Published in final edited form as:
Int J Immunogenet. 2013 February ; 40(1): 72–76. doi:10.1111/iji.12024.

16IHIW HLA typing by NGS: Workshop review

D De Santis[1], **D Dinauer**[7], **J Duke**[4], **HA Erlich**[3], **CL Holcomb**[3], **C Lind**[4], **K Mackiewicz**[4], **D Monos**[4], **A Moudgil**[6], **P Norman**[6], **P Parham**[6], **A Sasson**[4,5], and **R JN Allcock**[1,2]

[1] Department of Clinical Immunology, PathWest, Royal Perth Hospital, Wellington Street, Perth, 6000, Western Australia

[2] Lotterywest State Biomedical Facility Genomics, School of Pathology and Laboratory Medicine, University of Western Australia, Nedlands, 6009, Western Australia

[3] Human Genetics Dept. Roche Molecular Systems, Pleasanton, CA, USA

[4] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania and The Children's Hospital of Philadelphia

[5] Bioinformatics Core, The Children's Hospital of Philadelphia, Philadelphia, PA, United States

[6] Departments of Structural Biology, Microbiology and Immunology. Stanford University School of Medicine, Stanford. USA

[7] Life Technologies, Wisconsin, USA

Summary

Human leukocyte antigens (HLA) genes play an important role in the success of organ transplantation and are associated with autoimmune and infectious diseases. Current DNA based genotyping methods, including Sanger sequence-based typing (SSBT), have identified a high degree of polymorphism. This level of polymorphism makes high-resolution HLA genotyping challenging, resulting in ambiguous typing results due to an inability to resolve phase and/or defining polymorphisms lying outside the region amplified. Next generation sequencing (NGS) may resolve the issue through the combination of clonal amplification, which provides phase information, and the ability to sequence larger regions of genes, including introns, without the additional effort or cost associated with current methods. The NGS HLA sequencing project of the 16IHIW aimed to discuss the different approaches to; (i) template preparation including short and long range PCR amplicons, exome capture and whole genome; (ii) sequencing platforms, including GS 454 FLX, Ion Torrent PGM, Illumina MiSeq/HiSeq and Pacific Biosciences SMRT; (iii) data analysis, specifically allele calling software. The pilot studies presented at the workshop demonstrated that although individual sequencers have very different performance characteristics, all produced sequence data suitable for the resolution of HLA genotyping ambiguities. The developments presented at this workshop clearly highlight the potential benefits of NGS in the HLA laboratory.

Address for Correspondence Dianne De Santis, Department of Clinical Immunology, PathWest Laboratory Medicine WA, Royal Perth Hospital, Wellington Street, Perth 6000, Western Australia, dianne.desantis@health.wa.gov.au.

Introduction

The major histocompatibility complex (MHC) encodes the classical human leukocyte antigens (HLA) known to be important in solid organ and haematopoietic stem cell transplantation (Eapen M et al, 2007 and Lee SJ et al, 2007), infectious disease (eg HIV, Hep C, CMV) (Gao X et al, 2005), autoimmune disease (eg diabetes, rheumatoid arthritis, coeliac) (Lie BA et al, 2005) and drug hypersensitivity reactions (Mallal S et al, 2002). The process of identifying HLA alleles as evolved from serological based methods to recent DNA based methods such techniques as RFLP, PCR-SSO, PCR-SSP, and Sanger sequencing based typing (SSBT) (Erlich HA, 2012 and Dunn Paul, 2012). Today DNA sequencing is the gold standard for HLA typing worldwide and is a well optimised, reliable, and efficient technique able to detect and identify both alleles at the 6 major loci (HLA-A, -B, -C, -DRB1, -DQB1 and -DPB1).

Current Sanger-based sequencing techniques have allowed the identification of over 8,000 HLA class I and II alleles (IMGT/HLA Database release 3.9.0, http://www.ebi.ac.uk/imgt/hla/). However, less than 10% of these alleles have been fully sequenced and only around 10% of these alleles meet the criterion of “common, well-documented” (CWD) alleles (Cano P et al. 2007) and more than 40% of alleles have only been reported once (as reported at the 16IHWS HLA rare allele workshop). Only the most polymorphic regions of HLA class I (exon 2 and 3) and II (exon 2) encoding the peptide binding sites are routinely assessed. Clinical studies have shown that matching these regions of the 6 major loci provides the best clinical outcomes (ie. less rejection, least GvHD) in solid organ and haematopoietic stem cell transplantation. However, even in perfectly matched transplants, approximately 30% of recipients experience adverse events within 5 years (Ottinger HD et al, 2003). It is not clear whether this is the result of genetic or non-genetic factors, and whether it reflects mismatches in regions not analysed using current methods – either other regions of HLA genes, or other genes in other genomic regions. Moreover, in a proportion of individuals undergoing HLA typing, ambiguous combinations of alleles may be shown. These may be the result of cis-trans ambiguities or of particular allele combinations being identical over the regions commonly analysed.

