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

Brett N. Bowman[1], Patrick Marks[1], N. Lance Hepler[1], Kevin Eng[1], John Harting[1], Takashi Shiina[2], Shingo Suzuki[2], Swati Ranade[1]

[1]Pacific Biosciences of California, Inc., Menlo Park, United States of America

[2]Tokai 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

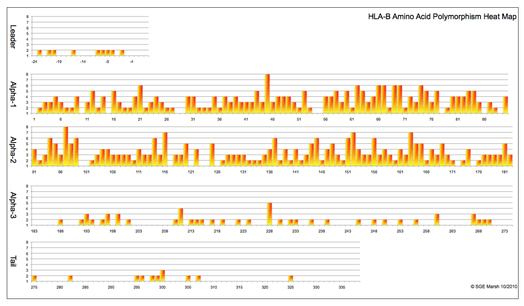


Figure 1. Variable Amino Acid Positions HeatMap of HLA-B

Across the ~2000-3500 known alleles for classical MHC Class I genes, over 2/3 of amino acid positions are polymorphic, making them among the most diverse loci in the Human Genome **[1]**.

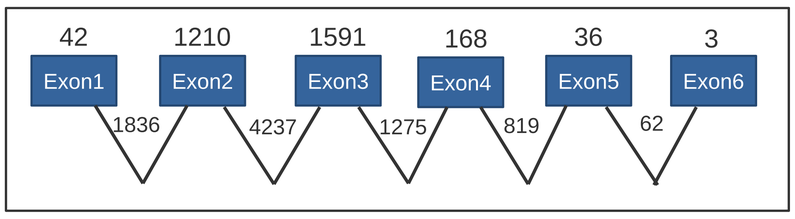


Figure 2. Diversity of Exons and Combinations in HLA-B

Numbers above Exons denote unique CDS exon sequences, while numbers between Exons denote the number of unique combinations with neighboring exons

Simple measures of sequence diversity significantly understate the complexity of sequencing and typing genes in the MHC, as an elevated rate of recombination has lead to further expansion of sequence diversity by recombining already polymorphic CDS sequences. Even these numbers likely understate the true diversity since only ~30-40% of reference alleles in the IMGT database have a full CDS sequence **[2]**. Thus even unambiguous, 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 or reduced expression levels. These biologically relevant differences can be caused by mutations occurring outside the CDS region, and by differences as small as a single nucleotide polymorphism.

Thus, the ideal method for Sequencing-Based-Typing should have the following characteristics:

- Highly accurate and unbiased consensus
- Capable of reliably phasing over spans of multiple kilobases
- Able to identify and separate alleles by differences as small as a single SNP

SMRT® Sequencing

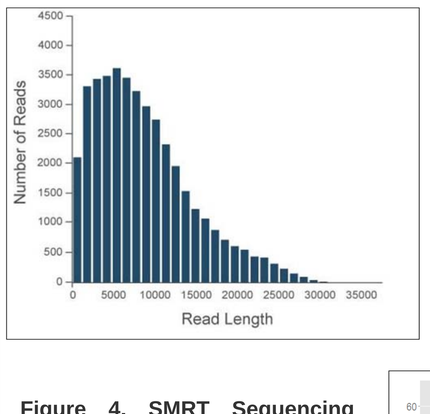


Figure 3. SMRT Sequencing Read-Length Distribution

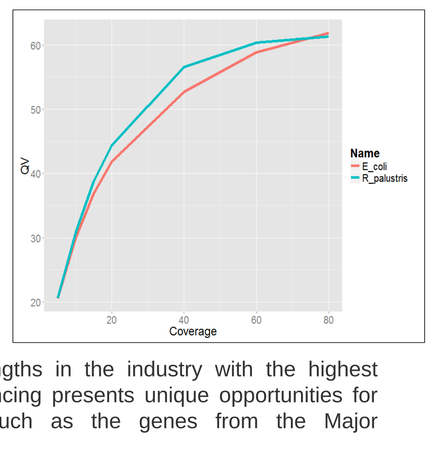
Distribution of read lengths from a typical SMRT Sequencing run on a PacBio® RS II using the P5/C3 chemistry.

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

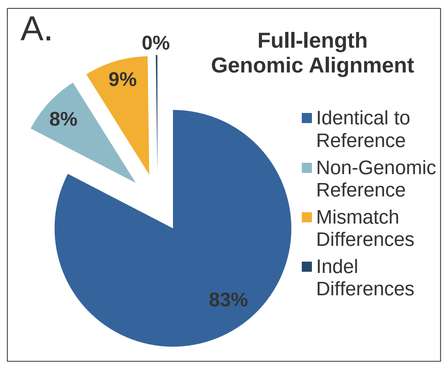
Concordance of consensus sequences by average genome coverage from SMRT Sequencing using the P5/C3 chemistry.

