Predicting an HLA-DPB1 Expression Marker Based on Standard DPB1 Genotyping: Linkage Analysis of Over 32,000 Samples

Bianca Schöne¹, Sabine Bergmann¹, Kathrin Lang¹, Ines Wagner¹, Alexander H. Schmidt¹,², Effie W. Petersdorf³, and Vinzenz Lange¹,§

¹DKMS Life Science Lab, Blasewitzer Str. 43, 01307 Dresden, Germany
²DKMS, Kressbach 1, 72072 Tübingen, Germany
³University of Washington, Fred Hutchinson Cancer Research Centre, Seattle, WA

Abstract

The risk of acute graft-versus-host disease (GvHD) after hematopoietic stem cell transplantation is increased with donor-recipient HLA-DPB1 allele mismatching. The single-nucleotide polymorphism (SNP) rs9277534 within the 3′ untranslated region (UTR) correlates with HLA-DPB1 allotype expression and serves as a marker for permissive HLA-DPB1 mismatches. Since rs9277534 is not routinely typed, we analyzed 32,681 samples of mostly European ancestry to investigate if the rs9277534 allele can be reliably imputed from standard DPB1 genotyping. We confirmed the previously-defined linkages between rs9277534 and 18 DPB1 alleles and established additional linkages for 46 DPB1 alleles. Based on these linkages, the rs9277534 allele could be predicted for 99.6% of the samples based on DPB1 genotypes (99.99% concordance). We demonstrate that 100% prediction accuracy could be achieved if the prediction utilized exon 3 sequence information. DPB1 genotyping based on exon 2 data alone allows no unambiguous rs9277534 allele prediction but was estimated to maintain 99% accuracy for samples of European descent. We conclude that DPB1 genotyping is sufficient to infer the DPB1 expression marker rs9277534 with high accuracy. This information could be used to select donors with permissive HLA-DPB1 mismatches without directly screening for rs9277534.

Keywords

HLA-DPB1; Gene Expression; SNP; Linkage; Genotyping; GvHD; rs9277534; Permissive Mismatch

§Corresponding author.

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.

Supplementary Files

Supplementary material available
1. Introduction

Hematopoietic stem cell transplantation (HSCT) from unrelated donors can cure various blood disorders [1]. A high level of donor-recipient HLA compatibility is crucial for the success of HSCT. Currently, matching for HLA-A, -B, -C, -DRB1 and -DQB1 alleles (10/10 match) is the gold standard for the selection of unrelated donors, while HLA-DPB1 is often not considered [2]. This leads to DPB1 mismatching in 80% to 86% of otherwise HLA-matched donor-recipient pairs [3,4]. Mismatching of HLA-DPB1 between donor and recipient is associated with a significantly higher risk of acute GvHD [5], which is a major impediment to successful transplantation.

Recent studies indicate that not all DPB1 mismatches confer equal risks after transplantation [6–9]. The presence of specific amino acid residues within hypervariable regions defined by DPB1 exon 2 has been shown to play a role in alloreactivity and risk of GvHD after transplantation [7,8]. In addition to a role for the recognition of T-cell epitopes, the level of HLA-DP expression in the patient influences the incidence of GvHD after HSCT, where mismatching against a highly-expressed patient allele leads to significantly higher GvHD risk compared to other combinations [9].

The expression level of an HLA-DPB1 allele was shown to be correlated with variant rs9277534 [10], an A→G SNP located within the 3′ UTR of DPB1 (Fig. 1A). This SNP therefore may serve as marker for HLA-DPB1 expression level even though the SNP itself is most likely not directly involved in DPB1 expression control. Conserved haplotypes of DPB1 exon 2 and rs9277534 were defined for common DPB1 alleles by direct phasing [9]. The rs9277534-A allele is associated with low DPB1 expression, whereas the rs9277534-G allele is associated with high DPB1 expression. When DPB1-matched donors are not available, this expression marker can be used to prospectively identify DPB1-mismatched donors who generate a permissive DPB1 mismatch against low-expression patient DPB1 alleles [9].

The 3′ UTR containing rs9277534 is currently not covered by routine genotyping assays for HLA-DPB1. An additional screening would increase the cost of HLA genotyping in the clinical setting for donor selection. Therefore, the aim of our study was to verify if the rs9277534 genotype can be accurately predicted from high-resolution DPB1 genotyping. For this purpose, we examined DPB1 exon 2, exon 3, and rs9277534 as separate amplicons in 32,681 individuals using our lab’s high throughput NGS workflow based on sequencing with Illumina MiSeq and HiSeq instruments (Fig. 1B) [11,12]. In addition, we applied full-length Single Molecule Real-Time (SMRT) sequencing for a small subset of 22 samples with unexpected linkage patterns after primary analysis. Our full-length sequencing approach covers the whole DPB1 gene including rs9277534 in the 3′ UTR in one amplicon (Fig. 1C) and therefore directly delivers the requisite information for phasing coding and non-coding polymorphisms.
2. Materials and methods

2.1. Samples

Samples were provided by DKMS donor centers in Germany (69.3%), USA (13.2%), UK (11.9%), and Poland (5.2%) and by the BMST (Bangalore Medical Services Trust) donor center in India (0.4%) between March 2016 and May 2016. The samples approximately reflect the ethnic diversity encountered in these countries. (Detailed information about the diversity of provenance for each donor center is listed in Supplementary Material 1.) The majority of samples (98%) were collected with nylon FLOQSwabs™ hDNA free (Copan Italia Spa, Brescia, Italy) and 2% as 150 μl whole blood in Venosafe 4 ml EDTA tubes (Terumo, Tokyo, Japan). During registration, donors signed an informed consent approving HLA genotyping and other analyses to facilitate or improve donor search for stem cell transplantation. No ethics committee approval was obtained as the described genotyping is within the scope of this consent form and performed as a genotyping service.

