Published in final edited form as:

Hum Immunol. 2015 December ; 76(12): 939–944. doi:10.1016/j.humimm.2015.07.235.

Concordance of Next Generation Sequence-based and Sequence Specific Oligonucleotide Probe-based *HLA-DRB1* genotyping

Julie A. Lane¹, Jameel R. Johnson^{1,2}, and Janelle A. Noble¹

Julie A. Lane: jlane@chori.org; Jameel R. Johnson: jameel.johnson@bcm.edu; Janelle A. Noble: jnoble@chori.org

¹Children's Hospital Oakland Research Institute, Oakland, CA

Abstract

Next generation sequencing (NGS) of clonally amplified DNA, using Roche 454 technology, was used to genotype *HLA-DRB1*, *DRB3*, *DRB4*, and *DRB5* loci (exon 2 only) from a set of 993 samples from newborns with maternally-reported African American ancestry. *DRB1* exon 2 was genotyped previously on the same sample set using sequence-specific oligonucleotide probe (SSOP) technology. Comparison of the genotype calls from both methods indicated concordance of 92.3%. Some discordance was expected due to the higher resolution of NGS data, compared to SSOP data. This resulted from selection of the incorrect allele from the ambiguity string produced by SSOP genotyping. Of 76 discordant genotypes, only three were due to resolution of ambiguity with the NGS method. The low percent of changes due to the increased resolution of the NGS method instills confidence in the overall value of previous data genotyped with moderate resolution methods, i.e., the vast majority of alleles present in a population are those that are detectable at moderate resolution. The remaining 73 discordant genotypes resulted from preventable errors in sample handling, data interpretation, and data entry. These results underscore the potential for error that can result from factors such as low quality genomic DNA, manual data entry, and interpretation of marginal genotyping results. Optimization of genomic DNA quality, automation of genotyping steps wherever possible, and use of the highest resolution technology available can lead to dramatic improvements in HLA genotype data quality. NGS-based methodology generated data of superior quality and accuracy compared to the SSOP system.

Corresponding author: Janelle A. Noble, PhD, Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr. Way, Oakland, CA 94609, jnoble@chori.org, 510-450-7657.

²Present address: Dr. Jameel R. Johnson, Ben Taub General Hospital, Dept of Internal medicine, 1504 Taub Loop, #2RM 81-001, Houston, TX 77030

AUTHOR CONTRIBUTIONS

J.A.L. generated data and wrote the manuscript. J.J. generated data. J.A.N. oversaw the project and wrote the manuscript.

CONFLICT OF INTEREST

No conflict of interest is reported for any of the authors.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

DRB genotyping; next-generation sequencing; linear array; HLA

1. INTRODUCTION

Beginning in 2003, we genotyped *DRB1*, *DQA1*, and *DQB1* in 998 DNA samples obtained from blood spot cards (Guthrie cards) from newborns in the state of California with maternally-reported African American ancestry. These data were generated to provide population-based frequencies in African Americans for use as controls for diseases with low prevalence in African Americans, and, in particular, for use in our study of type 1 diabetes in African Americans [1]. For T1D, prevalence estimates suggest that the cohort should include at most one future patient[2]. The original genotypes were generated with sequence-specific oligonucleotide probe (SSOP) technology. Beginning in 2011, the cohort was genotyped with next generation sequencing based genotyping to determine frequencies for *DRB3*, *DRB4*, and *DRB5* loci. The Roche 454 exon-based NGS method provides concurrent sequencing of those loci and *DRB1*; thus, the sample set was genotyped twice for *DRB1* (exon 2 only). These two data sets allowed evaluation of the concordance of *DRB1* data generated eight years apart on the same sample set with different technologies.

2. SUBJECTS AND METHODS

2.1 Subjects

Four thousand newborn blood spot cards, 1000 from each of Asian, African American, Hispanic, and White maternally-reported ethnic groups, were received from the State of California Department of Public Health, Genetic Disease Branch. Samples were used for population genetic studies with permission from the state of California and with approval of the Institutional Review Board from Children's Hospital and Research Center Oakland.

