Minimization of Chimera Formation and Substitution Errors in Full-Length 16S PCR Amplification

Steve Oh, Richard Hall, Kevin Eng, Lawrence Hon, Cheryl Heiner
PacBio, 1380 Willow Road, Menlo Park, CA 94025

Abstract

The constituents and intra-communal interactions of microbial populations have garnered increasing interest in areas such as water remediation, agriculture and human health. One popular, efficient method of profiling communities is to amplify and sequence the evolutionarily conserved 16S rRNA sequence. Currently, most targeted amplification focuses on short, hypervariable regions of the 16S sequence. Distinguishing information not spanned by the targeted region is lost and species-level classification is often not possible.

SMRT Sequencing easily spans the entire 1.5 kb 16S gene, and in combination with highly-accurate single-molecule sequences, can improve the identification of individual species in a metapopulation.

However, when amplifying a mixture of sequences with close similarities, the products may contain chimeras, or recombinant molecules, at rates as high as 20-30%. These PCR artifacts make it difficult to identify novel species, and reduce the amount of productive sequences.

We investigated multiple factors that have been hypothesized to contribute to chimera formation, such as template damage, denaturing time before and during cycling, polymerase extension time, and reaction volume. Of the factors tested, we found two major related contributors to chimera formation: the amount of input template into the PCR reaction and the number of PCR cycles.

Sequence errors generated during amplification and sequencing can also confound the analysis of complex populations. Circular Consensus Sequencing (CCS) can generate single-molecule reads with >99% accuracy, and the SMRT Analysis software provides filtering of these reads to >99.99% accuracies. Remaining substitution errors in these highly-filtered reads are likely dominated by mis-incorporations during amplification. Therefore, we compared the impact of several commercially-available high-fidelity PCR kits with full-length 16S amplification.

We show results of our experiments and describe an optimized protocol for full-length 16S amplification for SMRT Sequencing. These optimizations have broader implications for other applications that use PCR amplification to phase variations across targeted regions and to generate highly accurate reference sequences.

SMRTbell Library Prep Workflow

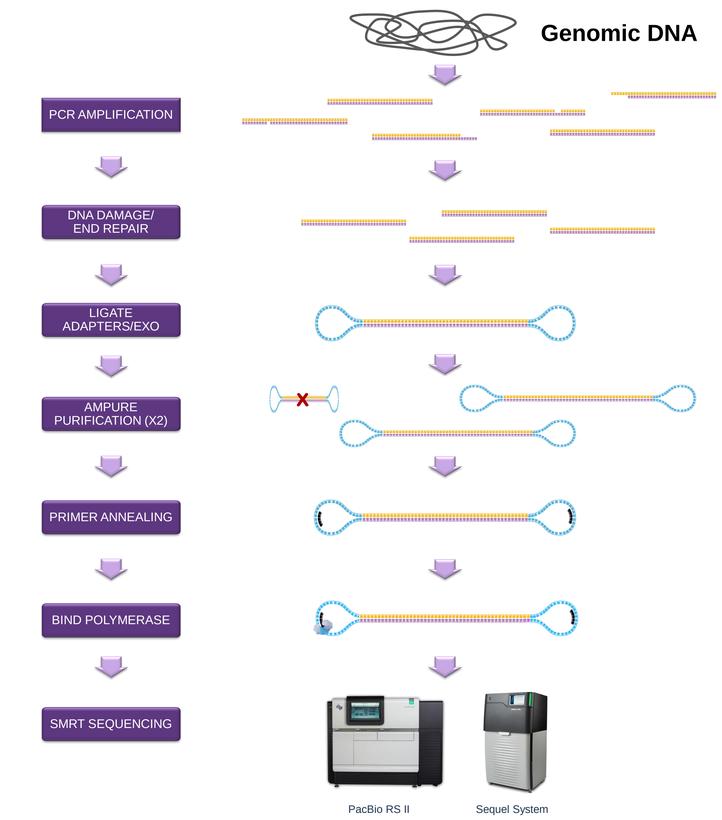
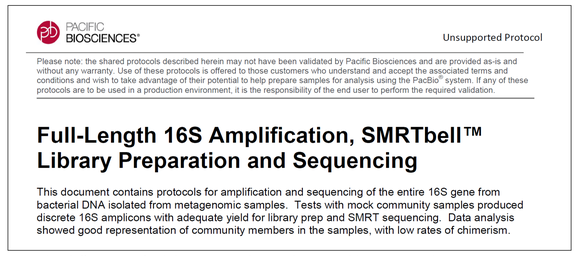


Fig 1. PCR amplification was optimized to minimize chimera formation and polymerase-dependent errors. The standard SMRTbell protocol was used for subsequent library preparation.

