Multiplexing Human HLA Class I & II Genotyping with DNA Barcode Adapters for High Throughput Research

Swati Ranade[1], Walter Lee[1], John Harting[1], Phil Jiao[1], Tyson Bowen[1], Kevin Eng[1], Lance Hepler[1], Brett Bowman[1], Nienke Westerink[2]

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

[2]GenDx, Utrecht, Netherlands

Abstract

Human MHC class I genes HLA-A, -B, -C, and class II genes HLA-DR, -DP and -DQ, play a critical role in the immune system as major factors responsible for organ transplant rejection. The have a direct or linkage-based association with several diseases, including cancer and autoimmune diseases, and are important targets for clinical and drug sensitivity research. HLA genes are also highly polymorphic and their diversity originates from exonic combinations as well as recombination events. A large number of new alleles are expected to be encountered if these genes are sequenced through the UTRs. Thus allele-level resolution is strongly preferred when sequencing HLA genes. Pacific Biosciences has developed a method to sequence the HLA genes in their entirety within the span of a single read taking advantage of long read lengths (average >10 kb) facilitated by SMRT® technology. A highly accurate consensus sequence (≥99.999 or QV50 demonstrated) is generated for each allele in a *de novo* fashion by our SMRT Analysis software. In the present work, we have combined this imputation-free, fully phased, allele-specific consensus sequence generation workflow and a newly developed DNA-barcode-tagged SMRTbell™ sample preparation approach to multiplex 96 individual samples for sequencing all of the HLA class I and II genes. Commercially available NGS-go® reagents for full-length HLA Class I and relevant exons of class II genes were amplified for hi-resolution HLA sequencing. The 96 samples included 72 that are part of UCLA reference panel and had pre-typing information available for 2 fields, based on gold standard SBT methods. SMRTbell™ adapters with 16 bp barcode tags were ligated to long amplicons in symmetric pairing. PacBio sequencing was highly effective in generating accurate, phased sequences of full-length alleles of HLA genes. In this work we demonstrate scalability of HLA sequencing using off the shelf assays for research applications to find biological significance in full-length sequencing.

HLA Sequencing on PacBio® RS II

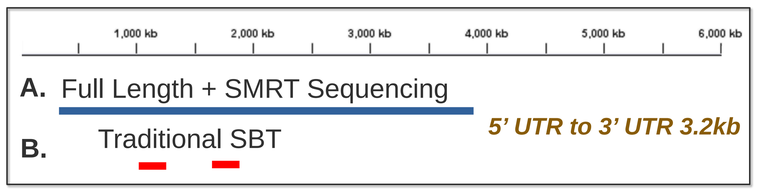

Figure 1. Allele Sequence Coverage Comparison

A)HLA-A amplified using NGS-go® Reagent & Sequenced on the PacBio RS II

B)Traditional Sanger sequencing (SBT) of exons #2 and #3.

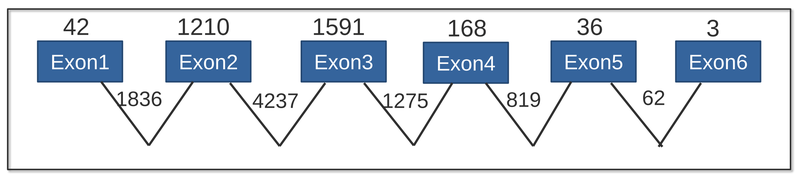

Figure 2. HLA-B Diversity due to Exon Combinations

A) Numbers above Exons denote unique coding sequences (CDS) or variants of exons,

B) Numbers between Exons denote the number of unique combinations with neighboring exons

Recombination events increase diversity in HLA genes, thus Exon-only sequencing is insufficient for resolving variation in new alleles. Additional variation is sometimes caused by mutations occurring in exons other than exon 2 & 3 CDS or sequences outside of the CDS. Thus, fully phased, allele-level genotyping with phasing across all exons and introns for accurate SNP determination in a single read span is highly advantageous for determining true heterozygosis between alleles.

