454-Sanger Joint Assemblies of Mammalian BACs

Christian J. Buhay, Michael E. Holder, Aniko Sabo, Xiang Qin, Huyen H. Dinh, Peter R. Blyth, Sandra L. Lee, Lynne V. Nazareth, Christie L. Kovar-Smith, Alicia C. Hawes, Huaiyang Jiang, Erica Sodergren, Donna M. Muzny, George M. Weinstock and Richard A. Gibbs
Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX 77030

We are constantly striving to improve the 454 sequencing pipeline at the BCM-HGSC. Advancements include optimization of the pipeline to achieve 100+ Mb per GS-FLX run, and the application of amplicon and paired-end sequencing methods. The 454 sequencing methods have been applied to numerous sequencing projects over the past year including 47 microbial genomes. We are developing strategies to integrate 454 assemblies of mammalian BAC pools with Sanger reads.

Microbial 454 sequencing and assemblies have resulted in assemblies comparable to, or better than that of Sanger sequencing, with N-50 contig sizes of 25kb. To evaluate how well the 454 sequencing technology would perform on mammalian genomes, we sequenced pools of 100 macaque and rat BAC clones using 454. Sequence data generated for the BAC pools were repeat masked and assembled using the 454 Newbler assembler. The assembled 454 contigs were then aligned to finished BAC sequence to assess contiguity and accuracy. We found that on average, 60% of finished BAC sequence was covered by 454 contigs in a 4x assembly; however, the N-50 contig length was low at roughly 3kb. Further investigations of strategies to combine whole genome shotgun Sanger reads with 454 data are in progress. We simulated whole genome shotgun assemblies, and contigs from the 454 assembly were mapped to each Sanger assembly. We evaluated ways to combine Sanger reads and 454 contigs and the influence of progressively increasing the 454 coverage (4x–8x–12x) on the quality of the combined assembly. Results indicate optimal gap closure at 5-6x 454 coverage, with N50 contig lengths increasing to 12kb or better. Conventional finishing tools such as Autofinish can then be used to close the remaining gaps.

454 Sequencing at BCM-HGSC

The 454 sequencing methods have been applied to numerous sequencing projects over the past year including bacterial genomes, mammalian BACs, genome mapping applications, concatenated cDNA libraries (CCS), functional mutation detection, micro RNAs and the larger *Acanthamoeba castellanii* genome (~45Mb). The pie graph illustrates the distribution of our sequencing projects by amount of 454 production bases. We have had success utilizing 454-Sanger joint assembly methods in finishing microbial organisms. The table below (fig. 2) is status of select microbial genomes in our finishing pipeline. Assemblies utilizing 454 paired-end reads are highlighted in yellow.

Fig. 2

Statistics for Microbial Genome in BCM Finishing Pipeline

Organism	Sanger Coverage	454 Coverage	Total Size(Mb)	Initial Scaffolds	Current Scaffold(s)	Status
S. aureus	9.3 X	16.3 X	2.8	5	1	complete
E. faecalis OG1RF	4 X	20 X	3	26	1	complete
P. stewartii	19 X	9.6 X	5	99	28	in progress
B. pumilus	10 X	15.5 X	3.8	14	4	in progress
T. paraluiscuniculi	0	59.1 X	1.1	77	1	in progress
M. bovis	1.8 X	23.9 X	2.4	59	33	in progress
S. iniae	1.5x	28 X	2	16	16	in progress

The implementation of 454 paired-end protocols and longer read lengths from the GS-FLX are shown to significantly improve microbial assemblies. The table on the right illustrates the impact of utilizing 454 paired-end reads and establishing contiguity in scaffold size on two microbial genomes. Also illustrated is the effect of longer read lengths on increased contig size in a comparison using S. aureus sequenced using the GS-20 and the GS-FLX instruments.

