SI: Approaches for Large Scale Metagenome Assembly

ACH, JMT, CTB

October 3, 2012

0.0.1 Summary of approaches used on mock community dataset

The HMP mock community dataset and its available draft reference genomes were used to evaluate our approaches towards data reduction and partitioning for *de novo* metagenomic assembly. Reads of the mock community dataset were initially digitally normalized to a coverage threshold of 20 (as previously described in Brown et al), reducing the total number of reads from 14 to 11 million. Additionally, to remove possible sequencing artifacts associated with high coverage sequences (previously described in Howe et al), highly-abundant sequences (20-mers present at coverage greater than 50-fold) were filtered and the dataset was further normalized to a coverage of 10, resulting in a total of 9 million reads (Figure **??**). Finally, the remaining reads were divided into disconnected sets of reads resulting in a total of 85,818 partitions containing greater than five reads (summarized in Table **??**).

1 Methods

1.1 Datasets

In this study, we examined two large soil metagenomes generated from soils collected from Iowa corn and native prairie soils. Sequencing was performed at the DOE Joint Genome Institute (Walnut Creek, CA). Reads were quality trimmed at where Phred scores indicated a score of '2'. The total quality-trimmed reads in the Iowa corn and prairie datasets were 1.8 million and 3.3 million, respectively. We also include a human gut mock community dataset (combined from SRA SRX055381 and SRX055380). For this mock community dataset, DNA from bacterial isolates originally recovered from within or on the human body was mixed together and sequenced. The mock community dataset originally contained 14.5 million reads.

To evaluate our approaches, we added simulated reads from either a single E. Coli (str. K-12 substr DH10B) or five E. coli strains (K-12 substr DH10B, E24377, O147:H7 str. EC4115, UMN026, SE15) into select metagenomes. We computationally generated 100 bp reads from each reference genome to a coverage of 10x and with a 2% error rate and subsequently randomly shuffled these reads with select datasets.

1.1.1 Digital normalization

Digital normalization was previously describe in X. For the mock community dataset, digital normalization was performed with the following parameters: K=20, coverage=20, and bloom filter size = 1 GB x 4. For Iowa corn metagenome, digital normalization parameters were as follows: K=20, coverage=20, and bloom filter size = 48 GB x 4. Similar parameters were used for the Iowa prairie metagenome, with the exception that the bloom filter size was 60 GB x 4.

1.1.2 Removal of high abundance sequences

To eliminate known sequencing artifacts in Illumina metagenomes (previously described in XXX), high abundance sequences (coverage greater than 50) were removed using the count-min-sketch datastructure used for digital normalization. For the relatively high coverage mock community dataset, filtered reads were subsequently normalized to a coverage of 10 (K=20, bloom filter size = 1 GB x 4).

1.2 Identication of *gyrB*, *recA*, and *rplB*

Well-curated sequences from the Ribosomal Database Project's (RDP) Functional Gene Repository were used to build HMM profiles for *gyrB*, *recA*, and *rplB*. These models are available upon request from RDP. The *gyrB*, *recA*, and *rplB* HMM models contained 809, 353, and 277 amino acids, respectively. Assembled contigs were translated in all six reading frames and searched against the HMM profiles using HMMER3 (v3.0). Results were filtered with an E-value threshold of 1e-5. The abundance of assembled contigs were previously calculated as described above and used to estimate total number of *gyrB*, *recA*, and *rplB* genes within soil metagenomes. We also evaluated our ability to identify these genes against the HMP mock reference dataset and found that in some cases, the genes were not identified for some isolates or multiple reading frames within a reference genome were found to contain the gene of interest. We thus corrected calculated abundances of these genes based on our ability to identify the 22 HMP isolates using

the described methods.