Multiplexing Strategy

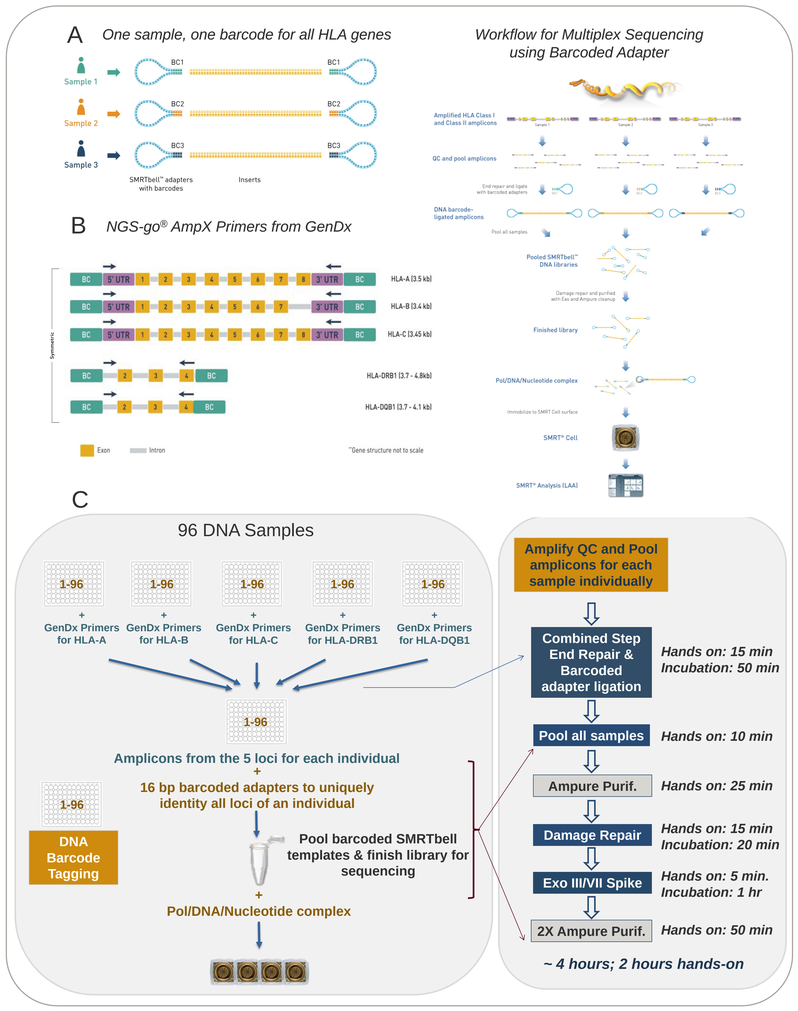

Figure 3. Multiplexing Strategy and Workflow

A] Per Patient Barcoding Scheme

B] Amplicons generated by GenDx HLA kits & symmetric SMRTbell™ adapters with barcodes, used for sample wise 16 bp barcode tagging. Barcodes are attached to the adapters for use with off-the-shelf assays

C] Barcoded adapters and template preparation set up for multiplexed sample sequencing

Automated SMRT® Analysis

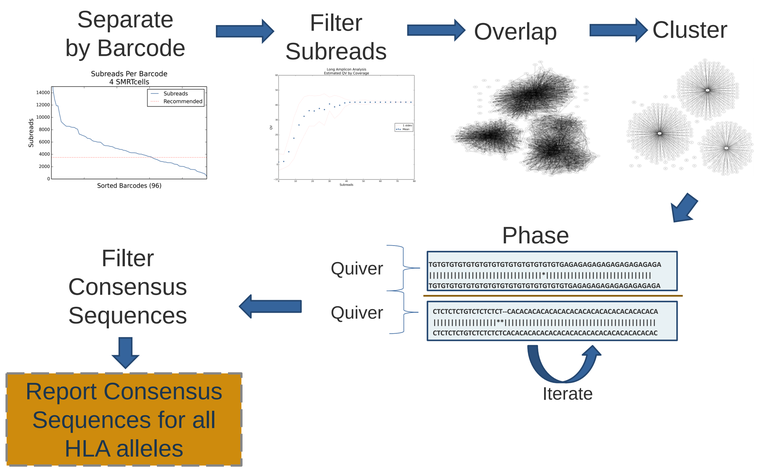

Figure 4. Diagram of Long Amplicon Analysis (LAA)

Subreads for 96 samples were grouped by barcode identity. Each group was independently processed and subreads were filtered based on user-definable criteria for read quality and length. All filtered subreads in a group are aligned and clustered based on the sequence similarity and iteratively "phased" by identifying and separating subreads according to high-scoring mutations. Each resulting sub-cluster from this *de novo* pipeline is polished with Quiver to generate high-quality consensus sequences which are filtered to remove artifacts.