2.2 Genomic DNA preparation

In 2002, DNA was extracted from punches from 4000 blood cards using QiaAmp 96 DNA Blood kits (Qiagen, Valencia, CA) in a 96-well block extractor per manufacturers instructions[3]. Cards were not sorted by ethnicity prior to DNA preparation; thus, African American samples were distributed at random among the total sample set[4]. DNA samples were stored in sets of 96 in boxes containing individual, unlabeled tubes. All steps of the DNA preparation and transfer were performed by a single individual. For *DRB1* genotyping of African American samples, DNA tubes from 1000 African American samples were identified, and DNA was manually transferred to 96-well plates. Due to limited sample availability and lack of availability of appropriate technology at the time to measure small volumes of DNA at low concentrations, DNA concentrations and yields were not measured. A target volume for PCR amplification was empirically determined. All SSOP linear array *DRB1* data were generated using these DNA preparations.

In 2007–2008, DNA was re-extracted from residual blood on blood spot cards. All residual dried blood was removed from each card, and DNA was prepared in individual tubes using

QIAamp blood kits (Qiagen, Valencia, CA) with a modified protocol. DNA samples were arrayed in 96-well plates, quantitated by spectrophotometry on a Nanodrop instrument (Thermo Scientific, Waltham, MA), and normalized to 10ng/μl. To ensure the greatest accuracy possible and minimize sample mix up, all steps of the DNA preparation process up to and including plating of the individual samples were performed in teams of two people.

2.3 DRB1 Genotyping by linear array

Initial *HLA-DRB1* genotyping was performed using SSOP linear array technology using noncommercially available reagents developed at and provided by Roche Molecular Systems [5,6]. Specifically, a two-step process was employed for DRB1 genotyping to minimize confounding by other *DRB* loci. Primers located at either end of DRB1 exon 2 (DRB “general” primers) were used for the initial amplification. These primers also amplify other *DRB* loci, including *DRB3*, *DRB4*, and *DRB5*. PCR products were hybridized to an array of probes specific to sequence motifs encoding the region around amino acid position 10 of the *DRB1*-encoded polypeptide. Allele-specific PCR primers corresponding to the sequence motifs in the position 10 probes (DRB “subtyping” primers) were then used, in combination with the *DRB* general downstream primer, to amplify the original sample for the allele group(s) indicated by the initial linear array. Secondary amplification was not performed for signals from probes specific for *DRB1*07:01*, *DRB1*09:01*, and *DRB1*10:01* due to the limited polymorphism of those allele groups known at the time. Up to two secondary amplification products per sample, specific to subgroups of *DRB1* alleles and not including other *DRB* loci, were hybridized to an array of probes corresponding to sequence motifs located downstream of the position 10-encoding region in DRB1 exon 2. Probe binding patterns were hand-entered into a customized version of Sequence COmpilation and REarrangement (SCORE™) software (v6.00.45)[7]. The combination of the *DRB1* general result with the *DRB1* subtyping result was used to generate the genotype call in SCORE™ with IMGT/HLA reference database HLADB2.2.0[8]. DNA concentrations were not determined for the initial set of DNA preparations due to low concentration and limited sample. The volume of DNA to include in the PCR reactions was empirically determined; thus, the amount of genomic DNA in each PCR reaction varied. In addition, because of the low concentration of the DNA samples, 40 to 50 cycles of amplification were used, rather than the usual 35.

2.4 DRB1 genotyping by NGS-based technology

For NGS genotyping, DNA samples from the second round of DNA preparation were used. These samples were of higher concentration and quality than those used for linear array genotyping. Amplification of exon 2 of the *DRB* genes was performed using “fusion” primers, which are locus-specific primers to which multiplex sample identifiers (MIDs) and platform-specific adaptors have been added[9]. All samples were sequenced with the Roche 454 GS Junior (Roche 454, Branford, CT); 130 samples were repeated on the Ion Torrent Personal Genome Machine (PGM) (Life Technologies, Carlsbad, CA) platform to confirm genotype calls. Fusion primers for both sequencing platforms differed only for the platform-specific adaptors; locus-specific sequences were identical and matched those used for the linear arrays. After initial amplification, individual PCR products were quantitated and pooled in equimolar amounts to create sequencing libraries. Libraries were mixed with

