HLA TYPING FROM NEXT GENERATION SEQUENCING DATA

E. Major, T. Hague, T. Nagy, K. Rigó, Sz. Juhos, A. Bérces*
Omixon Biocomputing – *www.omixon.com*

omixon BIOCOMPUTING SOLUTIONS
– get the right answers

Introduction

The purpose of this study was to evaluate the performance of the Human Leukocyte Antigen (HLA) typing algorithm implemented in **Omixon Target**. This software can handle any Next Generation Seqeuencing (NGS) genomic data ranging from Whole Genome Sequencing (WGS) through exome to targeted amplicon sequence data. **Omixon Target** uses different algorithms for the different NGS sequencers since their error models are different. The method is customizable to a specific sequencer and capture target.

Validation datasets

Validation datasets included public data sources such as the HapMap project as well as clinical data from the Hungarian National Blood Services. Whole genome and exome data was downloaded from the short read archive (ncbi.nlm.nih.gov/sra)
Whole genome datasets: Illumina HiSeq 2000 2x101 bp sequencing
Exome datasets: Whole exome sequencing datasets were selected from the 1000 genomes project (http://www.1000genomes.org/category/exome) and the HLA types were determined in the HapMap project (www.hapmap.org). Most exome sequencing of the reported samples were sequenced at the Broad Institute and use Agilent SureSelect exome v2 kit, and 2x76 bp Illumina sequencing.
Targeted datasets: Amplicon based amplification of three full loci and one part locus, sequenced Illumina MiSeq, 2x150 bp sequencing.
RainDance datasets: Illumia HiSeq, 2x100 bp sequencing.

RainDance HLAseq kit

	Allele 1		Allele 2		Concordance	
	Reference	Omixon	Reference	Omixon	Including Ambiguities	Unambiguous
ID: NA07348	A*02:01	A*02:01:01	A*24:02	A*24:02:01	2	2
	B*39:06	B*39:06:02	B*44:02	B*44:02:01	2	2
	C*05:01	C*05:09:01	C*07:02	C*07:02:01	1	1
ID: NA10835	A*02:01	A*02:01:01	A*23:01	A*23:01:01	2	2
	B*40:01	B*40:01:02	B*41:01	B*41:01	2	2
	C*03:04	C*03:04:01	C*17:01	C*17:01:01	2	2
ID: NA12750	A*02:01	A*02:01:01	A*31:01	A*31:01:02	2	2
	B*07:02	B*07:02:01	B*15:01	B*15:01:01	2	2
	C*03:03	C*03:03:01	C*07:02	C*07:02:01	2	2
ID: NA18504	A*36:01	A*11:50Q	A*36:01	A*36:01	1	1
	B*15:03	B*15:03:01	B*39:10	B*39:10:01	2	2
	C*02:10	C*02:10	C*12:03	C*12:03:01	2	2
TOTAL:					22	
					92%	

The analysis challenges

The HLA region of the human genome consists of significant segmental duplications, and includes a number of pseudogenes. Because this region carries the highest inherited genetic diversity in humans, the HLA genes are often more similar to their pseudogene counterparts than to the same gene in a different individual. Due to high diversity of the region and because NGS reads are shorter than the duplicated segments, correct mapping of NGS reads to a reference genome cannot be carried out accurately. For this reason most HLA typing applications map the reads directly to the genotype database.

