Poster No. 688F

Allele-Level Sequencing and Phasing of Full-length HLA Class I and II Genes Using SMRT® Sequencing Technology

Swati Ranade[1], Jason Chin[1], Brett Bowman[1], Kevin Eng[1], Patrick Marks[1], Shingo Suzuki[2] , Yuki Ozaki[2], and Takashi Shiina[2]

[1]Pacific Biosciences of California, Inc., 1380 Willow Road, Menlo Park, CA, 94025, USA

[2] Department of Molecular Life Sciences, Tokai University School of Medicine, Isehara, Kanagawa, Japan

Introduction

The three classes of genes that comprise the MHC gene family are actively involved in determining donor-recipient compatibility for organ transplant, as well as susceptibility to autoimmune diseases via cross-reacting immunization. Specifically, Class I genes HLA-A, -B, -C, and Class II genes HLA-DR, -DQ and -DP are considered medically important for genetic analysis to determine histocompatibility. They are highly polymorphic and have thousands of alleles implicated in disease resistance and susceptibility. The importance of full-length HLA gene sequencing for genotyping, detection of null alleles, and phasing is now widely acknowledged. While DNA-sequencing-based HLA genotyping has become routine, only 7% of the HLA genes have been characterized by allele-level sequencing, while 93% are still defined by partial sequences. The gold-standard Sanger sequencing technology is being quickly replaced by second-generation, high-throughput sequencing methods due to its inability to generate unambiguous phased reads from heterozygous alleles. However, although these short, high-throughput, clonal sequencing methods are better at heterozygous allele detection, they are inadequate at generating full-length haploid gene sequences. Thus, full-length gene sequencing from an enhancer-promoter region to a 3'UTR that includes phasing information without the need for imputation still remains a technological challenge. The best way to overcome these challenges is to sequence these genes with a technology that is clonal in nature and has the longest possible read lengths. We have employed Single Molecule Real-Time (SMRT®) sequencing technology from Pacific Biosciences for sequencing full-length HLA class I and II genes.

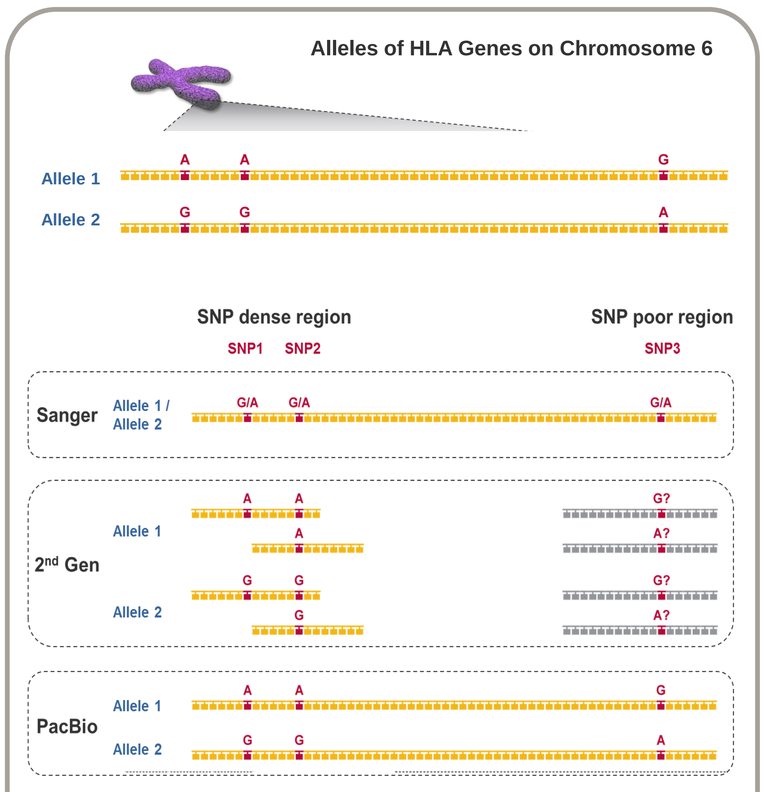
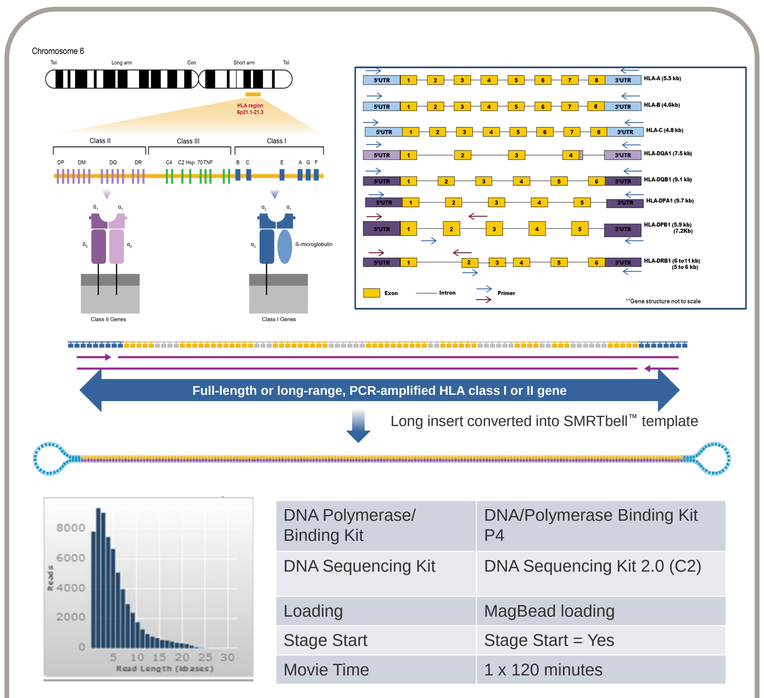


