Human Immunology 80 (2019) 461–464

Contents lists available at ScienceDirect

Human Immunology

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

Accuracy of NGS HLA typing data influenced by STR

Hanneke W.M. van Deutekom, Wietse Mulder, Erik H. Rozemuller*

GenDx, Yalelaan 48, 3584CM Utrecht, the Netherlands

ARTICLE INFO

Keywords:
STR
NGS
HLA typing
HLA-DRB5
HLA-DRB1
Repeat

ABSTRACT

Next Generation Sequencing (NGS) has become a major technology in HLA typing. The expectations are that highly accurate and unambiguous typing results will be obtained. However, HLA typing by NGS has some limitations caused by imperfections in the PCR amplification. The accuracy of NGS data is investigated by analyzing the Short Tandem Repeats (STR) regions. For this analysis HLA-DRB5 is used as the model. The repeat length in a sample highly influences the repeat length distribution present in the reads of NGS data. With a repeat length of 20 only 50% of all reads were of the estimated repeat length, seriously hampering distinguishing allelic differences in this region correctly. Our findings are confirmed by doing the same analysis in HLA-DRB1. Despite the uncertainty of determining the repeat lengths, several new HLA-DRB5 alleles have been identified in this paper.

1. Introduction

Next Generation Sequencing (NGS) has become a major technology in HLA typing. NGS is known for highly accurate and unambiguous HLA typing results. However, HLA typing by NGS has some limitations. Both the PCR amplification as well as the technique to determine the DNA sequence cause inaccuracies in the final reads. PCR errors may consist of randomly erroneous incorporation of single nucleotides and slippage, especially at homopolymer regions and regions with STRs [1–3]. Depending on the sequencing platform, additional errors may be encountered.

Since in NGS analysis each read originates from a single molecule, the aforementioned errors can be inspected in close detail. Using the detailed inspection the effect of those errors on the final analysis result is investigated.

To determine the influence of STR regions on the accuracy of the NGS data, the repeated region adjacent to the 3′ end of exon two in HLA-DRB5 is studied. This region consists of two flanking dinucleotide repeats. The HLA-DRB5 gene is a good model because many individuals only possess a single copy of this gene, which eliminates analysis challenges caused by heterozygous individuals. HLA-DRB5 is associated with DRB1 * 15 or DRB1 * 16 alleles [4]. Selecting samples from individuals with either a DRB1 * 15 or a DRB1 * 16 allele, and a second allele which does not belong to these groups, guarantees that only one copy of the HLA-DRB5 gene is present. Based on the STR length distribution in the reads, an estimate of the repeat length can be determined. Additionally the percentage of reads with this estimated repeat length can be identified. This percentage influences the reliability of the HLA typing, and could explain false-heterozygous calling in this region.

The analysis is extended to HLA-DRB1 gene, which contains a similar STR structure located in the beginning of intron two. As the repeat patterns within HLA-DRB1 are distinct, the alleles within the heterozygous samples can be separated.

2. Materials and methods

2.1. Sample selection

Samples were selected to contain a single copy of the HLA-DRB5 gene, or are homozygous for HLA-DRB1 and HLA-DRB5.

For DRB1, samples were selected to include samples of all DRB1 allele groups.

2.2. PCR

Human genomic DNA samples were used to amplify HLA-DRB5 (n = 18) and HLA-DRB1 (n = 23) by a Polymerase Chain Reaction (PCR) using NGSgo®-AmpX (GenDx) and the GenDx-LongRange PCR kit according to the Instructions for Use. The enzyme in this kit is a mixture of a Taq polymerase and a proofreading enzyme with 3′-5′-exonuclease activity for the removal of erroneously built-in, wrong bases. The processivity is primarily driven by the Taq polymerase.