Results

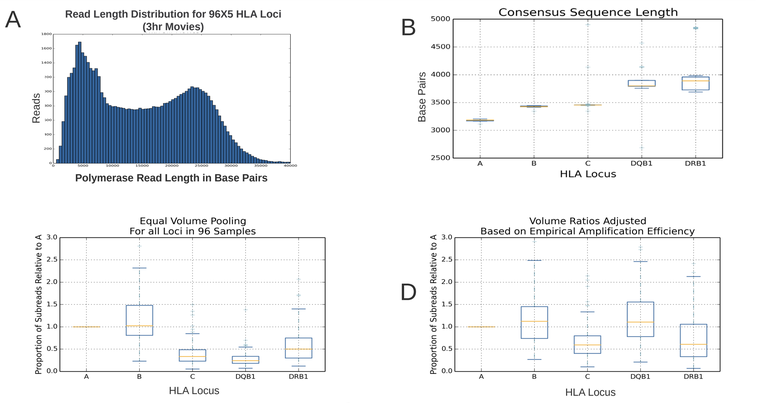

Figure 5. Qualitative analysis of SMRT® Sequencing data for HLA amplicons across all samples

A] Average polymerase read length distribution across all amplicons in all samples

B] Consensus sequence length for each HLA gene in concordance with expected full-length amplicon size

C] Unbalanced pool (without molar ratio considerations) created by equal volume pooling of HLA amplicons (HLA- A, -B, -C, -DQB1, -DRB1) across all samples

D] Balanced pooling by global adjustment of volume rations in comparison to HLA-A. Ratios were derived from empirical observations in sequencing data of unbalanced pools.

Individual amplification efficiencies on a per sample basis were not considered.

Locus	Total alleles expected	Data generated via automated LAA analysis	Alleles mis-identified as chimera	High-quality calls discordant to orthogonal typing	Missed alleles found with extra coverage	Missed calls due to severe allelic imbalance in PCR products	Adjusted rate for detection of all alleles from (4 cells)
A	177	172	1	4	0	0	100.0
B	180	177	0	3	0	0	100.0
C	175	167	0	4	4	0	97.7
DRB1	171	159	0	9	1	2	98.2
DQB1	178	170	0	2	4	2	96.6
Total	881	845	1	22	9	4	98.5

Table 1. 96 Samples interrogated at 5 loci
(Class I genes: HLA-A, -B, -C
Class II genes: HLA -DRB1, & -DQB1)

Comparison of PacBio allele calls generated using internal tools to pre-typing data from orthogonal method (Data based on unbalanced equal volume pools across amplicons and samples)

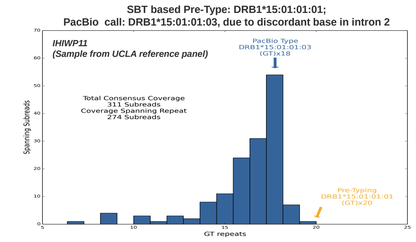

Figure 6. Example of discordant call for one allele from pre-typed reference sample IHIWP11

The PacBio call 15:01:01:03 is supported by consensus sequence derived from high-quality reads as well as high coverage. The second allele was concordant and called as DRB1*14:07:01

Conclusion

- High quality *de novo* consensus sequences (QV >50) were generated using reads >3 kb in automated LAA pipeline
- 868 of the 881 expected HLA alleles were detected in LAA output
- 13 alleles missed by automated analysis came from the 14 samples with <1500 subreads per sample
- 9 of the 13 alleles could be retrieved with manual intervention or additional SMRT Cell sequencing.
- Balanced pooling volume ratios improved subread coverage for all loci on a global level (Figure 5D)
- The low coverage samples (<800 reads) continued to stay undetectable in automated analysis even after balanced pooling
- Low coverage samples due to poor PCR amplification need systematic pooling using QC data
- 22 discordant calls will be evaluated in future for validation of discordant sequences.

References :

[1] Robinson, James, et al. "the IMGT/HLA database." *Nucleic acids research*41.D1 (2013): D1222-D1227.
[2] Chin, Chen-Shan, et al. "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data." *Nature methods* 10.6 (2013): 563-569.
[3] https://github.com/bnbowman/HlaTools