Finishing Methods for 454/Sanger Mixed Assemblies

Yan Ding, Donna M Muzny, Shannon P. Dugan-Rocha, Guan Chen, Alicia Hawes, Lesette M. Perez, Yih-Shin Liu, Zhangwan Li, Suzhen Wang, Judith Hernandez, Christian J. Buhay, Xiang Qin, George M. Weinstock and Richard Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX 77030

Summary

Several microbial genomes including *Staphylococcus aureus USA-300, Bacillus pumilus, Pantoea stewartii, Treponema paraluiscuniculi, Moraxella bovis* and *Enterococcus faecalis OG1RF*, *Streptococcus iniae* have recently been sequenced at BCM-HGSC using a combined 454 and Sanger sequencing platform. To date, the *S. aureus* and *E. faecalis OG1RF* genomes have been completed and the remaining genomes are in our finishing pipeline with varied levels of Sanger coverage. Although the implementation of 454 paired-end protocols significantly reduces the number of contigs without order and orientation, this combined assembly method carries some additional challenges and leaves traditional closure strategies limited by template availability and uncertainty for some low quality regions represented only by 454 reads. In addition, multiple ribosomal RNA regions within each genome increase the difficulty for finishing. Therefore we have implemented closure methods including direct genomic sequencing using GenomiPhi, multiplex PCR sequencing and long range PCR sequencing to deal with these new challenges. Direct genomic sequencing has been designed not only for closing or extending gap regions where template is not available but also for resolving areas of complicated genomic structure where traditional sequencing methods fail. This method has contributed to the completion of the *S. aureus* genome and results with the *B. pumilus*, *P. stewartii* and *S. iniae* genomes have shown success rate of 60%-88% with an average Phred20 of 580 bp. Multiplex PCR sequencing has also been used for bridging contigs without order and orientation in complicated microbes such as *P. stewartii.* The initial assembly of *P. stewartii* contained about 181 unscaffolded contigs. Here Autofinishing and additional primer walking were used to close over 100 gaps leaving 70 contigs without order and orientation. A series set of 43 multiplex PCR reactions were then used to close additional 20 gaps in the assembly. Finally, long range PCR has been applied for sequencing highly repetitive regions, such as ribosomal RNA, that have been identified with a BLAST based repeat tagging tool. The results and protocols of these strategies are presented.

Statistics for Microbial Genome in BCM Finishing Pipeline

	Contigs Before Autofinish	Contigs After Autofinish	Current Supercontigs	Additional Finishing Methods Applied	Percent Finishing with Additional Methods	Total Size(Mb)	Current Status
S. aureus	28	8	1	Direct Genomic Walk	10%	2.8	Completed
P. stewartii	181	100	45	Direct Genomic Walk Multiplex PCR Sequencing	30%	5	in progress
B. pumilus	55	50	4	Direct Genomic Walk Multiplex PCR Sequencing Long Range PCR Sequencing	15%	3.8	in progress
T. paraluiscuniculi	77	14	1	Direct Genomic Walk Multiplex PCR Sequencing Long Range PCR Sequencing	8%	1.1	in progress
E. Faecalis OG1	43	13	1	Long Range PCR Sequencing	5%	2.8	Completed
M. bovis	187	118	33	N/A	N/A	2.4	in progress
S. iniae	86	62	16	N/A	N/A	2	in progress

The chart to the left shows the status of assemblies before autofinish, after autofinish and supplemental finishing methods that have been applied for each microbial genome. Generally, more than 50% of contigs will be completed with order and orientation after autofinish. Closure techniques utilized during Autofinish include primer walking and regular PCR sequencing. Approximately 20-30% of the contigs will be linked through regular finishing techniques and the remaining contigs may be resolved through direct genomic sequencing using GenomiPhi, multiplex PCR sequencing and long range PCR sequencing. The results for each of these techniques vary depending on the complexity of the microbial genome sequence.

