SUNDAY-713 (June 19, 2016)

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

Steve Oh, Richard Hall, 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. PacBio 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 they 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 in the PCR reaction and the number of PCR cycles.
- A second problem that can confound analysis is sequence errors generated during amplification and sequencing. The SMRT Analysis Reads of Insert program provides filtering of single-molecule CCS+ reads to 99.99% predicted accuracy.  Substitution errors in these highly filtered reads may be dominated by mis-incorporations during amplification. Sequence differences in full-length 16S amplicons from several commercial high-fidelity PCR kits were compared.
- We show results of our experiments and describe our 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 generate highly accurate reference sequences.

## SMRTbell Library Prep Workflow

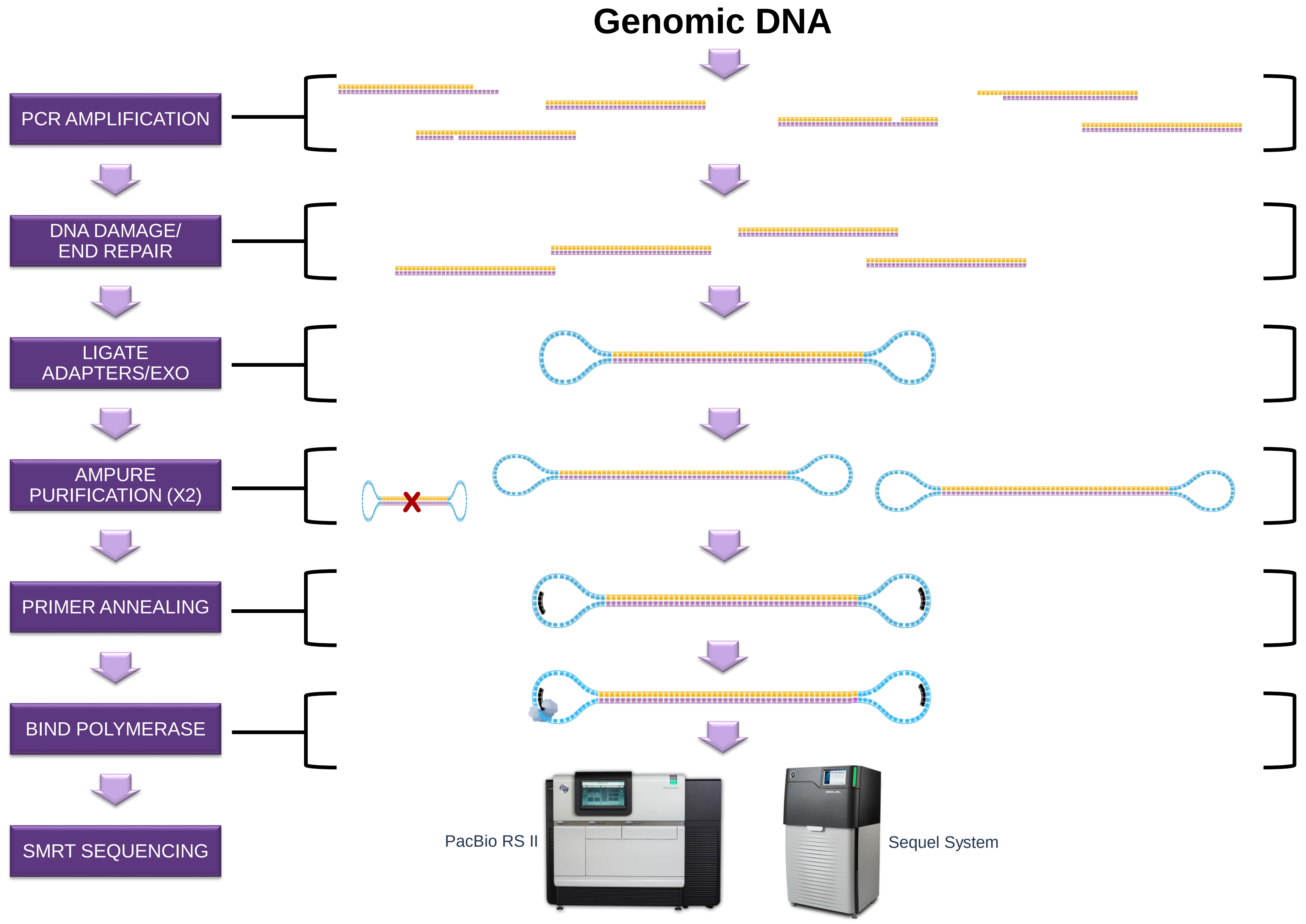
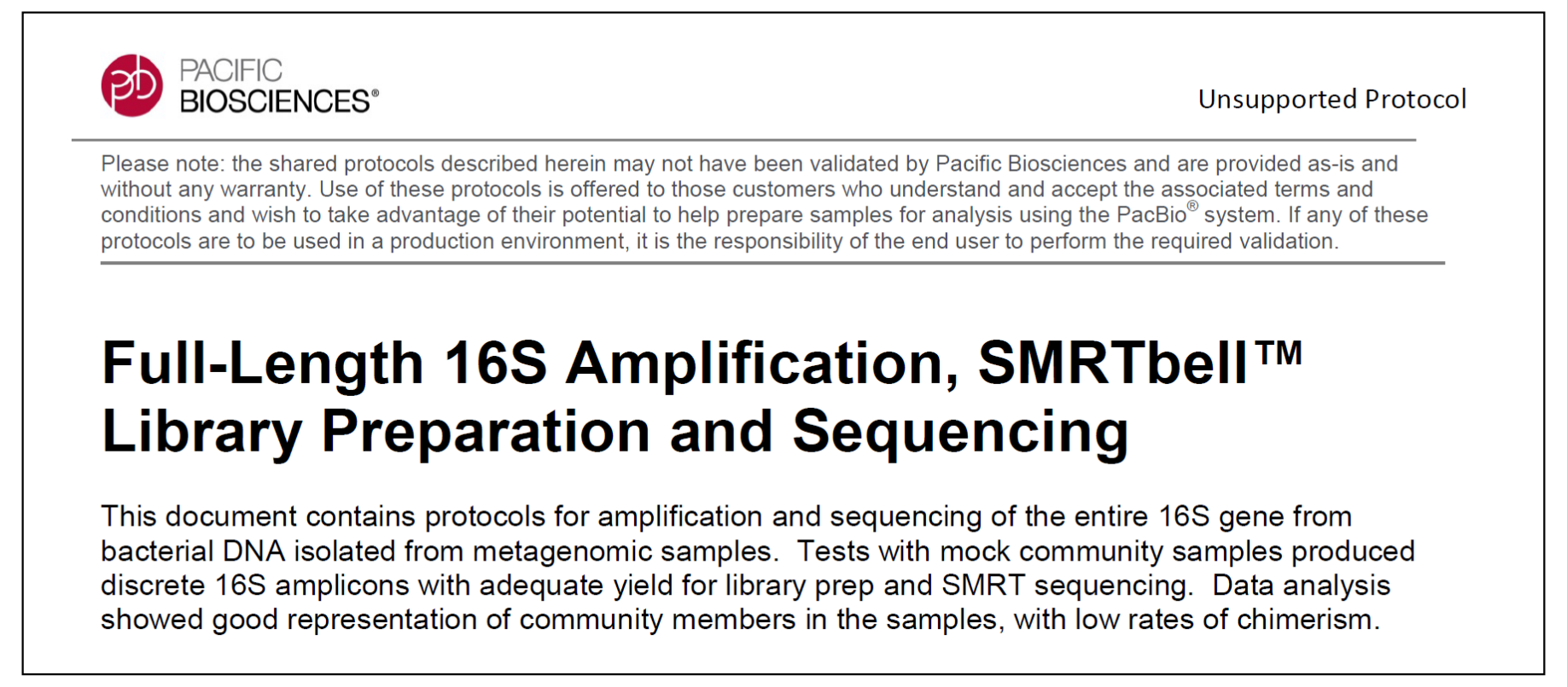


**Fig 1.  Optimized full-length 16S workflow.** PCR amplification conditions were developed to minimize chimera formation and polymerase-dependent errors. The standard SMRTbell library preparation protocol was then used.

