Published in final edited form as:

Hum Immunol. 2010 October ; 71(10): 1011–1017. doi:10.1016/j.humimm.2010.07.012.

A novel single cDNA amplicon pyrosequencing method for high-throughput, cost-effective sequence-based HLA class-I genotyping

Simon M Lank¹, Roger W Wiseman¹, Dawn M Dudley², and David H O'Connor^{1,2,3}

¹Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, Wisconsin, USA.

²Department of Pathology and Laboratory Medicine, University of Wisconsin-Madison, Madison, Wisconsin, USA

Abstract

HLA genotype influences the immune response to pathogens and transplanted tissues; accurate HLA genotyping is critical for clinical and research applications. Sequence-based HLA typing is limited by the cost of Sanger sequencing genomic DNA and resolving cis/trans ambiguities, hindering both studies correlating high-resolution genotype with clinical outcomes, and population-specific allele frequency surveys. We present an assay for sequence-based HLA genotyping by Titanium read length clonal Roche/454 pyrosequencing of a single, universally diagnostic PCR amplicon from HLA class-I cDNA that captures the majority of exons 2, 3 and 4 used for traditional sequence-based typing. The amplicon is predicted to unambiguously resolve 85% of known alleles. A panel of 48 previously HLA-typed samples was assayed with this method, demonstrating 100% non-null allele typing concordance. We show this technique can multiplex at least 768 patients per sequencing run with multiplex identifier sequence bar-coding. Unprecedented typing throughput results from a novel single cDNA-PCR amplicon strategy requiring only one PCR amplification per sample. This method dramatically reduces cost for genotyping of large cohorts.

Keywords

Pyrosequencing; HLA; sequence-based typing; multiplexing; high-throughput

1. Introduction

Gene products of the human leukocyte antigen (HLA) region shape the adaptive immune response and HLA genotype is associated with clinical outcome in many diseases, drug hypersensitivities, and autoimmune disorders, as well as for tissue transplantation where the degree of HLA mismatch between donor and recipient correlates strongly with tissue rejection and graft-vs-host disease [1-12].

³To whom correspondence should be directed: 555 Science Drive, Madison, WI 53711, Ph: 608-890-0845, Fax: 608-265-8084, doconnor@primate.wisc.edu .

Author Contributions

DHO, RWW, and SL designed the research, SL performed the research and analyzed the data, SL, RWW, DHO and DMD wrote the manuscript, DHO and RWW supervised the project.

Population level balancing selection in response to diverse pathogen epidemics has made HLA genes the most polymorphic of any known loci with over 2,300 classical HLA class-I alleles described to date [13-18], presenting a unique diagnostic challenge [19]. With practical limits on the number of potential PCR reactions per individual, single pass SSP/SSOP typing methods cannot detect all possible alleles, and can be complex or imprecise when exploring allele frequencies in poorly characterized populations. Problems include false negatives or positives in the case of rare or null alleles, and difficulty determining the presence of alleles not specifically assayed for by the panel, which can present misleading amplification or hybridization patterns [20-23]. Novel allele discovery and unambiguous first-pass genotyping requires sequence-based typing (SBT) of individual HLA loci, with current methods using locus- and exon-specific amplification primers, followed by Sanger capillary sequencing [24]. This method, while practical for allele discovery, is underutilized for genotyping due to limitations of Sanger SBT, primarily throughput restriction from multiple time- and cost-intensive exon-specific amplicon preparations each of which must be capillary sequenced [24, 25]. Reagent costs alone for high-resolution HLA class-I Sanger SBT can be upwards of \$300 per patient. Another difficulty results from bulk sequencing co-amplified alleles of heterozygous loci, resulting in cis/trans phase ambiguities [26, 27].

Previous studies have applied Roche/454 GS FLX pyrosequencing to HLA typing from genomic DNA (gDNA), extending Sanger-based methods while using inherently clonal pyrosequencing to specifically address the resolution of cis-trans ambiguities within exons. Physical subdivision of the picotiter sequencing plate (PTP) and multiplex identifier (MID) barcoding have been used to simultaneously sequence dozens of patient samples [28, 29]. With multiple amplicons per sample, these techniques have achieved, at most, 48 individuals per full sequencing run, which, due to the expense of amplification and sequencing reagents keeps the per-sample cost high and throughput low [30]. We recently described a technique for high-throughput and cost-effective macaque MHC class-I genotyping utilizing GS FLX pyrosequencing of a universal, 190 bp cDNA PCR product [31]. Recognizing that this method allowed comprehensive MHC class-I typing from a single amplification we have adapted it to the human HLA, utilizing increased read length with GS FLX Titanium chemistry to expand the size of the diagnostic amplicon, thereby increasing typing resolution. To this end we sought novel, conserved primer binding sites within the class I open reading frame, and evaluated the resolution and conservation of the resultant PCR products *in silico*. We established a validation panel of previously HLA-SSP typed cell lines to test the novel cDNA-derived, second-generation HLA sequence-based typing assay (cSBT) and used MID tagged primers to demonstrate an unprecedented level of sample multiplexing within a single Roche/454 pyrosequencing experiment.

