A novel approach to high throughput microbial genome finishing: Incorporation of 454 sequence data for gap closure in low quality Sanger data

FF009
LA-UR #07-1998

Olga Chertkov, Avinash Kewalramani, Riley Arnaudville, Cliff Han and Thomas Brettin
Los Alamos National Laboratory, Los Alamos NM; Joint Genome Institute, Walnut Creek CA

ABSTRACT

Background: The main goal of Joint Genome Institute's (JGI) Microbial Genome Program is to deliver finished microbial genomes faster, cheaper and better quality. Traditional method of producing multiple subclone libraries and Sanger reads is costly and also time consuming. 454 Life sciences has recently developed a highly scalable, highly parallel high throughput system which does not require traditional cloning based techniques and that produces high coverage, high quality genome data. In order to reduce the costs associated with library production and finishing it has become necessary to evaluate the incorporation of sequencing data produced by 454 into Sanger assembly. This evaluative study helps us in understanding how the 454 data can be used in an automated fashion to close the gaps in Sanger data and reduce the cost of finishing.

Procedure: In contrast to Sanger data, 454 data does not have a phrap score associated with every base. In order to assemble it with Sanger data using the phrap assembler, every base in this data has to be assigned a specific quality score. We split 454 contigs into pieces of 2000 bp and overlap of 100 bp and have employed the following two strategies:
Strategy1: Assign every 454 base a low quality of 20. Assemble fake 454 pieces with Sanger data. When they connect the end of the contigs, we pick primers on this low quality data to close the Sanger gaps. Since the end points of the gaps are known the number of reactions to pick compared to what is required if only 5X Sanger data is used is drastically reduced.
Strategy2: Assign every 454 base a good quality of 30, except the problematic areas, that we identified as homopolymeric regions. Assign quality to every base in homopolymeric regions based on a blast database between Sanger data and 454 data for few finished projects.

RESULTS:

Improvement in finishing efficiency

454 sequencing data was incorporated into the Sanger data using strategy 1 and strategy 2. This provided us with three different data sets on which we could run simulations to analyze the best possible way to improve and automate our finishing process. The first set was 454 data incorporated into 5X coverage Sanger data using Strategy 1. The second set of data was 454 data incorporated into 5X coverage Sanger data using Strategy 2 and the third set was the high quality Sanger data with 10 X coverage with no 454 data. Table 1. shows the number of primers after the first round of Autofinish primer picking for each dataset. It can be said that 454 data incorporation via any of the strategies does lead to improvement in the sense that lesser number of primers are picked as compared to when only Sanger data is used.

Achieving cost savings at the cost of quality of the finished genome is not acceptable. The second simulation involved picking three projects, of which one had lots of contigs and scaffolds and was a long way from being finished and the other two had small number of contigs and scaffolds and were closer to being finished. For all three projects we incorporated 454 data with low coverage (5X) Sanger data using both the strategies. These data sets were subjected to multiple rounds of AutoFinish primer picking and the results were analyzed using Consed.

Distribution of error rate based on motif length

We calculated the total error rate of homopolymers of particular size across 14 projects. Errors rates for Adenine and Thymine were grouped together as were the error rates for Cytosine and Guanine. The error rate for a particular homopolymer in strategy 2 is inversely proportional to the homopolymer length. However when error rate of Purine and Pyrimidine homopolymers were plotted against the size of the homopolymer we see from Fig.1, that as size increases error rate also increases. This clearly suggests that 454 homopolymer data is more erroneous for longer homopolymers. Also the high error rate of A,T homopolymers as compared to G,C homopolymers suggests that 454 data is much more reliable for genomes containing high incidence of G/C homopolymer regions rather than A/T homopolymer regions.

Analysis of error rate based on GC content of genome

Fig.2 illustrates the Error rate across all motifs in a particular project against the GC% (total GC content of the genome which is varying in the range of 27 to 68%) . We do not see any correlation across the whole genome between the GC% and the number of errors associated with the 454 data. Hence we do not feel that the 454 sequencing technology is biased in the prediction of bases in any way associated with the GC content of the genome within this range.

