Connectivity Analysis of Metagenomic Data

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

January 4, 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 to reduce the data 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). 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% (130,167,100/401,587,511) and 3.5% (1,731,348/50,000,000), respectively. For the small, medium, and large soil metagenomes, the coverage was estimated at 1.4% (686,435/50,000,000), 4.7% (4,652,502/100,000,000), and 5.60% (29,160,328/528,346,510), 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 76% (in the human gut metagenome) of the total reads.

@In a smaller subset of the human gut metagenome (50 million reads), the relative size of

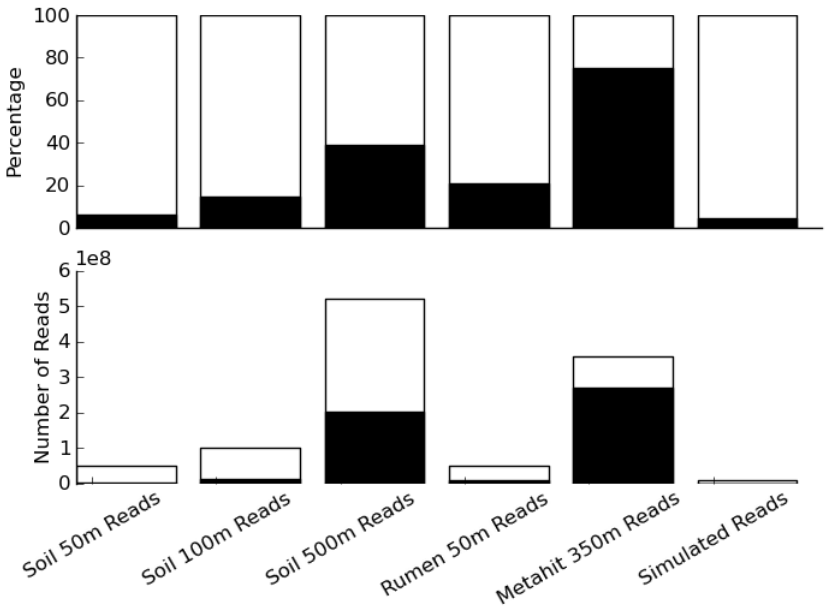

Figure 1: Within each metagenome, a dominant set of connected reads was identified which we refer as a lump.

the lump was similar and comprised over 75% of the sequencing reads (data not shown).

For the three soil datasets of increasing size, the size of the lump was not proportional to increases in the amount of 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 14-fold, respectively.

@This supra-linear increase of lump size relative to the increase of number of reads suggests that the connectivity of reads within a lump increases more rapidly than the contents of the lump. @Note that filtering out highly abundant k-mers effects - CTB, need to analyze

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. Thus, the local graph density of a linear sequence would be 2, and additional branches or repeats would increase graph density. 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