2. Materials and Methods

2.1 Reference samples

A panel of 48 comprehensively HLA SSP-typed samples was prepared to validate the assay [Supplementary Table 1]. 40 were reference B-lymphoblastoid cell lines provided by the International Histocompatibility Working Group cell bank. Additionally, five embryonic stem cell lines with previously determined HLA genotypes were included (two from duplicate cell cultures) for samples 41 through 47. Finally, an RNA sample isolated from whole blood, which had been previously typed clinically by SSP, was included.

2.2 RNA isolation & cDNA synthesis

cDNA from the reference samples was prepared analogously to methods used for macaque MHC class-I sequencing as previously described [31]. Briefly, total cellular RNA was isolated from cultured cell line samples using the MagNA Pure LC RNA Isolation Kit

(Roche Applied Sciences) and quantified with a NanoDrop 1000 Spectrophotometer V3.7 (Thermo Scientific). Total RNA was normalized to 10 ng/ul, and 50 ng was reverse transcribed using the SuperscriptTMIII First-Strand Synthesis System (Invitrogen) with oligo(d)T primer to generate cDNA templates.

2.3 HLA-A/B/C universal primer design

Primers capable of universally amplifying HLA class-I cDNA segments were found by aligning all 2,330 known HLA class-I sequences and searching for sites of high conservation. Allele sequences from the IMGT/HLA database were aligned using MUSCLE multiple-alignment software to identify regions of low polymorphism [Figure 1a] [18, 32]. Conserved primer binding sites common to HLA-A, B, & C alleles at bp 145-162 (5') and 703-726 (3') were used to design a 581 bp amplicon spanning the highly polymorphic regions of HLA class-I exons 2 and 3. We also identified a second potential 3' primer at bp 1053-1072, resulting in a 927 bp amplicon.

2.4 PCR amplification

DNA primers containing the HLA-specific sequences, along with one of 48 distinct 10 bp MID tags and amplicon library A or B sequencing adapters were used to amplify the 581 bp region identified by primer design [Figure 1b, Supplementary Table 2]. PCR was performed using high-fidelity Phusion polymerase (New England Biolabs) with the following reaction setup: 25 ul 2x Phusion polymerase, 2 ul combined 10 uM primers, 3 ul cDNA, and 20 ul H₂O. The following thermocycling conditions were used: 98°C for 3:00 min; 23 to 27 cycles of 98°C for 0:05 sec, 60°C for 0:10 sec, 72°C for 0:20 sec; 72 C for 5:00 min. PCR products were purified either by 1% agarose gel electrophoresis followed by MinElute gel extraction & purification (Qiagen), or using size-exclusion SPRI Ampure-XP DNA-binding paramagnetic beads (Agencourt) according to the manufacturer's protocol.

2.5 Product quantification, normalization & pooling

Purified PCR products were quantified in 96-well format using a LightCycler 480 (Roche) for fluorescence detection in a picogreen assay (Invitrogen). HLA-Ref32, which failed to produce detectable PCR products post-purification, was re-amplified with the MID1 primer in a subsequent sequencing experiment. Post-purification sample concentrations ranged between 1.7 and 15.1 ng/ul. All samples were normalized to 1 ng/ul and pooled at equimolar concentrations to create a highly multiplexed amplicon library. Using a total amplicon size of 651 bp (total size including adaptor and MID sequences) the library was re-quantified after pooling and diluted to 10⁸ molecules / ul for emPCR.

2.6 GS-FLX pyrosequencing

emPCR was performed on the amplicon library using a low volume emPCR amplicon kit according to the manufacturer's protocols and pyrosequenced on a Genome Sequencer FLX-Titanium instrument (Roche/454 Life Sciences) at the University of Illinois at Urbana-Champaign High-Throughput Sequencing Center. The library was sequenced in a single 1/16th gasketed region of a 70×75 mm Titanium PicoTiterPlate, and base calling was performed with the on-instrument amplicon filter settings. Confirmatory library re-sequencing was performed on a GS Junior instrument (Roche/454 Life Sciences) using Titanium reagents with analogous base calling.

2.7 Data analysis

Processed & quality-filtered reads from the on-instrument software (single sff file) were binned by MID into 48 separate sequence sff files using sfffile software (Roche/454). These files were further quality-trimmed to remove poor sequence at the end of the reads using

CLC Genomics Workbench v3.7 (CLC Bio) with a trim quality threshold of 0.005. Short sequences (less than 190 bp) were also removed from analysis. The trimmed and MID-binned sequences were aligned to a database of all known HLA-A, B & C alleles using Blast Like Alignment Tool (BLAT) [33]. Sequences that aligned perfectly, or were perfect matches except for single base insertions or deletions were considered for subsequent analysis. The aggregate pattern of BLAT alignment matches was used to determine the HLA-A, B & C genotypes for each MID-tagged sample.

