BC Cancer Agency CARE & RESEARCH · GENOME SCIENCES CENTRE

HPTASR /ˈapiˌtīzər/

René Warren ● Gina Choe ● Sarah Munro ● Mauro Castellarin ● Richard Moore ● Robert Holt

HLA class I Predictions by Targeted Assembly of NGS Shotgun Reads

Introduction

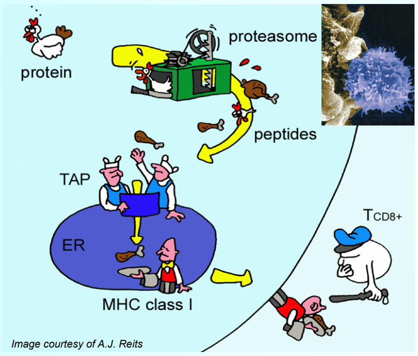


Image courtesy of A.J. Reits

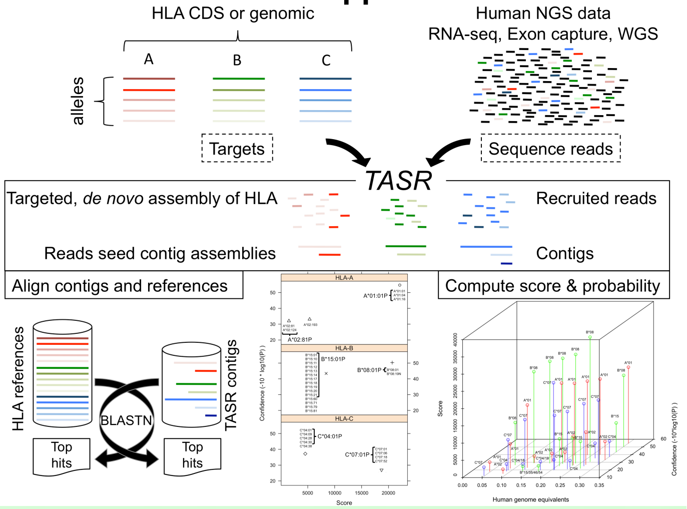
HLA class I (MHC-I)

- Human Leukocyte Antigen
- Most polymorphic alleles in the genome
- Expressed at surface of all nucleated cells
- Present altered & non-self peptides to T cells
- Major genes are A,B,C

allele
group
gene
A*02:01

- Knowing HLA is key to successful graft
- Potent vaccines ought to consider HLA
- Current HLA typing methods: $$$, laborious, time consuming
- *Can HLA alleles be predicted directly from NGS shotgun data?*

