Long Amplicon Analysis: Highly Accurate, Full-length, Phased, Allele-Resolved Gene Sequences from Multiplexed SMRT® Sequencing Data

Brett N. Bowman1, Patrick Marks1, N. Lance Hepler1, Kevin Eng1, John Hartling1, Takashi Shina2, Shingo Suzuki2, Swati Ranade1

1Pacific Biosciences of California, Inc., Menlo Park, United States of America
2Tokai University School of Medicine, Isehara, Japan

Introduction

The correct phasing of genetic variations is a key challenge for many applications of DNA sequencing. Allele-level resolution is strongly preferred for histocompatibility sequencing where recombined genes can exhibit different compatibilities than their parents. In other contexts, gene complementation can provide protection if deleterious mutations are found on only one allele of a gene. These problems are especially pronounced in immunological domains given the high levels of genetic diversity and recombination seen in regions like the Major Histocompatibility Complex. A new tool for analyzing Single Molecule, Real-Time (SMRT) Sequencing data – Long Amplicon Analysis (LAA) – can generate highly accurate, phased, and full-length consensus sequences for multiple genes in a single sequencing run.

Motivation

Several high-throughput sequencing methods have been applied to the Human Genome Project, including whole-genome sequencing and targeted sequencing. Whole-genome sequencing has been used to identify genetic variations and to infer haplotypes across the genome. However, whole-genome sequencing is not always sufficient to resolve genetic variations in regions with high levels of genetic diversity and recombination. Targeted sequencing, on the other hand, can provide high-resolution data for specific regions of the genome. However, it is limited by the number of regions that can be targeted and the cost of sequencing.

Figure 2. Diversity of Exons and Combinations in HLA-B

Numbers above boxes denote unique CDS exons, while numbers between boxes denote the number of unique combinations with neighboring exons. Simple measures of sequence diversity significantly underestimate the complexity of sequencing and typing genes in the MHC, as an elevated rate of recombination has led to further expansion of sequence diversity by recombining already polymorphic CDS sequences. Even these numbers likely underestimate the true diversity since only ~50-40% of reference alleles in the IMGT database have a full CDS sequence. Thus, even unambiguously, error-free sequencing of individual exons is insufficient to consistently resolve the typing of individual alleles. In addition, there are hundreds of known null-alleles, and many more with altered sequence diversity significantly understate the complexity of analyzing difficult genomic loci such as the genes from the Major Histocompatibility Complex.

SMRT® Sequencing

Figure 3. SMRT Sequencing Read-Length Distribution

Distribution of reads generated from a typical SMRT Sequencing run on a Pacific® RS II using the P5/C3 chemistry. With median read lengths over 8 kb and thousands of reads over 10 kb, it becomes possible to sequence and correctly phase full-length HLA genes without cloning or manual curation.

Figure 4. SMRT Sequencing Consensus Concordance

Consistency of consensus sequences by average genome coverage from SMRT Sequencing using the P5/C3 chemistry. When sequencing exons are truly random, consensus accuracy depends only on having sufficient coverage.

By combining the longest read lengths in the industry with the highest consensus accuracy, SMRT Sequencing presents unique opportunities for analyzing difficult genomic loci such as the genes from the Major Histocompatibility Complex.

Long Amplicon Analysis

Figure 8. Diagram of Long Amplicon Analysis

If the input sequences are barcoded, subreads grouped by barcode pair and overlapped as shown, each group, subreads are filtered based on user-definable criteria for read quality and length. Subreads that pass all filters are then clustered to each other and clustered based on the results. Each resulting sub-cluster is polished with Quiver to generate a high-quality consensus [3]. Finally, the consensus sequences are collected and filtered to remove PCR artifacts.

Conclusions

The large diversity and importance of phasing for some loci in the Human Genome make analysis with traditional sequencing methods difficult. The long read lengths and high consensus accuracy of SMRT Sequencing make it well suited for analyzing such difficult loci. In pilot studies on pooled HLA class I amplicons, consensus sequences produced by Long Amplicon Analysis show evidence of error in less than 1% of sequences with the push of a button. Long Amplicon Analysis can also generate perfect, phased consensus sequences for HLA class II amplicons over 9,000 bp. HaloTools, a Python tool that wraps Long Amplicon Analysis, can generate perfect consensus sequences from complex mixtures.