Samples HLA-Ref25 & HLA-Ref32 were underrepresented in the initial sequencing experiment and were re-sequenced in a follow-up study to determine the genotype presented. Samples 6, 16, 17, 21, 39, 44, 45 & 46 were also re-sequenced to provide validation of genotyping across different experiments, although the re-sequencing did not reveal any additional allele determinations for these samples (data not shown).

3. Results

3.1 Universal HLA-A, -B, and -C amplification primers

Building on previous work designing a diagnostic PCR amplicon for non-human primate MHC class-I genotyping, we sought an analogous strategy for the human HLA [31]. We aligned all HLA class-I allele sequences available as of July 2009 (749 HLA-A, 1159 HLA-B, and 422 HLA-C) and examined the alignment to determine potential primer sites. Despite the extraordinary variability of HLA class-I sequences [Figure 1a], several highly conserved regions were identified as targets for universal amplification primers. The amplicon chosen was designed to capture the most variable regions of HLA sequences, which contain diagnostic SNPs to distinguish related alleles, while staying within the 800 bp upper limit for unidirectional amplicon coverage, with a 400-500 bp read length using current Roche/454 GS FLX-Titanium pyrosequencing. We identified candidate forward and reverse primer-binding sites at bp 145-164, and 706-726 respectively [Figure 1b]. The resulting 581 bp amplicon primers universally target known HLA-A, B and C alleles; however, 57 of the sequences examined (2.4%) contain single base pair mismatches in the primer regions and may be amplified at reduced efficiency [Supplemental Table 3]. We were not able to comprehensively evaluate the primer conservation of the 3' primer, as only 1,317 of the 2,330 named alleles have full exon 4 sequences. Nevertheless, we believe the 3' primer to be more conserved than the 5' primer, since it has a lower percentage of mismatched allele sequences (0.6% vs 2.1%) when controlling for this variable. Importantly, the amplicon captures the main regions of HLA class-I sequence variability in the epitope-binding domains encoded by exons 2 and 3.

The allelic resolution of cSBT was estimated by the uniqueness of sequences within the diagnostic amplicon relative to all described HLA alleles. We again aligned the open reading frames of all 2,330 known HLA class-I alleles over these 581 bp and assembled them at 100% identity. The majority of alleles examined (85%) showed unique, unambiguous sequences across this section of the open reading frame and did not assemble with any other alleles [Table 1]. A further 1% of the alleles were predicted to be assayable to 4-digit resolution, while the remaining 14% of alleles were assayable to 2-digit resolution. Thus, for the majority of described alleles, the presented amplicon allows high-resolution typing, while for others it is medium/low resolution. Our diagnostic amplicon captures sequence from exons 2, 3, and 4, though it misses the first 71 bp of exon 2, and the last 170 bp of exon 4, resulting in genotypes somewhat lower in resolution than complete exon 2, 3 & 4 sequencing from genomic DNA. We provide a full list of all aligned HLA-A, HLA-B, and HLA-C alleles with their predicted resolution across the 581 bp amplicon [Supplemental Tables 4, 5, 6].

In addition to the 581 bp amplicon used in the current study, we identified a second 3' candidate primer site at bp 1053-1072, which, in conjunction with the presented 5' primer yields a longer 927 bp amplicon and captures additional diagnostic SNPs. The longer amplicon resolves all of exons 3, 4, 5, & 6, and the majority of exons 2 & 7 [Figure 1a]. Most of the 2,330 named class I alleles do not have complete sequence for exon 7, which contains the putative primer site, but of the 455 alleles that do, none contain base pair mismatches, making it the most conserved primer binding site we identified [Supplemental Table 3]. While an amplicon of this size is not currently compatible with GS FLX Titanium chemistry, read length increases in the future should allow single-pass, unidirectional or complete bidirectional sequencing of this cDNA-PCR amplicon for cSBT. We analyzed the resolution of the 927 bp candidate amplicon with methods analogous to those described for the 581 bp amplicon. Across the expanded region 95% of alleles were found to be completely unambiguous, allowing high-resolution typing [Table 1].