2.2. Isolation and quantification of DNA

DNA was isolated from buccal swabs or whole blood using the magnetic-bead-based “Chemagic DNA Buccal Swab kit Special” or “Chemagic DNA Blood Kit Special” (Perkin Elmer, Baesweiler, Germany), respectively. The isolated DNA was eluted in 100 μl elution buffer (10 mM Tris-HCl pH8.0), and its concentration was measured by fluorescence (SYBR Green, Biozym, Hessisch Oldendorf, Germany) using the TECAN infinite 200Pro plate reader (Tecan, Männedorf, Switzerland). Samples with DNA concentrations of < 2 ng/μl were excluded from genotyping.

2.3. Amplicon-based sequencing of HLA-DPB1 exons 2 and 3 and rs9277534

2.3.1. PCR—Exons 2 and 3 of the HLA-DPB1 gene were amplified in a multiplexed PCR as previously described [11], resulting in amplicons about 340 bp in length. The rs9277534 region was targeted by an additional primer group (forward: GAATGACTGATTCTCAGTGAGCTGCC, reverse: ACATGATTGCTTTGCCTTTTCCCAAC and ACGTATTGCTTTGTCTCTTTCCCCAG), resulting in PCR products about 425 bp in length. (The length and position of all three amplicons is depicted in Fig. 1B.)

The two DPB1 PCR reactions were performed alongside PCRs for several other loci in 10 μl volume each, using 384-well plates with FastStart™ Taq DNA Polymerase (Roche, Basel, Switzerland) and the associated buffer system [12]. All amplification products belonging to one sample were pooled using volumes appropriate to obtain balanced read coverage for each amplicon. A secondary PCR was performed on the pool to elongate the amplicons with indexes and sequencing adapters for Illumina sequencing. Target-specific primers and indexing primers were obtained from Metabion (Metabion International AG, Planegg, Germany).

2.3.2. Sequencing—After indexing PCR, 384 barcoded samples were pooled together. Pooled PCR products were purified with SPRselect beads (BeckmanCoulter, Brea, USA) with a ratio of 0.6:1 beads to DNA, and subsequently quantified by qPCR. Denaturation and
dilution of the sequencing library were executed as recommended by Illumina (MiSeq Reagent Kit v2-Preparation Guide). Libraries with commonly 384 or 3840 samples were loaded at 12.5 pM onto MiSeq or HiSeq flow cells, respectively, with 10% PhiX spiked in. Paired-end sequencing was performed for 2 times 249 cycles with MiSeq Reagent Kit v3 or HiSeq Rapid SBS Kit v2 (Illumina, San Diego, USA) on MiSeq or HiSeq, respectively.

2.3.3. Genotyping—DPB1 alleles were assigned from exons 2 and 3 sequences using the genotyping software neXtype as previously described [11]. rs9277534 alleles were interpreted with FasType, developed in Python. FasType uses the GEM mapper [13] for mapping reads against reference sequences. According to the allele sequences in the IPD-IMGT/HLA database, all currently described full-length DPB1 sequences converge to two distinct sequences at the rs9277534 amplicon region. These two sequences differ by seven individual SNPs and were used as reference for rs9277534. Samples with less than 50 reads coverage on any of the three amplicons (exon 2, exon 3, rs9277534) were not included in the analysis to avoid spurious results due to allelic drop outs [14]. The threshold for calling a heterozygous result for rs9277534 was set to 20% of reads mapped to the target region. All genotyping results and analyses contained in this study are based on the IPD-IMGT/HLA release 3.23, which was the current version at the time of genotyping.

2.4. Whole-gene sequencing of HLA-DPB1

Samples with unexpected linkage patterns or otherwise inconclusive results after primary analysis were selected for secondary analysis on a Pacific Biosciences RS II instrument (Pacific Biosciences, Menlo Park, USA) as previously described [15]. To this end, the entire DPB1 gene, inclusive of 5′ and 3′ UTRs, was amplified by long-range PCR (Fig. 1C). A barcoded sequencing library was prepared using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, USA) following standard protocols. Base calling and de-multiplexing was performed using Pacific Biosciences’ SMRT Portal. The whole-gene sequence data was analyzed with NGSengine (GenDx, Utrecht, The Netherlands).

2.5. Linkage verification of samples with phasing ambiguities

Due to the lack of physical phasing between exons 2 and 3 during primary analysis, some of the DPB1 genotyping results contained phasing ambiguities. These samples are characterized by having two distinct sequences in exons 2 and 3 each, of which each combination has been reported as allele in the IPD-IMGT/HLA database. Such genotyping results are described by NMDP codes leading to a loss of information. For example, the genotyping result DPB1*04:ADCGE (including DPB1*04:01 and DPB1*350:01)/DPB1*17:JECX (including DPB1*17:01 and DPB1*131:01) could represent a DPB1*04:01/DPB1*17:01; b) DPB1*04:01/DPB1*131:01; c) DPB1*350:01/DPB1*17:01, or d) DPB1*350:01/DPB1*131:01. However, DPB1*04:01 and DPB1*17:01 and likewise DPB1*350:01 and DPB1*131:01 share the same exon 3 sequences. Therefore, neither a) nor d) would correctly reflect the genotype of the sample. On the other hand, samples actually harboring these allele combinations would not result in NMDP codes, but in straightforward two-field resolution genotyping (DPB1*04:01/DPB1*17:01 or DPB1*350:01/DPB1*131:01). Therefore, all samples with phasing ambiguities were resolved and only the valid allele combinations were used to assign the expected rs9277534 alleles for the linkage
analysis. In the given example, all valid combinations (DPB1*04:01/DPB1*131:01 or 
DPB1*350:01/DPB1*17:01) led to an expected heterozygous rs9277534. In contrast, the 
invalid combinations (DPB1*04:01/DPB1*17:01 or DPB1*131:01/DPB1*350:01) would 
have led to an expected homozygous rs9277534 (AA or GG, respectively).