## 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.

**Table 1.  Recommendations for PCR Cycling and Template Input Amounts**

| Step | Temp | Time | # Cycles | Input DNA | # Cycles |
|---|---|---|---|---|---|
| Denaturation | 95°C | 30 sec | 20 to 27 (see right) | 5 ng | 20 |
| Annealing | 57°C | 30 sec |  | 0.5 ng | 23 |
| Extension | 72°C | 60 sec |  | 50 pg | 27 |

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

## PCR Parameters Affecting Chimera Formation

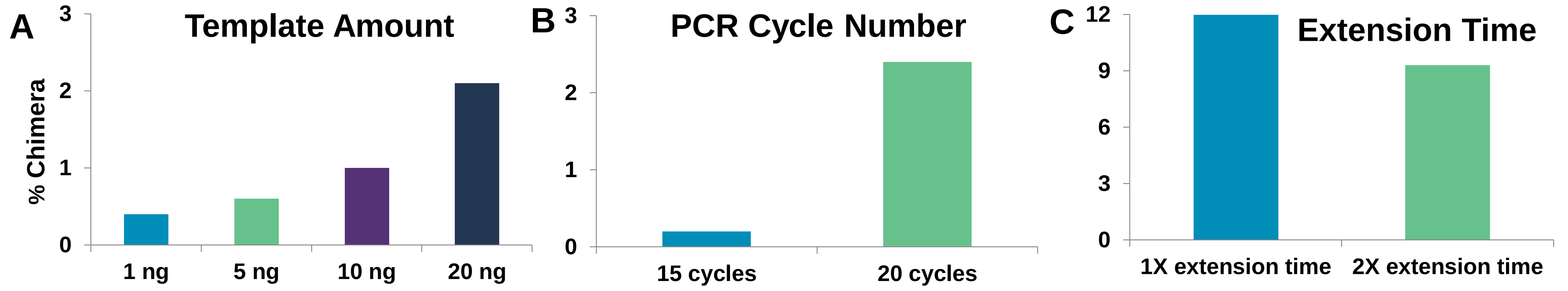


**Fig 2. Chimera rates by PCR conditions.  (A)** 1, 5, 10 and 20 ng of BEI even mock community metagenomic DNA were amplified  using  20 cycles.  **(B)** 15- and 20- cycle amplifications were compared with 20 ng input.  **(C)**10 ng of the same bacterial template was amplified for 25 cycles to exacerbate chimera rate. Chimera rates for 1X and 2X the recommended PCR extension times were determined.  For all experiments, SMRTbell libraries were constructed and sequenced on the PacBio RS II. The PCR conditions selected for the full-length 16S protocol above were selected to balance decreased chimera formation and yield of PCR product.