Direct Genomic Sequencing by Using GenomiPhi

The GenomiPhi DNA Amplification method amplifies linear genomic DNA with Phi29 DNA polymerase which has high proofreading activity. It can generate microgram quantities of DNA from nanogram amounts of starting material after an overnight incubation. GenomiPhi DNA Amplification Kit (GE Healthcare Life Sciences, product code: 25-6600-01 and 25-6600-02) has been applied to generate the DNA shown in Picture1 (right). Following amplification, the products are purified with EXO-SAP and direct sequencing then takes place. A basic flowchart of the direct genomic sequencing protocol can be found below. Picture 2 then shows an example of direct genomic sequence responsible for closing an intrascaffold gap in the *S. iniae* genome assembly. Success rates vary from 60%-88% for the different microbial genomes in which complexity of the target genome, uniqueness of the primers and purity of GenomiPhi product may play significant roles.

Protocol for GenomiPhi Amplification and Sequencing

Bacterial Genomic DNA 10ng/ul (Genome size: 2-5Mb)

↓

Sample buffer 9ul, 95°c X 3 minutes, then cool to 4°c

↓

Add mix reaction buffer 9ul, enzyme 1ul(all on ice)

Incubate at 30°c for 18 hours

↓

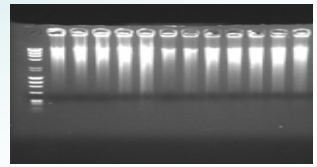
Inactivate enzyme at 65°c for 10 minutes, cool to 4°c

↓

EXO-SAP 2ul for 15 minutes

↓

Sequencing with Premix AB Big Dye 1:8, DNA 1:1 dilution



Picture 1: GenomiPhi amplified DNA for *P. stewartii* and *B. pumilus* with different initial concentration. The ideal starting concentration for amplification using GenomiPhi has been determined to be 10ng/ul, measured with Picogreen.

Picture 2: The direct genomic sequence shown above is longer than 600 bp (Phred20) and bridges two intrascaffold contigs. The top trace and bottom trace are PCR sequence and direct genomic sequence respectively.

A comparison between PCR sequencing and direct genomic sequencing for the same intrascaffold gap using primers picked by Autofinish shows that more than 90% of direct genomic sequence can be correctly placed within the assembly. Direct genomic sequencing allows us the ability to target and close gap regions without order and orientation information. Current efforts for direct genomic sequencing using GenomiPhi will focus on increasing sequence quality by incorporating different methods of purification of the GenomiPhi product, automated picking unique primers with higher stringency and optimization for microbial genomes of different size and composition.

Multiplex PCR Sequencing

Multiplex PCR sequencing using QIAGEN multiplex PCR master mix(2X) (lot No. 127128396 with RNase free water) has been applied in the closure team at BCM-HGSC. The kit provides robust and reliable performance for PCR amplicons with high fidelity enzyme and pre-mix formula in cost-effective and convenient way. This technique has been used in closing the gaps and to help order and orient contigs in assemblies of complicated microbial genomes such as *P. stewartii,* and *B. pumilus.* The *P. stewartii* genome is known to contain 12 plasmids ranging in size from 4 kb to 323 kb. This microbe also contains at least 9 copies of ribosomal RNA. Due to its difficult genome structure, the initial assembly containing 9.6X 454 coverage and 19X Sanger coverage was fragmented into 181 contigs. Autofinish and regular finishing decreased the total contig number to 100, many of which had no order or orientation information. Multiplex PCR sequencing was employed for this project in an attempt to close these difficult gaps. Application of a multiplex strategy enabled simultaneous amplification of up to 10 PCR products. Here multiple primers flanking gap regions were added to a single reaction and any products were isolated and sequenced with the corresponding primer set. The primers for multiplex PCR had been designed to be at least 22 bp or longer with a melting temperature near 60°c. Each primer was also BLATed against a local microbial database to guarantee uniqueness. Picture 1 (right) shows results of multiplex PCR for *P. stewartii.*

Protocol for Multiplex PCR Sequencing

primers premix (final concentration: 0.2μM for all primers)*	5μl
QIAGEN multiplex PCR master mix(2X)	25μl
microbial genomic DNA(10ng/μl)	1μl
variable water	19μl

*total 20 randomly pair primers will be mixed together

↓

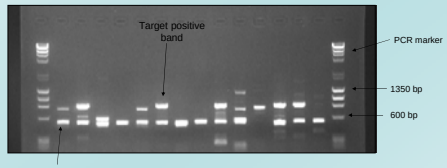
Multiplex Cycling Condition

Initial activation step:	95°c	15 min
3-step cycling step:		
Denaturation	94°c	30 s
Annealing	57-63°c*	90 s
Extension	72°c	90 s
No. of cycles	30	
Final extension	72°c	10 min