Fig 1. DNA Based HLA-genotyping scheme for various sequencing platforms

Methods

- PCR systems were developed to amplify entire HLA genes ranging between 4 kb and 10 kb as shown in Figure 2
- Equimolar mixtures of multiple amplicons of full-length HLA class I genes and long-range amplicons covering entire class II genes were converted into SMRTbell™ libraries
- Long reads ranging from ~3,500 to 20,000 bases were generated on the PacBio® RS II for all 10 samples



DNA Polymerase/ Binding Kit	DNA/Polymerase Binding Kit P4
DNA Sequencing Kit	DNA Sequencing Kit 2.0 (C2)
Loading	MagBead loading
Stage Start	Stage Start = Yes
Movie Time	1 x 120 minutes

Single SMRT Cell provides full information (all exons + all introns) for allele-level genotyping along with SNP phasing information

Fig 2. Sample preparation and sequencing

- Long amplicon analysis pipeline in SMRT Analysis vs 2.1 used to cluster and generate consensus for HLA Class I genes (4.6 to 5.5 kb).
- An alternative analysis pipeline in development for >6 kb amplicons was used for clustering and consensus generation of the Class II gene DRB1
  - Long subreads clustered using Markov clustering.
  - High-confidence phased clusters combine information from multiple SNPs instead of recursively splitting reads on each SNP individually.
  - Quiver error model used during phasing to generate consistent and high-quality consensus sequences.
- HLA genotypes were determined by comparing the consensus sequences to known Tokai University references using Sequencher ver.5.0.1 DNA sequence assembly software (Gene Code Co., MI). The genotyping was also confirmed using BLAT analysis of all exons with IMGT-HLA database.
- HLA genotypes were also determined via Conexio Genomics HLA Typing software

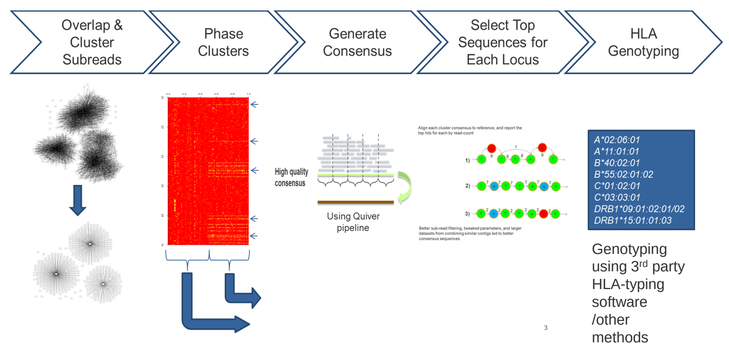


Fig 3. Long amplicon analysis followed by HLA genotyping: A pipeline for HLA allele level consensus sequence generation for unambiguous typing

Results

Fig 4. Consensus sequence generated from allele-specific cluster of continuous long subreads

- SMRT Sequencing was able to generate full-length continuous sequences for both alleles in all targeted loci for all samples.
- Unambiguous determination of genotype along with phasing information of homozygous and heterozygous alleles demonstrates the resolving power of this clonal sequencing method.
- Internal genotyping methods match Conexio results (data not shown).
- Analysis of HLA –DPA1, -DPB1,-DQA1 and –DQB1 still in process.

