HAPCAD: An open-source tool to detect PCR crossovers in next-generation sequencing generated HLA data

Short title: HAPCAD: Detecting HLA crossovers in NGS data

Shana L McDevitt^{1,3,*}
SLM: shana.mcdevitt@gmail.com

JessenV Bredeson²
JVB: jessenbredeson@berkeley.edu

Scott W Roy³
SWR: scottwroy@gmail.com

[Julie A Lane¹](mailto:jlane@chori.org)
JAL: jlane@chori.org

Janelle A Noble¹
JAN: jnoble@chori.org

¹Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr. Way, Oakland, California, 94609 USA

²Department of Molecular & Cell Biology, University of California, 142 LSA #3200, Berkeley, CA, 94720 USA

³San Francisco State University, Department of Biology, 1600 Holloway Avenue, San Francisco, California, 94132 USA

*Corresponding Author

Shana L McDevitt
Children's Hospital Oakland Research Institute
5700 Martin Luther King Jr. Way
Oakland CA 94609
shana.mcdevitt@gmail.com

831-247-2332

Keywords: human leukocyte antigen; IMGT/HLA Database; next-generation sequencing; PCR crossovers; open-source tools

Abbreviations: Hybrid Amplicon/PCR Crossover Artifact Detector (HAPCAD); Human Leukocyte Antigen (HLA); Next-Generation Sequencing (NGS); Sequence Specific Oligonucleotide (SSO); Common and Well-Documented (CWD); Sequence-Based Typing (SBT); Parent Allele 1 (P1); Parent Allele 2 (P2)

Abstract

Next-generation sequencing (NGS) based HLA genotyping can generate PCR artifacts corresponding to IMGT/HLA Database alleles, for which multiple examples have been observed, including sequence corresponding to the *HLA-DRB1*03:42* allele. Repeat genotyping of 131 samples, previously genotyped as *DRB1*03:01* homozygotes using probe-based methods, resulted in the heterozygous call *DRB1*03:01+DRB1*03:42*. The apparent rare *DRB1*03:42* allele is hypothesized to be a “hybrid amplicon” generated by PCR crossover, a process in which a partial PCR product denatures from its template, anneals to a different allele template, and extends to completion. Unlike most PCR crossover products, “hybrid amplicons” always corresponds to an IMGT/HLA Database allele, necessitating a case-by-case analysis of whether its occurrence reflects the actual allele or is simply the result of PCR crossover. The Hybrid Amplicon/PCR Crossover Artifact Detector (HAPCAD) program mimics jumping PCR *in silico* and flags allele sequences that may also be generated as hybrid amplicon.

1.Introduction

Amplicon-based Human Leukocyte Antigen (HLA) genotyping is most commonly performed using 454TM pyrosequencing chemistry, which offers the longest Next Generation Sequencing (NGS) single-read lengths available at over 800 base pairs [1]. 454TM-based NGS platforms are among the few NGS platforms capable of sequencing through most HLA amplicons, ranging from 300 to 750 base pairs, in one read. Amplicons are designed to contain the most polymorphic and functionally relevant regions of HLA, namely, the exons encoding the peptide-binding groove [2]. However, allelic polymorphism and sequence similarity among gene sequences limit PCR primer design. Improperly developed primers may amplify other loci in addition to the target exon, or they may fail to amplify all alleles of a particular locus. While difficult to design, primers specific to single HLA exons have been successfully developed for all classical loci except the *HLA-DRB* loci.