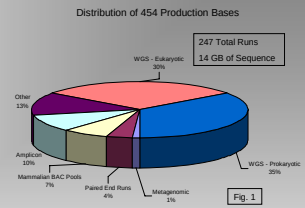


Fig. 1

454 Mate Pairs

454 Assembly	*F. tularensis* KO97	*E. faecalis* OG1RF
Instrument	GS-20	GS-20
Genome Size	1.8 Mb	2.8 Mb
Total reads	443,124	560,817
Coverage	25x	21.5X
Average read length	105	106
Avg.454 Quality Total trimmed contigs	26.4 126	26.5 128
Avg. contig length	14.0 Kb	21.1 Kb
N50 size	23.7 Kb	49.6 Kb
Longest contig	77,480	181,200
Paired-End Reads	212,571	225,208
Total Scaffolds	20	8
Average Scaffold Size	91 Kb	339 Kb
Scaffold N50	116 Kb	597 Kb

GS-20 vs. GS-FLX

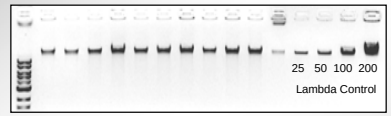
S. aureus USA-300	*S. aureus* USA-300
GS-20	GS-FLX
2.9 Mb	2.9 Mb
431,656	353,192
16X	32X
105	257
26.6	26
266	94
10.6 Kb	30.5 Kb
29.0 Kb	107.7 Kb
80,331	490,089

Fig. 3

454 Pooling & Mapping Strategy at BCM-HGSC

Mammalian BAC pools present a unique challenge in comparison to microbial genomes:

- Equal BAC representation in the pooling process.
- Possible DNA over-representation due to overlapping regions or duplications.
- Repeats– increased complexity in mammalian genomes yields 454 contig N50 lengths of approximately 1-3kb.
- Deconvolution of 454 reads/contigs – can we accurately map the 454 contig back to its donor BAC.



Pooling starts with a gel quantification of each BAC. Serial dilutions of a lambda control are made to 25ng – 200ng. Each well (BAC) is then compared to the control to estimate the concentration of the BAC. 50ug of each BAC is then added to the pool.

Primary 454-Sanger Mapping Strategy

Map BAC-ends to whole genome assembly.

↓

Excise sequence and qualities from whole genome assembly (WGA). Divide sequence into 1k lengths (or chunks) with 500 bp overlaps.

↓

Using Atlas BACfisher, fish for WGS reads using sequence chunks as bait. Import reads into each BAC set.

↓

Screen WGS for vector, and trim poor quality ends.

BLAST each BAC WGS set against the 454 contig set generated by Newbler.
Parameters: 94% identity with an end-to-end hit within 5 bases.

↓

Parse BLAST hits. Import all 454 contig sequence from mapping into the BAC set.

↓

Create .phd file for each 454 contig using unaltered qualities from Newbler assembler.

↓

Co-assemble 454 contig .phds and Sanger WGS using phredPhrap.

Joint Assembly Finishing Tools

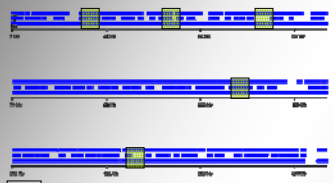


Fig. 4

Tools such as line-plots and Consed are utilized to evaluate the Sanger/454 mixed assemblies. The dot-plot (fig. 4) is an example of a BAC in which 454 contigs are closing several gaps in the 4x Sanger whole genome assembly (WGA). Several gaps in the WGA are closed by 454 contigs (shown in green). We can also visualize and edit mixed assemblies using Consed (fig. 5). The screen shot below shows a 454 contig resolving a difficult region. We suspect that the region is a stem-loop structure often susceptible to sequence dropouts. We were able to use a 454 contig to validate that the entire region was captured, and confirm problematic base calls.