One way to resolve the issue of genotyping ambiguity is to analyse additional exons or even entire genes. This is technically feasible using the current sequencing approaches, but is significantly more expensive and laborious to perform. Moreover, the amount of additional variation detected is likely to be a small fraction of what is currently detected and its value is as yet unknown. Hence, the combination of additional cost and limited increased information has meant that few laboratories have been willing or able to pursue the matter further.

The last 6 years has seen a dramatic revolution within the DNA sequencing world. Starting in 2005 with the release of the 454/GS-FLX sequencer (Roche Diagnostics), an entirely new approach has emerged, known as next-generation (NGS) or massively parallel sequencing (MPS). There has been a large growth in development and operation of these new DNA sequencers predominantly in the research world, and most recently in their potential role in the routine diagnostic laboratory.

Results

The 16IHIW in Liverpool, UK in May 2012 held the first workshop on the application of NGS to HLA genotyping. The main aims of the workshop were to identify methods of template preparation, examine the pros and cons of the different sequencing platforms and, most importantly, determine the requirements for data analysis and in particular allele calling software to assess ambiguity resolution. The workshop provided a broad overview of the field as it currently stands.

Henry Erlich and Cherie Holcomb (Roche Molecular Systems, USA) demonstrated the amplicon pooling approach on the Roche 454/FLX sequencer (Bentley G et al, 2009 and Holcomb et al, 2011). The Roche 454 system employs long sequence reads (avg. 500 bp) that allow for setting phase for linked polymorphisms within the sequenced DNA library fragment. Erlich and Holcomb et al demonstrated the use of amplicon sequencing using the 454 GS FLX and GS Junior instruments, with Conexio Assign ATF 454 genotyping software, and “fusion primers” to amplify informative exonic and intronic regions for the HLA-A,-B,-C, DRB1 (and DRB3/4/5), DQB1, DQA1, DPB1, and DPA1 loci (the amplicons range from 366 to 750 bp) and assign genotypes. The use of fusion primers, which include the 454 adaptors and multiplex identifiers (MIDs), means that the library is created during the PCR and, following purification from primer-dimer, quantification and limiting dilution can be used directly in emulsion PCR. A set of 14 primer pairs (Bentley G et al, 2009 and Holcomb et al, 2011) currently available commercially as the GS GType HLA Primer Sets, and more recently a 22 primer pair set (the Very High Resolution or VHR set, encompassing additional exon and intron regions) capable of distinguishing most “null” alleles have been developed. Using a simplified workflow, 99.4% concordance was achieved with the genotypes of 20 reference samples with minimal ambiguity (56% unambiguous at the 2-field level, with a median ambiguity string length of 2 for the loci that are not called unambiguously). In order to achieve high throughput HLA typing, the workflow was simplified further to incorporate up to 96 MIDs using the 4 primer approach and the Access Array (Fluidigm). In this micro-fluidics system, up to 48 different genomic DNA samples and up to 48 different primer pairs are used to generate 2,304 different PCRs. The results for 192 samples with 8 primer pairs and 96 samples with 14 primer pairs in a single GS FLX run achieved 100% genotyping concordance and an average of 166 reads and 173 reads per amplicon respectively. The Roche GS GType HLA Primer Sets (Life Sciences Reagents) for the 454 FLX and GS Junior platform is currently the only commercially available HLA typing kit for next generation sequencing.

David Dinauer (Life Technologies, USA) provided an update on the development of research assays for HLA typing on the Ion Torrent PGM platform. A simulated comparison of read length and HLA phase ambiguities across the antigen recognition exons for HLA-A, - B, - C, and -DRB1 was performed and as of IMGT release v3.9, a significant number of CWD cis/trans ambiguities remain at read lengths below 400 base pairs for class I genes due to recombination between exon 2 and exon 3. Life Technologies is developing long read chemistry and protocols to enable read lengths of 400 bases on the PGM platform. The company believes this will provide superior results for HLA sequencing on the PGM platform. The feasibility of library preparation approaches for whole-gene HLA sequencing are currently being compared. Different workflows have been evaluated, including long-range PCR followed by fragmentation and ligation of emulsion PCR adaptor sequences. Alternatively, a multiplexed short-range PCR approach is under consideration based on the current “AmpliSeq” workflow. An amplicon tiling approach may also be feasible and could include full coverage of HLA genes in as few as two PCR tubes per sample. These two approaches are being evaluated to determine the HLA community's preferred approach.