2.6. Estimation of DPB1 allele frequencies

From the observed genotypes, we formulated a Bayesian multinomial model and employed 
Gibbs sampling as described in [15] in order to estimate DPB1 allele frequencies. Since this 
algorithm assumes standard Hardy Weinberg equilibrium, we restricted the analysis to the 
subset of 19,301 samples of self-declared German ancestry.

3. Results

3.1. Amplicon-based genotyping of 32,681 samples for DPB1 exons 2 and 3 and rs9277534

We examined 32,681 samples meeting all quality criteria as defined in Section 2.3.3. 
Evaluation of the reads obtained for the rs9277534 amplicon showed a high and even 
coverage of both alleles (Fig. 2). Read numbers averaged at about 1200 reads per sample in 
both heterozygous and homozygous samples (Fig. 2A). In heterozygous samples, both 
alleles were amplified in a balanced fashion with each allele covered by 40 to 60% of the 
reads (Fig. 2B). Therefore, there was no indication of an amplification bias for either allele.

3.2. Establishment of novel linkages

Due to the lack of physical phasing from the amplicon-based genotyping approach, 
identification of the linkage between rs9277534 and DPB1 exons 2 and 3 was performed 
indirectly (Fig. 3). In samples with a homozygous rs9277534, the SNP allele found was 
declared to be linked to both DPB1 alleles. Conversely, samples with a heterozygous 
rs9277534 were only used to define a novel linkage if one of the two alleles had an already 
established linkage. In these cases, the known linkage was assumed to be correct, and the 
remaining DPB1 allele and observed rs9277534 allele were defined as linked. Subsequently, 
newly established linkages were used in an iterative analysis until no further novel linkages 
could be defined. Using this procedure, we confirmed the previously reported linkages 
(originally based only on exon 2) and established novel linkages for 46 DPB1 alleles, which 
we confirmed in at least three samples, each. Of these, 22 alleles are linked to rs9277534-A, 
whereas 24 alleles are linked to rs9277534-G (Table 1).

3.3. Verification of known and novel linkages

Based on the linkages defined in Table 1, we re-evaluated all 32,681 samples for unexpected 
linkages. Due to the lack of physical phasing between exons 2 and 3 during our primary 
analysis, some of the DPB1 genotyping results contained phasing ambiguities which 
prevented two-field genotyping resolution. Since those ambiguities may include alleles of 
both linkage groups, they required a more sophisticated analysis (for more details, see 
Methods, section 2.5). Therefore, we separated our sample pool into three subsets: 1) 
samples predicted to be homozygous for rs9277534, 2) samples predicted to be 
heterozygous without phasing ambiguities, and 3) samples with phasing ambiguities (Table 
2).
The first subset constituted the simplest case, in which a homozygous rs9277534 was expected based on the DPB1 genotyping. Some of those DPB1 genotyping results contained ambiguities due to the lack of phasing; however, those ambiguities were of no relevance to the rs9277534 linkage as all alleles belonged to the same linkage group. The homozygous subset contained 19,322 samples, of which 18 showed an unexpected linkage in the primary analysis. All but one of these unexpected linkages were disproved as artifacts during secondary analysis by full-length SMRT sequencing (artifact sources are explained below). The remaining unexpected linkage was a sample with DPB1*34:01 linked to rs9277534-A.

The second subset consisted of samples with an unambiguous DPB1 genotyping, for which a heterozygous rs9277534 was expected. Of the 7,481 samples in this group, four samples showed an unexpected linkage in the primary analysis. Two of these were disproved as artifacts during secondary analysis. Intriguingly, the remaining two samples both contained DPB1*34:01 linked to rs9277534-A, as already observed in the first subset.

The third subset consisted of samples with ambiguous DPB1 genotyping results due to the lack of phasing between exons 2 and 3. We resolved these codes into valid allele combinations and determined the expected rs9277534 alleles for all these combinations. We confirmed the expected heterozygous rs9277534 alleles in all of the 5,761 samples in this group.

A total of 117 samples were excluded from the linkage analysis at this point, including 31 samples harboring novel DPB1 alleles and 23 samples with rare alleles observed only once or twice. Additionally, 63 samples, heterozygous for rs9277534, were excluded because they encoded one allele with an undefined exon 3 sequence and a second allele with remaining phasing ambiguities. Due to the lack of exon 3 sequence information, the second allele could only be resolved to a group of alleles which linked to both SNP alleles.

In total, our primary analysis identified 22 samples with an unexpected linkage, 19 of which were disproved as artifacts during secondary analysis. The majority of these artifacts resulted from contaminations acquired during the workflow or sample acquisition (17 samples). The contaminated samples were flagged as such during routine genotyping analysis of DPB1 based on the detection of more than two alleles in at least one HLA locus. Independent samples from the same individuals were subjected to SMRT analysis, which confirmed the predicted linkages in all 17 cases. In the remaining two artifact samples, secondary analysis by SMRT sequencing revealed a PCR drop out during primary analysis. In both samples, one of the rs9277534 alleles failed to amplify for unknown reasons.