3.2 In silico resolution analysis of the assay validation panel

In total, 73 distinct class-I alleles were represented in the set of previously typed reference samples including 24 HLA-A, 29 HLA-B and 20 HLA-C alleles. We examined the predicted allelic resolution of the cSBT amplicon for these samples by matching the previously determined alleles with the expected cSBT resolution [Supplemental Tables 4, 5, 6]. Surprisingly, we found that many alleles in the panel were predicted to be ambiguous with other lineage-identical alleles across the 581 bp sequence, and only assayable to 2-digit resolution. Thus, although 85% of all known HLA class-I alleles are predicted to be completely unambiguous, our reference panel contained a high proportion of the 15% of alleles with non-unique sequences. In total 28 of the alleles were predicted to be completely unambiguous, 8 were assayable to 4-digit resolution, and the remaining 37 were assayable to 2-digit resolution [Supplemental Table 7].

3.3 Sample Multiplexing and Pyrosequencing

To facilitate sample multiplexing during sequencing we incorporated 10 bp, unique MIDs into the HLA-specific amplification primers used during primary PCR [Figure 1b & Supplemental Table 2], allowing us to uniquely tag each sample and identify sequence reads during analysis. Several specific MID-tagged HLA primer pairs were found to amplify samples with lower efficiency. Primers with MIDs 6, 25, 39, and 44 had reduced pyrosequencing efficiency, while MID 32 primers failed to amplify a detectable primary PCR product. Although all samples other than HLA-Ref32 gave usable sequencing data, these reduced-efficiency samples were all re-sequenced to confirm genotype using different MID-primer pairs in a separate experiment (data not shown). To further increase throughput, we used an automated 96-well format quantification assay and liquid-handling robotics to rapidly quantify, normalize and pool individual PCR products at equimolar concentrations, creating a single tube, multiplexed amplicon library. Bidirectional amplicon pyrosequencing by GS FLX-Titanium chemistry in a 1/16th 75×70 PTP gasketed region yielded 22,739 high-quality, filterpass sequence reads after image processing.

3.4 Sequence Analysis and BLAT Genotyping

Sequences were parsed by MID for subsequent analysis, with an average of 483 sequences (range 101-831) identified per sample, excluding MID 32 that failed to amplify [Table 2]. In total, 22,704 sequences (99.8%) contained an identifiable MID. After quality trimming and short sequence removal (<190 bp), an average of 398 sequences per MID (18,725 total) were retained. The average read length of forward and reverse sequences after trimming was 342 bp, resulting in complete sequence coverage of the entire 581 bp amplicon.

Trimmed sequences were aligned to a reference database containing all named HLA-A, B and C alleles using BLAT [33]. Alignments were considered valid if the query sequence presented a perfect match with any of the reference alleles, or if the sequence matched perfectly but contained single base insertion or deletion (indel) sequencing artifacts; indel-aligned reads comprised 27% of the quality-filtered total. An average of 316 sequences per MID (14,841 total) had valid alignments with known HLA alleles, 79% of the quality-filtered total. Non-aligned sequences predominantly contained single base pair sequencing or PCR artifacts, many associated with long homopolymer sequences within the open-reading frame, which commonly interfere with GS FLX pyrosequencing [34].

A mean average of 117, 118, and 81 reads were identified per sample at the HLA-A, B, and C loci respectively [Table 2]. This depth of coverage was adequate to identify the one or two alleles per locus as expected for each sample. Genotypes were determined by identifying alleles with perfect matches for both forward and reverse reads for each locus. Only 1 or 2 alleles per locus were positively identified for each sample, showing no signs of sample contamination or data cross-talk across multiple MIDs during analysis [Figure 2].

3.5 cSBT Concordance with Previous HLA Typing

We compared the comprehensive HLA-A, B and C genotypes generated by cSBT to previous typing results. In the first-pass sequence analysis, 41 samples were perfectly concordant with previous genotyping (all expected alleles found). Of the remaining seven samples, four were perfectly concordant except for the original SSP-typed allele B*390601, which was determined by cSBT to be B*390602. Similarly, two other samples were perfectly concordant except for B*4402, which was determined to be B*4427 by cSBT. These discrepancies reflect the limitations of the panels used to do original typing, which could not distinguish between highly similar alleles (e.g. B*390601, B*390602) with identical amplification patterns (IHWG, Michelle Rogers personal communication). Considering these alleles as concordant, 242 of the 243 previously SSP-typed alleles in the panel (99.6%) were identified by cSBT.

The only truly discordant sample was HLA-Ref39, which was typed by SSP to contain a null allele, A*2308N, that was only represented by a single sequence with cSBT. A*2308N contains a frameshift mutation leading to an abortive protein through a premature stop codon. Since our automated analysis procedure ignored allele matches with less than 2 sequences this allele was not detectable without a priori knowledge. Since no other null alleles were represented in the reference panel, it is unclear whether additional null alleles would also be missed by cSBT; however, their biological relevance remains unclear.

We also found 100% concordance between predicted allele resolution and observed resolution. In accordance with those predictions, a subset of alleles was typed at higher resolution than by SSP previously, whereas others were typed at lower resolution. The former occurred more frequently for HLA-B alleles, as expected for the most polymorphic locus, while the latter was particularly true for the A*02, A*24, C*04 and C*07 groups, which contain many highly similar alleles with unique polymorphisms outside the cSBT amplicon used.