beads (platform-specific) such that a maximum of one molecule should have been attached to any bead. Clonal amplification of PCR product on individual beads was performed by emulsion PCR, using emPCR (lib-A) kits for the 454 platform (Roche 454, Branford, CT) or template kit OT2 for the PGM platform (Life Technologies, Carlsbad, CA). Beads containing amplified DNA were purified away from empty beads, using magnetic separation with bead recovery reagents for the 454 platform (Roche 454, Branford, CT) or the OneTouch 2 system for the PGM platform (Life Technologies, Carlsbad, CA). Appropriate numbers of loaded beads were sequenced per manufacturers' protocols, using GS Junior titanium kits for the 454 platform (Roche 454, Branford, CT) and PGM 400v2 sequencing kits and 314v2 chips for the PGM platform (Life Technologies, Carlsbad, CA). Genotype calls were made using a combination of Conexio ATF Assign™ (Conexio Genomics, Freemantle, Western Australia) and Sequence COmpilation and REarrangement (SCORE™) software[7]. Data from the 454 platform were processed by onboard amplicon processing software and further processed with a custom Amplicon Variant Analysis (AVA) script. Resulting reads in FASTA format were imported into a customized version (v1.1.0.35) of Conexio ATF Assign™ (Conexio Genomics, Freemantle, Western Australia). Data output from the PGM platform (Torrent Suite version 4.0.2) was in FASTQ format and was converted to FASTA format before importing into Conexio ATF Assign™. To de-multiplex samples, SCORE™ (v6.00.45) was customized to generate a Conexio input .xml file containing each sample ID linked to a unique 10-base MID. Initial genotype calls were made in Conexio ATF Assign™, output files were imported into SCORE™ to create a data report consistent with those for SSOP data and to consolidate data into a single software. IMGT/HLA database release HLADB_3.10.0 was used for reference [8].

3. RESULTS

3.1 NGS sequencing results

HLA-DRB1 exon 2 sequence data were generated from 993 samples using NGS methods. All samples were sequenced on the Roche 454 platform. Sample plates included 92 DNA samples and four negative control wells per plate. Each of ten full plates and one partial plate were sequenced individually on the 454 platform. One hundred thirty samples, with either low read depth (<60× total coverage) or with genotype calls discordant with original SSOP data, were re-genotyped in two runs on the PGM platform for confirmation; samples successfully run on both platforms yielded identical genotypes. Read depths varied by sample and sequence run; most samples had greater than 30 reads per allele per direction (120× coverage). Manual editing of genotype calls was required for only 15% of genotypes using the Conexio ATF Assign™ software. No edits to the consensus sequence were necessary, and no novel polymorphisms were detected.

3.2 Genotyping and overall concordance

Data from 993 samples, genotyped by both SSOP and NGS methods, were examined for concordance. In total, 76 genotypes were discordant between the two data sets, giving an overall concordance rate of 92.3%. Notably, 18 samples were originally classified as homozygous but turned out to be heterozygous, which, in our experience, is not uncommon for HLA data. Apparent homozygosity can sometimes result from an allele dropout in the

initial PCR. In this case, however, the quality of the DNA and of the DRB1 “general” genotyping strips was suboptimal, leading to inaccuracies in selection of samples for subsequent “subtyping” reactions. In NGS based genotyping, the number of reads for a homozygous allele in a sample was usually higher than that for each allele of a heterozygote in that particular sequencing run, increasing the confidence that the homozygous call was correct.

Because the SSOP method produces calls with greater ambiguity (i.e., more possible calls consistent with the primary data) than the NGS-based method, some level of discordance is expected due to the inherent differences in the technologies. With NGS genotyping and two-field HLA resolution very little ambiguity is generated in the results. In most cases, the ambiguity string for a given genotype was either non-existent or contained one additional genotype. The highest observed ambiguity string in these NGS data was four genotypes. Discrepancies between the two data sets could be classified into three main categories, 1) ambiguity resolution, in which initial data were correct, but the common allele was chosen from the ambiguity string when the sample really contained a rare allele, 2) genotype calling errors, which can be sorted into data entry errors, data interpretation errors, and errors resulting from poor quality DNA or reagents, and 3) sample handling errors, including errors in sample identification or selection and sample mixups. These are summarized in Table 1 and addressed individually below.