In summary, the analysis of 32,564 samples revealed a concordance rate of 99.99% with the linkages proposed in Table 1. In only 117 samples (0.4% of all samples), a linkage could not be proposed and therefore neither confirmed nor disproved. These cases included very rare or novel alleles or particular phasing ambiguities. All samples with unexpected linkages confirmed by SMRT analysis carried a DPB1*34:01 allele linked to rs9277534-A instead of rs9277534-G.
3.4. Novel alleles stemming from evolutionary recombination events between exons 2 and 3

Most DPB1 alleles in the IPD-IMGT/HLA database lack sequence information for exon 3. One of those alleles with an unknown exon 3 sequence is DPB1*34:01. In our sample set, we identified 18 samples carrying a DPB1*34:01 allele. For nine of these, a linkage could not be proposed due to the phasing issue explained above (Section 2.5). Since in six of the remaining nine samples, DPB1*34:01 was linked to rs9277534-G, we defined the linkage accordingly in Table 1. In these six samples, the exon 3 sequences of the DPB1*34:01 alleles were identical to the DPB1*01:01 exon 3 sequence. In contrast, in the three samples with unexpected linkage, the exon 3 sequences of the DPB1*34:01-like alleles were identical to the DPB1*02:01 exon 3 sequence. Therefore, full-length sequencing revealed two distinct alleles sharing the same DPB1*34:01 exon 2 sequence but having distinct exon 3 sequences. Interestingly, the rs9277534 variant linked to each allele is in concordance with the linkage defined for other alleles with the same exon 3 sequence (Fig. 4).

The nine samples previously excluded because of the phasing issues confirmed this pattern (postulating the presence of the most frequent of the compatible alleles and the corresponding linkages; the odds of most frequent versus less frequent alleles ranged from 1:16 to 1:1200). Seven of these samples carried the DPB1*34:01-rs9277534-G allele with the DPB1*01:01 exon 3 sequence, while the two remaining samples harbored the DPB1*34:01-rs9277534-A allele with the exon 3 sequence of DPB1*02:01. Interestingly, according to self-declared ethnic origin, all 13 individuals carrying the DPB1*34:01-rs9277534-G allele were of German (11) or Turkish (2) descent, while the DPB1*34:01-rs9277534-A allele was identified in samples from a different range of ethnicities, i.e. three individuals of self-declared Iberian descent, one from Sub-Saharan Africa, and one from the UK/Ireland. This suggests that the frequency of the two DPB1*34:01 like alleles varies greatly between populations.

Additionally, ten of the 31 samples with novel alleles encountered during this study were characterized by hitherto unknown combinations of known exon 2 and 3 sequence features (Table 3). Interestingly, only for one sample, the detected rs9277534 allele matched the allele expected based on the exon 2 sequence. However, in all ten samples, the detected rs9277534 allele matched the expectation when only considering the observed exon 3 sequence. Therefore, we hypothesized that rs9277534 might be more accurately predicted from exon 3 than exon 2. The full sequences of all these novel alleles, including both variants of DPB1*34:01, are currently being submitted to the IPD-IMGT/HLA database. In addition, as part of a separate project, we are currently submitting full-length sequence data of many DPB1 alleles which are so far only partially defined.

3.5. Relevance of exon 2 or exon 3 for the linkage to rs9277534

In the previous section, we described several cases of alleles sharing the same exon 2 sequences but linked to different rs9277534 alleles. However, in all these cases, the alleles with alternative rs9277534 alleles also featured alternative exon 3 sequences. Furthermore, the DPB1 gene structure is characterized by a large intron between exons 2 and 3 (4 kb, Fig. 1A). Therefore, the distance between rs9277534 and exon 2 is 6 kb, whereas the distance...
between rs9277534 and exon 3 is only 1.8 kb. Taken together, this suggests that the exon 3-rs9277534 linkage might be more reliable than the exon 2-rs9277534 linkage.

On the basis of the linkages defined for 64 alleles (Table 1), we performed a theoretical sequence analysis to establish the linkage between exon 2 or exon 3 sequences and rs9277534 (Supplementary Material 2). This analysis revealed that several alleles sharing the same exon 2 sequences are linked to different rs9277534 alleles. Based on allele frequencies (as determined for the subset of samples with German origin), however, we estimated an error probability of 0.9% for predicting the rs9277534 variant based on exon 2 sequence information alone. In contrast, our sequence analysis revealed no alleles which share the exon 3 sequences but are linked to alternative rs9277534 alleles. However, this linkage analysis disregarded all alleles with undefined exon 3 sequences, which account for a large portion of known alleles (541 of 630 DPB1 alleles in the IPD-IMGT/HLA database), albeit with low frequencies.

We therefore re-analyzed the complete dataset of 32,681 samples for the linkage of exon 3 sequences to rs9277534. This analysis was performed purely on observed sequence features (ignoring allele naming and associated limitations) and included alleles with undefined exon 3, the 23 rare alleles with undefined linkage, and the 31 novel alleles with SNPs or uncommon exon 2/exon 3 combinations. This analysis revealed 100% concordance of rs9277534 linkage for all alleles sharing the same exon 3 sequence features in all 32,681 samples (Table 4, Supplementary Table 4). These results demonstrate that exon 3 alone provides sufficient information for inferring the association of rs9277534 alleles.

Furthermore, this analysis underscored the limited sequence variability in exon 3 in the analyzed sample set. The three most common sequence features (DPB1*02:01, DPB1*01:01 and DPB1*04:01) accounted for 99.15% of the observed alleles in our study. Inclusion of DPB1*15:01 increases coverage to 99.89% of the observed alleles. For each exon 3 sequence feature, Table 4 lists the alleles observed in this study and those that were not observed but share the same exon 3 sequence. Based on the perfect linkage between exon 3 and rs9277534 observed in our sample set, we assume that the linkage will extend to these unobserved alleles as well. In summary, we found 100% linkage between the exon 3 sequence and the rs9277534 allele, whereas a prediction of rs9277534 based on exon 2 sequences alone yields an error rate of about 0.9%.