HLA-DR molecules comprise an alpha chain, which is essentially monomorphic, and a highly polymorphic beta chain [3]. Unlike other class II molecules, the beta chain of the DR molecule can be encoded by one of four *DRB* loci [3]. Every human chromosome 6 contains a *DRB1* gene. Depending upon the *DRB1* allele present, the same chromosome may or may not also contain a *DRB3*, *DRB4*, or *DRB5* gene. Thus, the surface of antigen presenting cells may include up to four different HLA-DR molecules, with each of the different DR-beta chains paired with the same DR-alpha chain [3,4]. Like many HLA loci, these *DRB* loci share extensive sequence homology [4], which, as noted previously, creates challenges in primer sensitivity

(amplifying the peptide-binding region of all divergent alleles of the DRB loci) and specificity (amplifying only a single locus, e.g., *HLA-DRB1*, *DRB3*, *DRB4*, or *DRB5*). Current primers co-amplify all alleles of all *DRB* loci present in the same PCR [5]. One consequence of this co-amplification of these closely-related loci is that amplicons may be generated that do not reflect the genome. Specifically, spurious PCR products are generated that represent chimeras of different loci, through a phenomenon termed “jumping” PCR (or PCR crossover), in which a partial PCR product dissociates from its template and anneals to a template generated from a different allele (Figure 1)[5,6]. This produces an amplicon that combines sequences from two different HLA alleles into a hybrid sequence or PCR crossover product. For the purposes of clarification throughout this manuscript, a PCR crossover product is defined as a “hybrid amplicon” only if the PCR crossover product corresponds to an IMGT/HLA Database allele. This phenomenon can happen with amplification of two alleles of a single locus; however, amplification of multiple DRB loci with a single set of primers appears to increase the likelihood. PCR crossover products defined by Holcomb *et al.* [6], as a majority do not correspond to IMGT/HLA Database alleles and are easily detected due to mismatches between the crossover sequence and allele sequence [6]. The hybrid amplicon subgroup of PCR crossovers can however pass undetected because the sequences are exact matches to IMGT/HLA Database alleles.

Multiple examples of putative hybrid amplicons corresponding to reported HLA alleles have been observed (Table 1). A striking example was the apparent appearance of the *HLA-DRB1*03:42* allele after introducing NGS technology into the lab. One hundred and thirty-one samples, previously genotyped as *DRB1*03:01/DRB1*03:01* homozygotes using sequence specific oligonucleotide (SSO) methods, were re-genotyped using the original SSO method

primer sequences with Roche 454TM NGS-based technology. Fifty-nine of these samples (45%) resulted in the heterozygous genotype call *DRB1*03:01+DRB1*03:42*. In our data sets, the *DRB1*03:42* allele has only been seen in *DRB1*03:01/DRB1*03:01* homozygotes containing at least one copy of *DRB3*01:01:02:01*, which supports the hypothesis that the appearance of the *DRB1*03:42* allele in our samples, is attributable to a hybrid amplicon, generated by the mechanism outlined in Figure 1, and not reflective of the genome. Experiments designed to address the unexpected, frequent occurrence of *DRB1*03:42* and other rare alleles within NGS-based HLA data sets were performed using motif-specific primers (Table 1) to target the amplification of *DRB1*03:42* and *DRB1*03:01* exclusively in separate PCR reactions from a cell line genomic DNA previously resulting in both *DRB1*03:01/DRB1*03:01* homozygote and *DRB1*03:01+DRB1*03:42* heterozygote NGS-based genotypes, previously genotyped as *DRB1*03:01/DRB1*03:01* homozygote via SSO methods (Figure 2). Results displayed in Figure 3, show that under these conditions *DRB1*03:01* was amplified while *DRB1*03:42* was not amplified from the genomic DNA of the cell, validating that *DRB1*03:42* arose as a hybrid amplicon in previous genotypes with this cell line.

Because the artificial sequences generated by “jumping” PCR can create hybrid amplicons that correspond to alleles in the IMGT/HLA Database [7], they can easily be misinterpreted, particularly by an inexperienced technician, as true genomic sequence, leading to incorrect biological inferences. Importantly, whether hybrid amplicons are also reflective of the genome in some cases was tested by Holcomb *et al.* using experimental methods shown to reduce PCR crossover product generation [6], the most significant change reflecting a reduction in PCR cycle number from 35 to 28 cycles. Using samples originally used to characterize an allele within

the IMGT/HLA Database [6,7], also observed as PCR crossover in a 35 cycle PCR reaction, *Holcomb et al.* report that three of four allele sequences, here defined as hybrid amplicons, appeared to be true genomic sequence when the PCR protocol was modified to include only 28 cycles. The fourth PCR crossover allele sequence observed to differ from the IMGT/HLA Database allele by a single base validated a sequencing error in the originally reported sequence and was used to correct the IMGT/HLA Database.