The PCR consists of 35 cycles of the following steps: 15 s, 95 °C

* Corresponding author.
E-mail address: e.rozemuller@GenDx.com (E.H. Rozemuller).

https://doi.org/10.1016/j.humimm.2019.03.007
Received 31 May 2018; Received in revised form 6 March 2019; Accepted 6 March 2019
Available online 26 March 2019

denaturation, 30 s, 65 °C annealing, and 6 min, 68 °C elongation. To study the effect on PCR slippage, the number of PCR cycles in the amplification protocol increased in steps of 2 cycli from 25 to 35.

2.3. Library preparation and NGS analysis

HLA-DRB5 and HLA-DRB1 amplicons were processed in the NGSgo® workflow for Illumina using NGSgo-LibrX and NGSgo-IndX (GenDx). Libraries were paired-end sequenced (2x150bp) on a MiSeq platform (Illumina). FASTQ files were analyzed in NGSengine® HLA typing software V2.9 (GenDx), using IPD-IMGT/HLA database 3.31.0 [5] to determine the HLA-DRB5 typing based on exon 2. HLA-DRB1 typing was based exon 2, exon 3 and intron 3. The best match reported by NGSengine is the typing result which is best represented by the NGS data. Note that the repeat discussed in this paper is located in intron 2, and is thus excluded for obtaining the typing result.

2.4. Repeats

The DRB5 repeat can be represented as GGGAATCTGA $(GT)_m$ $(GA)_n$ GGAAGAGAGAG in which the factors *m* and *n* are numbers describing the lengths of the respective dinucleotide repeats.

A software tool was developed to analyze all reads from the NGS data which fully cover the repeat region, including the flanking sequences. For these reads, the *m* and *n* values are determined and the distribution of these factors is studied.

The estimated repeat length is defined as the value with the largest number of reads. This is based on the assumption that the majority of fragments that will be generated during the PCR reaction have the correct length. However, for longer repeat lengths, the majority of generated fragments have lengths which differ from the estimated repeat length and thus large numbers of fragments with an inaccurate length will be present (see Fig. 1B).

Data of HLA-DRB1 of 23 samples have been studied in a similar way. The DRB1 repeat is flanked by ATCTGA and CGCCAT. Within these two flanking regions there is a GT and a GA repeat, however the GA repeat is interrupted by other nucleotides. Hence, only the GT repeat has been analyzed, since this repeat is present in the same configuration in all alleles.

Note that, the GT repeat of DRB1 * 07:01:01 is interrupted by a G and a T; ATCTGAGTGGTGTTT$(GT)_m(GA)_n$GCGAGACCGCCAT. The GT repeat used for this analysis starts after the additional T, as indicated in the pattern before. This pattern was observed in all samples with a DRB1 * 07:01:01.

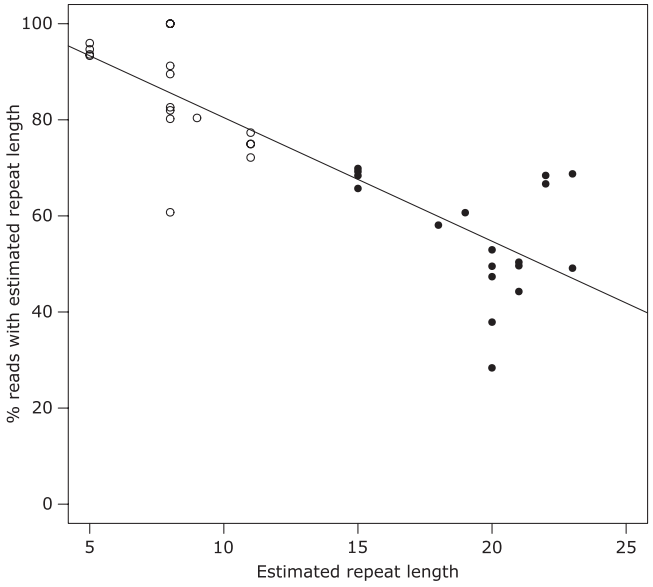

Fig. 2. Relation between the estimated repeat length and the percentage of reads with the estimated GA (open dots) and GT (closed dots) repeat length for HLA-DRB5.