3.6. Exon 3 sequence analysis

Four described exon 3 sequence features were not observed in this study (Table 4). However, sequence similarity analyses may suggest a linkage for these based on the observed correlations. Therefore, we aligned all known exon 3 sequence features of DPB1 to reveal common motifs (Fig. 4). The sequences define two separate groups that are consistent with regard to their linkage to rs9277534. We conclude that they probably represent phylogenetically separate clades for rs9277534-A and rs9277534-G. Interestingly, the two primary sequences, distinguished by seven SNPs, are also the two most common sequence features. All other known exon 3 sequences deviate only by one SNP from one of those two root sequences, including the four not observed alleles with unique exon 3 sequences. Based on this clear segregation, we postulate that DPB1*168:01 and DPB1*534:01 are linked to...
rs9277534-A while DPB1*498:01 and DPB1*533:01 are linked to rs9277534-G. We hypothesize that novel exon 3 sequence features will align clearly with one or the other rs9277534 allele. Indeed, the 16 samples with novel exon 3 sequences observed in this study support this hypothesis, confirming the perfect linkage between rs9277534 and exon 3.

4. Discussion

Upfront DPB1 genotyping of transplant patients and potential unrelated donors has the potential to decrease GvHD risk after HSCT by selecting HLA 12/12-matched donors or restricting DPB1-mismatched donors to permissive combinations [3,7–9]. In an expression model, when patients with one or two low-expression HLA-DP allotypes lack fully-matched donors, selecting a DPB1-mismatched donor that generates a mismatch against one low-expression recipient DPB1 allele may lower GvHD risk compared to mismatching against a high-expression allotype [9,10]. Since in our cohort, about 90% of donors encode either the rs9277534-AA or rs9277534-AG genotype, this matching schema considerably increases the number of eligible donors compared to a perfect 12/12 match schema.

In the current study, we defined rs9277534 linkages for 64 DPB1 allele groups including the 18 previously described linkages [9], and established that the linkage between rs9277534 and DPB1 genotype is highly reliable. In fact, we did not find a single discrepancy in over 32,000 samples. Since even direct analysis of rs9277534 could hardly be achieved with lower error rate, introducing experimental genotyping specifically for this SNP appears rarely warranted. Instead, high-resolution DPB1 genotyping including exon 2 and exon 3 information seems sufficient to predict the rs9277534 genotype based on Table 1. At least in Caucasian populations, this information should allow prediction of rs9277534 genotypes for more than 99.6% of individuals (Table 2). For the remaining alleles not included in Table 1 (very rare or novel alleles), our data suggest that rs9277534 genotype prediction based on the exon 3 sequence information is highly reliable (Table 4). Furthermore, we established that even for novel exon 3 sequences, the rs9277534 genotype should be predictable with reasonable accuracy, given the fact that all known DPB1 exon 3 sequences segment into two clearly separate phylogenetic clades corresponding to the two rs9277534 alleles.

We further estimated the error rate of predicting the rs9277534 genotype if DPB1 genotyping was based solely on exon 2 sequence analysis. Currently, five frequently observed DPB1 allele groups (DPB1*03:01, DPB1*04:01, DPB1*04:02, DPB1*17:01, and DPB1*19:01) include haplotypes with alternative linkage between exon 2 and the rs9277534 allele (Supplementary Material 2), but all of them are rare. Therefore, rs9277534 prediction is estimated to deliver correct results in more than 99% of the cases in Caucasian samples even when only relying on exon 2 sequence information. The linkage of exon 3 to rs9277534, on the other hand, seems to be highly conserved in Caucasian populations. Whether the same haplotype relationships can be found in non-Caucasian populations remains to be determined. In non-Caucasian populations where high-resolution DPB1 allele data are not available, rs9277534 prediction would ideally include sequence information on exon 2 and exon 3.
Our study is largely based on short amplicon sequencing in which the exon 2-exon 3-rs9277534 linkage was evaluated in an iterative process without direct phasing. However, the likelihood of incorrectly assigned linkages should be very low because of the high number of samples analyzed. Any sample containing one allele with an alternative linkage would have been detected as an unexpected result during primary analysis and would then have been confirmed during secondary analysis with SMRT sequencing, which included full physical phasing information. Encountering two haplotypes with alternative linkages in the same sample with rs9277534-AG genotype is anticipated to be a very rare event. Similarly, incorrectly assigned linkages due to allelic drop outs would have been caught as unclear results during primary analysis. Indeed, of the 32,681 samples in our study, 2 (0.006%) had incorrect linkage, which was resolved on secondary analysis. Therefore, the combination of technologies applied was appropriate to deliver the linkage information regarding rs9277534 with very high accuracy. This also applies to clinical practice: in some specific rs9277534-AG genotypes with phasing ambiguities, determining which allele carries the exon 3 sequence associated with high or low expression may require a long range NGS approach similar to our secondary.

Note that this work does not pinpoint the actual nucleotide variants responsible for the differences in allele expression between rs9277534-A linked and rs9277534-G linked alleles. As the 100% correlation between exon 3 and rs9277534 likely extends to other sequence variations located between exon 2 and the 3′ end of the gene, discovering which specific areas control DPB1 expression remains to be addressed in future research. The current study was also not designed to evaluate the relative differences between the T-cell epitope (TCE) or expression-based approaches for donor selection [8,9], which must await a comprehensive side-by-side analysis of the two models in large clinical cohorts.

We conclude that rs9277534 genotype and thereby DPB1 expression level can be reliably predicted from DPB1 genotyping results, preferably based on sequence information including exon 3. When DPB1-matched donors are not available, donors with one permissive DPB1 mismatch may be considered using DPB1 genotype information alone as a surrogate for direct typing of the rs9277534 allele. This should considerably simplify the translation of the recent research findings into clinical practice, and provide a means to lower the risk of GvHD after unrelated donor HSCT.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

We gratefully thank all the colleagues at DKMS Life Science Lab who participated in the success of this project by their daily dedicated work, in particular Carmen Schwarzelz and her laboratory team, Annett Heidl and her genotyping team, and Arend Große and his laboratory technology and automation team. We especially thank Lisa Hedrich and Johanna Andreas for adjusting the routine lab workflow for this project, Jan Heinicke for preparing the figures for this manuscript, and Jürgen Sauter from DKMS for providing ethnicity data. EWP is supported by grants AI069197, CA100019, CA162194 and CA18029 from the US National Institutes of Health.

