114 had longer alignment lengths after removing knot-causing sequences (Figure X). The overall increase in alignment length of gene-coding regions in filtered assemblies indicates that, at least in the case of our rumen metagenome, removing knot-causing sequences can significantly benefit assembly.

@ I leave this out because I dont know how to include it right now. In general, although there was a significant enrichment of knot-causing sequences at ORF edges, we did not observe that many of these were contributing to misassemblies (only 2 of 85 contigs with ORFs enriched at edges contained mismatches to best alignment in reference genomes). @In the 24 sample bigger metahit dataset. 122 hits with same aln length 307 hits worse with filtering (length decreased) 281 hits better with filtering (length inccreased), these are reference sequences taht are weaker than rumen - not generated from the data. from isolates - thus coudl have some strain variation etc. for now, i exclude this.

3 Conclusions

@This section needs a rewrite, its just choppy. I can work on this too.

Short-read sequencing technologies, such as the Illumina platform, are creating unprecedented opportunities to deeply study complex environments. Sequence assembly and annotation, rather than sequence generation, are now the major limitations for metagenomic studies of

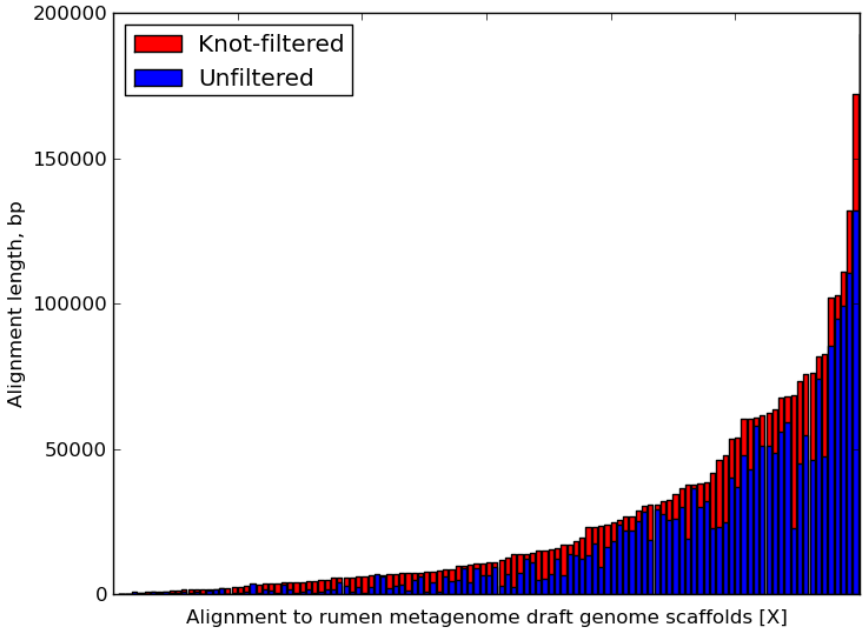

Figure 5: rumen orfs blah blah

environmental systems. Coincidently, an understanding of the nature of metagenomic sequences is of paramount importance to building specific tools for their analysis. In this study, we have demonstrated the presence of sequencing biases contributing to artificial sequence connectivity and suggest that removing these sequences does not significantly change and may improve the resulting assemblies. Our analysis also shows that the number of highly-connecting sequences increases significantly with increasing sequencing datasets. These artifactual sequences not only contribute to erroneous assemblies and gene annotations but also interfere with the ability to break apart the assembly graph which is a potential solution for scaling assembly for complex metagenomes. New approaches for de novo metagenomic assembly apply the properties of the assembly graph structure to resolve confounding assembly paths. These assemblers decompose the de Bruijn graph and subsequently perform iterative assemblies of isolated components [10] (Namike 2011 unpublished). A better understanding of the effects of sequencing biases on these approaches is necessary to effectively use them to generate accurate assemblies.