3.6 Calculation of Theoretical Multiplexing

The presence of 6 maximum distinct sequences per individual, and the large number of GS-FLX Titanium sequence reads allowed us to achieve unprecedented HLA-SBT throughput by multiplexing 48 samples in a 1/16th PTP region simultaneously, the equivalent of 768 samples per full run. This method yielded, on average, 483 raw MID-specific sequences per

sample, resulting in a mean of 53 HLA-aligned sequences per allele in fully heterozygous samples.

In the validation panel, all samples with greater than 140 raw sequences were correctly genotyped, thus the number of samples multiplexed into a single plate region could be increased significantly. Using a target of 200 MID-specific reads with 20,000 reads per 1/16th PTP region, it should be possible to genotype up to 96 samples per region, or 1,536 per PTP when using a larger panel of MIDs. If the assay were multiplexed in a 1/8th region of a PTP (which results in more than twice the high-quality read number as a 1/16th region and can yield ~60,000 reads) theoretical multiplexing is 2,304 samples per plate, utilizing 288 different MID primer pairs per region.

4. Discussion

Despite the importance of high-resolution HLA genotyping for clinical transplantation matching and infectious disease research, technological barriers of scalability and cost have prevented SBT from being routinely implemented [11]. cSBT overcomes these barriers, laying the groundwork for large-scale typing projects that have previously been impractical. GS-FLX pyrosequencing has dramatically decreased the cost per base of DNA sequencing, though the aggregate cost and time of individual sequencing experiments using this platform is high. Pyrosequencing-based HLA typing has been limited due to the low throughput and multiplexing levels achieved with amplicon pyrosequencing when using multiple locus- and exon-specific amplifications from gDNA. This has increased the depth of coverage per sample required for comprehensive typing and complicated analysis required to resolve phasing ambiguities due to assembly of multiple sequence contigs for each typed allele.

For cSBT we have identified highly conserved HLA class-I primers that amplify essentially all class-I alleles from cDNA, generating a single, high-resolution diagnostic amplicon that can be bidirectionally sequenced with current Titanium pyrosequencing technology. Although our initial validation panel used only 48 different MID primers, multiplexing with hundreds of unique MIDs has been previously demonstrated for other applications and could be applied [35]. In the current study we have translated the low cost per base of pyrosequencing into a low cost per sample, increasing HLA-SBT throughput by an order of magnitude over that previously reported [29, 36]. By using a single amplicon, cSBT is highly amenable to automation when each step is performed in 96-well plate format, making it particularly well-suited to typing in a core laboratory demanding high throughput, or for large disease association studies.

A small percentage (2.4%) of all described HLA alleles contain single base mismatches with our current 581 bp amplicon PCR primers. While it is unknown whether these alleles would be correctly identified with cSBT, in similar applications single base primer mismatches can often be overcome by the use of a high-fidelity polymerase with proofreading activity [31]. Although the current study included 242 distinct allele calls, none of the known primer-mismatched alleles were present in our reference panel. Further work validating our presented primers in samples with known mismatched alleles or on larger cohorts may be necessary to validate the assay in clinical typing settings demanding quantifiable false-positive and negative rates. Our failure to detect a null allele present in the validation panel highlights the differences between an RNA-derived cSBT and genomic DNA SBT method. While null alleles may be typed from gDNA, cSBT provides a clearer portrait of the biologically relevant transcripts that can be translated into functional proteins. It is likely that lower transcript abundance due to translational mediated decay [37] will result in some null alleles being missed at the depth of coverage described here. cSBT may in fact be more precise when genotyping null alleles that are co-amplified with functional alleles by SSP. In

such cases SSP can give a false-positive “non-null” genotype, whereas cSBT would fail to detect the non-functional null allele.

As Roche/454 pyrosequencing read lengths continue to improve - with a near-term target approaching 1,000 bp - cSBT may be accomplished with a longer 927 bp amplicon, which will improve the percentage of uniquely resolvable alleles to 95%. While our presented results demonstrate unprecedented typing throughput at medium resolution, investigators or labs considering cSBT for truly high-resolution typing may find it advantageous to use this longer amplicon in the near future. Other options to improve allele resolution with current pyrosequencing length include the use of a second overlapping diagnostic amplicon to capture additional regions of allelic diversity, and using population allele frequency data to infer alleles from ambiguous groupings [22].