Shared Protocol for Full-Length 16S Amplification


PACIFIC BIOSCIENCES®

Unsupported Protocol

Please note: the shared protocols described herein may not have been validated by Pacific Biosciences and are provided as-is and without any warranty. Use of these protocols is offered to those customers who understand and accept the associated terms and conditions and wish to take advantage of their potential to help prepare samples for analysis using the PacBio® system. If any of these protocols are to be used in a production environment, it is the responsibility of the end user to perform the required validation.

Full-Length 16S Amplification, SMRTbell™ Library Preparation and Sequencing

This document contains protocols for amplification and sequencing of the entire 16S gene from bacterial DNA isolated from metagenomic samples. Tests with mock community samples produced discrete 16S amplicons with adequate yield for library prep and SMRT sequencing. Data analysis showed good representation of community members in the samples, with low rates of chimerism.

For the full protocol, visit www.pacb.com/support/documentation

PCR Parameters Affecting Chimera Formation

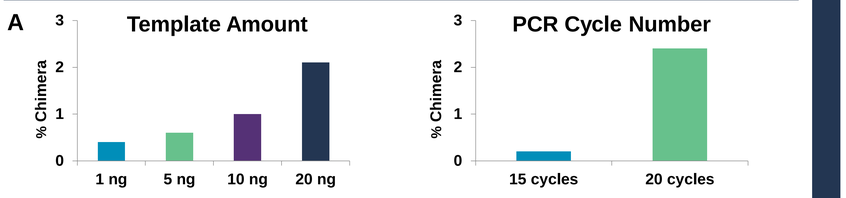


Fig 2A. 1, 5, 10 and 20 ng of BEI even mock community metagenomic DNA was amplified using 20 cycles, **left**. 15 and 20 cycle amplifications were compared with 20 ng input, **right.** 8 SMRTbell libraries were prepared from these amplicons and sequenced on the PacBio RS II. The chimera rates were compared and plotted above.

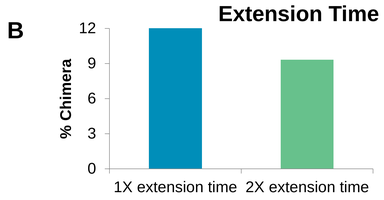


Fig 2B. 10 ng of the same bacterial template was amplified for 1X and 2X the recommended PCR extension times. Chimera rates are shown, **left**. Amplification was for 25 cycles in order to exacerbate chimera rate

Table 1. PCR Reagents

Reagent	Volume
water	50 – X µl
5X KAPA HiFi Buffer	10 µl
10mM dNTPs	1.5 µl
10 µM Forward primer	1.5 µl
10 µM Reverse primer	1.5 µl
Template DNA (5ng)	X
KAPA HiFi Hotstart DNA Polymerase	1 µl

Table 2. Cycling Protocol

Temperature	Time	Input Template 5 ng	0.5 ng	50 pg
95° C	30 sec			
57° C	30 sec	20 cycles	23 cycles	27 cycles
72° C	60 sec			
4° C	∞			

Forward Primer - AG**R**GTT**Y**GAT**Y**MTGGCTCAG
Reverse Primer - **R**GYTACCTTGTTA**C**GACTT

Table 1,2. PCR conditions were developed to balance decreased chimera formation and yield of PCR product.