3.3 Ambiguity resolution (0.3%)

Four percent (n = 3) of the 76 discordant genotype calls are attributable to ambiguity resolution issues. The DRB1 locus is extremely polymorphic, with 1638 named alleles (<http://www.ebi.ac.uk/ipd/imgt/hla/stats.html>)[10]. Thus, genotyping with a moderate-resolution method such as the linear array results in an ambiguity string that lists multiple genotype combinations consistent with the probe-binding data. For purposes of data analysis, we select the most likely genotype from the string, i.e., the genotype containing the most common set of alleles. In some cases, the real genotype contains one or two rare alleles and, while present in the ambiguity string, is not selected for the final data set because genotypes containing two common alleles are also present. Frequently, high-resolution, NGS-based genotyping can resolve the ambiguity and reveal the identity of the true allele. In these cases, the initial genotyping data are correct, but the wrong selection was chosen from the ambiguity string. This situation occurred for only three samples in the comparison and resulted from the fact that the SSOPs in the linear array do not cover every nucleotide position of exon 2. For one sample, the called genotype included the allele *DRB1*11:01* from the linear array data; NGS data revealed that the allele was actually *DRB1*11:37*. In a second sample, the genotype called from linear array data included *DRB1*11:02*, which was actually shown to be *DRB1*11:48* by NGS. Both of these distinctions were attributable to an A at nucleotide position 227, which was not tested by the linear array. For the third sample, a genotype that included a called *DRB1*12:01*- from linear array data turned out to include *DRB1*12:25* when genotyped by NGS, due to the presence of a C at position 280 that was not tested in the linear array. Figure 1 shows the polymorphisms that were not detectable by the linear array. In all three of these cases, the allele that was originally called appears to be much more frequent than the allele that was revealed by the NGS data. Frequencies for these

alleles, reported by the National Marrow Donor Program from genotyping 416,581 African American samples, are *DRB1*11:01* (0.0854), *DRB1*11:37* (0.0000096), *DRB1*11:02* (0.0388), *DRB1*11:48* (0.000018), and *DRB1*12:01* (0.0382); *DRB1*12:25* was not observed in African Americans (<https://frequency.nmdp.org/NMDPFrequencies2011/jsp/haploTypeFrequencies.jsp>)[11].

The set of reported HLA alleles is constantly expanding[1]. When the initial SSOP data were generated in 2003, the alleles *DRB1*11:48* and *DRB1*12:25* had not yet been reported. This illustrates the importance of storing primary genotyping data, rather than simply storing data in the form of allele calls. In addition, a standardized system of HLA genotyping data notation and storage would facilitate comparison of HLA genotyping data among researchers and across platforms, enabling meta-analyses. Finally, standardization of HLA notation has clinical implications for improved specificity in donor selection.

3.4 Genotype calling errors (5.1%)

The largest category for discrepancies was genotype calling errors. Sixty-seven percent (n = 51) of the 76 discordant genotype calls are attributable to genotype calling errors (Table 2). These can be sorted into data entry errors, data interpretation errors, and errors due to marginal quality raw data. For the linear array data, probe binding patterns were entered into SCORE™ manually. Nine instances of data entry errors were identified. For most, a probe adjacent to the probe with the positive signal was entered as positive, leaving the true positive probe to appear negative, e.g., probe 9 was positive but probe 10 was entered into SCORE™.

Eighteen instances of data interpretation errors were seen. These occurred for multiple reasons, but were mostly attributable to assessing whether a probe with a weak signal was positive or negative. The linear array assay, by its nature in having multiple DNA probes that must hybridize to target DNA under identical conditions, contained some probes that regularly cross-hybridized to similar alleles, and some probes that were faint compared to the other signals (Figure 2). Distinguishing cross-hybridization from true signals can be difficult and requires experience with the technology, both in general and with the specific reagent lot.