Open-source alignment tools optimized for large, second-generation sequencing data sets make it is possible to generate HLA-reference guided genotypes from raw sequence data for hundreds of individuals in less than a day. cSBT could also accelerate novel allele discovery by allowing large cohorts to be screened for putative novel alleles by identifying individuals presenting fewer than 6 known alleles (not fully heterozygous). The subset of non-aligned, high-quality sequences from such individuals could be examined for contigs representing putative novel sequences, which would then be fully characterized by standard novel allele discovery methods [16].

cSBT holds great promise for large cohort studies linking high-resolution HLA genotype to disease outcome. It could also be used to efficiently type large tissue-donor registries in large-scale, high-throughput genotyping experiments. The release of the Roche/454 GS-Junior instrument, which can generate up to 100,000 high quality Titanium-length reads, should allow genotyping of up to 192 patients simultaneously in a medium-sized laboratory at a per-sample sequencing reagent cost of around \$5. cSBT functionally eliminates the bottlenecks of time and cost associated with previously described Sanger and pyrosequencing-based HLA typing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by UWF-WP-MERC grant #20070709. We thank the International Histocompatibility Working Group and; WiCell Research Institute for providing reference cell lines; William Rehrauer for helpful suggestions; the University of Illinois at Urbana Champaign High Throughput Sequencing Unit for sequencing support and; 454 Life Sciences, a Roche Company for early access to the GS Junior.

Abbreviations

PCR	polymerase chain reaction
cSBT	cDNA-derived sequence based typing
bp	base pair
MHC	major histocompatibility complex
HLA	human leukocyte antigen
SSP	sequence specific primer assay
SBT	sequence-based typing

emPCR	emulsion polymerase chain reaction
MID	multiplex identifier
SNP	single-nucleotide polymorphism
PTP	picotiter plate

References

1. Lewinsohn DA, Winata E, Swarbrick GM, Tanner KE, Cook MS, Null MD, et al. Immunodominant tuberculosis CD8 antigens preferentially restricted by HLA-B. *PLoS Pathog.* 2007; 3:1240–1249. [PubMed: 17892322]
2. Sanchez A, Wagoner KE, Rollin PE. Sequence-based human leukocyte antigen-B typing of patients infected with Ebola virus in Uganda in 2000: identification of alleles associated with fatal and nonfatal disease outcomes. *J Infect Dis.* 2007; 196(Suppl 2):S329–36. [PubMed: 17940968]
3. Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, et al. Common west African HLA antigens are associated with protection from severe malaria. *Nature.* 1991; 352:595–600. [PubMed: 1865923]
4. Alexander L, Cuchura L, Simpson BJ, Andiman WA. Virologic and host characteristics of human immunodeficiency virus type 1-infected pediatric long term survivors. *Pediatr Infect Dis J.* 2006; 25:135–141. [PubMed: 16462290]
5. Guerini FR, Tinelli C, Calabrese E, Agliardi C, Zanzottera M, De Silvestri A, et al. HLA-A*01 is associated with late onset of Alzheimer's disease in Italian patients. *Int J Immunopathol Pharmacol.* 2009; 22:991–999. [PubMed: 20074462]
6. Mallal S, Phillips E, Carosi G, Molina JM, Workman C, Tomazic J, et al. HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med.* 2008; 358:568–579. [PubMed: 18256392]
7. Gao X, Nelson GW, Karacki P, Martin MP, Phair J, Kaslow R, et al. Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N Engl J Med.* 2001; 344:1668–1675. [PubMed: 11386265]
8. Thorsby E, Lie BA. HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. *Transpl Immunol.* 2005; 14:175–182. [PubMed: 15982560]
9. Baxter-Lowe LA, Maiers M, Spellman SR, Haagenson MD, Wang T, Fernandez-Vina M, et al. HLA-A disparities illustrate challenges for ranking the impact of HLA mismatches on bone marrow transplant outcomes in the United States. *Biol Blood Marrow Transplant.* 2009; 15:971–981. [PubMed: 19589487]
10. Morishima Y, Sasazuki T, Inoko H, Juji T, Akaza T, Yamamoto K, et al. The clinical significance of human leukocyte antigen (HLA) allele compatibility in patients receiving a marrow transplant from serologically HLA-A, HLA-B, and HLA-DR matched unrelated donors. *Blood.* 2002; 99:4200–4206. [PubMed: 12010826]
11. Shankarkumar U, Pawar A, Ghosh K. Implications of HLA sequence-based typing in transplantation. *J Postgrad Med.* 2008; 54:41–44. [PubMed: 18296807]
12. Zachary AA, Kopchaliiska D, Jackson AM, Leffell MS. Immunogenetics and immunology in transplantation. *Immunol Res.* 2010; 47(1-3):232–239. [PubMed: 20101475]
13. Hedrick PW. Pathogen resistance and genetic variation at MHC loci. *Evolution.* 2002; 56:1902–1908. [PubMed: 12449477]
14. Hughes AL, Yeager M. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet.* 1998; 32:415–435. [PubMed: 9928486]
15. Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol.* 2005; 15:1022–1027. [PubMed: 15936272]
16. Marsh, SG. Tissue Antigens. 2010. Nomenclature for factors of the HLA system, update February 2010. Epub ahead of print