In combination, Holcomb et al. and the motif-specific primer experiments show that while most sequences in the IMGT/HLA Database may represent true genomic alleles, some may represent hybrid amplicons. This underscores the need to reduce the occurrence of PCR crossover events and to detect hybrid amplicons should they occur. Although reducing the number of PCR cycles from 35 to 28 cycles severely reduced the presence of hybrid amplicons [6] the possibility still exists that PCR crossover events can occur in future experiments. Consequently, all NGS-based HLA datasets should be scrutinized for potential PCR crossover-generated sequences.

Many commercially-available HLA genotyping software programs, including Omixon Target HLATM and GenDxNGSEngine® software, currently include a PCR crossover product flagging function; however, not all laboratories use these fee-based programs. As a part of a movement in HLA genetics to develop open source tools to uniformly handle HLA data across the field, we have developed a tool to identify allele sequences within the current IMGT/HLA Database that can also arise as hybrid amplicons.

2. Materials and Methods

PCR primers (Integrative DNA Technologies, Coralville, IA) were developed to amplify regions

within target motifs common to DRB alleles of interest, *DRB1*03:01:01:01*, *DRB1*03:42*, *DRB3*01:01:02:01*, and *DRB3*01:14* (Figure 2). These motif-specific primers and Roche DRB general primers were used in twelve different combinations to amplify cell line DNA (HAR) of known homozygous *DRB1*03:01:01:01* and homozygous *DRB3*01:01:02:01* genotypes (Table 1). Each of twelve 25 µl PCR amplification reactions contained 20 ng of purified cell line genomic DNA (HAR), 1 unit of FastStart High Fidelity Polymerase (Roche Applied Sciences, Indianapolis, IN), 1X FastStart High Fidelity Reaction Buffer (1.8 mM MgCl included), 1.2 µM of PCR Grade Nucleotide Mix (Roche Applied Sciences, Indianapolis, IN), 10% Ameresco brand glycerol (Life Technologies, Carlsbad, CA) and 0.4 µM of each forward and reverse primer (Integrative DNA Technologies, Coralville, IA). All PCR amplification was performed in a gold plated 96-well Applied Biosystems 9700 thermal cycler (Life Technologies, Carlsbad, CA) and basic “two-step” cycling parameters were followed: primary denaturation 94°C/5 min., followed by 35 cycles for 94°C/15 s, and 62°C/45 s, and a final extension for 72°C/8 min. Post-PCR, target amplification products were visualized on a 2% Agarose Electrophoresis Gel (Life Technologies, Carlsbad, CA) (Figure 3).

Hybrid Amplicon/PCR Crossover Artifact Detector or HAPCAD, scripted using the Perl programming language, generates a list of potential hybrid amplicons that may arise from the co-amplification of alleles at a given locus and correspond to known HLA alleles in the IMGT/HLA reference database [7]. As previously noted, hybrid amplicons are often seen in NGS-based DRB genotyping. The HAPCAD output generated and discussed here is based on DRB exon 2 based sequencing. Nine previously recognized DRB hybrid amplicons were used as controls to validate HAPCAD’s accuracy (Table 2). HAPCAD can be used to generate a list

of potential hybrid amplicons within the exon sequences of all classical HLA loci; however, without previously recognized hybrid amplicons at these non-DRB loci program output validation is difficult and cannot be assumed.

