Contents lists available at ScienceDirect

Human Immunology

journal homepage: www.elsevier.com/locate/humimm

Patterns of non-ARD variation in more than 300 full-length HLA-DPB1 alleles

Steffen Klasberg[a], Kathrin Lang[a], Marie Günther[a], Grit Schober[a], Carolin Massalski[a], Alexander H Schmidt[a,b], Vinzenz Lange[a], Gerhard Schöfl[a,*]

[a] DKMS Life Science Lab, Dresden, Germany
[b] DKMS, Tübingen, Germany

ARTICLE INFO

Keywords:
HLA
DPB1
Full-length gene sequencing
Non-coding variation
Novel HLA alleles
HLA genotyping

ABSTRACT

Our understanding of sequence variation in the HLA-DPB1 gene is largely restricted to the hypervariable antigen recognition domain (ARD) encoded by exon 2. Here, we employed a redundant sequencing strategy combining long-read and short-read data to accurately phase and characterise in full length the majority of common and well-documented (CWD) DPB1 alleles as well as alleles with an observed frequency of at least 0.0006% in our predominantly European sample set. We generated 664 DPB1 sequences, comprising 279 distinct allelic variants. This allows us to present the, to date, most comprehensive analysis of the nature and extent of DPB1 sequence variation.

The full-length sequence analysis revealed the existence of two highly diverged allele clades. These clades correlate with the rs9277534 A → G variant, a known expression marker located in the 3′-UTR. The two clades are fully differentiated by 174 fixed polymorphisms throughout a 3.6 kb stretch at the 3′-end of DPB1. The region upstream of this differentiation zone is characterised by increasingly shared variation between the clades. The low-expression A clade comprises 59% of the distinct allelic sequences including the three by far most frequent DPB1 alleles, DPB1*04:01, DPB1*02:01 and DPB1*04:02. Alleles in the A clade show reduced nucleotide diversity with an excess of rare variants when compared to the high-expression G clade. This pattern is consistent with a scenario of recent proliferation of A-clade alleles.

The full-length characterisation of all but the most rare DPB1 alleles will benefit the application of NGS for DPB1 genotyping and provides a helpful framework for a deeper understanding of high- and low-expression alleles and their implications in the context of unrelated haematopoietic stem-cell transplantation.

1. Introduction

HLA-DP is one of the three classical class II immunoglobulin cell-surface molecules encoded within the hyperpolymorphic human leukocyte antigen (HLA) complex.

HLA-DP molecules are expressed on antigen-presenting cells and, like other HLA molecules, are centrally involved in the adaptive immune response. As heterodimers, the HLA-DP proteins consist of structurally homologous α- and β-chains, each contributing half of the antigen-binding groove. These chains are encoded by the HLA-DPA1 and -DPB1 genes, respectively.

HLA genes play a crucial role in solid organ and haematopoietic stem-cell transplantation (HSCT) [1,2], and the allelic concordance of HLA genes of stem-cell donors and patients is arguably one of the most relevant factors for HSCT outcome [3,4]. Currently, the class I genes HLA-A, -B and -C and the class II genes HLA-DRB1 and, to a lesser extent, HLA-DQB1 are considered critical for transplantation outcome [5,3,6]. The high resolution genotyping and complete allelic matching of these five genes is viewed as the gold standard for unrelated donor selection [7]. This view is being challenged by studies showing that the match status of DPB1 does significantly affect the risk of graft-versus-host-disease (GVHD), disease relapse, graft rejection and non-relapse mortality [1,8–11]. Additionally, DPB1 is suspected to play a role in predisposition to and clearance of hepatitis B virus infections [12,13] and other diseases [14–16].

In the context of donor selection for HSCT limited attention has been paid to variation outside of the antigen recognition domain (non-ARD), i.e.,

a) coding variation outside of exon 2 (exons 2 and 3 for class I

[*] Corresponding author.
E-mail addresses: klasberg@dkms-lab.de (S. Klasberg), lang@dkms-lab.de (K. Lang), guenther@dkms-lab.de (M. Günther), schober@dkms-lab.de (G. Schober), massalski@dkms-lab.de (C. Massalski), schmidt@dkms.de (A.H. Schmidt), lange@dkms-lab.de (V. Lange), schoefl@dkms-lab.de (G. Schöfl).

https://doi.org/10.1016/j.humimm.2018.05.006
Received 5 February 2018; Received in revised form 13 May 2018; Accepted 29 May 2018
Available online 04 June 2018

genes), and
b) non-coding variation in introns and the untranslated regions (UTRs).

Variation in the UTRs has been associated with disease susceptibility and pathogenesis [17–19]. UTRs often contain elements mediating transcriptional and posttranscriptional regulation of expression levels and may be involved in mRNA stability and microRNA binding [20–23]. Variation in such elements may thus modify the levels of immunogenicity of otherwise ARD-identical alleles.

For DPB1 it has been shown that, in addition to allelic mismatches, HLA-DP expression levels influence the incidence of GVHD after HSCT [9]. The expression level of a DPB1 allele has been found to correlate with variant rs9277534 [13,9], an A → G single nucleotide polymorphism (SNP) located in the 3′-UTR of DPB1. Alleles with the variant rs9277534-A (referred to as A clade) are associated with low expression of DPB1, while alleles with the variant rs9277534-G (G clade) are associated with high expression of DPB1 [13]. The 3′-UTR rs9277534 variants have been shown to be in perfect linkage with the known exon 3 sequences of DPB1 [24]. The extent of genetic differentiation between A-clade and G-clade alleles and the exact sequence motifs that cause the differences in expression are not currently known, however. So far, all studies associating functional differences in expression to non-ARD variation in DPB1 were limited by the lack of full-length sequences for most alleles.

The HLA system is the most variable region in the human genome with thousands of different observed alleles across the major HLA genes (https://www.ebi.ac.uk/ipd/imgt/hla/stats.html). By February 2018, 962 DPB1 alleles have been recorded in the central public database for HLA allelic variation (IPD-IMGT/HLA release 3.31.0) [25]. Although the number of described alleles is high and ever increasing, there is only a modest number of alleles for which full genomic sequences are available. Prior to the submission of the bulk of the sequences described here, the full-length sequences of only 13 DPB1 alleles were available (IPD-IMGT/HLA release 3.25.0, compare Fig. 1).

Structurally, DPB1 is composed of five exons, with exon 2 encoding the HLA antigen recognition domain (ARD). The gene encompasses 11,475 bp with a transcript length of 1560 bp (Fig. 2). Interestingly, functional genetic structures overlap with DPB1 both on the coding and the template strand (Fig. 2):

a) The 5′-UTR, exon 1 and parts of intron 1 of DPA1 overlap with DPB1 starting from the middle of exon 2 (template strand).
b) A processed pseudogene of the ribosomal protein L32 resides in intron 1 of DPB1 (coding strand) [26,27].
c) Two promotor regions overlap with the 5′-part of DPB1: One promotor overlaps with the 5′-UTR, exon 1 and a part of intron 1 (coding strand). A second promotor overlaps with a part of intron 1, exon 2 and intron 2 (template strand).
d) conserved binding motif for the CTCF transcriptional repressor is located in intron 2 at position 7199 to 7265 (template strand) [28].

For the present work, we applied a dual sequencing approach to obtain full-length sequences of all major DPB1 alleles. This approach combined Pacific Biosciences Single Molecule Real Time sequencing with Illumina shotgun sequencing to generate fully-phased highly accurate consensus sequences [29]. We analysed the sequences with respect to differences between A-clade and G-clade alleles as defined by the 3′-UTR variant rs9277534. We inferred patterns of nucleotide diversity and selection in different regions of DPB1, and described the number and location of polymorphic sites to identify and quantify noncoding variation within and between A-clade and G-clade alleles. Finally, we identified novel gene features in the non-coding part of DPB1.

