Immune Regions Are No Longer Incomprehensible with SMRT Sequencing

Swati Ranade[1], Jason Chin[1], Steve Kujawa[1], Paul Peluso[1], John Harting[1], Lawrence Hon[1], Richard Hall[1], Primo Baybayan[1], Kevin Eng[1], Julie Karl[2], Michael Graham[2], Roger Wiseman[2], David O'Connor[2], Daniel E.Geraghty[3], Chul-Woo Pyo[3], Jason Howard[4], Erich Jarvis[4] & Jonas Korlach[1]

[1]PacBio, 1380 Willow Road, Menlo Park, CA; [2] University of Wisconsin, Madison, WI; [3]Fred Hutchinson Cancer Research Center, Seattle, WA; [4]Howard Hughes Medical Institute & Duke University, Durham, NC

Abstract

Fine mapping of causal immune-gene variants for their known association with cancer, drug-induced hypersensitivity and autoimmune diseases, are of paramount importance[1]. As population-scale medical genomics initiatives' collection of short read whole genome and exome sequencing (WGS)/(WES) data and SNP array data for Genome Wide Association Studies (GWAS) has increased, the development of imputation software for analysis of immune gene variants has been rampant.

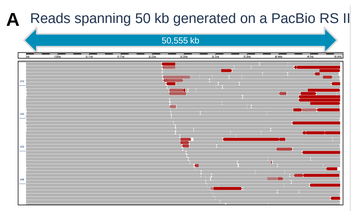
Despite increased efforts, fine mapping has yielded few discoveries[1]. This may be because extraction of haplotype information in these complex immune regions characterized by excessive copy number variants (CNVs), high-density hyper-polymorphic genes, pseudo genes, conserved extended haplotypes (CEH) with enormous linkage disequilibrium (LDs), and long stretches of repetitive DNA is challenging for any reference-reliant, short-read sequencing or GWAS methods[2].

We present detailed views of human and non-human Killer Immune Receptor (KIR) and Major Histocompatibility Complex (MHC) genes and/or haplotypes, captured by SMRT Sequencing. High accuracy, long reads and uniform coverage from SMRT Sequencing allowed allele-level segregation of sequences with full phasing across SNP-poor regions. Using either whole-genome or targeted sequencing approaches, we were able to carry out *de novo* analysis, including avian species.

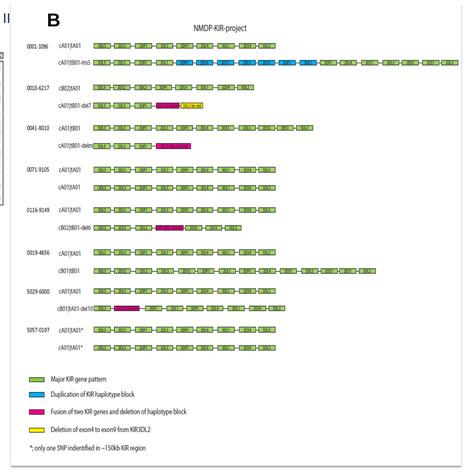
Human KIR Sequencing

DE NOVO ASSEMBLY OF COMPLETE KIR HAPLOTYPES

- Unambiguous assembly of KIR regions has been difficult to impossible.
- Long reads in combination with a recombineering sample-preparation approach for isolating a tiling of targeted fosmids containing contiguous genomic regions (~35 kb – 50 kb)[3] allowed *de novo* assembly of individual KIR haplotypes with high confidence[4].
- 16 such haplotypes were assembled from 8 individuals (available in NCBI)
- A 9 kb partial reference, available for homozygous sample 0071-9105 in the IMGT database, was 100% concordant with the PacBio sequence.



~50 full-length reads w/ many partials >20 kb per SMRT Cell from a linearized fosmid SMRTbell library



A: Subreads of a full-length fosmid library sequenced on PacBio RS II mapped to a 50 kb read spanning a known fosmid sequence

B: 16 completely phased KIR haplotypes from 8 individuals providing gene order, copy number variants, and genotypes