Finally, we identified twenty-four discrepancies that were a direct result of marginal raw data. Poor quality data may be attributable to a number of factors, including poor-quality reagents (DNA, linear arrays, or hybridization reagents), equipment issues, or operator error. For the SSOP *DRB1* genotyping, some lots of linear arrays may have exceeded their shelf life, resulting in faint or missing probes, especially when combined with low quality DNA or suboptimal hybridization conditions. In some cases, when results from the *DRB1* general assays were unclear, allele-specific priming and subtyping were attempted on samples that did not contain the sequence motif specified by the allele-specific primer. These primers occasionally amplify alleles with the wrong motif, although generally with low efficiency. Since the subtyping linear arrays are used for all allele groups, mis-amplified PCR product can lead to an incorrect assumption of the presence of an allele.

3.5 Sample handling errors (2.2%)

Twenty-two discrepant calls resulted from sample handling errors. This represents 29% of the total discrepancies. These are the most preventable of all and underscore the need for extreme care. For the original genotyping effort, DNA was selected from 998 unlabeled tubes in a set of 4000. A single individual marked the tubes containing African American samples and manually transferred those samples to 96-well plates. Sources of potential error include marking the wrong tube(s), retrieving DNA from the wrong tube(s), and replacing the 8-tube strip cap rotated 180 degrees, such that the wrong tubes appear marked.

For the second extraction of DNA from the cards, multiple safeguards were used to ensure that the DNA were assembled in 96-well plates in the same order as in the original sample set. For all steps leading to and including plating of DNA into 96-well plates, at least two people were present. Cards were first sorted into sets of 92 for each plate and then into sets of 11 (in four rows intended to include a negative control well) or 12 for each row. One person presented the cards, in order, to another person who punched the blood spots into tubes arranged in 96-well format. The cards were checked before and after punching each set, to make sure they were in the correct order. Two workers were also present when DNA aliquots were transferred from tubes to 96-well plates. Twenty-two of the HLA DRB1 genotypes generated by DNA sequencing are from a different sample than was in the corresponding location in the original sample set. Given the great care taken during assembly of the second set of extracted blood spot DNA, the most likely explanation is that errors were made in constructing the first sample set.

3.6 Data changes

The DRB1 data for 993 samples (1986 allele calls) were corrected for 3 “ambiguity resolution” and 51 “genotype calling” errors (54 corrections in 53 samples). The *DRB1* allele frequency distributions of the original data and the corrected data are shown in Table 3. After correction, three alleles (*DRB1**03:07, *DRB1**11:07, and *DRB1**12:03), called in the original data, were no longer present, and seven alleles (*DRB1**03:15, *DRB1**08:08, *DRB1**11:37, *DRB1**11:48, *DRB1**12:25, *DRB1**13:16, *DRB1**14:08), not called previously, were included. *HLA-DQA1* and *-DQB1* genotypes were available for 990 of the 993 samples in this sample set. Estimated *DRB1-DQA1-DQB1* haplotypes were generated using the “Bridging ImmunoGenomic Data-Analysis Workflow Gaps” (BIGDAWG) analysis system (<https://cran.r-project.org/web/packages/BIGDAWG/vignettes/BIGDAWG.html>). Correction of the genotyping errors resulted in decreased heterogeneity in the dataset. The uncorrected data included 179 unique *DRB1-DQA1-DQB1* haplotype combinations, of which 87 appeared only once. The corrected data included 156 unique haplotypes, of which 71 appeared only one time. Most of the haplotypes that were removed appeared unusual compared to established patterns of linkage disequilibrium, adding support to the assumption that the NGS data are correct. In addition, primary data for the SSOP were re-examined for the samples with discrepant calls, and the reason for each was revealed. Most of the samples reported here were part of the 1641 sample control data set in a study of HLA association with type 1 diabetes in African Americans. When the corrected genotypes were replaced in the context of the larger data set, the number of unique haplotypes changes from 251 to 228, with the removal of 32 and addition of 8 unique haplotypes. Genotype

diversity decreased as well, with the initial 958 unique genotypes reduced to 935 by the removal of 46 and addition of 23 genotypes. The unique haplotypes that were added to or removed from the data appeared only once or twice in the data set; thus, they would be in the group of haplotypes that were binned in the association analyses and would not change the overall association results.