**Funding:**
Abbreviations

GVHD: graft-versus-host disease
SNP: single nucleotide polymorphism
UTR: untranslated region
HSCT: hematopoietic stem cell transplantation
SMRT: single molecule real-time [sequencing]

References

1. Gratwohl A, Pasquini MC, Aljurf M, Atsuta Y, Baldomero H, Foeken L, Gratwohl M, Bouzas LF, Confer D, Frauenthoder K, Gluckman E, Greinix H, Horowitz M, Iida M, Lipton J, Madrigal A, Mohy M, Noel L, Novitzky N, Nunez J, Oudshoorn M, Passweg J, van Rood J, Szer J, Blume K, Appelbaum FR, Kodera Y, Niederwieser D. Worldwide Network for Blood and Marrow Transplantation (WBMT). One million haemopoietic stem-cell transplants: a retrospective observational study. Lancet Haematol. 2015; 2:e91–100. DOI: 10.1016/S2352-3026(15)00028-9 [PubMed: 26687803]

2. Tiercy JM. How to select the best available related or unrelated donor of hematopoietic stem cells? Haematologica. 2016; 101:680–687. DOI: 10.3324/haematol.2015.141119 [PubMed: 27252513]

3. Shaw BE, Mayor NP, Russell NH, Aperley J, Clark RE, Darbyshire P, Ethell ME, Goldman JM, Little AM, Mackinnon S, Marks DJ, Pagliuca A, Thomson K, Marsh SGE, Madrigal JA. Diverging effects of HLA-DPB1 matching status on outcome following unrelated donor transplantation depending on disease stage and the degree of matching for other HLA alleles. Leukemia. 2010; 24:58–65. DOI: 10.1038/leu.2009.239 [PubMed: 17726164]

4. Fernández-Viña MA, Klein JP, Haagenson M, Spellman SR, Anasetti C, Noreen H, Baxter-Lowe LA, Cano P, Flomenberg N, Horowitz MM, Oudshoorn M, Petersdorf EW, Saltijn-Kool E, Darbyshire P, Ethell ME, Goldman JM, Little AM, Mackinnon S, Marks DJ, Pagliuca A, Thomson K, Marsh SGE, Madrigal JA. Diverging effects of HLA-DPB1 matching status on outcome following unrelated donor transplantation depending on disease stage and the degree of matching for other HLA alleles. Blood. 2013; 121:4603–4610. DOI: 10.1182/blood-2012-09-481945 [PubMed: 23596045]

5. Shaw BE, Gooley TA, Mallki M, Madrigal JA, Begovich AB, Horowitz MM, Gratwohl A, Ringdén O, Marsh SGE, Petersdorf EW. The importance of HLA-DPB1 in unrelated donor hematopoietic cell transplantation. Blood. 2007; 110:4560–4566. DOI: 10.1182/blood-2007-06-095265 [PubMed: 17726164]

6. Crocchiolo R, Zino E, Vago L, Oneto R, Bruno B, Pollichienci S, Sacchi N, Sormani MP, Pancorbo J, Lamparelli T, Fanit R, Garbarino L, Mioti V, Bandini G, Bosi A, Ciceri F, Bacigalupo A, Fleischhauer K. Nonpermissive HLA-DPB1 disparity is a significant independent risk factor for mortality after unrelated hematopoietic stem cell transplantation. Blood. 2009; 114:1437–44. DOI: 10.1182/blood-2009-01-200378 [PubMed: 19515726]

7. Zino E, Frumento G, Martinetti M, Sormani MP, Picco F, Di Terlizzi S, Parodi AM, Sergeant R, Martinetti M, Bontadini A, Bonfazi F, Lisini M, Mazzi B, Rossini S, Servida P, Ciceri F, Bonini C, Lamino E, Bandini G, Locatelli F, Aperley J, Bacigalupo A, Ferrara GB, Bordignon C, Fleischhauer K. A T-cell epitope encoded by a subset of HLA-DPB1 alleles determines nonpermissive mismatches for hematopoeitic stem cell transplantation. Blood. 2004; 103:1417–1424. DOI: 10.1182/blood-2003-04-1279 [PubMed: 14576061]

8. Fleischhauer K, Shaw BE, Gooley T, Malik M, Bardi P, Bignon JD, Dubois V, Horowitz MM, Madrigal JA, Morishima Y, Oudshoorn M, Ringdén O, Spellman S, Velardi A, Zino E, Petersdorf EW. Effect of T-cell epitope matching at HLA-DPB1 in recipients of unrelated-donor haemopoietic-
1. Petersdorf EW, Malkki M, O’Huigin C, Carrington M, Gooley T, Haagenson MD, Horowitz MM, Spellman SR, Wang T, Stevenson P. High HLA-DP Expression and Graft-versus-Host Disease. N Engl J Med. 2015; 373:599–609. DOI: 10.1056/NEJMoai1500140 [PubMed: 26267621]

2. Thomas R, Thio CL, Apps R, Qi Y, Gao X, Marti D, Stein JL, Soderberg KA, Moody MA, Goedert JJ, Kirk GD, Hoots WK, Wolinsky S, Carrington M. A Novel Variant Marking HLA-DP Expression Levels Predicts Recovery from Hepatitis B Virus Infection. J Virol. 2012; 86:6979–6985. DOI: 10.1128/JVI.00406-12 [PubMed: 22496224]

3. Lange V, Böhme I, Hofmann J, Lang K, Sauter J, Schöne B, Paul P, Albrecht V, Andreas JM, Baier DM, Nething J, Ehninger U, Schwarze C, Pingel J, Ehninger G, Schmidt AH. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics. 2014; 15:doi: 10.1186/1471-2164-15-63 [PubMed: 24460756]