Dianne De Santis and Richard Allcock (PathWest Laboratory Medicine, Western Australia) showed the cost-effective generation of data from full gene HLA class I and exon 2 – 3 HLA Class II LR-PCR amplicons on the Ion Torrent PGM platform. The method which utilises long-range amplicon pooling and sample barcoding allows the analysis of relatively low numbers of samples for multiple loci. The results of the pilot study were consistent with the expected data output of a 314 chip of >10Mb, mean read lengths of 200bp and high coverage of all bases although some reductions in coverage were noted across exons 2 and 3 at HLA-A, -B and -C. Sequencing of a heterozygous sample revealed short-range haplotypes that will be important in resolving cis/trans genotyping ambiguities. The study concluded that although allele calling software was not yet available to determine whether NGS would resolve the genotyping ambiguities seen with Sanger SBT, preliminary results indicated that HLA genotyping on the PGM platform was economically and efficiently possible for a routine HLA laboratory.

Curt Lind and Dimitrios Monos (The Children's Hospital of Philadelphia and University of Pennsylvania, USA) showed a comparison of sequencing data quality and genotyping on the 4 major sequencers – the 454/FLX (and by extension the GS-Jnr at lower volume), the Ion Torrent PGM, the Illumina MiSeq (and by extension the HiSeq at very high sample numbers) and the Pac Bio SMRT sequencer (Table 1). The latter instrument was of particular interest as it claims extremely long read lengths. Several samples were amplified at the HLA-A, B, C, DRB1, and DQB1 loci. Full length HLA Class I genes were amplified using custom designed primers from 5'UTR to 3'UTR. HLA Class II loci were amplified from Exon 2 to Exon 4 using primers from GenDx (Utrecht, Netherlands). Samples were prepared for sequencing on the four different platforms according to manufacturers’ specifications. Overall, the comparison showed that the individual sequencers have very different performance characteristics in terms of error rates and that each produces data that needs to be processed appropriately. They showed that all the major sequencers had similar substitution error rates, but that the MiSeq sequencer had the lowest total observable error rate, with the 454/FLX and PGM more error-prone, especially around homopolymers. The SMRT had the highest error rate being mostly indels. Consensus sequence of all aligned reads was 99.9-100% accurate for all platforms. Definitive HLA genotypes were obtained without any ambiguity (alternative genotypes) and were concordant with the known alleles of these samples.

A significant area of concern within NGS has been the specific analytical pipelines and software that will be used and how these will be implemented in a routine diagnostic laboratory. Presentations from Conexio Genomics (Fremantle, Western Australia) and GenDx (Utrecht, The Netherlands) showed that both companies’ software was able to accept data input in FASTA or FASTQ format (a standard format output from most modern sequencers) and accurately determine HLA genotypes using test datasets.

While PCR-based amplification strategies are the dominant current paradigm, in future, the laboratory workflow for large-scale whole-genome sequencing may be far simpler than exists currently. The goal of the $1000 genome is likely to be achieved in the next 5 years and there may come a point at which routine WGS might be the most cost-effective and simplest way to generate primary sequence data. In that case, the remaining challenge will be the analytical mechanism by which the sequence data from HLA genes could be extracted from this larger dataset and used to accurately call HLA genotypes.

Presentations in the workshop by Paul J Norman and Arnav Moudgil from Peter Parham's Stanford laboratory showed how reads from the 1000 genomes project could potentially be extracted and used to determine HLA and KIR genotypes. Genome-wide de-novo assembly can be inaccurate and single reference sets are inadequate for generating alignment-based genotype calls. The Stanford group created a pipeline that automatically extracts HLA and KIR reads from whole genome data sets, attributes them to specific loci and calls the genotype. Using filters that consider all variants of each homologous gene, including pseudogenes, they ensured selection of only those read-pairs mapping exclusively to a locus. All procedures were performed using freely available software and Python-language programming, whilst local alignments for visual inspection were created in parallel. The final genotype resembled a sequence-based genotype that covered all exons, albeit with additional haplotype phase information obtained due to the high density of overlapping reads. The pipeline was tested empirically on whole-exome data from eight heterozygous individuals and there was complete concordance with standard genotyping results. Analysis of exome-sequencing also revealed 20 previously unknown KIR alleles, which were verified by standard sequencing and family studies. Whilst their method ensured precision, the coverage was dependant on read length and depth, and the nucleotide diversity of a given locus. Although the most popular current technique, exome targeting, remains the most efficient, all types of whole-genome sequence data can be harvested for HLA and KIR allele content. The implications of these experiments are quite profound and might change the way in which genetic testing could be performed in future.