HAPCAD also queries the most up-to-date Common and Well-Documented (CWD) allele list [8], and notes whether a potential crossover appears on the CWD list to enable a quick method of flagging rare alleles in the output. Alleles are defined as being “common” if they have been identified in multiple populations, have known frequencies, and are supported by sufficient data to verify their presence [8]. Alleles that are not as widely distributed as common alleles are categorized as being “well-documented”, these alleles must have been verified five times in unrelated individuals using Sequence Based Typing (SBT) methods or have been observed at least three times in the same haplotype in unrelated individuals for inclusion in the current CWD list [8]. Approximately 9% of the current IMGT/HLA allele list are CWD alleles, the remaining ~91% of alleles should be considered “rare alleles” [8]. *DRB1*03:42* falls into this category.

HAPCAD source code is open-source and available for download at <https://bitbucket.org/smcdevitt/hapcad>. The program is well documented with a detailed usage/help page, accessed using the “--help” command line option or simply by executing the program without any of the required options or inputs. A program “READ ME” is also posted and provides an even more detailed overview of program usage requirements. HAPCAD does not currently have a graphical user interface and is executed using the command line interface on either a UNIX or Linux operating system.

HAPCAD source code is executable and can be run globally by altering the environmental

variable or run locally by placing HAPCAD and its dependables in the working directory.

HAPCAD can be executed with a number of options, which will be detailed below as they relate to program functionality. All options should be defined prior to inputs on the command line.

The latest “DRB_nuc.txt” alignment file, in IMGT/HLA Database release version 3.20.0, (<ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/>), was used as primary input for the HAPCAD program. It is crucial that only an unaltered IMGT/HLA “X_nuc.txt” alignment file, where X is the HLA locus, be used as program input, any other file type will result in premature program termination. The second program input is generated using information from the published CWD allele list available at <http://cwd.immunogenomics.org>, the allele names with the “Common” or “Well-Documented” designation are provided to the program in comma separated value (CSV) format (example input per HLA locus available at <https://bitbucket.org/smcdevitt/hapcad>).

HAPCAD, by default, generates potential hybrid amplicons pairwise between all distinct sequences within an alignment file, generating a hybrid to represent a PCR crossover at each polymorphic position between the two sequences. The program creates the hybrid by appending the 5’ sequence before each polymorphism, from the first parent allele (P1), to the remaining 3’ sequence starting at the polymorphic position from the second parent allele (P2). The generated hybrids are then cross referenced back to the allele alignments to determine which IMGT/HLA alleles could also potentially be hybrid amplicons generated by PCR crossover (Figure 4).

HAPCAD can be optimized to match an individual PCR-based approach by selecting the

alignment range to be considered as an *in silico* amplicon. For example, the reported DRB general fusion primers anneal in exon 2, requiring that only some alignment positions be considered when generating the potential hybrid amplicons, which can be specified using the “-r” or “--range” option. Deletions in the alignment input file must be considered when specifying the program range; for example, the 454 DRB general fusion primers (Roche, Pleasanton CA) generate an amplicon including base 105-345 of the “DRB_nuc.txt” alignment; however, the range input should be 109-350 to take into account the four deletions between positions 21 and 22 and the single deletion between positions 162 and 163 of the alignment file, which offset the start base range by four positions and the end base range by five positions.

For full program usage refer to the “READ ME” or the extended help/usage page available in the source repository. The DRB program output analyzed here was generated with the following command:

```
% HAPCAD -o DRBv3.20.0_crossovers.out -r 109-350 DRB_nuc.txt DRB_CWD_v2.0.0.csv,
```

where % represents the command line prompt, HAPCAD is the program executable, “-o” redirects STDOUT to generate a program output file called “DRBv3.20.0_crossovers.out”, “-r” dictates the alignment range from bases 109 through 350 to be considered in hybrid generation, and “DRB_nuc.txt” and “DRB_CWD_v2.0.0.csv” are the input data files on which the program operates. If the “-o” option is not used, the program output will print to the screen and will not be captured in a file. Debug mode can also be activated by adding the “--debug” option after the program executable but before the input files, which will generate a verbose output to STDERR reporting program progress and detailed lists of the sequences created and compared by the program. The “-v” option can also be executed to simply print the IMGT/HLA

allele version from the alignment file to STDOUT. Activating the “-v” option will result in the premature termination of the program and should only be used to double check the alignment file version number prior to executing HAPCAD to generate the hybrid amplicon output.

