Access Full Spectrum of Polymorphisms in HLA Class I & II Genes, without Imputation for Disease Association and Evolutionary Research

Swati Ranade[1], John Harting[1], Walter Lee[1], Kevin Eng[1], Lance Hepler[1], Brett Bowman[1], Raul Kooter[2] , Claudia Rebel[2], Erik Rozemuller [2], Nienke Westerink[2]

[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025
[2]GenDx, Utrecht, The Netherlands

Abstract

MHC class I and II genes are critically monitored by high-resolution sequencing for organ transplant decisions due to their role in GVHD. Their direct or linkage-based causal association, have increased their prominance as targets for drug sensitivity, autoimmune, cancer and infectious disease research. Monitoring HLA genes can however be tricky due to their highly polymorphic nature. Allele level resolution is thus strongly preferred. However, most studies were historically focused on peptide binding domains of the HLA genes, due to technological challenges. As a result knowledge about the functional role of polymorphisms outside of exons 2 and 3 of HLA genes was rather limited. There is are also relatively few full-length gene references currently available in the IMGT HLA database. This made it difficult to quickly adopt high-throughput reference reliant methods for allele-level HLA sequencing. Increasing awareness regarding role of regulatory region polymorphisms of HLA genes in disease association[1], nonetheless have brought about a revolution in full-length HLA gene sequencing. Researchers are now exploring ways to obtain complete information for HLA genes and integrate it with the current HLA database so it can be interpreted used by clinical researchers. We have explored advantages of SMRT Sequencing to obtain fully phased, allele-specific sequences of HLA class I and II genes for 96 samples using completely *De novo* consensus generation approach for imputation-free 4-field typing. With long read lengths (average >10 kb) and consensus accuracy exceeding 99.999% (Q50), a comprehensive snapshot of variants in exons, introns and UTRs could be obtained for spectrum of polymorphisms in phase across SNP-poor regions. Such information can provide invaluable insights in future causality association and population diversity research.

Workflow

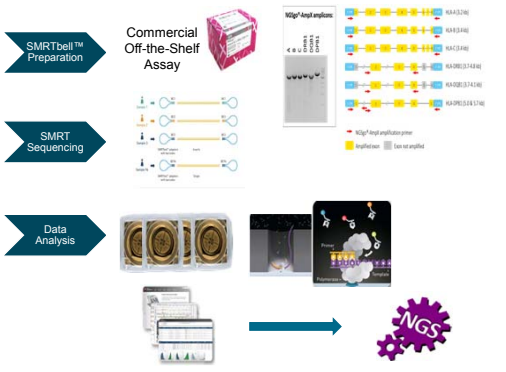

Figure 1: SMRT® Sequencing HLA genes
Multiplex Sequencing of HLA amplicons for 96 samples was carried out using Barcoded Adapters as per PacBio protocols[2]

Fully Phased Consensus Sequences Analyzed with NGSengine Software

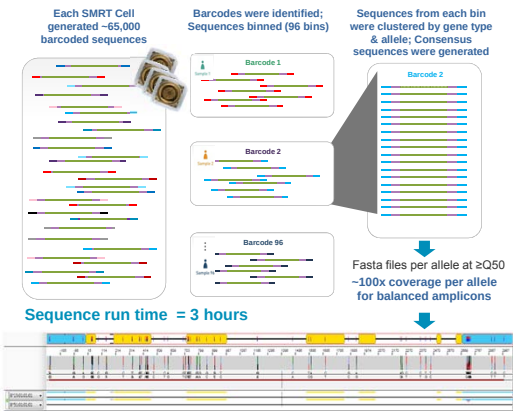

Figure 2: Consensus sequences generated in SMRT Analysis followed by HLA typing in NGSengine
Barcode separated reads are processed in a de *novo analysis* pipeline with user-definable criteria. "Phased" reads are polished with Quiver to generate high-quality consensus sequences which are then typed using NGSengine typing software.