Discussion

The explosion of NGS technologies means that a variety of different approaches are possible, with some approaches being better suited to particular sequencing platforms. The choice of approach and platform for a given laboratory will depend not only on required genotyping resolution, but also sample throughput, automation capacity, cost and turnaround time. Ideally, the goal of HLA genotyping is very-high (or even absolute) resolution of alleles. This may be provided by NGS because of the dramatic decrease in sequencing cost (hence sequencing entire genes is feasible), as well as the inherent nature of sequencing individual DNA strands (hence identifying genetic phase absolutely over 100-400bp lengths). Analytical software will need to be able to take account of the need to analyse entire genes and sequence from individual DNA strands in determining allele assignment.

The first NGS HLA genotyping workshop identified three important areas for consideration for the development and implementation of NGS platforms in the routine HLA laboratory; template preparation, sequencing platform and data analysis.

A number of different approaches for the preparation of DNA libraries can be adopted:

1. Highly multiplex PCR amplifying either individual exons or regions of interest (including introns) as implemented by Roche 454 sequencing system or Life Technologies. This approach has the advantage that no subsequent fragmentation and ligation is required as the sequencing adaptors are generally amplified onto target fragments. However, it also means that sequencing is generally unidirectional and that primer design is constrained by the many polymorphisms which are found even in the introns of HLA genes. Hence, multiple primer pools may still be necessary.

2. Long-range PCR of individual loci, followed by pooling and fragmentation. This approach allows the fragments to be treated as a small whole-genome (Shiina et al, 2012).

3. Oligonucleotide-based hybridization/capture techniques analogous to exome sequencing, targeting large specific regions of >50Mb are well established on all sequencing platforms and it has been proposed that a similar approach might be feasible for HLA genes or even the MHC as a whole. The efficacy of hybridisation capture on short regions is certainly less effective than larger regions.

4. Unbiased whole-genome sequencing has been suggested as a long-term aim for routine healthcare. The effectiveness of unbiased WGS as a primary data generation technique needs to be established, but the potential to extract sequences from genomic regions of interest is large.

The four major NGS platforms, Roche GS 454 FLX, Illumina MiSeq/HiSeq, PGM-Ion Torrent and Pacific Biosciences, were all represented at the workshop and all were shown to produce sequence data suitable for the resolution of HLA genotyping ambiguity. This is an intriguing result as it suggests that in the long term, the differences between specific sequencers used will be less important. It will allow individual laboratories to choose and implement laboratory protocols and hardware that suit their specific needs in terms of sample throughput and workflow, with the knowledge that reliable, robust genotyping at a sufficient quality can be achieved. There has been debate over the merits of short- and long-read sequencing and a specific difference between the platforms is read length (Table 1). However, the data shown at the workshop suggests that these differences were not critical in the accurate identification of alleles. It is generally agreed that longer reads are desirable, as they will allow for the resolution of ambiguous loci spaced widely across a particular gene. The challenge for the future remains in the data analysis process and how best to utilise the data generated. It remains to be seen how mapping algorithms handle the shorter sequence read lengths obtained with the newer benchtop sequencers.

The ability to sequence full genes at a comparable or lower cost than Sanger SBT will allow complete HLA gene sequences to be determined, and potentially the identification of many novel alleles. In a concerted effort, the aim of the NGS HLA Working Group is to work towards filling in the gaps left by the Sanger SBT of HLA alleles. The additional information generated will expand the current knowledge of sequence variation in the HLA system and is of use in the clinical laboratory in the development of new assays, in population genetics, in disease association studies and in clinical transplant studies. However, the ability to generate this new data raises potential problems with the naming of newly identified alleles. A number of questions remain, such as whether novel alleles will be accepted if they cannot be completely phased? Will already named alleles need to be re-named due to new polymorphisms in currently un-sequenced exons/introns? Will novel alleles need to be confirmed using Sanger SBT before being accepted by the nomenclature committee? These questions will be answered in subsequent workshops and most importantly in collaboration with the HLA nomenclature committee.