## Mock Community Representation

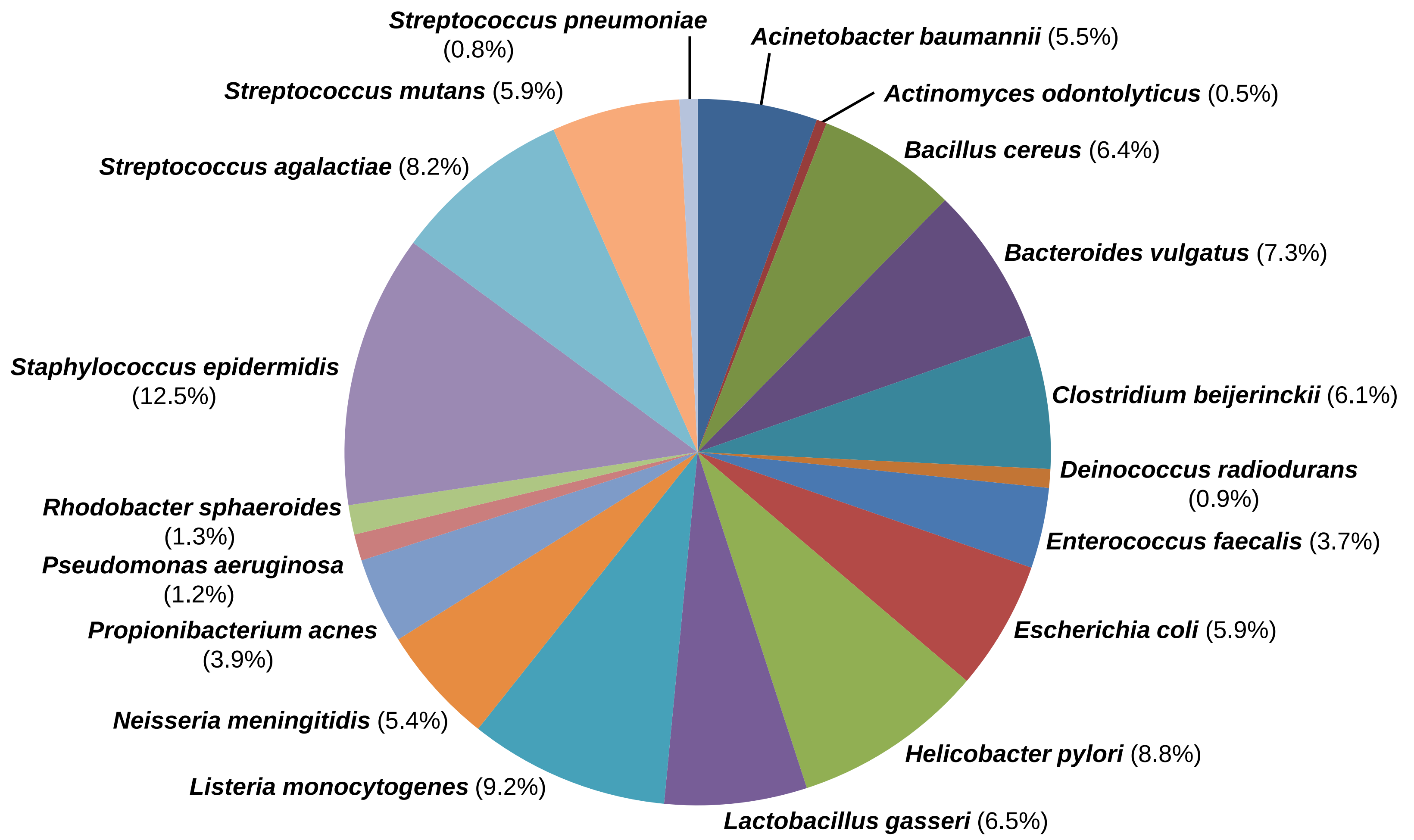
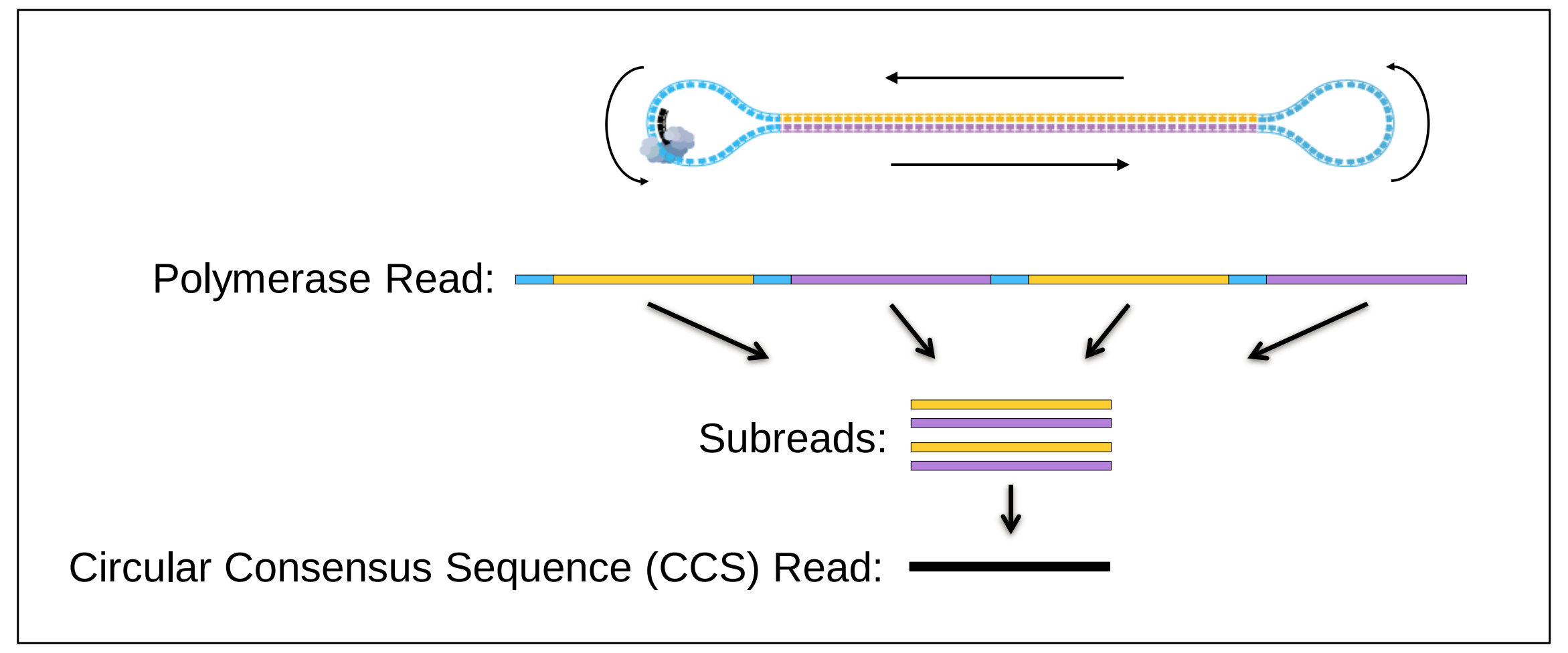