HAPCADoutput consists of a single comma-separated file (CSV-formatted file) including 1) the name of the allele matching the generated hybrid amplicon (hybrid); 2) the name of the allele that contributes the 5' exon 2 sequence (first parent allele or P1); 3) the name of the allele that contributes the 3' exon 2 sequence (second parent allele or P2); 4) the polymorphic positions flanking the region of sequence homology between P1 and P2, within which the PCR crossover must occur between P1 and P2 sequences to generate the hybrid; 5) a notation of whether that allele is a CWD allele and at which field; and 6) a notation distinguishing whether an allele is C, WD, or “.” as a place holder for non-CWD alleles.

For example, a single output line would read:

“DRB1*03:42,DRB3*01:01:02:01,DRB1*03:01:01:01,22-97,.”,

where *DRB1*03:42* is the allele name corresponding to the hybrid amplicon produced by any combination of *DRB3*01:01:02:01* (P1) and *DRB1*03:01:01:01* (P2) with a junction between the 127th and 202nd alignment positions, and “.” indicates that *DRB1*03:42* is not a CWD HLA allele. The HAPCAD output also includes the IMGT alignment version and used to generate the program output and the date and time the output was generated.

3. Results

A number of DRB generated hybrid amplicons have been observed in our data sets (Table 2) for which the nucleotide position range in which the PCR crossover must have occurred was

manually determined. This combined set of manually-curated putative crossovers (from Table 1) was used to validate the accuracy of the HAPCAD program in reporting hybrid amplicons. As noted in Table 1, HAPCAD correctly flagged 9 of 9 previously observed hybrid amplicons and their expected crossover ranges. However, because the number of possible hybrid amplicons was not known prior to executing HAPCAD, one cannot determine whether the output includes all potential alleles.

HAPCAD results indicate that of the 1,822 DRB alleles in the IMGT/HLA Database version 3.20.0, at least 1,175 of those alleles could be generated as a hybrid amplicon in 7,318,367 discrete DRB allele combinations, substantially increasing the observed repertoire of possible hybrid amplicons reported in Table 1 and by Holcomb *et al.* [6]. A caveat is that HAPCAD inflates the potential number of likely hybrid amplicons due to the inability to rule out the use of hybrid amplicons as parent alleles as well. For example, *DRB1*03:42* and *DRB3*01:14* can be generated by PCR crossover between *DRB1*03:01:01:01* and *DRB3*01:01:02:01*; however, if these hybrid amplicon corresponding alleles are paired as parent alleles, *DRB1*03:01:01:01* and *DRB3*01:01:02:01* will be flagged as potential hybrid amplicons generated by *DRB1*03:42* and *DRB3*01:14*. CWD information has been included as a tool to identify alleles that may be flagged as hybrid amplicons due to this phenomenon, the most likely hybrid amplicons will not be included in the CWD column but will be flagged as rare. However, the designation of an allele as CWD is not sufficient for exclusion as a potential crossover. For example, *DRB1*16:01:01:01* is a common allele that is also noted in our data sets as the PCR crossover product of *DRB1*15:01:01:01*, *DRB1*15:02:01*, or *DRB1*15:04* with *DRB5*01:01:01* or *DRB5*01:02*. CWD alleles appearing on the HAPCAD output can potentially be ruled out or deemed less likely to exist as a hybrid amplicon if the

parent alleles are not CWD alleles.