The newly developed high-throughput sequencing systems provide a very exciting development for HLA genotyping. Although there is still much work to be done for implementation of these technologies into an accredited HLA laboratory, it is clear that the use of these technologies for HLA genotyping is fast approaching.

References

Bentley G, et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens. 2009; 74:393. [PubMed: 19845894]

Cano P, Klitz W, Mack SJ, Maiers M, Marsh SG, Noreen H, et al. Common and well-documented HLA alleles: report of the Ad-Hoc committee of the american society for histocompatiblity and immunogenetics. Hum Immunol. May.2007 68(5):392. [PubMed: 17462507]

Dunn P. Human leucocyte antigen typing: techniques and technology, a critical appraisal. Int J Immunogenet. Dec.2012 38(6):463. 2011. [PubMed: 22059555]

Eapen M, Rubinstein P, Zhang MJ, Stevens C, Kurtzberg J, Scaradavou A, et al. Outcomes of transplantation of unrelated donor umbilical cord blood and bone marrow in children with acute leukaemia: a comparison study. Lancet. Jun 9.2007 369(9577):1947. [PubMed: 17560447]

Erlich HA. HLA DNA typing: past, present and future. Tissue Antigens Jul. 2012; 80(1):1.

Holcomb, et al. A multi-site study using high-resolution HLA genotyping by next generation sequencing. Tissue Antigens. 2011; 77:206. [PubMed: 21299525]

Gao X, Bashirova A, Iversen AK, Phair J, Goedert JJ, Buchbinder S, et al. AIDS restriction HLA allotypes target distinct intervals of HIV-1 pathogenesis. Nat Med. Dec.2005 11(12):1290. [PubMed: 16288280]

Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood. Dec 15. 2007; 110(13):4576.

Lie BA, Thorsby E. Several genes in the extended human MHC contribute to predisposition to autoimmune diseases. Curr Opin Immunol. Oct.2005 17(5):526. [PubMed: 16054351]

Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C, et al. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet. Mar 2.2002 359(9308):727. [PubMed: 11888582]

Ottinger HD, Ferencik S, Beelen DW, Lindemann M, Peceny R, Elmaagacli AH, et al. Hematopoietic stem cell transplantation: contrasting the outcome of transplantations from HLA-identical siblings, partially HLA-mismatched related donors, and HLA-matched unrelated donors. Blood. Aug 1.2003 102(3):1131. [PubMed: 12689945]

Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, et al. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens. Oct.2012 80(4):305. [PubMed: 22861646]

Table 1

A summary of results from the University of Pennsylvania and The Children's Hospital of Philadelphia showing sequencing platform characteristics.

Metric	Spec/ CHOP	454 GS FLX	Illumina MiSeq	Ion Torrent PGM	Pacific Biosciences RS (C2 Chemistry)
Output	Spec	240-440Mb(8 region)	1 - 2Gb	20 Mb (314 Chip) 200 Mb (316 Chip)	60-140 Mb/SMRT-Cell
	CHOP	200-250 Mb (8 region)	0.91 – 1.75Gb	23-97 Mb (314 Chip) 132-540 Mb (316 Chip)	61-141 Mb/SMRT-Cell
Read Length (bases)	Spec	~400	2× 150	~200	~3,000
	CHOP	300-380b (Avg) 758b (Max)	2× 150b (Avg & Max)	130-217b (Avg) 398b (Max)	1,907-2,244b (Avg) 14,159b(Max)
Total Error (indels, substitutions (%)*	CHOP	1.20%	0.32%	1.71%	14.1%
Flexibility (config, # samples)		++	+	+++	++
		multiple region config	Read length choices	Different chips	Addt'l SMRT Cells
Pre-Sequencing Time (Hours)	Library Prep	3	4 - 6	2 - 4	2 - 3
	emPCR → Enrich	6 - 8		4 - 8	
	Run Prep	2 - 3		2	
Sequencer Run Time (Hours)		10	27 (2×150 bp)	2.5 – 4.5	1 ½
Equipment Cost		$500,000 (GS FLX) $100,000 (GS Junior)	$125,000	$66,000 (PGM & Server) $15,000 (OneTouch)	$700,000

[*] Error rate was calculated using DNA from homozygous sample(s) amplifying the five HLA loci and aligning the reads against the full length of genomic sequences matching the alleles of the sample(s). All the major sequencers had similar substitution error rates, but that the MiSeq platform had the lowest total observable error rate, with the 454/FLX and PGM more error-prone, especially around homopolymers. The SMRT had the highest error, mostly indels.