Whole genome sequencing

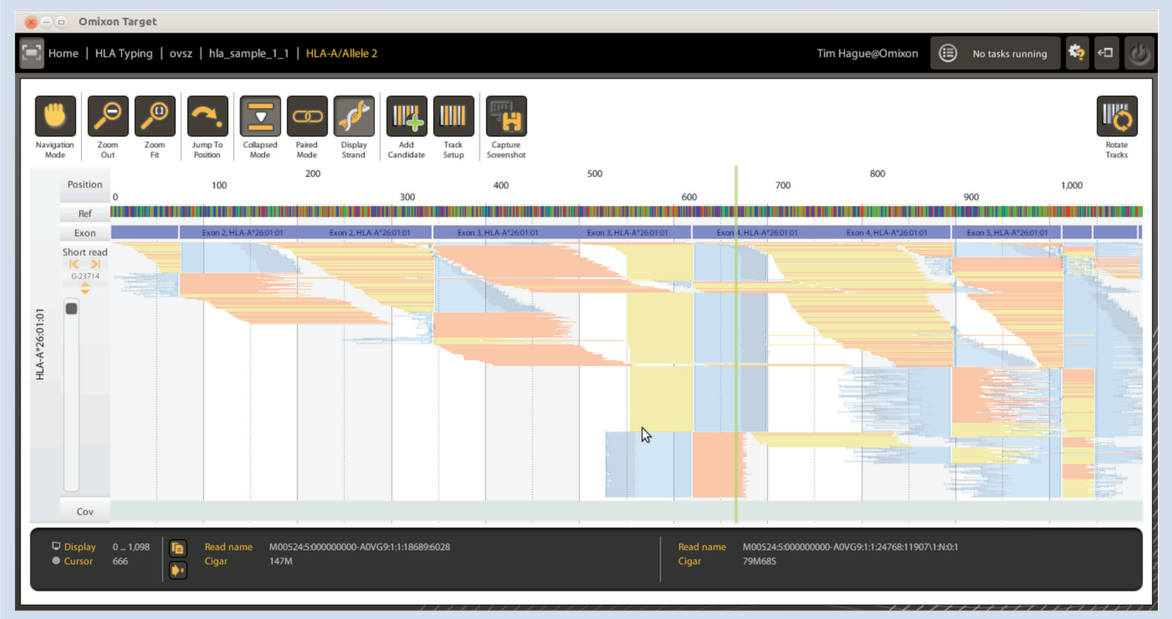
	Allele 1		Allele 2		Concordance	
	Reference	Omixon	Reference	Omixon	Including Ambiguities	Unambiguous
ID: NA06985/SRR400039	A*02:01	A*02:01:01	A*03:01	A*03:02:01, A*03:01:01	2	1
	B*07:02	B*07:02:01	B*57:01	B*07:33	1	1
	C*06:02	C*06:02:01	C*07:02	C*07:66, C*07:02:01, C*07:46	1	1
ID: NA18504/SRR350142	A*36:01	A*36:01	A*36:01	A*36:01	2	2
	B*15:03	B*15:03:01	B*39:10	B*38:01:01, B*38:02:01, B*39:08, B*39:07, B*39:13:01, B*39:13:02, B*67:01:01, B*67:01:02, B*67:02, B*39:05:01, B*38:14	1	1
	C*02:10	C*06:02:01, C*12:05, C*12:13, C*12:04:02	C*12:03	C*12:03:01, C*06:02:01, C*12:05, C*12:13, C*12:04:02	0	0
ID: NA18913/SRR350153	A*34:02	A*34:02:01	A*74:01	A*74:02:01, A*74:01, A*74:11	2	1
	B*53:01	B*51:02:02	B*53:01	B*53:01:01, B*51:02:02	1	0
	C*04:01	C*04:01:01	C*04:01	C*04:01:01	2	2
ID: NA18966/SRR359062	A*11:01	A*11:02:01, A*11:01:01	A*31:01	A*31:01:02	2	1
	B*40:01	B*40:10:01	B*51:01	B*51:04, B*51:01:01	2	1
	C*07:02	C*07:02:04	C*14:02	C*14:02:01	2	2
ID: NA18971/SRR359095	A*02:06	A*02:06:01	A*24:02	A*24:02:01	2	2
	B*40:02	B*40:02:01	B*40:02	B*40:02:01	2	2
	C*03:04	C*03:04:01	C*07:02	C*07:02:01	2	2
ID: NA18974/SRR360136	A*02:01	A*02:01:01	A*26:01	A*26:01:01	2	2
	B*39:01	B*39:01:01	B*40:01	B*40:01:02, B*40:79	2	1
	C*07:02	C*07:02:01	C*07:02	C*07:02:01	2	2
ID: NA18975/SRR359070	A*02:01	A*02:01:01	A*02:06	A*02:06:01	2	2
	B*07:02	B*07:02:01	B*15:01	B*15:01:01	2	2
	C*04:01	C*04:01:01	C*07:02	C*07:02:01	2	2
ID: NA18976/SRR359110	A*33:03	A*33:03:01	A*33:03	A*33:03:01	2	2
	B*44:03	B*44:03:01	B*44:03	B*44:03:01	2	2
	C*14:03	C*14:03	C*14:03	C*14:03	2	2
ID: NA18981/SRR359083	A*02:01	A*02:01:01	A*31:01	A*02:01:01	1	1
	B*51:01	B*51:01:01	B*51:01	B*51:01:01	2	2
	C*14:02	C*14:02:01	C*14:02	C*14:02:01	2	2
ID: NA19092/SRR189830	A*36:01	A*36:01	A*68:02	A*68:71	1	1
	B*42:01	B*42:02, B*42:01:01	B*81:01	B*81:01	2	1
	C*08:04	C*08:15:01	C*17:01	C*08:15:01	0	0
ID: NA19119/SRR359106	A*23:01	A*23:01:01	A*36:01	A*36:01	2	2
	B*35:01	B*35:01:01	B*49:01	B*49:01:01	2	2
	C*04:01	C*04:01:01	C*07:01	C*07:01:02	2	2
ID: NA19131/SRR359096	A*66:03	A*66:02	A*68:02	A*68:02:01	1	1
	B*42:01	B*42:01:01	B*53:01	B*53:01:01	2	2
	C*04:01	C*04:01:01	C*17:01	C*04:33	1	1
ID: NA19171/SRR359061	A*02:02	A*02:02	A*23:01	A*23:38N, A*23:01:01	2	1
	B*52:01	B*52:01:02	B*53:01	B*53:01:01	2	2
	C*04:01	C*04:01:01	C*16:01	C*16:01:01	2	2
ID: NA19204/SRR359064	A*02:05	A*02:05:01	A*30:02	A*30:02:01	2	2
	B*18:01	B*18:01:01	B*58:01	B*58:01:01	2	2
	C*07:01	C*07:18	C*08:02	C*08:02:01	1	1
TOTAL:					71	63
					85%	75%