2. Material and methods

2.1. Samples

A total of 332 whole blood or buccal swab samples were selected to represent all common and well-documented (CWD) DPB1 alleles [30], as well as alleles that are not currently defined as CWD but with an allele frequency of at least 0.0006% in our predominantly European donor population (i.e., alleles occurring > 50 times in 4 million heterozygous samples). During registration, donors signed an informed consent

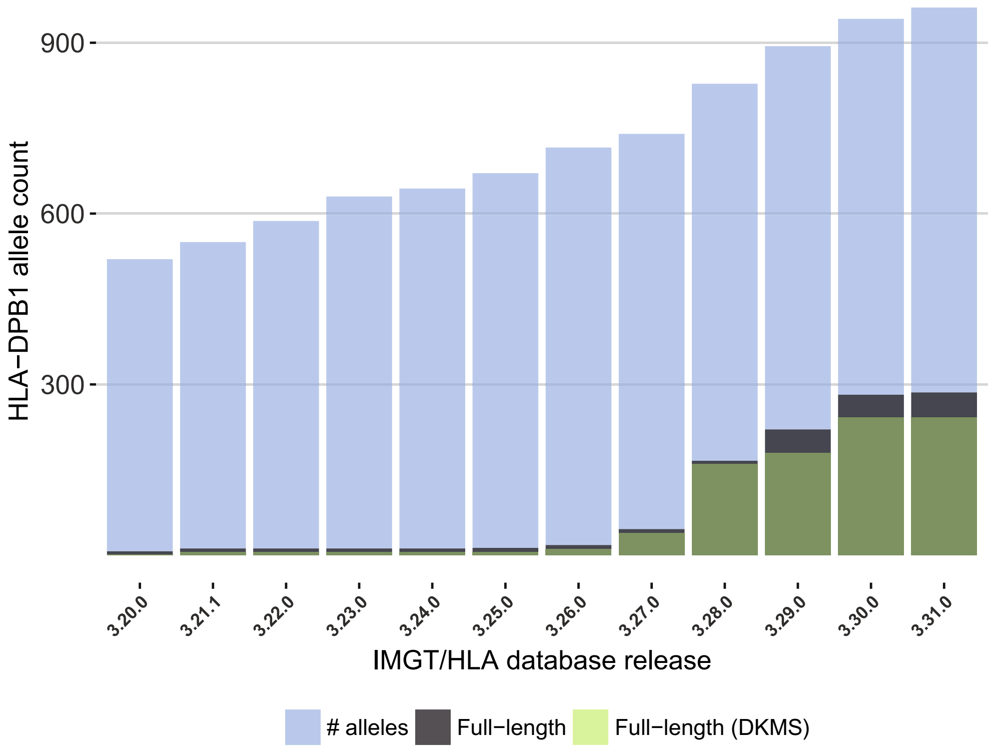

Fig. 1. Cumulative numbers of DPB1 alleles in the IPD-IMGT/HLA database: The cumulative number of known alleles from release 3.20.0 onwards. Only 13 alleles were known in full length when the first sequences reported here were generated (release 3.25.0). The current database release (3.31.0) contains 286 full-length alleles, 252 of which were generated as part of the present study. 27 additional alleles are being processed and not yet included in IPD-IMGT/HLA. Full length is defined here as including both 5′- and 3′-UTR sequence information.

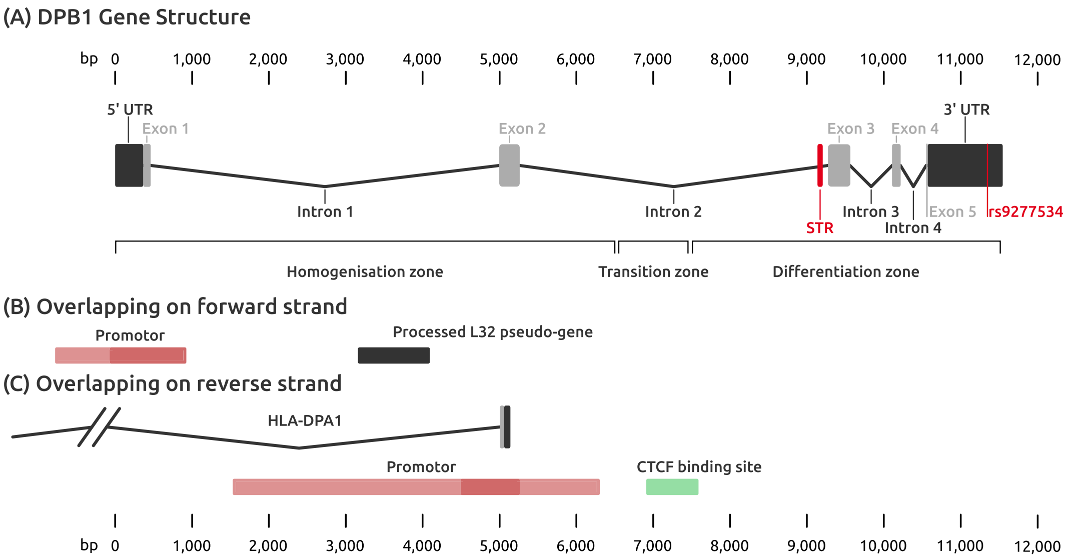
(A) DPB1 Gene Structure
bp 0 1,000 2,000 3,000 4,000 5,000 6,000 7,000 8,000 9,000 10,000 11,000 12,000
5' UTR
Exon 1
Exon 2
Exon 3
Exon 4
3' UTR
Intron 1
Intron 2
STR
Intron 3
Intron 4
Exon 5
rs9277534
Homogenisation zone
Transition zone
Differentiation zone
(B) Overlapping on forward strand
Promotor
Processed L32 pseudo-gene
(C) Overlapping on reverse strand
HLA-DPA1
Promotor
CTCF binding site
bp 0 1,000 2,000 3,000 4,000 5,000 6,000 7,000 8,000 9,000 10,000 11,000 12,000

Fig. 2. HLA-DPB1 gene structure and overlapping features: (A) Gene model of DPB1; black boxes denote 3′- and 5′-UTR, grey boxes denote exons. The variant rs9277534 and the microsatellite (Short tandem repeat, STR) are shown in red. *Homogenisation zone*, *transition zone*, and *differentiation zone* describe sequence blocks with contrasting patterns of variation. See the text for details. Positions are relative to the start of the 5′-UTR. (B) Features that overlap with DPB1 on the coding strand. (C) Features that overlap with DPB1 on the template strand.

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 analyses are within the scope of this consent form. DNA was isolated using the “Chemagic DNA Blood or Buccal Swab Kit Special” according to the manufacturer’s instructions (PerkinElmer, Baesweiler, Germany). The isolated DNA was eluted in 10 mM Tris-HCl pH8.0 elution buffer. DNA concentration was measured with SYBR Green fluorescence (Biozym, Hessisch Oldendorf, Germany).

2.2. Amplification, library preparation and sequencing strategy

Lab procedures and the sequencing strategy have been described elsewhere in detail [29]. Briefly, for each sample we performed two independent long-range PCR reactions targeting the whole DPB1 gene using primers flanking the UTR-regions. The 12 kb amplicons were sequenced by two strategies on different platforms:

a) shotgun sequencing on Illumina MiSeq instruments (Illumina, San Diego, California) generating 250 bp paired-end reads, and
b) Single Molecule Real Time (SMRT) sequencing on PacBio RS II instruments (Pacific Biosciences, Menlo Park, California).

Phase-defined allelic consensus sequences were generated from the long-read data and polished with the high-fidelity short reads using the software package DR2S (https://github.com/gschofl/DR2S). The data were submitted to the EMBL-ENA and the IPD-IMGT/HLA databases using TypeLoader (https://github.com/DKMS-LSL/typeloader) [31].

2.3. Statistical analysis

All analyses were carried out using the R software environment for statistical computing, version 3.4.3 [32]. A multiple sequence alignment (MSA) of all resulting full-length DPB1 alleles was constructed using the R package DECIPHER [33]. The MSA was manually checked and gap-site adjusted. Measures of nucleotide diversity, π [34], divergence, *d* [35], and deviation from expectations from the neutral model (*Tajima’s D* [36]) were estimated using the R package Popgenome [37]. All statistics were computed in a sliding window with a width of 100 nt and jumps of 25 nt. The genealogical relationship amongst haplotypes was explored by calculating undirected minimum spanning trees (MSTs) using the igraph package [38] and visualised using the neato graph layout algorithm implemented in the graphviz software package [39]. Feature positions in DPB1 are provided using HLA-DPB1*01:01:01:01 as reference allele.