When sequencing errors 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.

Full-Length HLA Class I



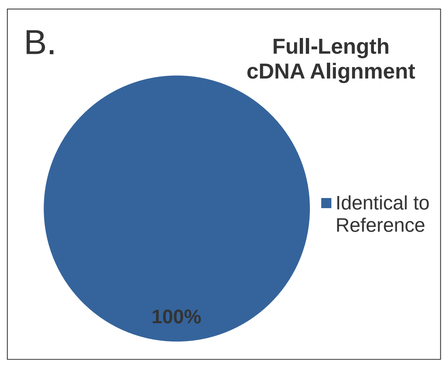


Figure 5. HLA Class IA Sequence Concordance with IMGT Database (A) Genomic (B) cDNA.

Data:

- 86 tissue samples
- Shiina et al. primers **[4]**.
- 3 loci per sample (258 total loci analyzed)
- 461 identified alleles

Results:

- 382 identical to IMGT genomic reference
- 37 had no genomic ref.
- 41 novel intronic variants
- 1 sequencing error
- All 460 full-length cDNAs identical to IMGT ref.

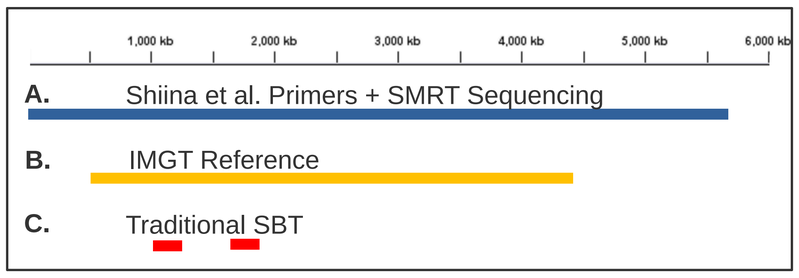


Figure 6. Allele Sequence Coverage Comparison for HLA-A

Allele sequences for HLA-A as generated by three different sequencing approaches: (A) Amplification with the Shiina et al. primers and sequenced with the PacBio RS II, (B) IMGT reference sequences, generated by amplification, cloning and tiled Sanger sequencing, (C) Traditional SBT approach, using Sanger sequencing of exons #2 and #3.

Full-length HLA Class II

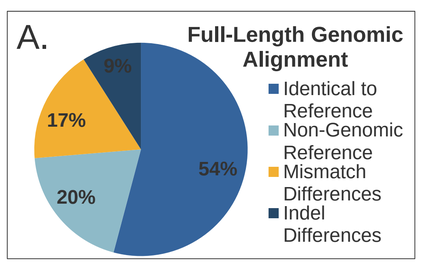
Figure 7. HLA Class II Sequence Concordance with IMGT Database (A) Genomic (B) cDNA.

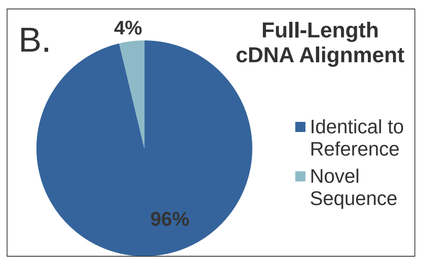
Data:

- 81 tissue samples
- 9 kb full-length DQB1
- 133 identified alleles

Results:

- 72 identical to IMGT genomic reference
- 128 full-length cDNAs identical to IMGT ref.
- 5 novel cDNA sequences