The number of potential hybrid amplicons in the HAPCAD output is even further increased by the inability to assume which partial PCR products are likely to form during PCR crossover; for instance, it is unlikely that a partial PCR product would be very short or represent a nearly complete amplicon. HAPCAD produces potential hybrid amplicons using all, but the first, polymorphic positions between both sequences (Figure 2), even if the initial partial product is only a handful of bases long. Future HAPCAD iterations will include a hybrid amplicon score, ranking the probability that a flagged allele is likely to be a hybrid amplicon, enabling the user to more efficiently use the HAPCAD output to predict the nature of potential hybrid amplicons within their data.

In its current form, a HAPCAD output can be opened in a spreadsheet based program, such as Microsoft Excel, sorted based on the hybrid amplicon containing column, and searched using the “find” function in excel. Although not ideal, while a second output parsing program is still in development, in this fashion the current program output can be used to search for hybrid amplicons within a data set. Further, to help weed out hybrid amplicons that may only be listed as a product of the above noted inflationary factors, users are advised to 1) scrutinize hybrid amplicons that have junction positions at the very beginning or end of the sequence and 2) query whether CWD designated hybrid amplicons also have two parent alleles listed as hybrid amplicons. Hybrid amplicons reported by HAPCAD meeting these criteria are less likely to reflect an actual hybrid amplicon.

HAPCAD executed with an Apple MacBook Pro, Darwin 14.3.0 (x86-64) UNIX operating system with 16GB RAM and an Intel(R) Core(TM) i7 dual-core, 3.1 GHz (4MB L3 cache)

processor completed in roughly 18 minutes. HAPCAD is intended to be optimized and executed for each type of PCR-based HLA locus specific genotyping experiment. HAPCAD is still being validated for use with HLA Class I loci, characterized by high allele numbers, but is still expected to complete within an hour on the system described above.

4. Discussion

Putative hybrid amplicons generated from partial HLA sequences can be determined from the IMGT/HLA Database using simple bioinformatics techniques. HAPCAD compares alleles in the IMGT/HLA Database and determines if any combination of two allele sequences can produce a putative hybrid amplicon that corresponds to a HLA allele. The program mimics jumping PCR *in silico*, systematically generating hybrid sequences, and comparing these putative hybrid amplicons to allele sequences.

HAPCAD is currently optimized to identify potential hybrid amplicons generated during exon-specific amplicon-based HLA genotyping methods. These methods were once the most cost-efficient NGS-based genotyping methods, but whole gene amplification-based NGS genotyping methods are quickly becoming more affordable and may soon supplant the previous exon-based methodologies [9]. Since exon-based and whole gene-based methods, as well as any foreseeable HLA genotyping methods, start with PCR amplification, the potential for PCR crossover remains a reality. Currently, HAPCAD requires a coding sequence alignment file but future program implementations will be adapted to function with any IMGT HLA alignment files to enable the search for hybrid amplicons subsequent to long-range PCR required for whole gene-based HLA genotyping strategies.

Subsequently, we hope to develop HAPCAD to be a web-based program, to facilitate the

comparison between the HAPCAD output and user-specific HLA datasets, directly reporting potential hybrid amplicons in user data and eliminating the necessity for user familiarity with operating on a command line. Further, the HAPCAD tool, as it stands, could be included as a part of other HLA NGS genotype analysis pipelines. Currently, an individual performing HLA genotyping would have to recognize the need to seek out tools to reduce the impact of hybrid amplicons in their data. Ideally, similar to fee-based HLA genotype programs, tools like HAPCAD will be included with open-source HLA genotyping software in the future. Future goals of this work include incorporation of user feedback and collaborations with other groups generating open-source HLA genotyping data analysis tools. Users are encouraged to download and execute HAPCAD tailored to their specific applications.

The full DRB HAPCAD output, source code, and CWD inputs are posted on BitBucket[™] and can be retrieved at <https://bitbucket.org/smcdevitt/hapcad>.

Acknowledgments

Funding for this work was partially supported by the National Institute of Health (DK61722 to J.A.N.), the National Science Foundation Science Masters Program Fellowship (DGE-1011717), and the California State University Program for Education and Research in Biotechnology Travel Grant Award program. We also thank Steven J. Mack and the entire Noble Lab, from Children's Hospital Oakland Research Institute, for thoughtful discussion and theoretical guidance.