Mock Community Representation

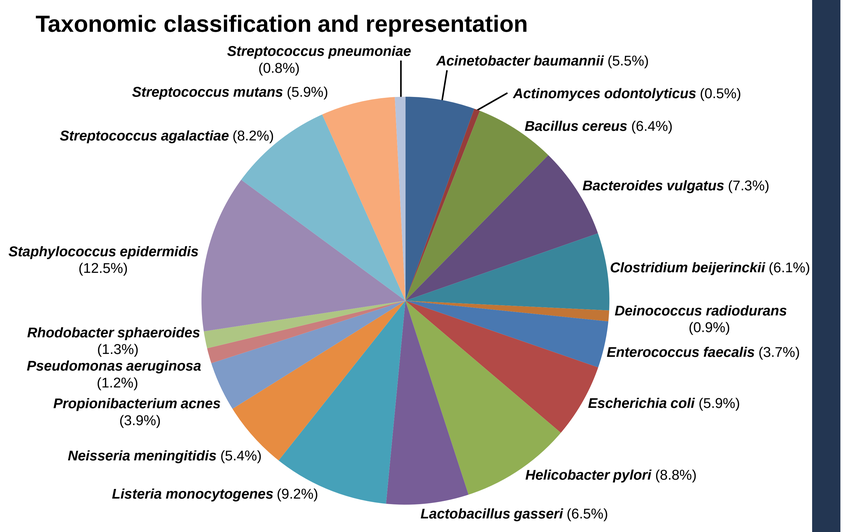
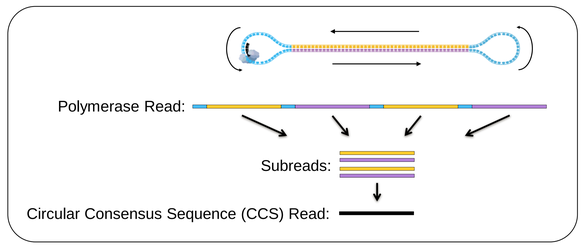


Fig 3. A mock community of 20 different species at near-equimolar concentration was tested using the optimized 16S PCR protocol. 19 species were detected at levels ranging from 0.8 – 12.5% of the in the total population.

Highly Accurate, Single Molecule Sequencing

Multiple Reads from a Single Molecule

As a function of the SMRTbell adapters, multiple single-pass reads are generated from an individual molecule. Combining these subreads corrects for random errors and results in a highly accurate single-molecule consensus sequence. Data can be filtered to an accuracy of 99.99% with the latest analysis version using CCS2.



Sequencing Error Rates Vary with PCR Polymerase

	Predicted Accuracy	Mean empirical accuracy	100% empirical accuracy	# Reads	Indel error rate	MM error rate
KAPA SYBR Fast qPCR	99%	99.73%	25.64%	38,889	0.09%	0.18%
	99.9%	99.78%	28.93%	31,938	0.04%	0.18%
	99.99%	99.79%	31.34%	19,678	0.02%	0.18%
KAPA HiFi Hotstart	99%	99.91%	69.33%	19,263	0.05%	0.04%
	99.9%	99.95%	75.70%	16,809	0.01%	0.04%
	99.99%	99.96%	79.23%	12,167	0.00%	0.04%
Unamplified *E. coli*	99%	99.75%	45.93%	12,112	0.27%	0.007%
	99.9%	99.95%	69.57%	7,763	0.056%	0.0031%
	99.99%	99.9855%	91.28%	4,471	0.0157%	0.0026%

Table 3. Amplicons from a single 16S sequence were produced using DNA polymerase with and without proofreading function. Higher indel and mismatch (MM) errors were detected when using a non-proofreading polymerase (top) compared to a proofreading enzyme (middle), significantly affecting the fraction of reads with 100% accuracy. Unamplified *E. coli* sequences were analyzed as a control (bottom); the reference sequence used was not from the same strain.

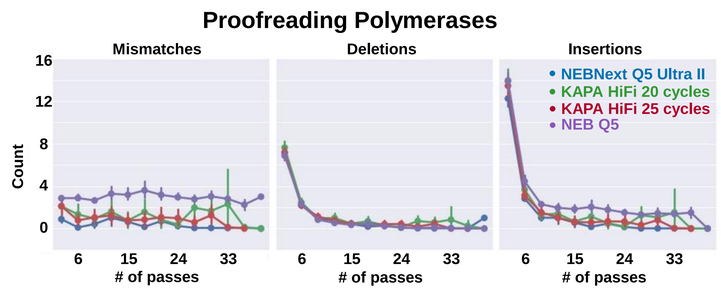


Fig 4. Mismatch, deletion, and insertion errors were analyzed for different polymerases with proofreading capability. For KAPA HiFi, conditions promoting chimera formation (25 cycles) was also examined. Note that there were few reads with a high number of passes for some conditions.

Results and Recommendations

A protocol for 16S amplification has been developed with minimal chimeras and high fidelity, based on the results described here:

- **Limiting PCR cycles** and the amount of **template DNA** reduced chimera formation most significantly compared to other parameters of 16S amplification.
- **Increasing the cycling extension time** also decreased chimera formation.
- **Proofreading polymerases** minimize base errors during PCR.

These findings are applicable when amplifying other mixtures of closely related sequences.

Acknowledgements

The authors would like to thank Dr. Tanja Woyke, the Microbial Genomics Program Lead at the DOE Joint Genome Institute, for the PCR primer sequences and initial amplification conditions.