Critical analysis of sequencing biases and errors in metagenomic datasets would extend beyond de novo metagenome assembly. For example, very little is known about the true diversity in environments and the depth of sequencing needed to obtain appropriate coverage of an environment. Available metagenomic datasets are often used to estimate these parameters, and

without accounting for sequencing errors and biases, these estimates will be largely inaccurate. The ability to make conclusions from deep metagenomic sequencing depends on efforts and tools to understand the data itself. Our efforts highlight the usage of connectivity analysis within an assembly graph representation to identify and evaluate potential sequencing artifacts. Future efforts to better understand the origin of these sequencing artifacts and their effects on other metagenomic analysis would be valuable. For assistance, we provide the sequencing datasets, knot-causing sequences, and unfiltered and filtered assemblies used in this study at lyorn:/scratch/adina/artifacts-datasets.

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) from samples MH0001 through MH0010 were obtained from ftp://public.genomics.org.cn/BGI/gutmeta/Raw_Reads (Qin et al, 2010). The agricultural soil metagenomes were generated from a sequencing effort (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. After quality-trimming, the rumen datasets contained a total of 50 reads, the human gut datasets contained 358 million reads, and the agricultural soil dataset contained a total of 520 million reads from which 50 and 100 million reads were randomly sampled as subsets. The simulated high complexity, high coverage dataset containing 9 million reads was previously published (Pignatelli, 2011) and was randomly selected from a set of 112 complete genomes.

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 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 the agricultural corn, human gut, and rumen datasets, and 4 x 1e9 bit bloom filters for the simulated datasets. 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.

4.3 Identifying highly-connected k-mers

We implemented a systematic traversal algorithm to identify highly connected k-mers, that is k-mers that are reachable from many locations in the graph. Waypoints are 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 L. Such excursions that cover more than N k-mers are identified as "big excursions", and k-mers that are present in more than five big excursions are labelled as knot-causing k-mers. Local graph density (G) is defined as the number of k-mers within a specified region, or N/L. For this study, L = 40 k-mer nodes, N = 200 k-mer nodes, and G ¿ 5 is considered a big excursion. To study the effects of knots on metagenomic assembly, these k-mers were filtered from reads by truncating the reads at the region the initial knot was identified.

@Address Titus's concern about local graph desntiy measurement not being randomly started at any nodes - where is that?

We examined the position of these knot-causing k-mers in reads contributing to the lump. Each sequence in the lump was broken into its constituent k-mers, and each k-mer was identified as either a knot-causing k-mer or a non-knot-causing k-mer. The total fraction of k-mers within each dataset lump which were identified as knot-causing are shown in Figure X.

4.4 Assembly of reads with and without filtering of knots

Independent de novo metagenomic assembly of knot-containing and knot-filtered reads were completed with Velvet (v1.1.02, cite Zerbino) with the following parameters: velveth 33 -short -shortPaired (if applicable to the dataset) and velvetg -exp_cov auto -cov_cutoff 0 -scaffolding

no -min_contig_lgth 500. Assemblies were also performed with ABYSS (v1.2.0, cite) with the following parameters: ABYSS -k 33 (include these results/put in supplementary?). Only contigs longer than 500 bp were considered in further analyses. Assemblies were evaluated by comparing number of contigs, number of base pairs, longest contig size, and number of shared constitutive k-mers.

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 ORFs shared within assemblies, ORFs were aligned to reference genes (simulated dataset: original 112 genomes (genes), rumen dataset: 15 draft genomes (scaffolds) from [cite]) using using BLASTN (v2.2.25). ORFs in both the unfiltered and filtered assemblies were identified as sharing the same reference annotation as its best alignment (E-value ¡ 1e-10, percent identity ¿ 90

References

[1] 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.

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

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

[4] 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.

[5] 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.

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

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

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

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

[10] 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.

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

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

[13] 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.

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

[15] J.C Venter, K Remington, J.F Heidelberg, A.L Halpern, D Rusch, J.A Eisen, D Wu, I Paulsen, K.E Nelson, and W Nelson. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66, 2004.

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