3. Results

3.1. Allele-level variation

3.1.1. Number of new full length alleles

The dual sequencing strategy enabled us to characterise 664 DPB1 alleles in full length while solving all phasing ambiguities. These comprised 279 distinct sequences of which 269 were not previously described in full length in the IPD-IMGT/HLA database (Version 3.25.0).

The 279 distinct alleles contain all 40 *common* and 14 *well-documented* alleles of the CWD catalogue (CWD release 2.0) [30]. A recently published complementary CWD catalogue based on European data [40] recognised 77 CWD-defined DPB1 alleles in total, of which 30 are defined as *common* and 47 are defined as *well-documented*. We still cover all alleles defined as *common* by the European catalogue, but lack data on 20 of the alleles defined as *well-documented* there.252 alleles have successfully been submitted to the IPD-IMGT/HLA database using the TypeLoader software [31]. The remaining 27 alleles are in the process of submission. This substantially increases the number of fully characterised alleles in IPD-IMGT/HLA and the knowledge of DPB1 sequences (see Fig. 1).

3.1.2. A clade and G clade

We augmented our dataset with 34 additional recently published full-length DPB1 sequences retrieved from IPD-IMGT/HLA [25], resulting in a total of 313 alleles for downstream analysis at the sequence level. Where possible, alleles were assigned to either the A clade or the G clade based on variant rs9277534. 176 alleles are a member of the A clade and 122 alleles are a member of the G clade. 15 of the 34 additional alleles from IPD-IMGT/HLA could not be assigned due to missing 3′-UTR sequence data.

3.1.3. Gene length

After dismissing the 15 alleles from IPD-IMGT/HLA due to

Table 1
Number of alleles, length and STR variation of HLA-DPB1 alleles grouped into A and G clades. Length in nucleotides. STR in number of repeat units

Clade	Distinct alleles	Length			STR
		Minimum	Median	Maximum	
A	176	11,510	11,526	11,544	9–17
G	122	11,455	11,466	11,475	4

incomplete UTR sequences, DPB1 alleles range from 11,455 to 11,544 nucleotides in length, including the UTRs. Alleles in the A clade (minimum 11,510 nt) are always longer than alleles in the G clade (maximum 11,475 nt), see Table 1.

3.2. Variation

3.2.1. Short tandem repeats

We identified a tetranucleotide microsatellite (repeat motif “GGAA”) located at the end of intron 2 after an A-homopolymer of 6–9 residues (position 9157; Fig. 3). Interestingly, in the G clade the microsatellite has a fixed number of four repeat units, while A-clade alleles show a highly variable number of 9–17 repeat units. Additionally, the A clade exhibits two repeat variations:

a) The penultimate repeat is conserved as “AGAA” in all A-clade sequences.
b) One allele shows a substitution of one repeat unit with “GGAG”.

This microsatellite length polymorphism partially explains the length difference of A clade and G clade. Whilst G-clade alleles are still never longer than A-clade alleles, there are alleles of the same length from both clades if the microsatellite is masked.

3.2.2. CTCF binding site

A CTCF binding site has been annotated inside intron 2 from position 7199 to 7265 [28] (compare Fig. 2). We found three polymorphic sites within the CTCF binding site. These polymorphic sites separate the A and G clades to a large degree, but not completely. Two positions (at 7208 nt and 7225 nt) separate the clades, except for four G-clade alleles (DPB1*19:01:01:01, DPB1*19:01:01:02, DPB1*34:01:01:01 and DPB1*34:01:01:02) showing the nucleotide of the A clade and one A-clade allele (DPB1*162:01:02) containing the nucleotide of the G clade. The third position (at 7237 nt) separates the clades, except for two G-clade alleles (DPB1*19:01:01:01 and DPB1*19:01:01:02) showing the nucleotide of the A clade and the A-clade allele DPB1*162:01:02 showing the nucleotide of the G clade. Two of the positions, 7225 nt and 7237 nt, are located within the binding motif.

3.2.3. Nucleotide diversity

We estimated the nucleotide diversity, $\hat{\pi}$, separately for the two clades and the divergence, $\hat{d}$, between the two clades in a sliding window over the whole gene sequence (Fig. 4). Overall, nucleotide diversity and divergence show a bipartite pattern: The 5′-part of the gene up to intron 2 at position 6533 nt exhibits a similarly patterned nucleotide diversity within the clades and little divergence between the clades. In this stretch of the gene, the highest diversity is observed in exon 2, which is coding for the highly variable ARD. In contrast, for most of the 3′-part of the gene after position 6533 nt, within-clade diversity is low, whilst between-clade divergence rises dramatically. The highest level of divergence between the clades are observed in the 3′-UTR.

This pattern suggests that few to no fixed structural differences exist between the clades throughout the 5′-part of DPB1. The 3′-part of DPB1, however, is characterised by a clear structural separation of the clades. Overall, alleles in the G clade exhibit higher levels of nucleotide diversity then sequences in the A clade.

3.2.4. Polymorphic positions

In total, the 313 DPB1 alleles feature 669 polymorphic sites of which 142 carry a deletion in at least one sequence (Table 2). Only 182 (27.2%) of the polymorphic sites are shared between the clades, i.e., the position is polymorphic in both clades (S sites). 312 sites are polymorphic only in either the A clade (168 A sites) or the G clade (144 G sites). 175 (26.2%) of the sites are polymorphic between the clades, but not within either of the clades itself, i.e., these are nucleotide positions that perfectly differentiate the clades (D sites).

Similar to the pattern of nucleotide diversity and divergence, the distribution of the categories of polymorphic sites differs along the gene (Table 2). In absolute terms, most polymorphisms reside in the non-coding regions. The introns 1 and 2 contain 164 and 281 polymorphic sites (percentage of substitutions: 3.61% and 6.96%), respectively. The 3′-UTR contains 88 polymorphic sites (percentage of substitutions: 9.14%). The coding regions with the highest number of variable positions are exons 2 and 3 with 45 and 27 polymorphic sites (percentage of substitutions: 17.5% and 9.57%), respectively. Most polymorphic sites in exon 2 are shared between clades, maintaining high variability in the ARD. No site in exon 2 is strictly differentiating the two clades. Exon 3 is the only coding region of DPB1 that contains variants that clearly differentiate the clades. Our results show the same seven segregating sites differentiating the clades in exon 3 that were previously described by Schöne et al. [24].

If we consider only high-frequency (HF) variation, here defined as polymorphisms with a minor allele frequency of more than 10% in either of the two clades, DPB1 still retains 351 HF polymorphic sites distributed similarly to all polymorphic sites (Fig. 5 and Table 3). The restriction to HF variation has the highest impact on shared sites and sites specific to the A clade, with only 17% and 18% of polymorphic sites showing HF variation, respectively. Once minor alleles are excluded, 70 sites originally classified to be A-specific, G-specific, or shared between clades become reclassified as differentiating the two clades, thereby increasing the number of clade-differentiating sites to 245.