Methods

Omixon Target uses a purpose-built algorithm to align the short reads to the IMGT/HLA database (1,2) and score the aligned reads in order to determine the HLA types. This algorithm is fast enough to process whole genome and whole exome data within 1 to 2 hours and targeted sequencing data under one minute on a desktop computer. The software runs on Windows, Linux and Mac OS X.

HLA typing sample results

Visualisation of HLA results



Agilent SureSelect exome kit

	Allele 1		Allele 2		Concordance	
	Reference	Omixon	Reference	Omixon	Including Ambiguities	Unambiguous
ID: NA12878/SRR098401	A*01:01	A*01:01:01	A*11:01	A*11:01:01	2	2
	B*08:01	B*08:01:01	B*56:01	B*56:01:01	2	2
	C*01:02	C*01:02:01	C*07:01	C*07:01:01	2	2
ID: NA12891/SRR098359	A*01:01	A*01:01:01	A*24:02	A*24:02:01	2	2
	B*07:02	B*42:01:01	B*08:01	B*08:79	0	0
	C*07:01	C*07:01:01	C*07:02	C*07:02:01	2	2
ID: NA18526/ERR031854	A*24:02	A*24:02:01	A*33:03	A*33:03:01	2	2
	B*40:01	B*40:01:02	B*58:01	B*58:02 B*58:01:01	2	1
	C*03:02	C*03:02:02	C*07:02	C*07:02:04	2	2
ID: NA18912/SRR111960	A*34:02	A*34:02:01	A*74:01	A*74:01	2	2
	B*53:01	B*53:01:01	B*53:01	B*53:01:01	2	2
	C*04:01	C*04:01:01	C*04:01	C*04:01:01	2	2
ID: NA18964/SRR098539	A*26:01	A*26:01:01	A*26:03	A*26:03:01	2	2
	B*35:01	B*35:01:01	B*40:02	B*35:31	1	1
	C*03:03	C*03:03:01	C*03:04	C*03:04:01	2	2
ID: NA19000/SRR099528	A*24:02	A*24:02:01	A*24:02	A*24:02:01	2	2
	B*54:01	B*54:01:01	B*67:01	B*67:01:01	2	2
	C*01:02	C*01:02:01	C*07:02	C*07:02:04	2	2
ID: NA19130/SRR107026	A*33:03	A*33:03:01	A*36:01	A*36:01	2	2
	B*13:02	B*13:02:01	B*35:01	B*35:01:01	2	2
	C*04:01	C*04:01:01	C*04:01	C*04:01:01	2	2
ID: NA19172/SRR111962	A*30:02	A*30:02:01	A*68:01	A*68:01:01	2	2
	B*35:01	B*35:01:01	B*44:03	B*44:03:01	2	2
	C*04:01	C*04:01:01	C*04:01	C*04:01:01	2	2
TOTAL:					45	44
					94%	92%

Results

We report concordance for both ambiguous and unambiguous results. (3,-Erlich)
Targeted amplicon sequencing gave 100% concordance for the A, B, C loci and all of these were unambiguous. Some ambiguity was found in DRB1 due to lack of coverage, as only exon2 was amplified for this locus, however all results were concordant.
As the datasets become more complex and cover larger regions the data analysis becomes more challenging, and the concordance is reduced. Using the data from a 260-gene RainDance HLAseq kit, Agilent exome kit and whole genome sequencing the concordance reached 92%, 94% and 85%, respectively. For both the RainDance kit and the exome data most results were unambiguous. Whole genome on the other hand showed 10% ambiguous assignments.

Exome sequencing is not designed for HLA typing, however as the MHC region and the HLA genes are usually captured it is possible to filter out reads matching for this region. With this approach the HLA typing can be as accurate as using an MHC kit, although the results can vary according to which exome kit is used.