Sample ID	HLA-A		HLA-B		HLA-C	
	Allele1	Allele2	Allele1	Allele2	Allele1	Allele2
TU01	*A*02:06:01*	*A*11:01:01*	*B*40:02:01*	*B*55:02:01:02*	*C*01:02:01*	*C*03:03:01*
TU02	*A*02:01:01:01*	*A*31:01:02*	*B*51:02:01*	*B*56:01:01:02*	*C*01:02:01*	*C*03:04:01:02*
TU03	*A*24:02:01:01*	*A*31:01:02*	*B*07:02:01*	*B*35:01:01:02*	*C*03:03:01*	*C*07:02:01:03*
TU04	*A*02:06:01*	*A*02:07:01*	*B*40:02:01*	*B*44:03:01*	*C*03:03:01*	*C*14:03*
TU05	*A*26:01:01*	*A*31:01:02*	*B*15:01:01:01*	*B*35:01:01:02*	*C*03:04:01:02*	*C*07:02:01:04*
TU06	*A*26:03:01*	*A*33:03:01*	*B*15:11:01*	*B*44:03:01*	*C*03:03:01*	*C*14:03*
TU07	*A*02:03:01*	*A*24:02:01:01*	*B*38:02:01*	*B*54:01:01*	*C*01:02:01*	*C*07:02:01:05*
TU08	*A*24:02:01:01*	*A*33:03:01*	*B*44:03:01*	*B*48:01:01*	*C*08:03:01*	*C*14:03*
TU09	*A*02:01:01:01*	*A*02:06:01*	*B*40:06:01:01*	*B*48:01:01*	*C*08:01:01*	*C*15:02:01*
TU10	*A*11:01:01*	*A*31:01:02*	*B*40:01:02*	*B*51:01:01*	*C*07:02:01:01*	*C*15:02:01*
TU21	*A*03:02:01*	*A*24:02:01:01*	*B*07:02:01*	*B*13:02:01*	*C*06:02:01:01*	*C*07:02:01:03*

Sample ID	*HLA-DRB1*	
	Allele1	Allele2
	Allele name	Allele name
TU01	*DRB1*09:01:02:01/02*	*DRB1*15:01:01:03*
TU02	*DRB1*09:01:02:02*	*DRB1*14:05:01:02*
TU03	*DRB1*01:01:01*	*DRB1*14:05:01:02*
TU04	*DRB1*04:10:03:01*	***DRB1*14:54:01:02***
TU05	*DRB1*09:01:02:01*	*DRB1*13:02:01:02*
TU06	*DRB1*04:05:01:01*	*DRB1*13:02:01:02*
TU07	*DRB1*04:03:01:02*	*DRB1*08:03:02:02*
TU08	*DRB1*13:02:01:02*	*DRB1*16:02:01:02*
TU09	*DRB1*14:05:01:02*	–
TU10	*DRB1*09:01:02:01*	*DRB1*12:01:01:02*
TU21	*DRB1*01:01:01*	*DRB1*07:01:01:01*

- 100% concordance with the cDNA references
- One mismatch in intron 2 of TU04 *DRB1*14:54:01:02 compared to SS-SBT generated reference*
- *Alleles were correctly assigned based on PacBio consensus sequences resolving all the ambiguities in the PCR-SSO typing of these samples*

Fig 5. HLA class I (A, B and C) and class II gene (DRB1) typing by comparison to Tokai University reference

Conclusion

- SMRT Sequencing of full-length HLA genes on the PacBio RS II opens a new era for allele-level high-resolution HLA typing.
- The ability to generate long haploid reads for accurate genotyping along with phasing makes the data well suited for detection of novel alleles.
- Consensus sequences generated from the PacBio analysis pipelines can be input into third party HLA-genotyping software from Conexio Genomics to make HLA-typing calls.

Reference:

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 sequencersTissue Antigens. 2012 Oct;80(4):305-16. doi: 10.1111/j.1399-0039.2012.01941.x. Epub 2012 Aug 4.

Acknowledgements:

The authors thank Damian Goodridge and David Sayer, Conexio Genomics