Contrasting low-frequency and high-frequency variation allows a

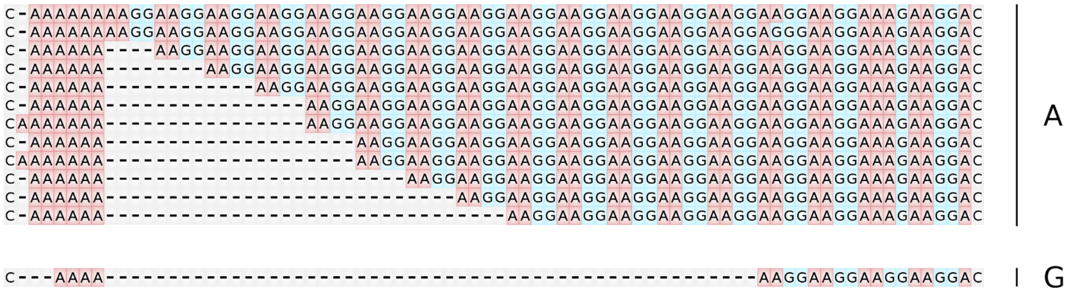
C-AAAAAAAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
C-AAAAAAAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAGGGAAGGAAAGAAGGAC
C-AAAAAA----AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
C-AAAAAA--------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
C-AAAAAA------------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
C-AAAAAA----------------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
CAAAAAAA----------------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
C-AAAAAA--------------------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
CAAAAAAA--------------------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
C-AAAAAA------------------------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
C-AAAAAA----------------------------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
C-AAAAAA--------------------------------AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAC
A
C---AAAA--AAGGAAGGAAGGAAGGAC
G

Fig. 3. Short tandem repeat in intron 2: The microsatellite is conserved in the G clade (4 repeat units). The A clade shows between 9 and 17 repeat units and two variations of the repeat unit.

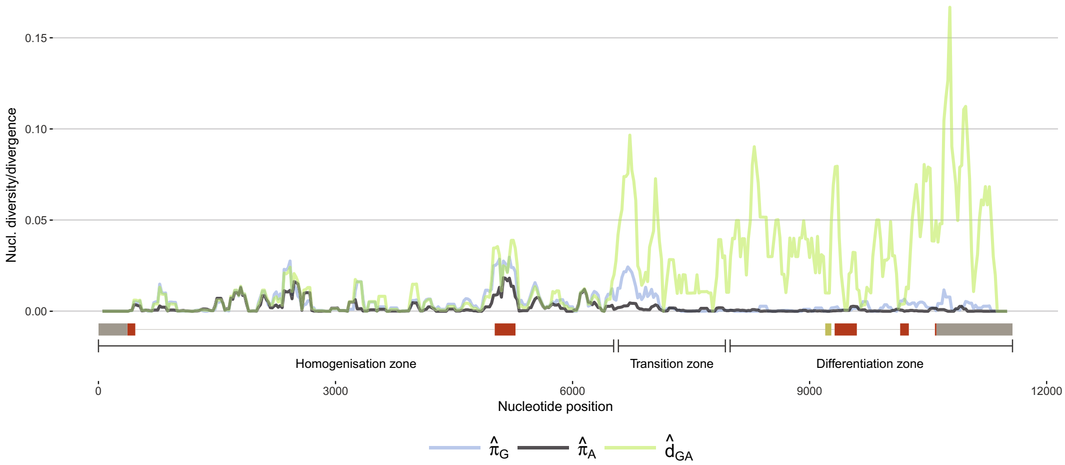

Fig. 4. Nucleotide diversity within the A and G clades and divergence between the clades: Computed in a sliding window of 100 bp. $\hat{\pi}_G$: nucleotide diversity in the G clade; $\hat{\pi}_A$: nucleotide diversity in the A clade; $\hat{d}_{GA}$: divergence between clades. A schematic representation of the gene structure is shown below with two grey boxes representing 5′- and 3′-UTR, red boxes representing exons 1–5 and a yellow box showing the position of the microsatellite. The ranges below the gene model show the homogenisation zone with no differentiating variants, the transition zone with increasing differentiation between the clades and the differentiation zone harbouring variants which fully separate the two clades.

Table 2
Distribution of polymorphic sites within HLA-DPB1: Number of sites and nucleotide diversity, $\hat{\pi}$, as the number of polymorphic sites relative to feature length (in 10^{-3}). H zone, T zone and D zone denote the homogenisation zone, transition zone and differentiation zone.

Feature	Length	Total		Shared		Only A		Only G		Differentiating	
		#	$\hat{\pi}$	#	$\hat{\pi}$	#	$\hat{\pi}$	#	$\hat{\pi}$	#	$\hat{\pi}$
5′-UTR	366	2	0.55	2	0.55	0	0.00	0	0.00	0	0.00
Exon 1	100	1	1.00	1	1.00	0	0.00	0	0.00	0	0.00
Intron 1	4548	164	3.61	67	1.47	49	1.08	48	1.06	0	0.00
Exon 2	264	45	17.05	27	10.23	10	3.79	8	3.03	0	0.00
Intron 2	4038	281	6.96	82	2.03	76	1.88	28	0.69	95	2.35
Exon 3	282	27	9.57	3	1.06	9	3.19	9	3.19	6	2.13
Intron 3	547	24	4.39	0	0.00	8	1.46	6	1.10	10	1.83
Exon 4	111	4	3.60	0	0.00	1	0.90	3	2.70	0	0.00
Intron 4	329	31	9.42	0	0.00	5	1.52	10	3.04	16	4.86
Exon 5	20	2	10.00	0	0.00	1	5.00	1	5.00	0	0.00
3′-UTR	963	88	9.14	0	0.00	9	0.93	31	3.22	48	4.98
H zone	6549	273	4.17	138	2.11	66	1.01	69	1.05	0	0.00
T zone	1415	59	4.17	41	2.90	13	0.92	5	0.35	0	0.00
D zone	3604	337	9.35	3	0.08	89	2.47	70	1.94	175	4.86
Total	11,568	669	5.78	182	1.57	168	1.45	144	1.24	175	1.51

classification of DPB1 sequence blocks with differential patterns of shared and differentiating polymorphic sites (compare Fig. 2). When considering all polymorphic sites, the first strictly clade-differentiating polymorphic site can be found at position 7383 nt. This defines the start of the differentiation (D) zone where the A and G clades become completely separated. When considering only HF variation, the first position which differentiates the two clades is located at 6533 nt. This position constitutes the end of the homogenisation (H) zone where A- and G-clade alleles are not distinguishable based on their sequence. Between these two positions lies a transition (T) zone, defined by an increasing segregation of A and G clades.

Several patterns can be observed for HF polymorphic sites along the gene (Table 3). The majority of differentiating sites (208 sites; 84.2%) are located in the differentiation zone, the remaining 39 are located in the transition zone. In contrast, shared polymorphic positions are exclusively found in the homogenisation zone. Clade-specific HF polymorphic positions are, again, not evenly distributed along the gene: Most positions specific to A (23 sites, 74%) are observed in the differentiation zone, whilst positions specific to G are found more often in the homogenisation zone (56 sites, 66%).

Haplotype networks visualise the contrasting genetic relationship between DPB1 alleles throughout the homogenisation zone versus the transition and differentiation zones (Fig. 6), and highlight the weak potential for an evolutionary transition of A-clade alleles to G-clade alleles and vice versa. Throughout the homogenisation zone of the gene, where DPB1 overlaps with DPA1 on the template strand, an evolutionary transition seems feasible as shown by the lack of a clear separation of the clades at the sequence level and modest edit distances between sequences belonging to different clades (Fig. 6 A). The network for the 3′-part of the gene (transition and differentiation zone) reflects the number of clade-differentiating sites. Here, the A clade and the G clade are separated by a minimum edit distance of 214, rendering a complete transition extremely unlikely (Fig. 6 B).