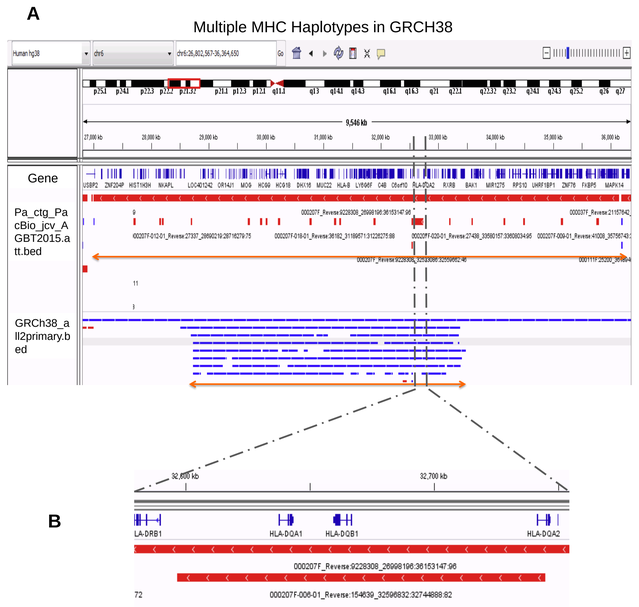
References

1. Trowsdale J. and Knight JC., (2013) Major histocompatibility complex genomics and human disease. Annual Review Genomics Human Genetics. 14, 301-323.
2. Brandt, D. Y. C., et al., (2015) Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data. 5(5), 931-941.
3. Pyo, Chul-Woo et al.; AGBT (2015) Complete resequencing of extended genomic regions using fosmid target capture and single molecule real-time (SMRT) long read sequencing technology.
4. Hall, Richard J. et al.; AGBT (2015) Assembly of complete KIR haplotypes from a diploid individual by the direct sequencing of full-length fosmids.
5. Chin, J. et al. AGBT (2015) Toward comprehensive genomics analysis with *de novo* assembly.
6. Ranade, S. et al. AGBT (2015) Access full spectrum of polymorphisms in HLA class I & II genes, without imputation for disease association and evolutionary research.
7. DNA oligo hybridization with Nimblegen SeqCap EZ System (Application Note)
8. Zhang et al Science (2015) Comparative genomics reveal insights into avian genome evolution and adaptation. Science 346(6215), 1311-1320.

Human MHC Sequencing

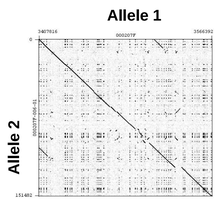
DE NOVO PHASED MHC COMPLEXES

- The MHC locus is difficult to extract from traditional human genome assemblies
- Inspected contigs corresponding to MHC in PacBio *de novo* human genome assembly, (85x coverage, FALCON assembler)[5]
- The 4 Mb MHC region of chromosome 6 was located within a single 9 Mb contig
- Haplotigs separated by diploid aware software 'Falcon', unzips and identifies 20 significant structural variations between the homologous MHC chromosomes



A: 4 Mb MHC in 9 Mb contig of PacBio assembly, compared to multiple MHC haplotypes in GRCh38

B & C: 151 kb HLA-DQA1/DQB1 heterozygous CEH haplotype block automatically phased by FALCON. The gene HLA-DQB1 is homozygous (02:02:01), however the heterozygosity in HLA-DQA1 (01:05 and 02:01) is sufficient to phase haplotypes



PHASING MHC GENES BY TARGETED SEQUENCING

- MHC regions from the above sample were interrogated with two orthogonal targeted sequencing approaches
- 8 HLA genes were amplified and sequenced using GENDX NGS-go® workflow for PacBio[6]
- Roche NimbleGen's SeqCap EZ Enrichment strategy[7] was used to capture and sequence MHC regions
- HLA amplicon and whole genome data were 100% concordant, with a single difference in homopolymeric region of one allele

D

Locus	Alleles 1	Alleles 2	% cDNA Identity to IMGT-HLA Database	Gene Length Allele 1 / Allele 2		% Base Identity with Whole-Genome Assembly
A	HLA-A*01:01:01:01	HLA-A*02:01:01:01	100%	3138	3152	100%*
B	HLA-B*13:02:01	HLA-B*37:01:01	100%	3394	3395	100%
C	HLA-C*06:02:01:01	Homozygous	100%	3424	-	100%
DRB1	HLA-DRB1*07:01:01:01	HLA-DRB1*10:01:01	100%	3904	3654	100%
DQB1	HLA-DQB1*02:02:01	Homozygous	100%	4114	-	100%
DPA1	HLA-DPA1*01:03:01:02	Homozygous	100%	4857	-	100%
DPB1	HLA-DPB1*04:01:01	Homozygous	100%	5867	-	100%
DQA1	HLA-DQA1*01:05	HLA-DQA1*02:01	100%	5746	5657	100%