4. Schöfl G, Lang K, Quenzel P, Böhme I, Sauter J, Hofmann JA, Pingel J, Schmidt AH, Lange V. 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. BMC Genomics. 2017; 18:161.doi: 10.1186/s12864-017-3575-z [PubMed: 28196473]

5. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 2012; 9:1185–1188. DOI: 10.1038/nmeth.2221 [PubMed: 23103880]

6. Schöfl G, Schmidt AH, Lange V. Prediction of spurious HLA class II typing results using probabilistic classification. Hum Immunol. 2016; 77:264–272. DOI: 10.1016/j.humimm.2016.01.012 [PubMed: 26826450]

7. Lang K, Wagner I, Schöne B, Schöfl G, Birkner K, Hofmann JA, Sauter J, Pingel J, Böhme I, Schmidt AH, Lange V. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. BMC Genomics. 2016; 17:doi: 10.1186/s12864-016-2687-1
Figure 1. HLA-DPB1 gene structure and regions analyzed in this study (schematic, to scale)
(A) Gene structure of DPB1, including both UTRs (black boxes), exons (grey boxes), and introns (black lines). SNP rs9277534 (red) is located towards the end of the 3′ UTR. Positions are given relative to the start of the 5′ UTR of HLA-DPB1*02:01:02.
(B) Selected gene areas were targeted with short amplicons (black boxes) for the primary analysis by Illumina sequencing.
(C) The whole gene was targeted with a long amplicon (black box) used for secondary analysis by SMRT sequencing.
Figure 2. Read distribution for the rs9277534 amplicon in 32,681 successfully analyzed samples
(A) Read count per rs9277534 allele. Samples were grouped by the rs9277534 allele and genotype.
(B) Read balance between rs9277534-A and rs9277534-G. Bins were formed on the percentage of reads for rs9277534-A. The threshold values for being considered homozygous vs. heterozygous in rs9277534 are marked with dashed lines.
Figure 3. Workflow to establish and verify linkages between DPB1 alleles and SNP rs9277534
After amplicon-based genotyping, samples were grouped by DPB1 and rs9277534 zygosity. First, all samples homozygous for DPB1 and rs9277534 (A) or heterozygous for DPB1 and homozygous for rs9277534 (B) were used to define linkages. Subsequently, for rs9277534 heterozygous samples (C) carrying one DPB1 allele with a linkage established in (A) and (B), the linkage of the other allele was defined. This process was repeated, taking into account the newly established linkages, until no further linkages could be defined (C). Any samples with linkages not consistent with those observed in other samples were subjected to secondary analysis by full-length sequencing (D).
Figure 4. The two variants of HLA-DPB1*34:01 observed in this study

Both DPB1 alleles share the same exon 2 sequence but have distinct exon 3 sequences, which link to different alleles of rs9277534. The SNPs distinguishing both exon 3 sequences are highlighted at their relative positions within the exon.

Variant 1 has meanwhile been entered into the IPD-IMGT/HLA database as DPB1*34:01.

Variant 2 is in submission and will receive a different name.
Figure 5. Alignment of all described sequence features for HLA-DPB1 exon 3
The linked rs9277534 alleles are denoted in brackets. Nucleotides not identical in all sequences are color coded: SNPs corresponding to rs9277534-A-linked alleles (green), rs9277534-G-linked alleles (red), and other SNPs (yellow). (Nucleotide positions within exon 3 are denoted on top.)
Table 1
Linkage between rs9277534 and HLA-DPB1 alleles based on exons 2 and 3.
The 18 linkages previously reported based on exon 2 alone [9] were all confirmed. In addition, we defined the
linkage for 46 further DPB1 allele groups. At full resolution, these 64 allele groups represent 155 of the 630
described alleles, with a combined frequency of 99.2% in this study.

| rs9277534 | Previously defined linkages (confirmed) | Novel linkages |
|-----------|-----------------------------------------|----------------|
| **A (Low Expression)** | 02:01, 02:02, 04:01, 04:02, 17:01 | 21:01, 24:01, 30:01, 31:01, 39:01, 40:01, 41:01, 46:01, 51:01, 55:01, 59:01, 71:01, 81:01, 105:01, 109:01, 124:01, 126:01, 133:01, 138:01, 416:01, 464:01, 535:01 |
| **G (High Expression)** | 01:01, 03:01, 05:01, 06:01, 09:01, 10:01, 11:01, 13:01, 14:01, 15:01, 16:01, 19:01, 20:01 | 18:01, 21:01, 25:01, 26:01, 29:01, 34:01, 35:01, 36:01, 45:01, 50:01, 52:01, 57:01, 63:01, 85:01, 88:01, 90:01, 104:01, 130:01, 131:01, 136:01, 296:01, 350:01, 414:01, 463:01 |
Table 2

HLA-DPB1-rs9277534 linkage analysis of 32,681 samples. Categories were formed on the zygosity of rs9277534 expected from the DPB1 genotyping based on exons 2 and 3 and the presence of a phasing ambiguity between those two exons. Secondary analysis was performed only for samples with an unexpected linkage or otherwise inconclusive results after primary analysis.

| Expected rs9277534 | Ambiguities due to unresolved phasing between exons 2 and 3 | Samples | Unexpected linkages |
|-------------------|-------------------------------------------------------------|---------|---------------------|
|                   |                                                             |         | Primary analysis (amplicons) | Secondary analysis (SMRT) |
| Homozygous        | Not relevant                                                | 21,322 (59%) | 18 (0.09%)         | 1<sup>a</sup> (0.005%) |
| Heterozygous      | No                                                          | 7,481 (23%)  | 4 (0.05%)          | 2<sup>a</sup> (0.03%) |
| s                 | Yes                                                         | 5,761 (18%)  | 0                  | 0                   |
| Unknown           | –                                                           | 117 (0.4%)   | –                  | –                   |
| total             |                                                             | 32,681    | 22 (0.07%)         | 3<sup>a</sup> (0.01%) |