3.2.5. *Neutral evolution*

We inferred selective forces acting on the A and G clades and between the clades by estimating Tajima's *D* in a sliding window over the whole gene sequence (Fig. 7). Negative values of Tajima's *D* indicate an

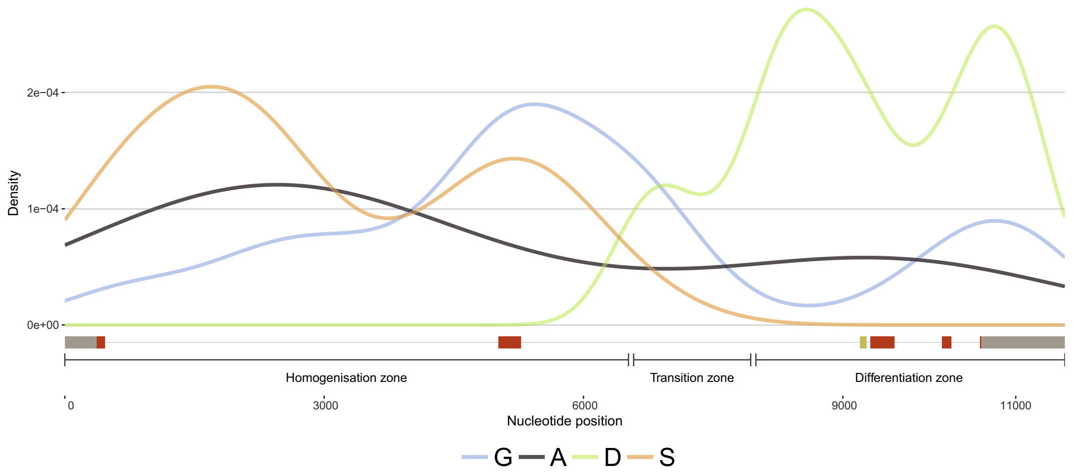

Fig. 5. Kernel density estimation of the distribution of polymorphic sites along HLA-DPB1. Polymorphic sites are classified as A-clade specific (A), G-clade specific (G), shared between A and G (S) and differentiating between A and G (D). Only sites are considered that are polymorphic in at least 10% of the sequences. A schematic representation of the gene structure is shown below with two grey boxes representing 5′- and 3′-UTR, red boxes representing exons 1 to 5 and a yellow box showing the position of the microsatellite. The ranges below the gene model show the homogenisation zone with no differentiating variants, the transition zone with increasing differentiation between the clades and the differentiation zone harbouring variants which fully separate the two clades.

excess of rare variants, suggesting the action of purifying selection on the sequences, a recent genetic sweep, or a recent population expansion. Positive values of Tajima's *D* indicate an excess of intermediate variants suggesting balancing selection.

Overall, only trends can be observed when analysing the clades separately (Fig. 7). *D* values for the G clade fall between 2 and −2, pointing to near neutral evolution. The A clade shows a similar distribution of Tajima's *D*, albeit at overall slightly lower values. This may be a reflection of a recent proliferation of high frequency A-clade alleles like DPB1*04:01, DPB1*04:02 and DPB1*02:01.

When estimating Tajima's *D* over all alleles without separating the clades, we observed a pattern of *D* values similar to the separate clades throughout the 5′-part of the gene. Throughout the 3′-part of DPB1, *D* estimates within A and G are mostly below zero, and as such point towards directional (possibly purifying) selection acting within clades. *D* values over all alleles, however, are highly positive indicating balancing selection acting on the sequence that most clearly separates the A and G clades.

4. Discussion

With only 13 known full-length sequences, HLA-DPB1 was until recently (IPD-IMGT/HLA release 3.25.0; July 2016) the least well characterised gene amongst the six major human MHC class I and class II genes. This was in part a consequence of the fact that DPB1 was long considered largely inconsequential for HSCT outcome, but also likely due to the relative difficulties in obtaining reliable genomic sequences for this gene in particular. The length of the gene (11.5 kb), several difficult to resolve regions (long homopolymeric stretches and a major microsatellite), and sometimes large distances between heterozygous positions, present challenges even for state-of-the-art NGS sequencing technologies. Here, we circumvent the difficulties by applying a redundant sequencing approach based on two independent PCR reactions analysed by complementary sequencing technologies (long-read SMRT sequencing and short-read shotgun sequencing) [29]. This approach not only guards against the systematic biases of different sequencing technologies, but also against the possibility of reporting PCR artefacts

Table 3

Distribution of high-frequency polymorphic sites within HLA-DPB1: Number of sites and nucleotide diversity, $\hat{\pi}$, as the number of polymorphic sites relative to feature length (in 10^{-3}). Only polymorphic sites occurring in at least 10% of alleles are considered. H zone, T zone and D zone denote the homogenisation zone, transition zone and differentiation zone.

Feature	Length	Total		Shared		Only A		Only G		Differentiating	
		#	$\hat{\pi}$	#	$\hat{\pi}$	#	$\hat{\pi}$	#	$\hat{\pi}$	#	$\hat{\pi}$
5′-UTR	366	1	0.00	0	0.00	1	0.00	1	0.00	0	0.00
Exon 1	100	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Intron 1	4548	45	0.01	21	0.00	6	0.00	26	0.01	0	0.00
Exon 2	264	20	0.08	8	0.03	0	0.00	13	0.05	0	0.00
Intron 2	4038	185	0.05	2	0.00	22	0.01	28	0.01	153	0.04
Exon 3	282	7	0.02	0	0.00	1	0.00	0	0.00	7	0.02
Intron 3	547	14	0.03	0	0.00	1	0.00	2	0.00	13	0.02
Exon 4	111	1	0.01	0	0.00	0	0.00	1	0.01	0	0.00
Intron 4	329	22	0.07	0	0.00	0	0.00	3	0.01	21	0.06
Exon 5	20	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
3′-UTR	963	56	0.06	0	0.00	0	0.00	11	0.01	53	0.06
H zone	6549	72	0.01	31	0.00	8	0.00	56	0.01	0	0.00
T zone	1415	48	0.03	0	0.00	0	0.00	11	0.01	39	0.04
D zone	3604	231	0.06	0	0.00	23	0.01	18	0.00	208	0.06
Total	11,568	351	0.03	31	0.00	31	0.00	85	0.01	247	0.02

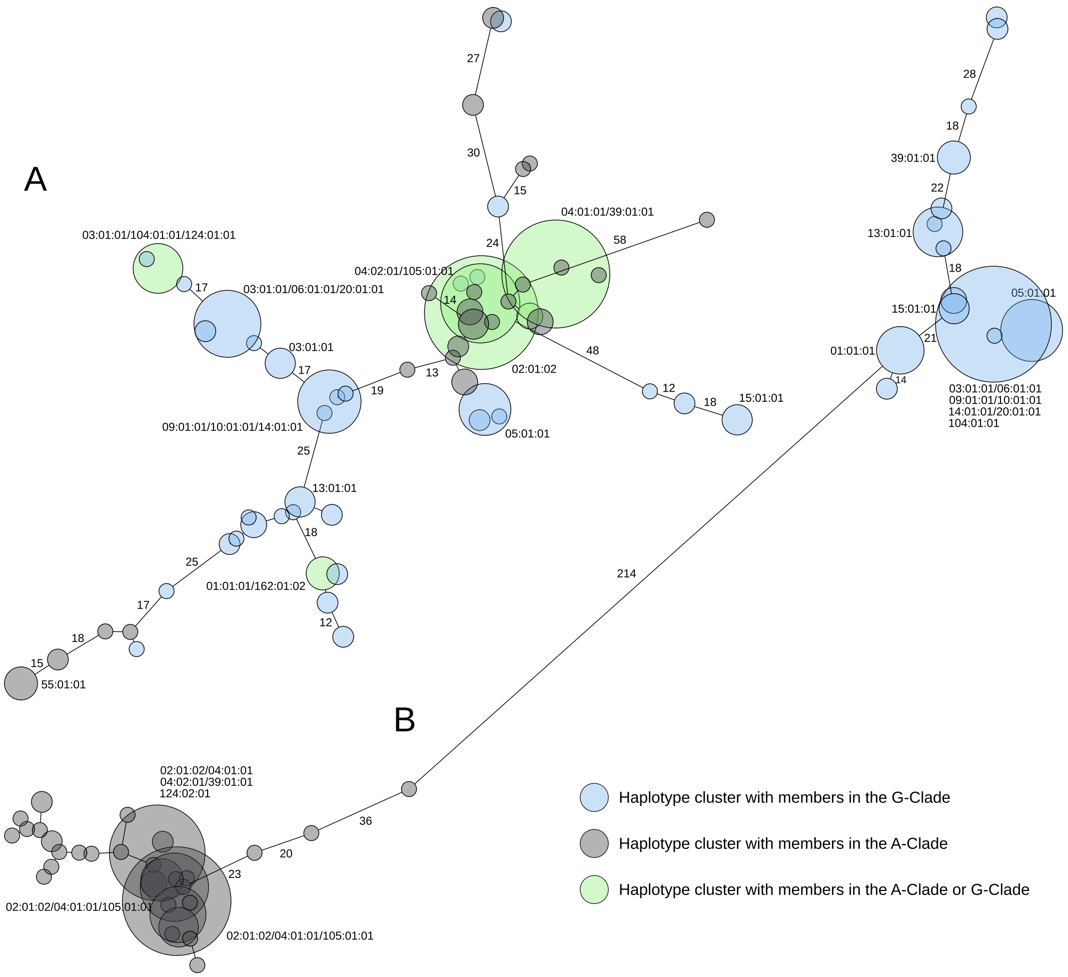

Fig. 6. Haplotype networks for HLA-DPB1: Nodes denote clusters of alleles with an edit distance ⩽ 2. Node size is proportional to the number of alleles contained within a cluster. For nodes with ⩾ 4 members, 6-digit codes are provided indicating the major member alleles. Edge length is proportional to the edit distance between the least distant member alleles in two nodes. Edit distances ⩾ 12 are indicated in the figure. A: Network based on the DPB1 sequence stretch from the 5′-UTR to the end of the homogenisation zone. B: Network based on the DPB1 sequence stretch from the start of the transition zone to the 3′-UTR.