References

- [1] Niklas N, Pröll J, Danzer M, Stabentheiner S, Hofer K, Gabriel C. Routine performance and errors of 454 HLA exon sequencing in diagnostics. *BMC Bioinformatics* 2013;14:176. doi:10.1186/1471-2105-14-176.
- [2] Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, et al. High-throughput, high-fidelity HLA genotyping with deep sequencing. *Proc Natl Acad Sci U S A* 2012;109:8676–81. doi:10.1073/pnas.1206614109.
- [3] Noble JA, Erlich HA. Genetics of Type 1 Diabetes. *Cold Spring Harb Perspect Med* 2012;2:a007732–a007732. doi:10.1101/cshperspect.a007732.
- [4] Andersson G. Evolution of the human HLA-DR region. *Front Biosci* 1998;3:d739–45.
- [5] Erlich H. HLA DNA typing: past, present, and future. *Tissue Antigens* 2012;80:1–11. doi:10.1111/j.1399-0039.2012.01881.x.
- [6] Holcomb CL, Rastrou M, Williams TC, Goodridge D, Lazaro AM, Tilanus M, et al. Next-generation sequencing can reveal in vitro-generated PCR crossover products: some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. *Tissue Antigens* 2014;83:32–40. doi:10.1111/tan.12269.
- [7] Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 2014;43:D423–31. doi:10.1093/nar/gku1161.
- [8] Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, et al. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* 2013;81:194–203. doi:10.1111/tan.12093.
- [9] De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16(th) IHIW : review of HLA typing by NGS. *Int J Immunogenet* 2013;40:72–6. doi:10.1111/iji.12024.

Figure Legends

Figure 1. “Jumping” PCR/PCR Crossover Schematic. Figure 1 displays the “Jumping” PCR mechanism used to explain the *in vitro* production of DRB1*03:42 from partial DRB3*01:01:02:01 and DRB1*03:01:01:01 alleles often found on a DR52 haplotype[4].

Figure 2: Motif-Specific Primer Schematic. Figure 2 displays a schematic of the amino acid sequence and annealing location the motif-specific primer targets in comparison to the Roche DRB general primers. Relevant alleles are also listed with their appropriate forward and reverse motif-specific primers, which are designed to exclusively amplify the target allele.

Figure 3: Motif-Specific Primer Experiment Results. *Figure 3 displays the amplification results from the Motif-Specific Primer Experiment. 2% Agarose electrophoresis was used to visualize the amplification products for the twelve noted primer pair combinations. If DRB1*03:42 were amplified, as detailed in Table 1, amplification (~350 bp) would be expected in gel lane A4.*

Figure 4. HAPCAD *in silico* PCR Crossover Mechanism. Figure 4 shows the basic mechanism by which HAPCAD generates putative hybrid amplicons with the pair-wise manipulation of IMGT/HLA Database allele sequences. Two parent alleles are selected for which the polymorphic positions are determined. Hybrid sequences are then constructed using the 5' sequence of parent allele 1 (P1) up to the base before each polymorphism and the remaining sequence of parent allele 2 (P2), including the polymorphic position through the end of the sequence. A second hybrid is then generated corresponding to the same polymorphic position consisting of the 5' P2 sequence up to the base before each polymorphism and P2 sequence

starting at the polymorphism until the end of the sequence. Parent allele sequences contributing to each hybrid are differentiated by a “|” before the polymorphic position in the figure. HAPCAD does not consider the first polymorphic position, as the resultant hybrids will always regenerate a parent allele. The putative hybrid sequences are then compared against the remaining alleles in the HLA alignment file to determine if any of the hybrid sequences correspond to an HLA allele. If a match is identified, the PCR crossover range will be reported between the polymorphic position at which the matched hybrid was generated and the previous polymorphic position, indicating the region of sequence homology between the two parent alleles.