18 of the DRB1 samples are heterozygous, due to the pattern around the GA repeat the two alleles of these samples could be distinguished reliably. For a list of the patterns of the repeats in DRB1 see Table 2.

3. Results

The NGS data of 18 DRB5 samples was used to determine the different lengths of the repeat per sample. A typical example of the length distribution of the GA repeat is shown in Fig. 1A. Here, 75% of all reads have a repeat length of 8 times GA, which is thus the estimated repeat length for this sample. The length distribution of the GT repeats in the

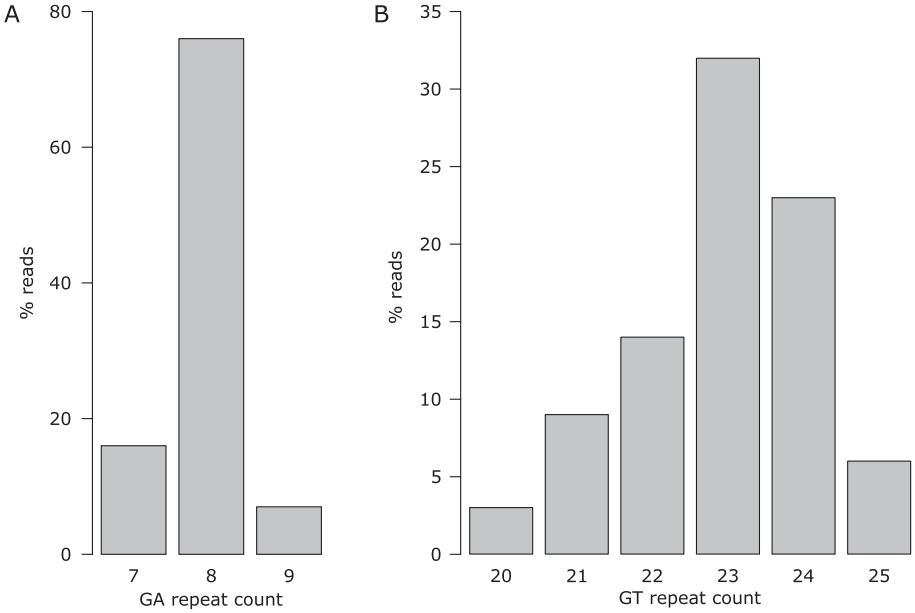

Fig. 1. Distribution of the length of the GA repeat and GT repeat in a sample typed as DRB5 * 01:02. 1A: The estimated GA repeat length is 8. 75% of all reads contained 8 times GA. 1B: The estimated GT repeat length is 23. 32% of all reads have 23 times GT.

Table 1
The estimated repeat length for the DRB5 samples. For DRB5 * 01:01, four different configurations of the factors m and n are found. For DRB5 * 01:02 four different configurations are identified. For DRB5 * 02:02 a single configuration is found. The number of alleles identified for each combination is indicated between brackets. The *represents the configuration of the repeat in the genomic sequence of DRB5 * 01:01:01 within the IPD-IMGT/HLA database 3.34, which is the only repeat sequence in this database.

	GA				
GT	*m\n*	5	8	9	11
	15				DRB5 * 02:02 (4)
	18		DRB5 * 01:01 (1)		
	19	DRB5 * 01:01 (1)			
	20		DRB5 * 01:01 (3)	DRB5 * 01:02 (1)	
	21	DRB5 * 01:01 (3)*	DRB5 * 01:02 (1)		
	22		DRB5 * 01:02 (2)		
	23		DRB5 * 01:02 (2)		

same sample is shown in Fig. 1B. The estimated repeat length is 23 repeats, which occurs in only 32% of all reads. 29% of all reads have 1 or 2 repeats more, 26% have less repeats.