due to the extremely low probability of the same artefact occurring twice in independent reactions [29].

HLA-DPB1 is increasingly gaining traction as an additional parameter for upfront HLA matching for unrelated haematopoietic stem-cell transplantation [11]. Especially a previously identified expression marker in the 3′-UTR, rs9277534-A or -G, is predictive of increased risk for graft-versus-host-disease in otherwise 10/10-matched donor-recipient pairs [9]. This specific variant has further been linked to several other conserved polymorphisms in exon 3 of DPB1 [24], but the gene has not been studied comprehensively throughout its length. Here we show that rs9277534 is a marker for two highly divergent allele groups (clades). We found that as many as 175 single nucleotide polymorphisms, a microsatellite, and mutations in a conserved CTCF binding site, clearly separate the two distinct clades.

Interestingly, the SNPs are not uniformly distributed over the gene structure of DPB1, but exhibit a strongly bipartite pattern. Variants which distinguish the two clades are exclusively found downstream of exon 2 starting at position 7524 in intron 2 throughout a region we name differentiation zone. Upstream of this region a stretch of ~1 kb shows strong differentiation between the clades but no completely fixed

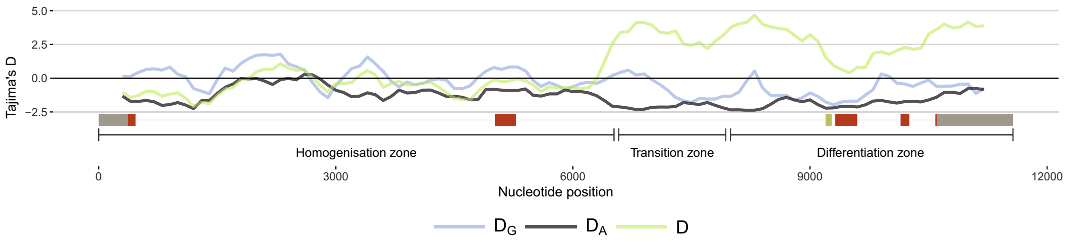

Fig. 7. Neutral evolution within the A and G clade and in all alleles: Estimation of Tajima’s D in a sliding window over the DPB1 gene. Negative values indicate an excess of rare variants, possibly caused by purifying selection, a recent genetic sweep or a recent population expansion. Positive values indicate an excess of intermediate variants, possibly caused by balancing selection. A schematic representation of the gene structure is shown below with two grey boxes representing 5′- and 3′-UTR, red boxes representing exons 1–5 and a yellow box showing the position of the microsatellite. The ranges below the gene model show the homogenisation zone with no differentiating variants, the transition zone with increasing differentiation between the clades and the differentiation zone harbouring variants which fully separate the two clades.

differences (transition zone). Further upstream, roughly corresponding with the onset of the DPA1 promotor sequence on the template strand, DPB1 is dominated by shared variation between the two clades (homogenisation zone).

The bipartite pattern may be explained by functional genomic features overlapping with DPB1. Especially because DPA1 overlaps DPB1 up to exon 2 on the template strand (Fig. 2), conserved positions required to maintain functionality of DPA1 as well as DPB1 may prevent the accumulation of clade-separating polymorphisms at certain sites.

Another key feature that differentiates the two clades is a tetranucleotide microsatellite in intron 3 (Fig. 3). At least 55 of the differentiating positions can be explained by the microsatellite, which is of fixed length in G-clade alleles, but of variable length in A-clade alleles. Still, 120 variable sites separate the two allele groups when masking the microsatellite. The haplotype networks, as a representation of the edit distance between alleles, exhibit a minimal edit distance of 214 nucleotide substitutions between the closest alleles of the two groups, indicating that a fast transition of A-clade alleles to G-clade alleles is not possible.

The differential influence of A- and G-clade alleles on GVHD risk is hypothesised to be driven by their difference in expression levels: A alleles are lowly expressed and G alleles are more highly expressed. We propose three not mutually exclusive scenarios to explain the difference in DPB1 expression and the concomitant changes in GVHD risk: First, the seven differentiating SNPs in exon 3, including two non-synonymous positions, may have a direct influence on HSCT outcome by altering the extracellular domain of the resulting protein. Possible mechanisms may include binding sites for other molecules or functional motifs like ubiquitination sites or signalling motifs. Although the majority of studies focus on ARD features, a comparison of known non-ARD structures in class II genes revealed no separating mutations in such motifs [41]. Second, conserved point mutations in the non-coding sequences of DPB1 may impact regulation of DPB1 expression by disturbing regulatory elements, especially in the 3′-UTR. Thus, a CTCF binding site located on the template strand in intron 2 is almost completely differentiated by three polymorphic positions between the two clades. CTCF binding sites have been shown to be involved in regulation of MHC-class II expression [28]. It is not possible, however, to directly infer the effects of mutations in the binding motif, as CTCF can function both as a transcriptional activator or as a repressor. In addition, we do not know which gene is impacted by this CTCF binding site, although both, an impact on DPA1 or DPB1 would affect the dynamics of HLA-DP heterodimer formation. Third, the size, i.e., the repeat count, of the microsatellite may serve as a regulator of DPB1 expression [42,43]. Under this scenario it is conceivable that expression of DPB1 is correlated to the size of the microsatellite: The more repeat units an allele contains, the lower is the expression level. Further studies are needed to examine a potential correlation between the length of the microsatellite and expression by combining full-length sequencing with determination of protein expression.

Phylogenetically, the MHC is an old system and MHC gene families can be found in all vertebrates, but it is also an extremely variable and evolutionarily dynamic region. This leads to the question how the two highly distinct DPB1 allele groups might have arisen. Four possibilities come to mind:

a) A recent emergence by accumulating point mutations with a subsequent loss of intermediate alleles. This seems extremely unlikely given the high edit distance between the two groups.

b) An ancient split between the two groups. This hypothesis also seems difficult to reconcile with the complete lack of intermediate alleles and is rendered unlikely by two additional observations: First, the A clade contains far fewer clade-specific polymorphisms than the G clade, suggesting an evolutionarily younger origin. Second, A-clade alleles were not present in all human populations and are, for instance, suspected to have entered the Japanese population only ~2300 years ago [44]. Given that signals of positive selection have been inferred for A-clade alleles in such populations, it is unlikely that the A clade was simply lost.

c) A gain of A-clade alleles by recombination of DPB1 with another HLA class II gene, followed by a period of positive selection on A-clade alleles. This scenario is in accordance with the few clade-specific polymorphisms amongst A-clade alleles and also the massive preponderance of A-clade alleles in populations of European origin, but no clear source for the A-clade sequence can be identified in the human genome.

d) Finally, introgression of DNA from an archaic sister species, possibly Neandertal hominids, as has been shown for HLA class I genes [45], may explain the observed pattern. However, more studies are required to ascertain this evolutionary scenario, including sensitive sequencing of ancient sister species and common ancestors.