<sup>a</sup> All remaining unexpected linkages related to DPB1 *34:01.
**Table 3**

HLA-DPB1 alleles with novel exon 2/exon 3 combinations. Sequence features (SF) represent unique exon sequences that may be shared among several alleles. All sequence features are named after the most common allele carrying this sequence. In addition, we listed the rs9277534 variants associated with the exon 3 SFs.

| Exon 2 SF | Expected exon 3 SF (rs9277534) based on exon 2 SF | Detected exon 3 SF (rs9277534) | # of samples |
|----------|---------------------------------------------------|--------------------------------|--------------|
| 01:01    | 01:01 (G)                                         | 02:01 (A)                      | 2            |
| 11:01    | 01:01 (G)                                         | 02:01 (A)                      | 2            |
| 39:01    | 02:01 (A)                                         | 01:01 (G)                      | 2            |
| 04:01    | 04:01 (A)                                         | 141:01 (A)                     | 1            |
| 05:01    | 01:01 (G)                                         | 02:01 (A)                      | 1            |
| 06:01    | 01:01 (G)                                         | 04:01 (A)                      | 1            |
| 09:01    | 01:01 (G)                                         | 02:01 (A)                      | 1            |

*This allele has meanwhile been assigned as DPB1*162:01:02.

*b* This allele has meanwhile been assigned as DPB1*584:01.
Table 4
rs9277534 linkages based on HLA-DPB1 exon 3 sequence features

The numbers of observations for each exon 3 sequence feature are reported together with the linked rs9277534 allele. Contradictory linkages were not observed for any sequence feature (D' = 1.0 in all cases). Additionally, the individual alleles observed for each sequence feature are listed. Alleles were considered “observed” if they were unambiguously typed in at least one sample. Furthermore, many alleles with thus far undefined exon 3 sequences are reported with their observed exon 3 sequence features. For alleles not distinguishable in exons 2 and 3, NMDP codes are listed to represent the whole allele group. Novel exon 3 sequences were not further distinguished from each other; however, all of them showed rs9277534 linkage consistent with the most similar known exon 3 sequence feature.

| rs9277534 Exon 3 sequence feature (SF) | # Observations | Exon 3 SF known | Alleles observed | Exon 3 SF unknown | Alleles not observed |
|----------------------------------------|----------------|----------------|----------------|----------------|-------------------|
| A 02:01 | 38,578 | 02:01:02, 02:01:08, 02:02, 04:01:01, 17:01, 23:01:01, 30:01, 39:01, 55:01, 105:01, 106:01, 124:01, 133:01 | 02:01:03, 02:01:04, 02:01:15, 04:01:04, 04:01:07, 04:01:08, 24:01, 31:01, 40:01, 41:01:01, 46:01:01, 51:01, 71:01, 81:01, 109:01, 151:01, 178:01, 214:01, 215:01, 281:01 | 04:01:15, 28:01, 33:01, 121:01, 123:01, 134:01, 137:01, 162:01 | 225:01, 398:01 |
| 04:02 | 7,723 | 04:02:01, 126:01, 138:01, 416:01 | 68:01, 190:01, 526:01 | 257:01, 351:01, 460:01, 462:01 |
| 464:01 | 7 | 464:01 | | |
| 535:01 | 7 | 535:01 | | |
| 04:01:31 | 2 | 04:01:31 | | |
| 415:01 | 2 | 415:01 | | |
| 141:01 | 2 | 141:01 | | |
| 459:01 | 1 | 459:01 | | |
| G 01:01 | 18,486 | 01: AETTA, 01:01:02, 03: FNXVX, 05: RGPW, 06:01, 09:01:01, 10:01:01, 11:01:01, 13: AKXDE, 14:01:01, 16:01:01, 18:01, 19:01, 20:01:01, 21:01, 26:01:02, 45:01, 50:01, 85:01, 88:01, 131:01, 136:01, 350:01, 414: AYMY, 463:01 | 25:01, 29:01, 34:01, 35:01:01, 36:01:01, 37:01, 52:01, 57:01, 63:01, 67:01, 90:01, 91:01, 130:01 | 27:01, 35:01:02, 132:01, 152:01, 353:01, 440:01, 536:01 |
| 15:01 | 508 | 15:01 | | |
| 03:01:08 | 30 | 03:01:08, 296:01 | 139:01 | | 05:01:02, 352:01 |
| ? 168:01 | 0 | | | 168:01 |
| 498:01 | 0 | | | 498:01 |
| rs9277534 Exon 3 sequence feature (SF) | # Observations | Alleles observed | Alleles not observed |
|-----------------------------------|----------------|-----------------|---------------------|
|                                   |                | Exon 3 SF known | Exon 3 SF unknown  |
| 533:01                            | 0              |                 | 533:01              |
| 534:01                            | 0              |                 | 534:01              |
| new                               | 16             |                 |                     |

*a* Rare alleles only observed in one or two samples.

*b* Exon 3 sequence features observed in less than 3 samples. Hence, their alleles were not included in Table 1.

*c* DPB1*34:01 is listed only with the exon 3 sequence feature “01:01” corresponding to the full length sequence we submitted to IPD-IMGT/HLA release 3.29. The alternative 34:01 allele with exon 3 sequence feature “02:01” has been submitted but not yet been named.

*d* Alleles contained in NMDP codes with phasing ambiguity, so possibly included in our sample set. However, they were never observed individually.

*e* Alleles whose exon 3 SF were unknown in HLA 3.23 but have meanwhile been submitted in full length (IPD-IMGT/HLA release 3.29).