Long Amplicon Analysis

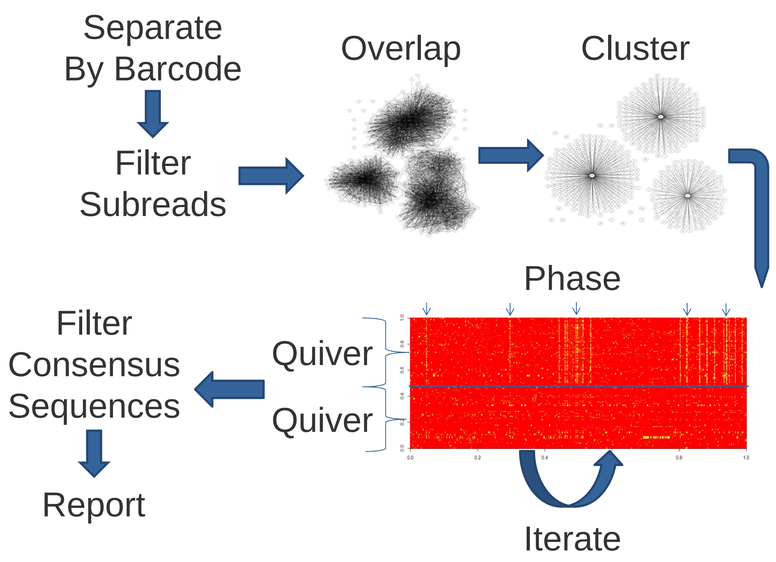


Figure 8. Diagram of Long Amplicon Analysis

If the input sequences are barcoded, subreads grouped by barcode pair and processed independently. Within each group, subreads are filtered based on user-definable criteria for read quality and length. Subreads that pass all filters are then aligned to each other and clustered based on the results. Each cluster is iteratively "phased" by identifying and separating subreads based on high-scoring mutations. 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.

Mixed Long Amplicon Analysis

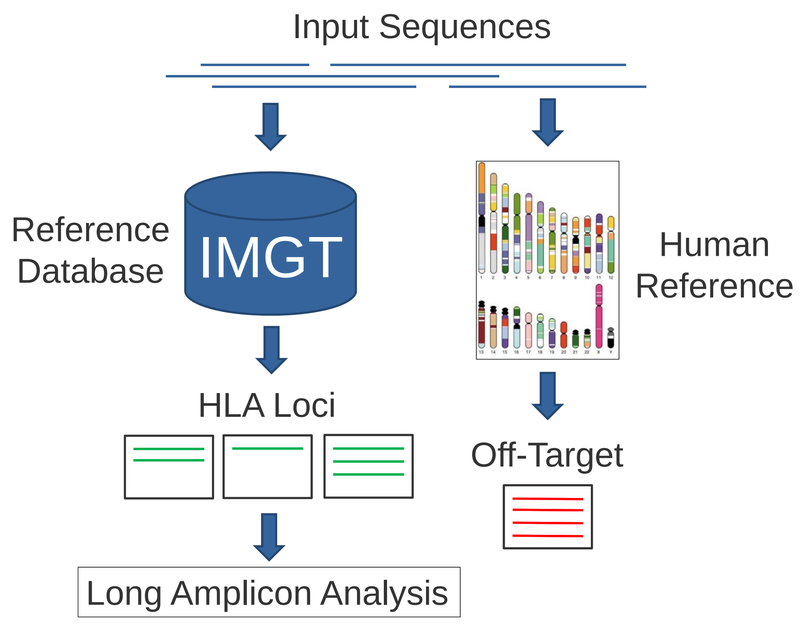


Figure 9. Diagram of HLA Tools Workflow

Experimental workflow engine implemented in HlaTools[5] for analyzing HLA Class II and complex HLA mixtures with Long Amplicon Analysis.

Locus	Size	TU01	TU02	TU03	TU04	TU05
HLA-A	~5,500bp	A*02:06:01	A*02:01:01:01	A*24:02:01:01	A*02:06:01	A*26:01:02
		A*11:01:01	A*31:01:02	A*31:01:02	A*02:07:01	A*31:01:02
HLA-B	~4,600bp	B*40:02:01	B*51:02:01	B*07:02:01	B*40:02:01	B*15:01:01:01
		B*55:02:01	B*56:01:01	B*35:01:01:02	B*44:03:01	B*35:01:01:02
HLA-C	~4,800bp	C*01:02:01	C*01:02:01	C*03:03:01	C*03:03:01	C*03:04:01:02
		C*03:03:01	C*03:04:01:02	C*07:02:01:03	C*14:03	C*07:02:01:04
HLA-DPA1	~9,700bp	DPA1*01:03:01	DPA*02:02:02	DPA1*02:01:01	DPA1*02:02:02	DPA1*01:03:01
		DPA1*01:03:01:05	*N/A*	DPA1*01:03:01:05	*N/A*	DPA1*01:03:01
HLA-DPB1	~12,300bp	DPB1*02:01:02	DPB1*02:01:02	DPB1*14:01	DPB1*05:01:01	DPB1*02:01:02
		DPB1*03:01:01	DPB1*05:01:01	DPB1*04:02:01:02	DPB1*03:01:01	DPB1*04:01:01
HLA-DQA1	~7,500bp	DQA1*01:02:01	DQA1*01:04:01	DQA1*01:01:01	DQA1*01:04:01:02	DQA1*01:02:01
		DQA1*03:02	DQA1*03:02	DQA1*01:04:01	DQA1*03:03:01	DQA1*03:02
HLA-DQB1	~9,100bp	DQB1*03:03:02	DQB1*05:03:01:02	DQB1*05:01:01	DQB1*04:02:01	DQB1*03:03:02:02
		DQB1*06:02:01	*N/A*	DQB1*05:03:01:02	DQB1*05:03:01:01	DQB1*06:04:01
HLA-DRB1	~11,500-16,000bp	DRB1*09:01:02	DRB1*09:01:02	DRB1*01:01:01:01	DRB1*04:10:NEW	DRB1*09:01:02
		DRB1*15:01:01	DRB1*14:05:01	DRB1*14:05:01	DRB1*14:54:01	DRB1*13:02:01