4. DISCUSSION

Creating a large data set that is 100% error free, while the ultimate goal, is, in practice, very difficult to achieve. For HLA, consideration of genotyping method depends on the available methodology at the time of the genotyping and generally represents a tradeoff between cost and resolution. The fact that the blood card samples described here were genotyped twice for DRB1 provided a forum to test the importance of the improved resolution. The fact that only 3 samples (0.3% of the total) had genotype changes due to the higher resolution of the NGS system suggests that most alleles seen in a given population, even one as diverse as African Americans, are common. This result is encouraging in regard to the validity of results from other studies where genotyping was performed with SSOP or other genotyping systems. The fact that the remaining 73 discrepant calls (7.4% of the total) were due to preventable errors was unexpected and somewhat concerning. The 22 sample handling errors are not related to the genotyping systems, but the 51 genotype calling errors directly reflect the quality of the genotyping system and the performance of the technical staff at the time of the genotyping. Tracing the source of those errors illuminated ways in which future errors might be avoided. Having experienced staff perform experiments is one factor that can affect the quality of the data. Even for experienced staff, having strict procedures in place to minimize errors cannot be undervalued. For example, when a large number of individual tubes must be handled, having a second person watching to ensure that samples remain in the correct order can help avoid sample mix ups. Use of liquid handling equipment, whenever feasible, is preferable to manual sample manipulation. Bar coding of samples is always preferable to hand writing on sample tubes. Hand entry of data should be avoided whenever possible. Ideally, sample identifiers and any data associated with them should be imported and exported electronically, or, at the very least, copied and pasted to avoid keystroke entry. The original genotyping of these samples was performed with low quality DNA and less than optimal reagents, using systems that required manual data entry. Mistakes were made, despite attempts to create an error-free data set. However, the fact that the changes resulting from correction of these errors occurred mostly in rare haplotypes, combined with the fact that the magnitude of the changes was small enough not to change the association data, instills confidence in previous HLA and disease association studies.

The choice of HLA genotyping is dependent on two major factors, cost and resolution. However, another factor to consider is the reliability of the data that are generated. Based on our experience, we recommend NGS-based genotyping as the optimal choice.

Acknowledgments

The authors wish to thank Carolyn Hoppe for procuring and sharing the blood cards from the State of California Department of Public Health, Genetic Disease Branch; Lara Vigil, David Nguyen, Shaily Arora, and Uche Eke for technical assistance; Shana McDevitt for technical assistance and data formatting; Derek Pappas for assistance with

haplotype estimation; and Nelle Cronen for assistance with manuscript preparation. This work was supported by grant DK61722 to J. A. N.

Abbreviations

HLA	Human Leukocyte Antigen
NGS	Next Generation Sequencing
SSOP	sequence-specific oligonucleotide probe
T1D	type 1 diabetes
PGM	Personal Genome Machine
MID	Multiplex Identifier