Locus	No. of Expected Alleles	Alleles Discordant to Pre-types	Alleles Proven Correct by Orthogonal Validation		Alleles Validated for 3 Fields	% Concordance to Known Sanger Pre-types
A	175	4	4/4	100%	0	100.0
B	180	3	3/3	100%	0	100.0
C	175	4	4/4	100%	0	100.0
DRB1	161	9	7/7	100%	2	98.8%
DQB1	177	2			2	98.3%
Total	**868**	**22**	**18/18**	**100%**	**4**	**99.4%**

Table 1: Comparison of PacBio sequences typed with NGSengine software for 96 samples with pre-typed data

- *863 of 868 correctly identified unique allele types*
- *Homozygous types were considered single calls*
- *Orthogonal sequencing with second-generation or Sanger sequencing validated 18 discordant PacBio calls as correct calls. Two of the DQB1 alleles were missed due to PCR allele drop-out issue*

Advantages of Long Reads

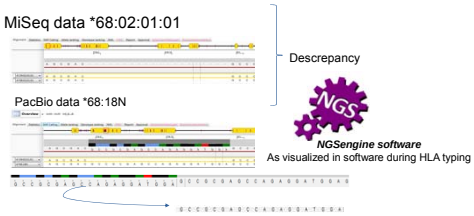

Figure 4A: Unambiguous phasing of alleles containing tandem duplications within exon 2
Long read lengths make unambiguous allele segregation and typing through tandem duplications stress free

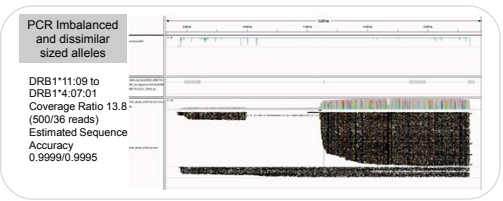

Figure 4B: **Heterozygous Phased Alleles of DRB1*04**
De novo consensus of alleles from individual reads spanning the amplicon length allows accurate detection of differentially sized alleles

SMRT HLA Sequence Characterization

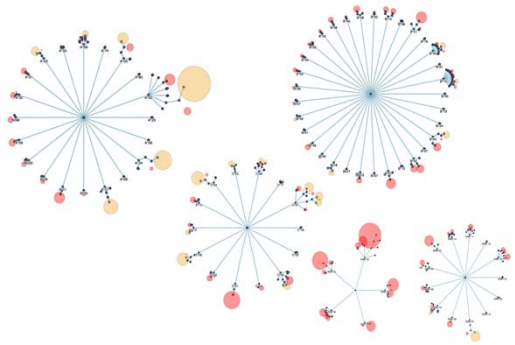

Figure 5: 96 Sample Dendrograms – HLA Class I & II
IMGT Type (best level we could reach)
4-Field Type (genomic reference available in database)
<4-Field Type (no genomic data in database)

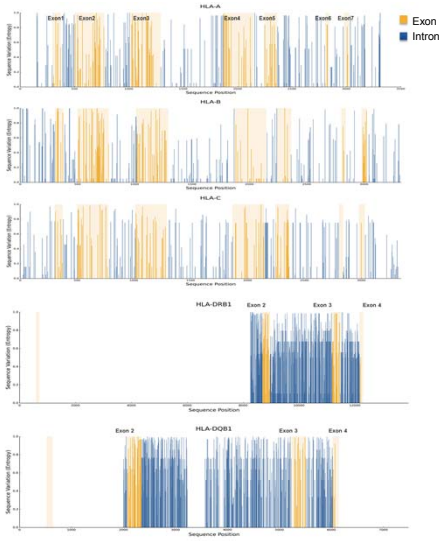

Figure 6: Polymorphisms across HLA Class I full-length genes and class II exon 2 to 4 in 96 samples

Conclusion

- *De novo* consensus sequences (QV >50) from LAA pipeline contain 4-field typing information
- DRB alleles validated only up to 3rd field due to lack of references in IMGT data base and limitations of orthogonal sequencing methods
- Sample dendrograms illustrate the state of gold standard reference sequences available for each HLA gene interrogated
- Allele-level sequencing will elucidate new polymorphisms in regulatory regions that may provide insights in causal variant analysis
- Reference-free analysis is key requirement for error-free phased HLA data without imputation

References :

[1] Kazuyoshi Hosomichi, et.al. The impact of next-generation sequencing technologies on HLA research., 2015, Journal of Human Genetics advance online publication 27 August 2015; doi: 10.1038/jhg.2015.102

[2] Ranade S., et al. (2015) Multiplexing Human HLA Class I & II Genotyping with DNA Barcode Adapters for High Throughput Clinical Research. Advances in Genome Biology & Technology Conference (2015 AGBT) Poster Presentation (Abstract 178)