Approach



HLA CDS or genomic

Human NGS data
RNA-seq, Exon capture, WGS

Targets

Sequence reads

TASR

Targeted, *de novo* assembly of HLA

Recruited reads

Reads seed contig assemblies

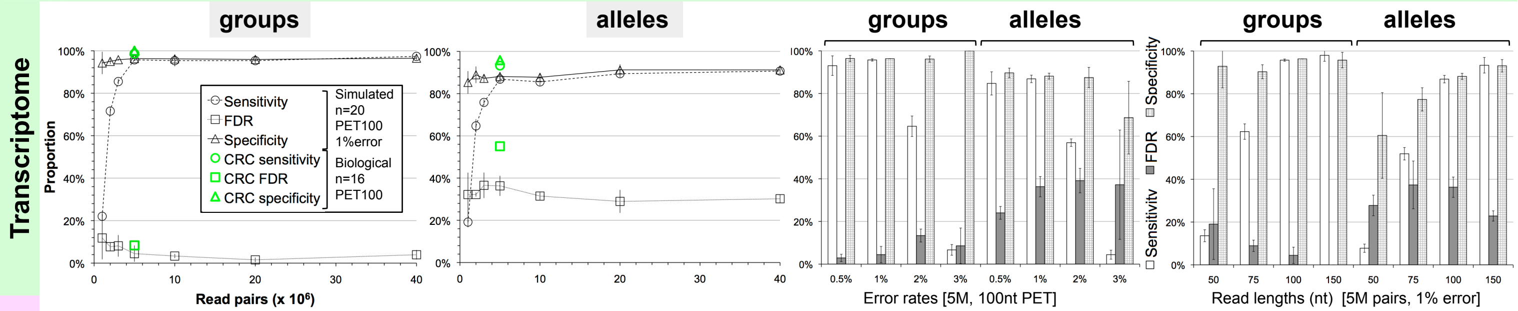
Contigs

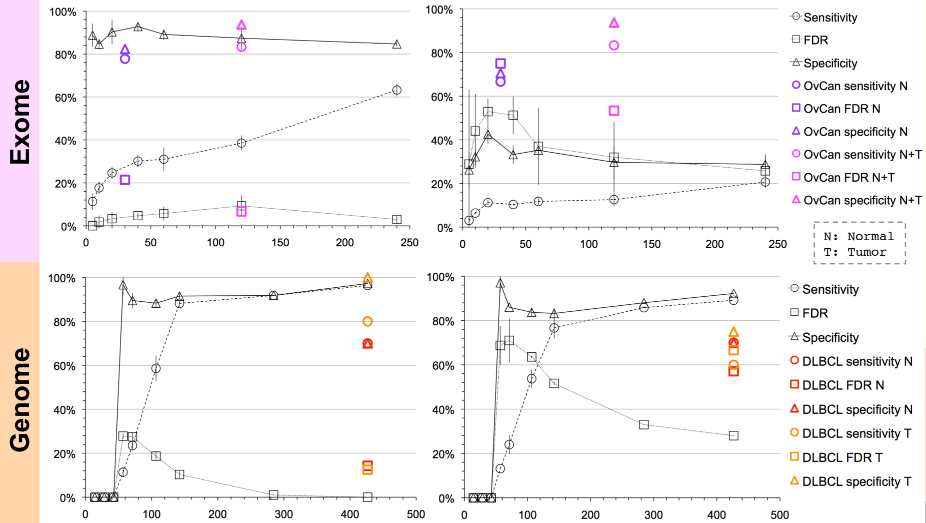
Align contigs and references

Compute score & probability

Results

Comparisons between simulated and biological NGS transcriptome, exome & genome shotgun data. The effect of read depth, length & base error on computational HLA group and alleles predictions are assessed.





Methods

SIMULATED DATA

- In separate experiments, we removed HLA CDS, exonic regions and genes from 15K randomly selected *Ensembl* (Flicek et al. 2011) transcripts, ~220K exon regions (Agilent) and the HuRef genome (Levy et al. 2007)
- For each data type, we randomly generated 20 sets of six (2xA, 2xB, 2xC) HLA alleles (Σ=120)
- In triplicate experiments, we merged each set of six sequences with HLA-less CDS, exon regions or HuRef, respectively and simulated 50, 75, 100 and 150nt PET with varied error rates using samtools' wgsim (Li et al. 2009), ran TASR (Warren and Holt 2011) and HPTASR

BIOLOGICAL DATA

- ~5M (normal+tumor samples), 100nt RNAseq PET from 65 **CRC** patients and verified the HLA alleles of 16 individuals by PCR-based typing
- ~30M (normal samples) and 120M (normal+tumor samples) 100nt exon capture PET from 3 **ovarian cancer** patients whose HLA alleles were verified by PCR-based methods
- ~430M (~30x depth coverage HG19) 100nt whole-genome shotgun (WGS) PET from normal and tumor samples of 4 **DLBCL** patients (Morin et al. 2011) validated by PCR-based HLA typing

Discussion & Conclusions

- Simulated & biological data sets both indicate HPTASR's potential for accurate HLA group and allele predictions from PET100 RNAseq data sets, using as little as 5M read pairs
- In contrast, to obtain a similar efficiency with simulated WGS data, 430M pairs are necessary
- HLA is highly expressed and its signal-to-noise ratio higher in RNAseq vs. WGS / EC data
- WGS and EC data are noisy and limit HPTASR's overall performance
- HLA pseudogenes may contribute to the noise in WGS and EC (not shown)
- As NGS read lengths continue to increase, HPTASR's HLA predictions are expected to improve

Glossary

HLA	Human Leukocyte Antigen	Name of the major histocompatibility complex (MHC) in humans, locus that contains immune system function genes
FDR	False Discovery Rate	Counted for each ambiguous HPTASR group/allele predictions (same score and probability as PCR-verified allele)
PET	Paired-End Tags	Alternate name for paired reads (both ends of a sequencing template sequenced)
EC	Exon Capture	Strategy to selectively sequence the coding regions of the genome (known as exome sequencing or targeted EC)
CRC	ColoRectal Cancer	Third most commonly diagnosed cancer in the world and second leading cause of cancer death in Canada
DLBCL	Diffuse Large B Cell Lymphoma	Aggressive non-Hodgkin lymphoma that accounts for approximately 40% of lymphomas among adults
OvCan	Ovarian Cancer	Most serious of all gynecological cancers

References

- Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. 2011. Ensembl 2011. Nucleic Acids Res. 39:D800-806.
- Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. 2007. The Diploid Genome Sequence of an Individual Human. PLoS Biol 5(10): e254. doi:10.1371/journal.pbio.0050254
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078-9.
- Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, et al. 2011 Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature. 476:298-303
- Warren RL, Holt RA. 2011 Targeted Assembly of Short Sequence Reads. PLoS ONE 6(5): e19816. doi:10.1371/journal.pone.0019816

BC Cancer Agency ● Genome Sciences Centre ● 675 W 10th Ave ● Vancouver BC V5Z 1L3 ● www.bcgsc.ca