17. Robinson J, Waller MJ, Fail SC, McWilliam H, Lopez R, Parham P, et al. The IMGT/HLA database. *Nucleic Acids Res.* 2009; 37:D1013–7. [PubMed: 18838392]
18. Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, et al. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.* 2003; 31:311–314. [PubMed: 12520010]
19. Bunce M, Young NT, Welsh KI. Molecular HLA typing--the brave new world. *Transplantation.* 1997; 64:1505–1513. [PubMed: 9415548]
20. Chen MJ, Chu CC, Shyr MH, Lin CL, Lin PY, Yang KL. Detection of a rare Caucasoid HLA-DRB1*0337 in a Taiwanese bone marrow donor using sequence-based typing method. *Int J Immunogenet.* 2010; 37:1–3. [PubMed: 19758311]
21. Yang KL, Chu CC, Huang CH, Lin SZ, Hsu WL, Liu DW, et al. Discovery of HLA-DRB1*0331 in a Taiwanese marrow donor and the importance of sequence-based typing in a rare or previously unrecognized allele. *Int J Immunogenet.* 2007; 34:91–95. [PubMed: 17373933]
22. Listgarten J, Brumme Z, Kadie C, Xiaojiang G, Walker B, Carrington M, et al. Statistical resolution of ambiguous HLA typing data. *PLoS Comput Biol.* 2008; 4:e1000016. [PubMed: 18392148]
23. Wu GG, Cheng LH, Li Z, Deng ZH, Zou HY, Wei TL, et al. Identification of a new HLA allele, A*1114, in a Chinese family. *Tissue Antigens.* 2003; 61:253–255. [PubMed: 12694575]
24. Adams SD, Barracchini KC, Simonis TB, Stroncek D, Marincola FM. High throughput HLA sequence-based typing (SBT) utilizing the ABI Prism 3700 DNA Analyzer. *Tumori.* 2001; 87:S40–3. [PubMed: 11401225]
25. Abbott WG, Tukuitonga CF, Ofanoa M, Munn SR, Gane EJ. Low-cost, simultaneous, single-sequence genotyping of the HLA-A, HLA-B and HLA-C loci. *Tissue Antigens.* 2006; 68:28–37. [PubMed: 16774537]
26. Paul P, Thomas D, Kawczak P, Good D, Cook DJ, Ball EJ. Resolution of cis-trans ambiguities between HLA-DRB1 alleles using single-strand conformation polymorphisms and sequencing. *Tissue Antigens.* 2001; 57:300–307. [PubMed: 11380938]
27. Adams SD, Barracchini KC, Chen D, Robbins F, Wang L, Larsen P, et al. Ambiguous allele combinations in HLA Class I and Class II sequence-based typing: when precise nucleotide sequencing leads to imprecise allele identification. *J Transl Med.* 2004; 2:30. [PubMed: 15363110]
28. Gabriel C, Danzer M, Hackl C, Kopal G, Hufnagl P, Hofer K, et al. Rapid High-throughput HLA Typing by Massively-parallel Pyrosequencing for High-Resolution Allele Identification. *Hum Immunol.* 2009; 70(11):960–964. [PubMed: 19706315]
29. Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, Trachtenberg EA, et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens.* 2009; 74:393–403. [PubMed: 19845894]
30. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. *Nat Biotechnol.* 2008; 26:1117–1124. [PubMed: 18846085]
31. Wiseman RW, Karl JA, Bimber BN, O'Leary CE, Lank SM, Tuscher JJ, et al. Major histocompatibility complex genotyping with massively parallel pyrosequencing. *Nat Med.* 2009; 15:1322–1326. [PubMed: 19820716]
32. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004; 5:113. [PubMed: 15318951]
33. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res.* 2002; 12:656–664. [PubMed: 11932250]
34. Meyer M, Stenzel U, Hofreiter M. Parallel tagged sequencing on the 454 platform. *Nat Protoc.* 2008; 3:267–278. [PubMed: 18274529]
35. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods.* 2008; 5:235–237. [PubMed: 18264105]
36. Lu Y, Boehm J, Nichol L, Trucco M, Ringquist S. Multiplex HLA-typing by pyrosequencing. *Methods Mol Biol.* 2009; 496:89–114. [PubMed: 18839107]
37. Hentze MW, Kulozik AE. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell.* 1999; 96:307–310. [PubMed: 10025395]

Figure 1.