References

1. Noble JA, Johnson J, Lane JA, Valdes AM. HLA class II genotyping of African American type 1 diabetic patients reveals associations unique to African haplotypes. *Diabetes*. 2013; 62:3292. [PubMed: 23801574]
2. Mayer-Davis EJ, Beyer J, Bell RA, Dabelea D, D'Agostino R Jr, Imperatore G, et al. Diabetes in African American youth: prevalence, incidence, and clinical characteristics: the SEARCH for Diabetes in Youth Study. *Diabetes Care*. 2009; 32(Suppl 2):S112. [PubMed: 19246576]
3. Hoppe C, Watson RM, Long CM, Lorey F, Robles L, Klitz W, et al. Prevalence of HFE mutations in California newborns. *Pediatr Hematol Oncol*. 2006; 23:507. [PubMed: 16849282]
4. Lane JA, Noble JA. Maximizing deoxyribonucleic acid yield from dried blood spots. *J Diabetes Sci Technol*. 2010; 4:250. [PubMed: 20307384]
5. Erlich H, Valdes AM, Noble J, Carlson JA, Varney M, Concannon P, et al. HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. *Diabetes*. 2008; 57:1084. [PubMed: 18252895]
6. Mychaleckyj JC, Noble JA, Moonsamy PV, Carlson JA, Varney MD, Post J, et al. HLA genotyping in the international Type 1 Diabetes Genetics Consortium. *Clin Trials*. 2010; 7:S75. [PubMed: 20595243]
7. Helmberg W, Lanzer G, Zahn R, Weinmayr B, Wagner T, Albert E. Virtual DNA analysis--a new tool for combination and standardised evaluation of SSO, SSP and sequencing-based typing results. *Tissue Antigens*. 1998; 51:587. [PubMed: 9694350]
8. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG. The IMGT/HLA database. *Nucleic Acids Res*. 2013; 41:D1222. [PubMed: 23080122]
9. 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. [PubMed: 19845894]
10. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. 2015; 43:D423. [PubMed: 25414341]
11. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum Immunol*. 2013; 74:1313. [PubMed: 23806270]

Figure 2. Examples of probe cross-hybridization in linear array data

The figure shows a portion of DRB1 linear array subtyping “strips” from 6 samples. Probe 28 is specific to extremely rare alleles; the faint signals seen on the strips are all due to cross-hybridization. Probe 30 illustrates the difference in signal strength that can be observed in these assays. In this case, strong signals indicated an exact match to sample DNA, and faint signals reflected cross-hybridization. On strip 6, the faint signal for probe 30 was incorrectly called positive, which led to an incorrect genotype call.

Table 1

Categories for discordance between data sets.

Discrepancy type	Number
Ambiguity resolution	3
Genotype calling	51
Data entry	9
Data interpretation	18
Low quality DNA	24
Sample handling	22

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

SSOP genotypes compared to NGS genotypes for 53 samples

SSOP DRB1_1	SSOP DRB1_2	NGS DRB1_1	NGS DRB1_2	Discrepancy Type
01:01	04:04	01:01	04:10	2
01:01	13:02	01:03	13:02	2
01:02	01:02	01:02	11:01	3
01:02	01:02	01:02	11:01	4
01:02	01:02	01:02	14:01	4
01:03	13:01	01:03	15:03	3
03:01	11:02	03:15	11:02	3
03:01	11:02	11:02	11:02	4
03:01	11:07	03:01	11:02	4
03:01	13:02	03:01	13:16	3
03:01	13:02	03:01	13:01	3
03:02	08:02	03:02	08:08	2
03:02	08:04	08:04	08:04	4
03:07	13:01	07:01	13:01	4
04:01	12:03	04:01	12:01	3
07:01	07:01	07:01	11:01	4
07:01	13:01	07:01	13:03	3
07:01	13:01	07:01	07:01	4
07:01	13:02	07:01	11:01	2
07:01	13:02	07:01	13:01	2
07:01	13:04	07:01	11:02	2
08:04	08:04	03:02	08:04	4
09:01	09:01	04:05	09:01	4
09:01	09:01	09:01	13:02	4
09:01	09:01	09:01	15:03	4
09:01	12:03	09:01	12:01	3
10:01	10:01	10:01	13:01	4
11:01	11:01	04:04	04:11	3
11:01	11:04	11:01	11:02	3
11:01	11:04	11:01	11:02	4
11:01	13:02	11:37	13:02	1
11:01	15:03	11:04	15:03	2
11:02	11:02	04:01	11:02	4
11:02	11:07	09:01	11:02	3
11:02	15:03	11:48	15:03	1
11:02	15:03	04:04	15:03	4
11:04	11:04	11:02	11:04	3
11:04	13:03	11:01	13:03	2
12:01	12:01	03:01	12:01	4