Conclusions

- Even reads as short as 72bp can be used for high accuracy (>92% concordance) HLA typing.
- **Omixon Target** reached 100% concordance for amplicon-based sequencing for the Class 1 loci using Illumina 2x150 bp sequencing.
- **Omixon Target** can assign HLA types for targeted amplicon data accurately without manual sequence editing.

Validation efforts

Validation efforts for **Omixon Target** are still ongoing and we are currently inviting submission of selected datasets with known types, which will we type for free.

Amplicon-based sequencing of 3 full loci (HLA-A, B, C) and one exon of HLA-DRB1

	Allele 1		Allele 2		Concordance	
	Reference	Omixon	Reference	Omixon	Including Ambiguities	Unambiguous
ID: 1	A*02:01	A*02:01:01	A*26:01	A*26:01:01	2	2
	B*38:01:01	B*38:01:01	B*41:01	B*41:01	2	2
	C*12:03	C*12:03:01	C*17:01	C*17:01:01	2	2
	DRB1*03:01	DRB1*03:01:08, DRB1*03:01:01	DRB1*14:54	DRB1*14:01:01, DRB1*14:54:01	2	0
ID: 2	A*03:01:01	A*03:01:01	A*26:01:01	A*26:01:01	2	2
	B*07:02	B*07:02:01	B*38:01	B*38:01:01	2	2
	C*07:02	C*07:02:01	C*12:03	C*12:03:01	2	2
	DRB1*11:04:01	DRB1*11:04:01	DRB1*15:01	DRB1*15:01:17, DRB1*15:01:01	2	1
ID: 3	A*01:01	A*01:01:01	A*68:01	A*68:01:01	2	2
	B*08:01:01	B*08:01:01	B*44:03	B*44:03:01	2	2
	C*04:01	C*04:01:01	C*07:01	C*07:01:01	2	2
	DRB1*03:01	DRB1*03:01:08, DRB1*03:01:01	DRB1*07:01	DRB1*07:01:01	2	1
ID: 4	A*02:01	A*02:01:01	A*32:01:01	A*32:01:01	2	2
	B*40:01	B*40:01:02	B*51:01	B*51:01:01	2	2
	C*03:04:01	C*03:04:01	C*14:02:01	C*14:02:01	2	2
	DRB1*07:01:01	DRB1*07:01:01	DRB1*11:01	DRB1*11:01:01, DRB1*11:01:08, DRB1*11:97	2	1
ID: 5	A*01:01	A*01:01:01	A*02:01	A*02:01:01	2	2
	B*38:01	B*38:01:01	B*40:01	B*40:01:02	2	2
	C*12:03	C*12:03:01	C*17:01	C*17:01:01	2	2
	DRB1*01:02:01	DRB1*01:02:01	DRB1*13:02:01	DRB1*13:02:01	2	2
ID: 6	A*02:05:01	A*02:05:01	A*26:01	A*26:01:01	2	2
	B*35:01	B*35:01:01	B*50:01:01	B*50:01:01	2	2
	C*04:01	C*04:01:01	C*06:02	C*06:02:01	2	2
	DRB1*12:01	DRB1*12:01:01, DRB1*12:10, DRB1*12:06, DRB1*12:17	DRB1*13:01	DRB1*13:117, DRB1*13:01:01	2	0
ID: 7	A*29:02:01	A*29:02:01	A*68:01:01	A*68:01:01	2	2
	B*40:01	B*40:01:02	B*44:02:01:01	B*44:02:01	2	2
	C*03:04:01	C*03:04:01	C*16:01:01	C*16:01:01	2	2
	DRB1*04:01:01	DRB1*04:01:01	DRB1*13:02:01	DRB1*13:02:01	2	2
ID: 8	A*01:01	A*01:01:01	A*02:01	A*02:01:01	2	2
	B*08:01:01	B*08:01:01	B*51:01:01	B*51:01:01	2	2
	C*07:01	C*07:01:01	C*15:02	C*15:02:01	2	2
	DRB1*03:01	DRB1*03:01:08, DRB1*03:01:01	DRB1*11:01	DRB1*11:01:01, DRB1*11:01:08, DRB1*11:97	2	0
TOTAL:					64	55
					100%	86%

References:

1, Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SGE
The IMGT/HLA Database
Nucleic Acids Research (2011) 39 Suppl 1 1171-6
2, Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE:
IMGT/HLA - a sequence database for the human major histocompatibility complex
Tissue Antigens (2000), 55:280-287
3, C. L. Holcomb, B. Höglund, , H.A. Erlich.
A multi-site study using high-resolution HLA genotyping by next generation sequencing
Tissue Antigens (2010), ISSN 0001-2815
http://www.1000genomes.org/category/exome
www.hapmap.org

Contact:

Email: attila.berces@omixon.com
Address: Petzval J. utca 56, Budapest, H-1119, Hungary