Locus	Size	TU06	TU07	TU08	TU09	TU10
HLA-A	~5,500bp	A*33:03:01	A*02:03:01	A*24:02:01:01	A*02:01:01:01	A*11:01:01
		A*26:03:01	A*24:02:01:01	A*33:03:01	A*02:06:01	A*31:01:02
HLA-B	~4,600bp	B*15:11:01	B*38:02:01	B*44:03:01	B*40:06:01:01	B*40:01:02
		B*44:03:01	B*54:01:01	B*48:01:01	B*48:01:01	B*51:01:01
HLA-C	~4,800bp	C*03:03:01	C*01:02:01	C*08:03:01	C*08:01:01	C*07:02:01:01
		C*14:03	C*07:02:01:05	C*14:03	C*15:02:01	C*15:02:01
HLA-DPA1	~9,700bp	DPA1*01:03:01	DPA1*02:NEW	DPA1*01:03:01	DPA1*01:03:01	DPA1*02:02:02
		DPA1*02:02:02	DPA1*02:01:01	DPA1*02:02:02	DPA1*02:02:02	DPA1*02:02:02
HLA-DPB1	~12,300bp	DPB1*02:01:02	DPB1*19:01	DPB1*04:01:01	DPB1*47:01	DPB1*05:01:01
		DPB1*02:01:02	DPB1*13:01:01	DPB1*05:01:01	DPB1*05:01:01	DPB1*05:01:01
HLA-DQA1	~7,500bp	DQA1*01:02:01	DQA1*01:03:01	DQA1*01:02:01	DQA1*01:04:01:01	DQA1*03:02
		DQA1*03:03:01	DQA1*03:01:01	DQA1*01:02:02	DQA1*01:04:01:02	*N/A*
HLA-DQB1	~9,100bp	DQB1*04:01:01	DQB1*03:02:01	DQB1*05:02:01	DQB1*05:03:01:02	DQB1*03:03:02:02
		DQB1*06:04:01	DQB1*06:01:01	DQB1*06:04:01	*N/A*	*N/A*
HLA-DRB1	~11,500-16,000bp	DRB1*04:05:01	DRB1*04:03:01	DRB1*13:02:01	DRB1*14:05:01	DRB1*09:01:02
		DRB1*13:02:01	DRB1*08:03:02	DRB1*16:02:01	*N/A*	DRB1*12:01:01

Table 1. Sequence Comparison for Mixed HLA Sequences.

Comparison of typing results for 10 published samples amplified and originally sequenced on other platforms[4]. Green – Matches existing genomic reference; Dark Blue – Matches existing cDNA reference; Light Blue – Differences in intronic homopolymer with existing genomic reference; Purple – Novel allele sequence missed by short-read sequencing.

- 10 mixed samples with 10 HLA amplicons from 8 HLA loci
- 153 alleles were compared (117 genomic, 36 cDNA only)
- All sequences matched at the cDNA level, while ~70% (80/117) of sequences and 96% (54/57) of Class IA matched at the genomic level
- Unresolved differences were entirely composed of indels in intronic single- and di-nucleotide repeat regions.
- SMRT Sequencing also identified 7 new alleles previously missed, including 1 at the 4-digit level.

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 IA amplicons, consensus sequences produced by Long Amplicon Analysis show evidence of error in <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
- HlaTools, a Python tool that wraps Long Amplicon Analysis, can generate perfect consensus sequences from complex mixtures

References

[1] http://hla.alleles.org/alleles/heat_maps.html

[2] Robinson, James, et al. "the IMGT/HLA database." *Nucleic acids research*41.D1 (2013): D1222-D1227.

[3] Chin, Chen-Shan, et al. "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data." *Nature methods* 10.6 (2013): 563-569.

[4] Shiina, T., et al. "Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers." *Tissue antigens* 80.4 (2012): 305-316.

[5] https://github.com/bnbowman/HlaTools