A: Domain structure and nucleotide variability of HLA-Class-I alleles across the entire open reading frame. 2,330 HLA-A, B and C alleles were aligned and examined for regions of high sequence conservation (low variability). Percent variability represents the proportion of sequences that deviated from majority consensus across the open reading frame. cDNA-PCR primers capable of amplifying a universal 581 bp PCR product that captures most of the allelic variability were identified in regions of the $\alpha 1$ and $\alpha 3$ domains (exons 2 & 4). A second, putative 3' primer capable of generating a longer 927 bp amplicon was identified in the cytoplasmic-tail domain (exon 7). The region of apparently high variability in the trans-membrane domain is an artifact of three conserved sequence motifs that are divergent between the A, B and C loci.

B: Design and structure of 581 bp 5' and 3' cSBT primers. The primers consist of four parts: the GS-FLX adapter (which differs for forward and reverse primers which bind to A and B capture beads respectively), the key sequence used to determine legitimate sequence reads, the 10 bp MID tag which varies between samples and is used to multiplex many samples in the same library, and the HLA-specific sequence. Forward and reverse primers are shown for MID1.

Allele determinations shown in bold, with the number of ambiguous allele possibilities in parentheses. Alleles/group were determined by BLAT analysis of cSBT sequence data. Loci with grey boxes were observed to be homozygous. All alleles, except those highlighted, matched previous typing results. Alleles in green are discordant, highly similar alleles to those previously typed (indicated in yellow) and reflect non-unique amplification or hybridization patterns in the original SSP/SSOP assay. The previously typed null allele in pink was not detected by cSBT.

Table 1
Number/percentage of named HLA class I alleles assayable to various typing resolution levels across 581 bp and 927 bp diagnostic cSBT amplicons

581 bp Amplicon				
HLA Locus	Alleles Aligned	Unambiguous	4-Digit	2-Digit
A Alleles	749	619	4	126
B Alleles	1159	1021	15	123
C Alleles	422	339	13	70
All Alleles	2330	1979	32	319
% Total		84.9%	1.4%	13.7%

927 bp Amplicon				
HLA Locus	Alleles Aligned	Unambiguous	4-Digit	2-Digit
A Alleles	749	706	8	35
B Alleles	1159	1114	14	31
C Alleles	422	393	11	18
All Alleles	2330	2213	33	84
% Total		95.0%	1.4%	3.6%

Table 2

Sequence reads identified by sample/MID at each stage of analysis

Sample	Sequences Binned	Sequences Trimmed	Indel Reads	Total Aligned	Sequences Aligned		
					A Aligned	B Aligned	Cw Aligned
MID1:	754	633	185	483	156	226	101
MID2:	597	474	96	370	126	161	83
MID3:	628	503	138	413	113	204	96
MID4:	748	630	158	535	208	209	118
MID5:	831	697	171	578	138	315	125
MID6:	140	103	19	80	29	25	26
MID7:	428	343	89	277	83	128	66
MID8:	402	327	66	258	91	126	41
MID9:	547	467	96	356	103	149	104
MID10:	581	479	99	394	159	174	61
MID11:	656	551	152	451	166	181	104
MID12:	293	212	62	159	48	75	36
MID13:	396	328	77	273	108	111	54
MID14:	399	343	90	275	110	116	49
MID15:	578	500	138	427	152	123	152
MID16:	423	358	76	263	123	85	55
MID17:	543	444	138	356	155	124	77
MID18:	415	343	83	268	104	119	45
MID19:	656	577	131	467	187	181	99
MID20:	645	534	117	421	115	219	87
MID21:	524	440	100	342	117	150	75
MID22:	554	455	91	319	109	147	63
MID23:	781	657	186	526	246	182	98
MID24:	642	509	119	381	146	161	74
MID25:	101	66	13	48	7	23	18
MID26:	546	471	83	350	87	146	117
MID27:	478	391	89	318	72	141	105

Sample	Sequences Binned	Sequences Trimmed	Indel Reads	Sequences Aligned		
				Total Aligned	A Aligned	B Aligned
MID28:	361	302	80	231	48	104
MID29:	419	346	91	278	83	103
MID30:	346	284	63	234	94	85
MID31:	352	261	52	196	58	96
MID32:	0	n/a	n/a	n/a	n/a	n/a
MID33:	437	365	75	254	82	102
MID34:	562	451	88	351	109	142
MID35:	299	254	50	203	62	104
MID36:	348	272	64	212	86	88
MID37:	341	275	55	197	85	72
MID38:	425	353	73	284	117	111
MID39:	148	122	26	97	20	45
MID40:	427	344	80	279	95	98
MID41:	461	389	94	328	169	20
MID42:	564	462	112	375	200	42
MID43:	397	333	77	277	143	16
MID44:	155	126	14	91	41	19
MID45:	394	318	82	263	145	23
MID46:	689	580	157	478	254	96
MID47:	674	539	108	440	224	27
MID48:	619	514	116	385	144	146
Average	473	398	94	316	117	118
Total:	22,704	18,725	4,419	14,841	5,517	5,540
						81
						3,784