**Fig 3.  Taxonomic classification and representation**.  A mock community of 20 species at near-equimolar concentration was tested using the optimized 16S PCR protocol. 19 species were detected at levels ranging from 0.8 – 12.5% of the total population.

## Highly Accurate Single Molecule Sequencing

**Fig. 4.  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 SMRT Analysis CCS2.



## Sequencing Error Rates Vary with PCR Polymerase

| Enzyme | Proofreading Capability | Predicted Accuracy | Mean Empirical Accuracy | 100% Empirical Accuracy | # Reads | Indel Error Rate | MM Error Rate |
|---|---|---|---|---|---|---|---|
| KAPA SYBR Fast qPCR | No | 99%<br>99.9%<br>99.99% | 99.73%<br>99.78%<br>99.79% | 25.64%<br>28.93%<br>31.34% | 38,889<br>31,938<br>19,678 | 0.09%<br>0.04%<br>0.02% | 0.18%<br>0.18%<br>0.18% |
| KAPA HiFi Hotstart | Yes | 99%<br>99.9%<br>99.99% | 99.91%<br>99.95%<br>99.96% | 69.33%<br>75.70%<br>79.23% | 19,263<br>16,809<br>12,167 | 0.05%<br>0.01%<br>0.00% | 0.04%<br>0.04%<br>0.04% |
| Unamplified *E. coli* | No | 99%<br>99.9%<br>99.99% | 99.75%<br>99.95%<br>99.9855% | 45.93%<br>69.57%<br>91.28% | 12,112<br>7,763<br>4,471 | 0.27%<br>0.056%<br>0.0157% | 0.007%<br>0.0031%<br>0.0026% |

**Table 2. Sequencing error rates from single target 16S amplicon 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).

## Results and Recommendations

- **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 amplification.

We have developed recommendations to mitigate chimera formation by decreasing PCR cycle number and template input as well as exceeding the polymerase manufacturer's extension time.  In addition, use of proofreading polymerases must be ensured to maximize the accuracy of sequenced reads.

## 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.