It seems likely that the rs9277534-A clade has been acquired after the speciation of *Homo sapiens*. The introgression hypothesis is further supported by patterns of natural selection. We observed strong signs of balancing selection across the two allele groups, indicating that both groups are beneficial and maintained in populations, which has, to a lower extent, been described previously [46,47]. Within the groups we observe a near neutral evolution, but with tendencies towards lower-than-expected diversity. This trend is greater in the A clade, possibly indicating a selective sweep of the more recently acquired group.

Overall, the additional full-length alleles of DPB1 greatly extend the knowledge of DPB1 which enables further studies towards the specification of DPB1 as an important factor for unrelated haematopoietic stem-cell transplantations. This is, to our knowledge, the first study to analyse non-coding variation in an HLA gene on such a comprehensive basis of full-length sequences. We found that DPB1 exhibits two deeply divergent conserved groups of alleles, which has not been described for any other HLA gene yet. Two likely paths of evolution for the divergent groups of DPB1 alleles, recombination with other HLA class II genes or introgression of DNA from archaic humans, are presented and should be targeted by further studies.

References

[1] B.E. Shaw, T.A. Gooley, M. Malkki, J.A. Madrigal, A.B. Begovich, M. Horowitz, A. Gratwohl, O. Ringdén, S.G.E. Marsh, E.W. Petersdorf, M.M. Horowitz, O. Ringde, The importance of HLA-DPB1 in unrelated donor hematopoietic cell transplantation, Blood 110 (113) (2007) 4560–4566.

[2] J. Trowsdale, J.C. Knight, Major histocompatibility complex genomics and human disease, Annu. Rev. Genomics Hum. Genet. 14 (1) (2013) 301–323.

[3] P. Loiseau, M. Busson, M.L. Balere, A. Dormoy, J.D. Bignon, K. Gagne, L. Gebuhrer, V. Dubois, I. Jollet, M. Bois, P. Perrier, D. Masson, A. Moine, L. Absi, D. Reviron, V. Lepage, R. Tamouza, A. Toubert, E. Marry, Z. Chir, J.P. Jouet, D. Blaise, D. Charron, C. Raffoux, HLA association with hematopoietic stem cell transplantation outcome: the number of mismatches at HLA-A, -B, -C, -DRB1, or -DQB1 is strongly associated with overall survival, Biol. Blood Marrow Transpl. 13 (8) (2007) 965–974.

[4] Y. Morishima, K. Kashiwase, K. Matsuo, F. Azuma, S. Morishima, M. Onizuka, T. Yabe, M. Murata, N. Doki, T. Eto, T. Mori, K. Miyamura, H. Sao, Biological significance of HLA locus matching in unrelated donor bone marrow transplantation, Blood 125 (7) (2015) 1189–1198.

[5] S.J. Lee, J. Klein, M. Haagenson, L.A. Baxter-lowe, D.L. Confer, M. Fernandez-vina, N. Flomenberg, M. Horowitz, C.K. Hurley, M. Oudshoorn, E. Petersdorf, M. Setterholm, S. Spellman, T.M. Williams, C. Anasetti, M. Eapen, H. Noreen, D. Weisdorf, High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation, Blood 110 (13) (2007) 4576–4583.

[6] D. Fürst, C. Müller, V. Vucinic, D. Bunjes, W. Herr, M. Gramatzki, R. Schwerdtfeger, R. Arnold, H. Einsele, G. Wulf, M. Pfreundschuh, B. Glass, H. Schrezenmeier, K. Schwarz, J. Mytilineos, High-resolution HLA matching in hematopoietic stem cell transplantation: a retrospective collaborative analysis, Blood 122 (18) (2013) 3220–3229.

[7] J.-M. Tiercy, How to select the best available related or unrelated donor of hematopoietic stem cells? Haematologica 101 (6) (2016) 680–687.

[8] K. Fleischhauer, B.E. Shaw, T. Gooley, M. Malkki, P. Bardy, J.D. Bignon, V. Dubois, M.M. Horowitz, J.A. Madrigal, Y. Morishima, M. Oudshoorn, O. Ringden, S. Spellman, A. Velardi, E. Zino, E.W. Petersdorf, Effect of T-cell-epitope matching at HLA-DPB1 in recipients of unrelated-donor haemopoietic-cell transplantation: a

retrospective study, Lancet Oncol. 13 (4) (2012) 366–374.

[9] E.W. Petersdorf, M. Malkki, C. O'hUigin, M. Carrington, T. Gooley, M.D. Haagenson, M.M. Horowitz, S.R. Spellman, T. Wang, P. Stevenson, High HLA-DP expression and graft-versus-host disease, New Engl. J. Med. 373 (7) (2015) 599–609.

[10] M. Burek Kamenaric, M. Maskalan, Z. Grubic, M. Mikulic, R. Serventi Seiwerth, N. Durakovic, R. Vrhovac, K. Stingl Jankovic, R. Zunec, HLA-DPB1 matching in unrelated hematopoietic stem cell transplantation program contributes to a higher incidence of disease relapse (2017). https://doi.org/10.1016/j.humimm.2017.08.008.

[11] K. Fleischhauer, B.E. Shaw, HLA-DP in unrelated hematopoietic cell transplantation revisited: challenges and opportunities, Blood 130 (9) (2017) 1089–1096.

[12] Y. Kamatani, S. Wattanapokayakit, H. Ochi, T. Kawaguchi, A. Takahashi, N. Hosono, M. Kubo, T. Tsunoda, N. Kamatani, H. Kumada, A. Puseenam, T. Sura, Y. Daigo, K. Chayama, W. Chantratita, Y. Nakamura, K. Matsuda, A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians, Nat. Genet. 41 (5) (2009) 591–595.

[13] R. Thomas, C.L. Thio, R. Apps, Y. Qi, X. Gao, D. Marti, J.L. Stein, K.A. Soderberg, M.A. Moody, J.J. Goedert, G.D. Kirk, W.K. Hoots, S. Wolinsky, M. Carrington, A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection, J. Virol. 86 (12) (2012) 6979–6985.

[14] G.M. Taylor, S. Dearden, P. Ravetto, M. Ayres, P. Watson, A. Hussain, M. Greaves, F. Alexander, O.B. Eden, Genetic susceptibility to childhood common acute lymphoblastic leukaemia is associated with polymorphic peptide-binding pocket profiles in HLA-DPB1*0201, Human Mol. Genet. 11 (14) (2002) 1585–1597.

[15] J. Field, S.R. Browning, L.J. Johnson, P. Danoy, M.D. Varney, B.D. Tait, K.S. Gandhi, J.C. Charlesworth, R.N. Heard, G.J. Stewart, T.J. Kilpatrick, S.J. Foote, M. Bahlo, H. Butzkueven, J. Wiley, D.R. Booth, B.V. Taylor, M.A. Brown, J.P. Rubio, J. Stankovich, S.A. Broadley, B.L. Browning, W.M. Carroll, L.R. Griffiths, A.G. Kermode, J. Lechner-Scott, P. Moscato, V.M. Perreau, R.J. Scott, M. Slee, A polymorphism in the HLA-DPB1 gene is associated with susceptibility to multiple sclerosis, PLoS One 5 (10). https://doi.org/10.1371/journal.pone.0013454.

[16] J.A. Hollenbach, S.D. Thompson, T.L. Bugawan, M. Ryan, M. Sudman, M. Marion, C.D. Langefeld, G. Thomson, H.A. Erlich, D.N. Glass, Juvenile idiopathic arthritis and HLA class I and Class II interactions and age-at-onset effects, Arthritis Rheum. 62 (6) (2010) 1781–1791.

[17] G. Martelli-Palomino, J.A. Pancotto, Y.C. Muniz, C.T. Mendes-Junior, E.C. Castelli, J.D. Massaro, I. Krawice-Radanne, I. Poras, V. Rebmann, E.D. Carosella, N. Rouas-Freiss, P. Moreau, E.A. Donadi, Polymorphic sites at the 3' untranslated region of the HLA-G gene are associated with differential hla-g soluble levels in the brazilian and French population, PLoS One 8 (10) (2013) 1–10.