1.3 Figures and Tables

Reference Genome	Coverage	No. Partitions	Length (bp)	Unfiltered Coverage (bp)	Filtered Dignorm Coverage (bp)	Unfiltered Assembly Coverage (% bp)	Filtered Assembly Coverage (% bp)
gi|32470588|ref|NC_005008.1|	2,412	9	4,439	4,439 100%	1,058 (24%)	97%	27%
gi|32470581|ref|NC_005007.1|	549	16	4,679	4,679 100%	4,585 (98%)	93%	72%
gi|32470520|ref|NC_005003.1|	533	21	6,585	6,585 100%	6,441 (98%)	92%	59%
gi|32470572|ref|NC_005006.1|	253	2	8,007	8,004 100%	7,953 (99%)	100%	100%
gi|32470532|ref|NC_005004.1|	112	52	24,365	24,358 100%	24,291 (100%)	96%	80%
gi|126640109|ref|NC_009084.1|	85	3	11,302	11,295 100%	11,270 (100%)	96%	96%
gi|32470555|ref|NC_005005.1|	74	12	17,261	17,202 100%	17,180 (100%)	96%	91%
gi|10957398|ref|NC_000958.1|	71	73	177,466	177,261 100%	174,614 (98%)	99%	92%
gi|10957530|ref|NC_000959.1|	52	37	45,704	44,974 98%	43,557 (95%)	100%	92%
gi|126640097|ref|NC_009083.1|	48	2	13,408	13,405 100%	13,383 (100%)	99%	99%
gi|15807672|ref|NC_001264.1|	40	63	412,348	410,970 100%	403,553 (98%)	99%	98%
gi|15805042|ref|NC_001263.1|	32	546	2,648,638	2,634,512 99%	2,589,566 (98%)	99%	98%
gi|27466918|ref|NC_004461.1|	30	476	2,499,279	2,498,081 100%	2,492,248 (100%)	98%	96%
gi|125654693|ref|NC_009008.1|	29	14	37,100	36,585 99%	33,250 (90%)	99%	96%
gi|161508266|ref|NC_010079.1|	29	442	2,872,915	2,298,758 80%	2,157,196 (75%)	93%	91%
gi|77404776|ref|NC_007490.1|	27	27	100,828	99,385 99%	93,550 (93%)	100%	96%
gi|125654605|ref|NC_009007.1|	24	92	114,045	108,526 95%	97,860 (86%)	96%	92%
gi|77404693|ref|NC_007489.1|	18	12	105,284	102,212 97%	96,169 (91%)	100%	99%
gi|24378532|ref|NC_004350.1|	16	131	2,030,921	2,029,376 100%	2,025,544 (100%)	99%	98%
gi|77404592|ref|NC_007488.1|	13	30	114,178	103,351 91%	93,637 (82%)	97%	96%
gi|77461965|ref|NC_007493.1|	13	628	3,188,609	2,919,441 92%	2,681,855 (84%)	98%	98%
gi|77464988|ref|NC_007494.1|	13	262	943,016	862,781 91%	788,626 (84%)	99%	96%
gi|126640115|ref|NC_009085.1|	11	683	3,976,747	3,939,190 99%	3,936,208 (99%)	98%	98%
gi|148642060|ref|NC_009515.1|	9	552	1,853,160	1,828,231 99%	1,826,639 (99%)	94%	93%
gi|150002608|ref|NC_009614.1|	7	7,751	5,163,189	4,899,622 95%	4,896,808 (95%)	81%	82%
gi|15644634|ref|NC_000915.1|	6	2,888	1,667,867	1,581,502 95%	1,581,024 (95%)	84%	85%
gi|194172857|ref|NC_003028.3|	6	4,123	2,160,842	2,047,832 95%	2,037,347 (94%)	76%	77%
gi|49175990|ref|NC_000913.2|	6	5,913	4,639,675	4,080,605 88%	4,074,119 (88%)	78%	78%
gi|50841496|ref|NC_006085.1|	6	6,459	2,560,265	2,169,547 85%	2,169,056 (85%)	58%	64%
gi|77358697|ref|NC_003112.2|	4	9,269	2,272,360	1,655,023 73%	1,626,301 (72%)	28%	33%