Table 1 Primer count for different finishing strategies for various microbial genomes

Microbe	454+Sanger Strategy 1(5X coverage)	454+Sanger Strategy 2 (5X coverage)	Sanger only (10X coverage)
Candidatus Desulfococcus oleovorans Hxd3(65% GC rich)	312	290	645
Herpetosiphon aurantiacus DSM 785 (51%GC rich)	1203	1221	1032
Fervidobacterium nodosum RT17-B1 (35% GC rich)	333	316	277
Pseudomonas putida GB-1 (62% GC rich)	105	93	94
Shewanella baltica OS223 (46% GC rich)	895	866	718
Petrotoga mobilis SJ95t (34% GC rich)	197	191	158
Geobacter lovleyi SZ (54% GC rich)	414	361	498
Thermosinus carboxydivorans Nor1 T (52% GC rich)	982	1070	996

Table 2 Finishing efficiency improvement for microbe Herpetosiphon aurantiacus DSM 785.

	S2 scaffold#	*S1 scaffold#*	*S2 contig#*	*S1 contig#*	*S2 low qual regions*	*S1 low qual regions*	*S2 primers*	*S1 primers*
Round 1	12	12	59	55	106	1278	1372	1359
Round 2	9	9	17	18	36	427	1459	1720
Round 3	6	6	12	11	30	182	1483	1825
Round 4	6	6	12	11	26	26	1649	1948

Table 3 Finishing efficiency improvement for microbe Petrotoga mobilis SJ95.

	S2 scaffold#	*S1 scaffold#*	*S2 contig#*	*S1 contig#*	*S2 low qual regions*	*S1 low qual regions*	*S2 primers*	*S1 primers*
Round 1	2	2	6	6	41	198	200	199
Round 2	2	2	4	4	16	58	257	253
Round 3	2	2	4	4	7	22	273	264
Round 4	2	2	4	4	7	7	289	280

Table 4 Finishing efficiency improvement for microbe Candidatus Desulfococcus oleovorans Hxd3

	S2 scaffold#	*S1 scaffold#*	*S2 contig#*	*S1 contig#*	*S2 low qual regions*	*S1 low qual regions*	*S2 primers*	*S1 primers*
Round 1	1	1	8	7	62	314	350	337
Round 2	1	1	4	3	11	68	385	362
Round 3	1	1	4	3	7	52	400	362
Round 4	1	1	4	3	5	5	400	409

S1 :Strategy 1
S2 :Strategy 2

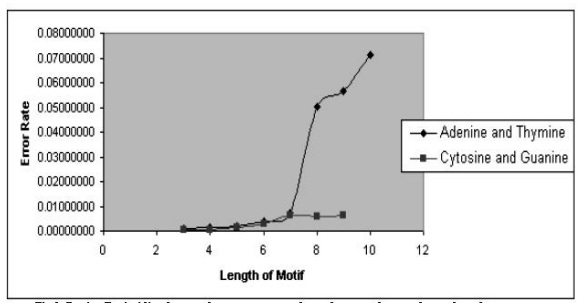

Fig 1. Purine/Pyrimidine homopolymer error rate dependency on homopolymer length.

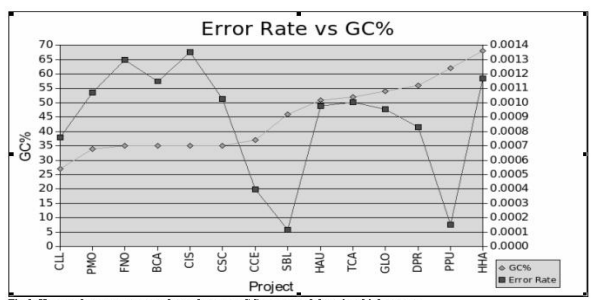

Fig 2. Homopolymer error rate dependency on GC content of the microbial genome.