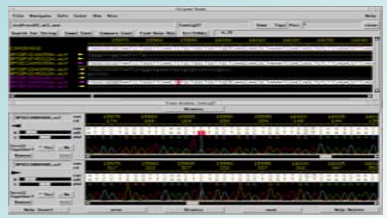
*a gradient PCR performed with annealing temperature incremented 1°c for each cycle up to 8 cycles

↓

EXO-SAP purification , sequencing with variable dilution of DNA AB BigDye mix 1:8



Picture 1 shows an example of multiplex PCR for *P. stewartii*. The bottom band is the control for each reaction. Any well with more than 2 bands was considered to yield a positive target and passed to sequencing. Here 43 multiplex PCR reactions for *P. stewartii* generated 28 positive target. Sequencing with individual primers leads to closure of 20 gaps.



Picture above shows sequence from multiplex PCR products used to close an interscaffold gap in the *P. stewartii assembly*. These two scaffolds have been merged based on the overlapping sequence from multiplex PCR. Generally, the entire length of the PCR product can be generated with high Phred quality. The same strategy has been applied for *B. pumilus* and *T. paraluiscuniculi.* Results have been consistent with those produced for *P. stewartii* genome, and have aided in contig extension and gap closure. Future developments will center on optimization of primer pairs for each PCR reaction, eliminating duplicated sequencing reactions and decreasing total PCR reaction volume.

Long Range PCR Sequencing

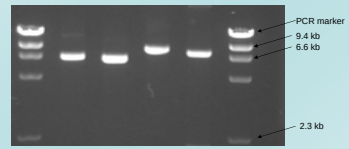
Some complicated microbes such as *P. stewartii* and *E. faecalis* contain multiple copies of ribosomal RNA (rRNA). With variable levels of Sanger coverage, these regions are poorly covered or the available reads stack together causing misassembles. The strategy for finishing and verifying these regions thus entails capturing complete copies of these ribosomal regions. To accomplish this, long range PCR sequencing has been applied. Initially, all rRNA regions are identified with an rRNA repeat tagging tool. Unique primers which are placed at least 1kb away from these tagged regions are designed for long range PCR. The figure below is an example of a repeat region identified with the tagging tool. Two different long range PCR kits including ABI GeneAmp XL PCR kit (Part No. N808-0192, Part No. N808-0193) and GE FideliTaq PCR Master Mix (2X) (product number 71182) have been tested. Results vary depended on different microbial genomes. A straightforward adjusted protocol (also listed below) are currently employed based on cost efficiency and handing simplification.

Tagging tool identifies and tags the rRNA region with a detailed repeat tag. Information including coordinates and ribosomal subunit (16S, 23S or 5S) are included.

Protocol and condition for long range PCR

Reaction:	Cycling conditions:
25 ul mix	1. 92°c 5s
1 ul 10pmol primerA	2. 92°c 5s
1 ul 10pmol primerB	3. 61°c 30s
1 ul 10ng DNA	4. 68°c 8 min
22 ul water	5. goto 2 X 29 cycles
Total volume = 50ul	6. 68°c 5 min
	7. 4°c forever

The protocol for long range PCR has been optimized to yield band sizes between 6kb and 15kb. Optimization includes decreasing the initial denaturation temperature and increasing extension time in order to obtain larger products. Picture 1 shows the gel image of long range PCR products for the rRNA regions found within the *E. faecalis* assembly.



E. faecalis has four copies of rRNA regions. Above gel image shows long range PCR products for each rRNA region.

In addition to sequencing with internal primers on the individual long range PCR product to capture each copy of complete rRNA sequence, the rRNA sequence linkage is also confirmed by comparing the fragment pattern of electronic digestion of an entire rRNA sequence assembled with restriction enzyme digestion of corresponding long range PCR product in the *E. faecalis* genome. Picture 2 shows an example of digestion image of PCR products for four individual rRNA with different enzymes including HindIII (H), EcorRI (E) and BamHI (B) respectively.