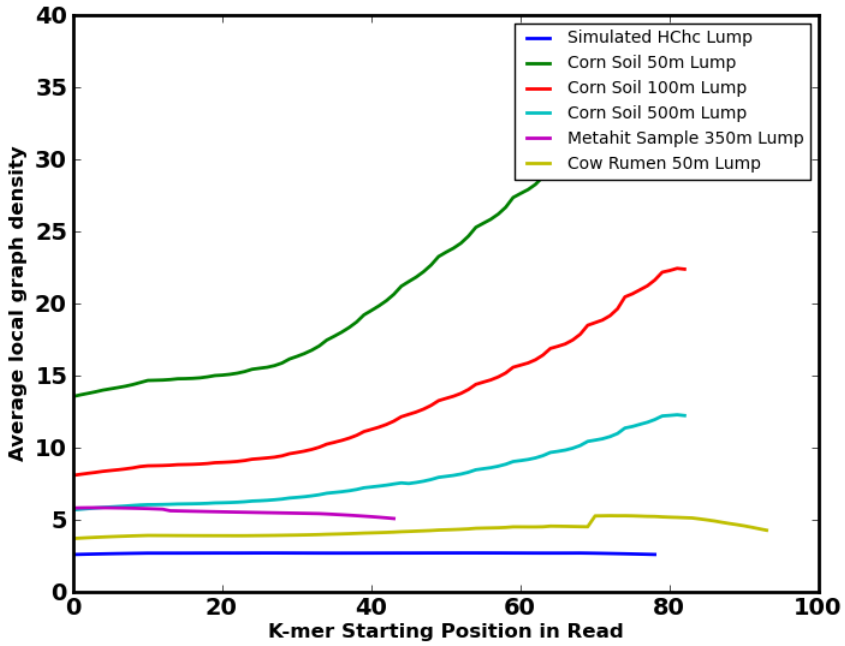

Figure 2: Graph density indicates the degree of connectivity between sequences. For the lump of the simulated metagenome, average graph density did not vary much by position along the read (2.68 +/- 0.03). In contrast, average graph density of sequences in metagenomic lumps varied significantly by position along the read. The average graph density for the small, medium, and large soil metagenome lumps were 22.3 +/- 7.8, 12.9 +/- 4.4, 7.9 +/- 2.0, respectively. The average graph density for the human-gut and rumen metagenome lumps were 5.5 +/- 0.2 and 4.3 +/- 0.5, respectively.

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%.

@Will insert table here

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 (average = 2.68 +/- 0.03) was stable for all positions along reads. In contrast, metagenome datasets had increased average local graph densities which 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 reads 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.

@Insert table of stoptag % here

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 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 and metagenomic 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 8-fold enrichment of these sequences in the reads compared to the final assembly. In metagenomic lumps, there were 5x more highly-connecting k-mers (this number will likely change with minimus) 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 studied.

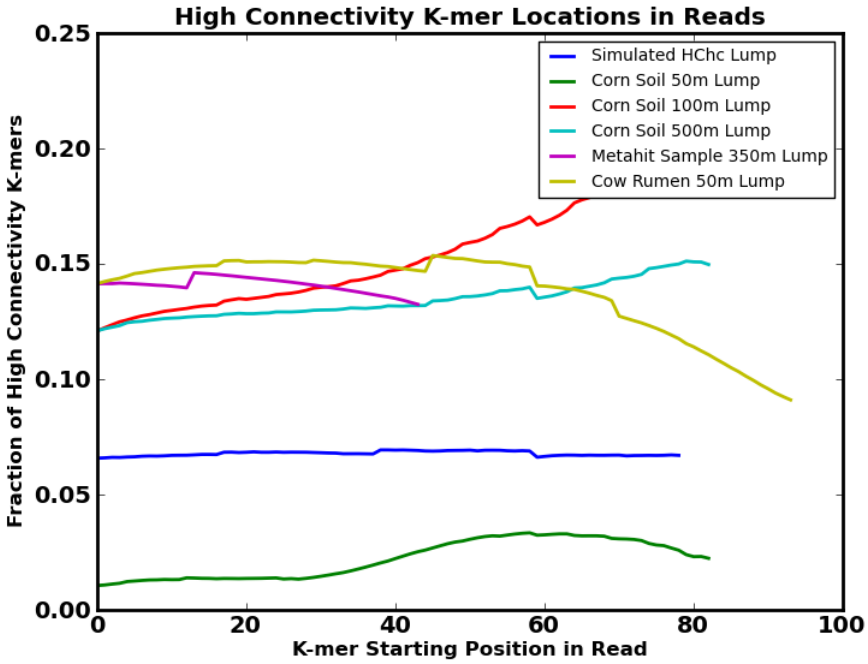

Figure 3: Highly-connecting sequences (k-mers) were present with position-specific bias within sequencing reads in metagenomic data.

@This table will be changed with minimus results - need to do minimus for metagenomic datasets.

@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.

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. To evaluate the accuracy of each assembly, we compared the predicted ORFs to the 112 genomes from which the simulated dataset orginated.

	% 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

Figure 4: Highly-connecting k-mers were more highly enriched in sequencing reads compared to assembled contigs, suggesting poor incorporation by assembly.