Table 5 Finishing progress with Sanger and 454 data

project	size, Mb	Sanger ct	sanger scaf	sanger covera	454 contigs	final ct	final scaf	removed 454 reads
fti	1.8	157	26	4.6	74	20	15	0
bsb	3.47	233	147	15	101	37	15	0
pmo	2.14	29	9	11.4	102	14	6	0
cil	3.92	69	41	10.8	150	15	8	0
clg	2.7	60	35	7	155	30	18	0
clo	4.11	367	242	4.6	161	33	18	0
gto	3.9	47	42	10.5	166	24	10	0
vha	4.03	197	97	6	178	34	25	0
voa	4.03	195	109	6.35	181	27	18	0
dpr	3.8	41	26	11.5	182	8	3	0
vcd	4.13	229	75	6.5	262	45	12	0
ppu	6	43	6	15.6	265	18	3	2
dal	6.37	369	144	8	335	51	17	0
vcc	4.03	283	80	6	377	43	16	0
vda	4.96	464	104	5	400	77	27	2
clp	4.6	285	203	6.2	548	75	58	0
tca	2.8	159	80	9.3	562	42	17	4
mpa	5.8	92	10	8.4	633	29	2	5
clj	4.25	135	35	9	812	41	12	8
hau	6.6	155	51	11	1012	59	12	24
cce	4	101	57	10.9	1178	38	16	9
mra	6.9	91	10	9	1236	30	11	2
bca	5.47	76	53	15.6	2443	75	37	59

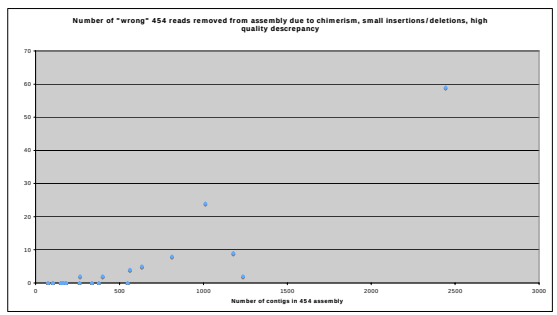

Fig 3. Number of 454 'wrong' reads dependency on quality of original 454 assembly

JGI-LANL has finished approximately 105 microbial genomes, some of them done with 454 and Sanger reads; all finished sequences have been uploaded into public databases. We have seen dramatic improvement in finishing efficiency when we incorporated 454 data in our assemblies. We also tried to identify other types of problematic regions in 454 data, but their number was very low. Only with the growing number of projects we found that the number of chimeric reads and reads with high quality discrepancies due to short insertion/deletions and/or mutations depends on a 454 coverage and number of 454 contigs in assembly.

Our recommendations for ideal 454 assembly: The number of contigs should be less than 500 for microbial genomes less than 8 Mb. It may require more than one 454 run per 2 Mb of sequence (especially if genome is more than 4 Mb). It will reduce the number of chimeric reads, reads with high quality discrepancies due to short insertion/deletions and mutations. These wrong 454 reads appear to be the result of low coverage in 454 assembly due to the lack of data. These wrong 454 reads tend to mess up assembly even with very good (10x) coverage of Sanger data.

Conclusion: Our primary objective of incorporating 454 sequencing data to close low quality gaps in Sanger data has been successful as presented in the results. We have been able to pick comparable number of primers, if not less, in an automated fashion during multiple cycles of AutoFinish. The strategy to assign low quality scores in only homopolymeric regions of 454 data is much more efficient in terms of primer picking and automated finishing than assigning low quality scores to all 454 bases. Also this strategy is more intuitive since the error rate in the homopolymeric region is dependent on the size of the homopolymer. We have also shown that the 454 sequencing data is fairly reliable and consistent in terms of errors for a variety of genomes, since errors are not dependent on the GC content of the genomes.

This poster is based on a Regular Research Paper (RRP) submitted to The 2007 International Conference on Bioinformatics and Computational Biology (BIOCOMP'07: June 25-28, 2007)
Paper ID #: BIC4379

US Department of Energy's Office of Science, Biological and Environmental Research Program and Los Alamos National Laboratory under contract No. W-7405-ENG-36