SSOP DRB1_1	SSOP DRB1_2	NGS DRB1_1	NGS DRB1_2	Discrepancy Type
12:01	12:01	12:01	13:02	4
12:01	12:03	03:01	12:01	4
12:01	13:02	08:03	13:02	2
12:03	15:03	12:25	15:03	1/3
13:01	13:03	13:02	13:03	3
13:02	13:02	03:02	13:01	4
13:02	14:01	13:01	14:08	3
13:02	15:03	15:03	15:03	3
13:02	15:03	13:16	15:03	3
13:02	15:03	13:16	15:03	3
13:03	13:03	13:03	15:03	4
15:01	15:01	13:01	15:01	4
15:03	15:03	13:01	15:03	4
15:03	15:03	11:04	15:03	4

Discrepancy types: 1 = ambiguity resolution; 2 = data entry error; 3 = data interpretation error; 4 = marginal quality raw data. Note that one sample was included in both the set of 3 ambiguity resolution samples and the set of 51 genotype calling errors.

Table 3

DRB1 frequencies from data generated by SSOP (Original count) and after changes to 53 samples, based on NGS data.

DRB1*	Original count (freq)	Revised count (freq)
01:01	52 (2.62%)	51 (2.57%)
01:02	81 (4.08%)	78 (3.93%)
01:03	7 (0.35%)	8 (0.40%)
03:01	139 (7.00%)	139 (7.00%)
03:02	114 (5.74%)	115 (5.79%)
03:07	1 (0.05%)	0 (0.00%)
03:15	0 (0.00%)	1 (0.05%)
04:01	33 (1.66%)	34 (1.71%)
04:02	4 (0.20%)	4 (0.20%)
04:03	7 (0.35%)	7 (0.35%)
04:04	16 (0.81%)	17 (0.86%)
04:05	25 (1.26%)	26 (1.31%)
04:07	17 (0.86%)	17 (0.86%)
04:08	1 (0.05%)	1 (0.05%)
04:10	1 (0.05%)	2 (0.10%)
04:11	8 (0.40%)	9 (0.45%)
07:01	215 (10.83%)	216 (10.88%)
08:01	8 (0.40%)	8 (0.40%)
08:02	5 (0.25%)	4 (0.20%)
08:03	3 (0.15%)	4 (0.20%)
08:04	97 (4.88%)	97 (4.88%)
08:06	7 (0.35%)	7 (0.35%)
08:08	0 (0.00%)	1 (0.05%)
08:11	1 (0.05%)	1 (0.05%)
09:01	72 (3.63%)	70 (3.52%)
10:01	39 (1.96%)	38 (1.91%)
11:01	127 (6.39%)	128 (6.45%)
11:02	73 (3.68%)	76 (3.83%)
11:03	2 (0.10%)	2 (0.10%)
11:04	19 (0.96%)	17 (0.86%)
11:07	2 (0.10%)	0.00 (0%)
11:10	2 (0.10%)	2 (0.10%)
11:37	0 (0.00%)	1 (0.05%)
11:48	0 (0.00%)	1 (0.05%)
12:01	78 (3.93%)	77 (3.88%)
12:02	8 (0.40%)	8 (0.40%)
12:03	4 (0.20%)	0 (0.00%)
12:25	0 (0.00%)	1 (0.05%)

DRB1*	Original count (freq)	Revised count (freq)
<i>13:01</i>	109 (5.49%)	112 (5.64%)
<i>13:02</i>	138 (6.95%)	131 (6.6%)
13:03	67 (3.37%)	67 (3.37%)
<i>13:04</i>	24 (1.21%)	23 (1.16%)
<i>13:16</i>	0 (0.00%)	3 (0.15%)
14:01	40 (2.01%)	40 (2.01%)
14:02	2 (0.10%)	2 (0.10%)
14:04	1 (0.05%)	1 (0.05%)
<i>14:08</i>	0 (0.00%)	1 (0.05%)
14:13	1 (0.05%)	1 (0.05%)
<i>15:01</i>	62 (3.12%)	61 (3.07%)
15:02	6 (0.3%)	6 (0.3%)
<i>15:03</i>	232 (11.68%)	234 (11.78%)
16:01	2 (0.1%)	2 (0.1%)
16:02	34 (1.71%)	34 (1.71%)

* alleles with changed counts after revision are indicated in bold italic