E

D: Genotyping for 8 HLA genes. All alleles were 100% concordant to PB-Whole Genome Sequencing data, except one homopolymeric (7C > 6C) base in intron of HLA A*01:01:01:01 (*)

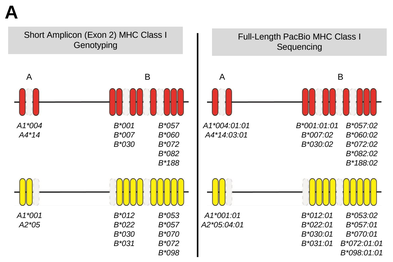
E1: Nimblegen SeqCap EZ MHC target-enriched PacBio reads aligned to HLA-DQA1 (hg19).

E2: PacBio sequences with predicted accuracy >97% phased using SAMtools (blue/pink). Consensus sequence captures larger region than PCR amplicon and useful for validating haplotype structures

Non-Human MHC Sequencing

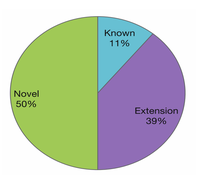
MHC ALLELE-DISCOVERY IN MACAQUES

- Macaque monkey MHC Class I regions are important for infectious disease research and more complex than humans due to gene duplications and variable number of pseudogenes
- Differing number of alleles express per haplotype, but only a few impact disease susceptibility
- SMRT-CCS PacBio reads provide full-length MHC class I cDNA sequences, which is a clear advantage for establishing robust MHC databases for Macaques



B	April 2015			November 2015		
	# Total Alleles	# Full-Length Alleles	% Full-Length	# Total Alleles	# Full-Length Alleles	% Full-Length
Cynomolgus Macaques	851	238	28%	1,020	523	51%
Pig-Tailed Macaques	297	110	37%	463	362	78%
Rhesus Macaques	765	469	61%	765*	469*	61%*

*Analysis of 64 Chinese rhesus and 32 Indian rhesus macaque samples currently underway



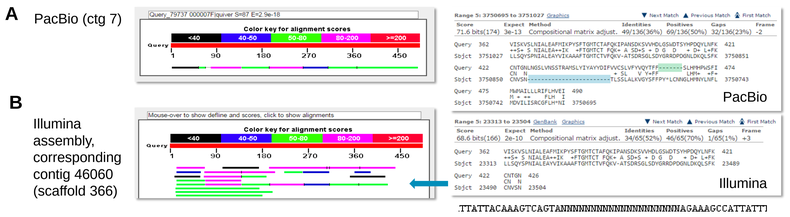
A: Short-read sequencing provide MHC types at 2 field resolution; SMRT-CCS reads give allele-level 4 field resolution

B: PacBio facilitates rapid increases in total number and % full-length alleles in the macaque databases

C: Discovered allele types within 322 full-length MHC alleles sequenced from 72 cynomolgus macaques in the Chinese breeding facilities

RESOLUTION OF AVIAN MHC REGIONS FROM WHOLE-GENOME ASSEMBLIES

- Avian immune regions are also important for understanding their disease resistance mechanism and can ultimately support avian conservation[8]
- Short-read assemblies have limitations for MHC analysis
- PacBio whole genome assembly of Anna's hummingbird using FALCON assembly method resolved several gaps in the short-read assembly



A: Full gene reconstruction of BLB2 homolog in PacBio assembly.

B: Illumina scaffold 366 with 20 'N's in the last exon. This truncates the protein prediction, and thereby also misses that in this exon, a region in the protein is deleted, and another region is inserted, relative to the perregrin falcon gene homolog.

Summary

- SMRT Sequencing provides imputation-free access to complex immune regions.
- Complete views of genotypes and haplotypes of immune regions captured in human and non-human species.
- *De novo* allows for reference-free analysis without imputation.
- These methods can be adopted to a multitude of immunology applications, including:
  - Establishing databases of gene alleles and/ or haplotypes
  - Disease association in biomedical research
  - Conservation biology

The authors acknowledge and thank everyone involved with the data generation for this poster. We also thank Martin Mayers & Cynthia Vierra-Green, National Bone Marrow Donor Program, Minneapolis, MN , for the 8 cell line samples used in the KIR haplotype sequencing project as well as for guidance provided during the project