[18] G. Amodio, V. Canti, L. Maggio, S. Rosa, M.T. Castiglioni, P. Rovere-Querini, S. Gregori, Association of genetic variants in the 3'UTR of HLA-G with Recurrent Pregnancy Loss, Hum. Immunol. 77 (10) (2016) 886–891.

[19] M.E. Jansen, I. Branković, J. Spaargaren, S. Ouburg, S.A. Morré, Potential protective effect of a G > A SNP in the 3'UTR of HLA-A for Chlamydia trachomatis symptomatology and severity of infection, Pathogens Disease 74 (2) (2016) 1–5.

[20] E.C. Castelli, L.C. Veiga-Castelli, L. Yaghi, P. Moreau, E.A. Donadi, Transcriptional and posttranscriptional regulations of the HLA-G gene, J. Immunol. Res. (2014) .

[21] A. Curinha, S. Oliveira Braz, I. Pereira-Castro, A. Cruz, A. Moreira, Implications of polyadenylation in health and disease, Nucleus 5 (6) (2014) 508–519.

[22] M. Ferizi, C. Leonhardt, C. Meggle, M.K. Aneja, C. Rudolph, C. Plank, J.O. Rädler, Stability analysis of chemically modified mRNA using micropattern-based single-cell arrays, Lab. Chip 15 (17) (2015) 3561–3571.

[23] S. Kulkarni, V. Ramsuran, M. Rucevic, S. Singh, A. Lied, V. Kulkarni, C. O'hUigin, S. Le Gall, M. Carrington, Posttranscriptional regulation of HLA-A protein expression by alternative polyadenylation signals involving the RNA-binding protein syncrip, J. Immunol. (2017) .

[24] B. Schöne, S. Bergmann, K. Lang, I. Wagner, A.H. Schmidt, E.W. Petersdorf, V. Lange, Predicting an HLA-DPB1 expression marker based on standard DPB1 genotyping: linkage analysis of over 32,000 samples, Hum. Immunol. 79 (1) (2018) 20–27.

[25] J. Robinson, J.A. Halliwell, J.D. Hayhurst, P. Flicek, P. Parham, S.G. Marsh, The IPD and IMGT/HLA database: allele variant databases, Nucl. Acids Res. 43 (D1) (2015) D423–D431.

[26] J.A.T. Young, J. Trowsdale, A processed pseudogene in an intron of the HLA-DPβ1 chain gene is a member of the ribosomal protein L32 gene family, Nucl. Acids Res. 13 (24) (1985) 8883–8886.

[27] S. Balasubramanian, D. Zheng, Y.-J. Liu, G. Fang, A. Frankish, N. Carriero, R. Robilotto, P. Cayting, M. Gerstein, Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes, Genome Biol. 10 (1) (2009) R2.

[28] P. Majumder, J.M. Boss, CTCF Controls expression and chromatin architecture of the human major histocompatibility complex Class II locus, Mol. Cell. Biol. 30 (17) (2010) 4211–4223.

[29] V. Albrecht, C. Zweiniger, V. Surendranath, K. Lang, G. Schöfl, A. Dahl, S. Winkler, V. Lange, I. Böhme, A.H. Schmidt, Dual redundant sequencing strategy: full-length gene characterisation of 1056 novel and confirmatory HLA alleles, HLA 90 (2) (2017) 79–87.

[30] S.J. Mack, P. Cano, J.A. Hollenbach, J. He, C.K. Hurley, D. Middleton, M.E. Moraes, S.E. Pereira, J.H. Kempenich, E.F. Reed, M. Setterholm, A.G. Smith, M.G. Tilanus, M. Torres, M.D. Varney, C.E. Voorter, G.F. Fischer, K. Fleischhauer, D. Goodridge, W. Klitz, A.M. Little, M. Maiers, S.G. Marsh, C.R. Müller, H. Noreen, E.H. Rozemuller, A. Sanchez-Mazas, D. Senitzer, E. Trachtenberg, M. Fernandez-Vina, Common and well-documented HLA alleles: 2012 update to the CWD catalogue, Tissue Antigens 81 (4) (2013) 194–203.

[31] V. Surendranath, V. Albrecht, J.D. Hayhurst, B. Schöne, J. Robinson, S.G. Marsh, A.H. Schmidt, V. Lange, TypeLoader: a fast and efficient automated workflow for the annotation and submission of novel full-length HLA alleles, HLA 90 (1) (2017) 25–31.

[32] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.

[33] E.S. Wright, Using DECIPHER v2.0 to analyze big biological sequence data in R, R J. 8 (1) (2016) 352–359 doi:V12242009.

[34] R.R. Hudson, M. Slatkin, W.P. Maddison, Estimation of levels of gene flow from DNA sequence data, Genetics 132 (2) (1992) 583–589 PMC1205159.

[35] J. Wakeley, The variance of pairwise nucleotide differences in two populations with migration, Theor. Popul. Biol. 49 (1) (1996) 39–57.

[36] F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, Genetics 123 (3) (1989) 585–595 PMC1203831.

[37] B. Pfeifer, U. Wittelsbürger, S.E. Ramos-Onsins, M.J. Lercher, PopGenome: an efficient swiss army knife for population genomic analyses in R, Mol. Biol. Evol. 31 (7) (2014) 1929–1936.

[38] G. Csardi, T. Nepusz, The igraph software package for complex network research, InterJournal Complex Sy, 2006, 1695.

[39] E.R. Gansner, S.C. North, An open graph visualization system and its applications to software engineering, Software Practice Exp. 30 (11) (2000) 1203–1233.

[40] A. Sanchez-Mazas, J.M. Nunes, D. Middleton, J. Sauter, S. Buhler, A. McCabe, J. Hofmann, D.M. Baier, A.H. Schmidt, G. Nicoloso, M. Andreani, Z. Grubic, J.M. Tiercy, K. Fleischhauer, Common and well-documented HLA alleles over all of Europe and within European sub-regions: a catalogue from the European Federation for Immunogenetics, HLA 89 (2) (2017) 104–113.

[41] J. Harton, L. Jin, A. Hahn, J. Drake, Immunological functions of the membrane proximal region of MHC Class II molecules, F1000 Res. 5 (0) (2016) 368 http://f1000research.com/articles/5-368/v1.

[42] M.V. Rockman, G.A. Wray, Abundant raw material for cis-regulatory evolution in humans, Mol. Biol. Evol. 19 (11) (2001) 1991–2004.

[43] M. Gymrek, T. Willems, A. Guilmatre, H. Zeng, B. Markus, S. Georgiev, M.J. Daly, A.L. Price, J.K. Pritchard, A.J. Sharp, Y. Erlich, Abundant contribution of short tandem repeats to gene expression variation in humans, Nat. Genet. 48 (1) (2015) 22–29.

[44] M. Kawashima, J. Ohashi, N. Nishida, K. Tokunaga, Evolutionary analysis of classical HLA Class I and II genes suggests that recent positive selection acted on DPB1*04:01 in Japanese population, PLoS One 7 (10) (2012) 1–11.

[45] L. Abi-Rached, M.J. Jobin, S. Kulkarni, A. McWhinnie, K. Dalva, L. Gragert, F. Babrzadeh, B. Gharizadeh, M. Luo, F.A. Plummer, J. Kimani, M. Carrington, D. Middleton, R. Rajalingam, M. Beksac, S.G. Marsh, M. Maiers, L.A. Guethlein, S. Tavoularis, A.M. Little, R.E. Green, P.J. Norman, P. Parham, The shaping of modern human immune systems by multiregional admixture with archaic humans, Science 334 (6052) (2011) 89–94.

[46] B.D. Bitarello, C. de Filippo, J.C. Teixeira, J. Schmidt, P. Kleinert, D. Meyer, A.M. Andrés, Signatures of long-term balancing selection in human genomes, bioRxiv Preprint. https://doi.org/10.1101/119529.

[47] D. Meyer, V.R. Vitor, B.D. Bitarello, D.Y. Débora, K. Nunes, A genomic perspective on HLA evolution, Immunogenetics (2017) 1–23.