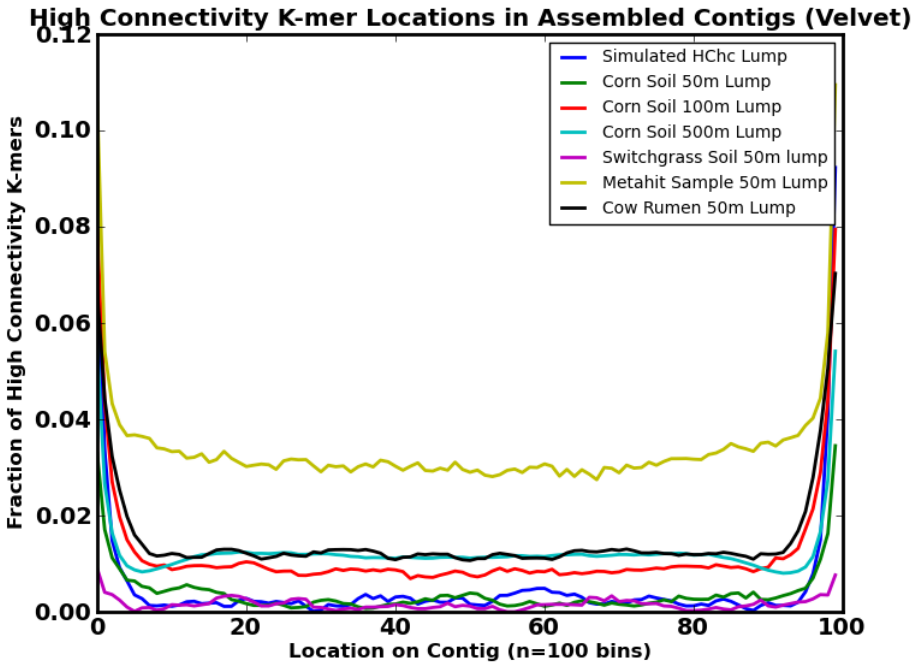

Figure 5: When incorporated into an assembly, highly-connecting sequences (k-mers) were disproportionately present at the ends of contigs.

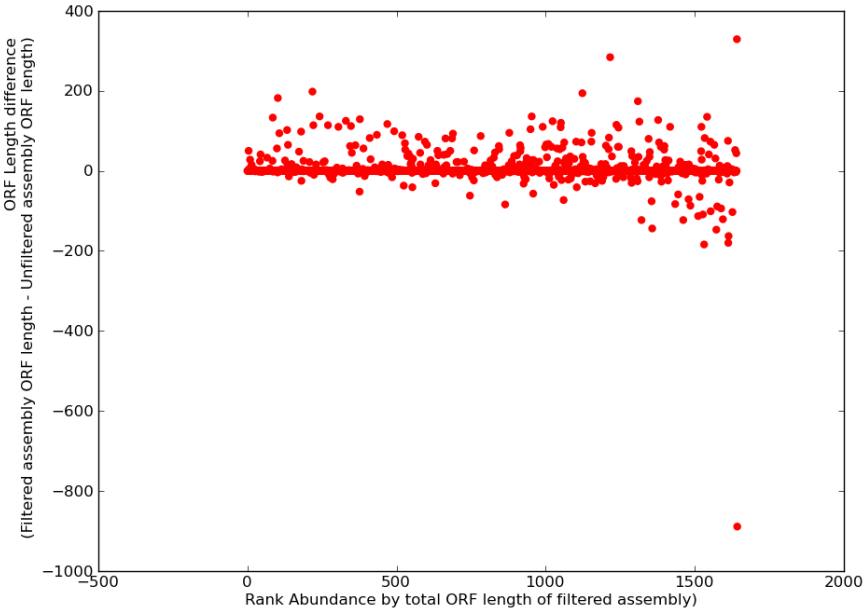

Figure 6: Alignment lengths of knot-filtered compared to unfiltered ORFs to reference genome proteins. The difference of the alignment length of filtered and unfiltered ORFs is shown for increasing predicted ORF lengths (ranked by filtered assembly), average=3.7 bp, stdev=35.8 bp, median=0 bp.

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 31% different. 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).

@@need to fix axes

@@are the unique orfs shorter? - small signal, not sure if good enough if strong enough to include here

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 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.