Table Legends

Table 1 Motif-Specific Primer Pair Combinations and Expected Amplifications. Table 1 displays twelve primer pair combinations used to amplify HLA, cell line DNA. Motif-specific targets and the allele sequences that may be amplified (DRB1*03:01, DRB1*03:42, DRB3*01:01, and DRB3*01:14) with specific combinations are noted.

Table 2. HAPCAD Output Validation. Table 2 shows the HAPCAD output values reported for the nine previously observed DRB PCR crossovers. The presence of these sequences within the output with the correct reported crossover ranges was used as a control to validate the accuracy of the program. The “expected” ranges are the ranges manually-curated from the IMGT alignment file and the “observed” ranges are the ranges reported in the program output. The expected and observed ranges are correct for all nine previously reported DRB hybrids.

Figure 1

Figure 2

Forward Primer	Reverse Primer	Target Allele	Putative Hybrid Amplicon
YSTS	V	DRB1*03:01	No
LRKS	V	DRB1*03:42	Yes
LRKS	G	DRB3*01:01	No
YSTS	G	DRB3*01:14	Yes

Figure 3

Figure 4

PCR crossover is modeled *in silico* with IMGT/HLA allele pairs

Parent Allele 1 -----A---G-----
Parent Allele 2 -----G-----A---

A pair of hybrid sequences are generated for each polymorphic position

Position 5 (P1+P2) -----|⁵G-----A--- (Regenerates P2)
Position 5 (P2+P1) -----|---A---G----- (Regenerates P1)
Position 8 (P1+P2) -----|⁸-----A---
Position 8 (P2+P1) -----G---|A---G-----
Position 13 (P1+P2) -----A---|¹³-----A---
Position 13 (P2+P1) -----G-----|G-----
****Position 39 (P1+P2)** -----A---G-----|A--- ****** ³⁹ ←
Position 39 (P2+P1) -----G-----|-----

Sequences are compared to IMGT/HLA Database alleles and matching sequences are flagged

The crossover range is determined by the polymorphic positions directly before and referencing the matched sequence (i.e. 13 – 39)

Table 1

Table 1: Motif-Specific Primer Pair Combinations and Expected Amplifications

Combination	Forward Primer	Reverse Primer	Expected Allele Amplification	
1	YSTS	V	DRB1*03:01	
2	YSTS	G	DRB3*01:14	
3	YSTS	DRB general	DRB1*03:01, DRB3*01:14	
4	LRKS	V	DRB1*03:42	
5	LRKS	G	DRB3*01:01	
6	LRKS	DRB general	DRB1*03:42, DRB3*01:01	
7	LLKS	V	None	
8	LLKS	G	None	
9	LLKS	DRB general	None	
10	DRB general	V	DRB1*03:42, DRB1*03:01	
11	DRB general	G	DRB3*01:01, DRB3*02:02, DRB3*01:14	
12	DRB general	DRB general	All	

Table 2: HAPCAD Output Validation

Hybrid	First Parent Allele (5' Sequence)	Second Parent Allele (3' Sequence)	Crossover Range	
			Expected	Observed
DRB1*03:42	DRB3*01:01	DRB1*03:01	127-202	127-202
DRB3*01:14	DRB1*03:01	DRB3*01:01	127-202	127-202
DRB3*02:04	DRB3*02:02	DRB1*03:01	245-318	245-318
DRB1*16:09	DRB1*15:02	DRB5*01:02	233-293	233-293
DRB1*16:01	DRB1*15:02	DRB5*01:02	203-233	203-233
DRB1*16:09	DRB1*15:01	DRB5*01:01	233-293	233-293
DRB1*16:01	DRB1*15:01/15:04	DRB5*01:01	205-233	205-233
DRB1*16:09	DRB1*15:04	DRB5*01:01	233-301	233-301
DRB5*01:02	DRB5*02:02	DRB1*16:01	203-293	203-293