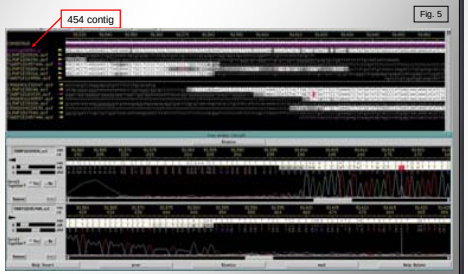


Fig. 5

454-Sanger Assembly and Error Metrics

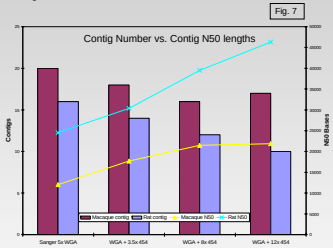
454 Newbler assemblies of mammalian BAC pools differ from assemblies of microbial genomes. This is the result of relatively large number of contigs and shorter N50 contig lengths (<5kb vs ~25kb) due to the higher frequency and complexity of repetitive elements, such as Alus, in mammalian genomes. Figure 6 (right) compares the 454 run and assembly metrics of rat and macaque pools of 100 BACs. Increasing 454 coverage was combined with existing Sanger whole genome assemblies. 454 contigs close gaps in the Sanger assembly at the rate of 2-3 gaps per 4x 454 coverage. Figure 7 (below) further illustrates the effects of 454 on a Sanger WGS assembly. A decrease in the number of contigs is accompanied by an increase in contig N50 lengths.

454 Metrics

Organism	Assemblies	Total Bases (Mb)	454 Reads Assembled (kb)	Average Read Length (bp)	Average Quality	454 Contig N50 (kb)	Actual Coverage
Macaque	WGA + 3.5x 454	80	259	251	27	1.0	5
	WGA + 8x 454	142	481	226	27	1.8	9
	WGA + 12x 454	218	748	250	26	2.9	13
Rat	WGA + 3.5x 454	115	309	254	27	1.2	5
	WGA + 8x 454	234	694	252	26	1.7	10
	WGA + 12x 454	341	1022	251	27	3.7	15

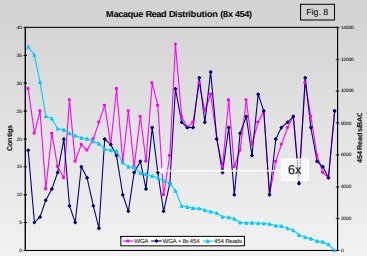
Fig. 6

Fig. 7



We evaluated the level of 454 coverage needed to see significant improvement in the joint assemblies (fig 8, right). We compared the number of 454 reads placed in each BAC against the number of contigs in the starting WGA assembly (pink line) and the number of contigs in the co-assembly of WGA and 454 (blue line) for both macaque and rat BACs. We estimate that for macaque we need approximately 6x 454 coverage to achieve the decrease in number of contigs. This indicates that the 454 reads are closing gaps; therefore, improving assemblies. The similar trend was observed for the rat BACs (data not shown).

Fig. 8



We estimated the error rate for 454 and 454-Sanger joint assemblies by comparing them to finished BAC sequence. As expected, the 454 contig error rates decreased with increased coverage for both macaque and rat. Macaque 454 contigs on average had higher error rates than rat contigs (fig. 9); 19-25 $x10^{-4}$ for macaque, compared to 5-16 $x10^{-4}$ for rat. The error rate for the joint WGA and 454 assemblies was dominated by WGA data and therefore followed the WGA error rate closely. It is important to note is that 454 contigs, while closing gaps, are not inflating the error rate. To this end, we have implemented 454-Sanger co-assemblies into our finishing pipeline, and have added 5 additional rat BAC pools to date.

Error Rate
Evaluated against finished BACs

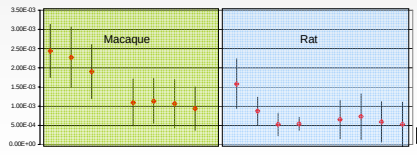


Fig. 9

Alternate Methods In Development

Alternate methods for pooling and mapping are being investigated. The following strategy for mapping is currently being investigated:

Map WGS reads from entire pool against 454 run using 454 mapping software.
- Default stringency: 90% identity and minimum 40 bp hit.
- Output: comprehensive text file and large .sff file

↓

Parse text file for list of reads with full hits. Create smaller .sff files containing reads that are local to the individual BAC.

↓

Assemble local 454 reads and Sanger WGS using Newbler assembler.

↓

Create .phd files for each contig from the resultant local assembly.

↓

Co-assemble 454 contigs with Sanger WGS using Phred and Phrap. Visualize assembly with Consed.

ALTERNATE POOLING

We are also assessing other pooling strategies in an attempt to more equally represent each BAC in the pool:

- Creating a sub-pool for library construction, and combining them into a larger pool thereafter.
- Pooling after the emPCR process