For all 18 samples the relation between the estimated repeat lengths and the percentage of reads with this length is shown in Fig. 2. The estimated repeat length and the percentage containing this length are correlated ($p < 0.001$, $R^2 = 0.71$). Remarkably, with an estimated repeat length of 20, the percentage of reads with the estimated repeat length is only 50%.

Typing the samples considering only exon 2, i.e. using the core region, revealed 8 samples to be DRB5 * 01:01, 6 samples DRB5 * 01:02 and 4 samples DRB5 * 02:02. Analysis of the repeats in the 8 DRB5 * 01:01 samples showed 4 different repeat configurations, as is shown in Table 1. The number of occurrences of these configurations in the analyzed samples is indicated between brackets. Analysis of the repeats of the 6 DRB5 * 01:02 samples showed 4 repeat configurations. All DRB5 * 02:02 samples have the same repeat configuration of 15 times GT followed by 11 times GA.

In all samples the sequences flanking the repeats were identical to the sequences described in the Materials and Methods.

The analysis has also been applied to the HLA-DRB1 gene. However, the repeat in DRB1 is more complicated than that in DRB5. DRB5 has a clear $(GT)_m$ $(GA)_n$ repeat, while the $(GA)_n$ repeat in DRB1 is interrupted by other nucleotides, creating patterns like "ATCTGA$(GT)_m$AA$(GA)_x$AA$(GA)_y$GCGCGCCAT". A longer list of patterns is provided in Table 2. Because these patterns are different between allele groups, they can be used to separate the two alleles in a heterozygous sample. Therefore, it is possible to determine the estimated GT repeat length in heterozygous DRB1 samples. Of the 23 samples that were used for analysis, 5 were homozygous and 18 were

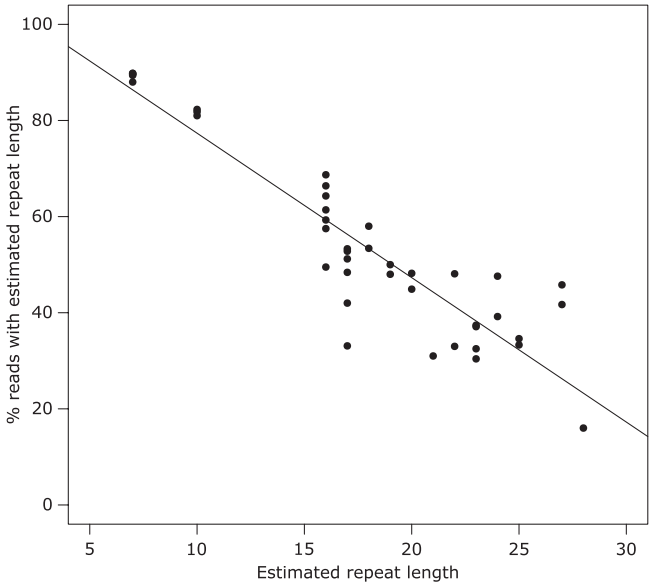

Fig. 3. Relation between the estimated GT repeat length in the HLA-DRB1 samples and the percentage of reads with the estimated repeat length.

Table 2
The DRB1 repeat patterns and estimated repeat lengths for the (3rd field) best matching alleles found in the panel. The sequence motifs flanking the repeats are underlined. Number of estimated repeat lengths are encoded by m, n1, n2 and n3.

Allele	Pattern	*m*	IPD-IMGT/HLA 3.34
DRB1 * 01:01:01	ATCTGA$(GT)_m$AAGAAA$(GA)_{n1}$GCGCGCCAT	16	16
DRB1 * 01:02:01	ATCTGA$(GT)_m$AAGAAA$(GA)_{n1}$GCGCGCCAT	19	15
DRB1 * 03:01:01	ATCTGA$(GT)_m(GA)_{n1}$CAGAGAGACA$(GA)_{n2}$GCGCCAT	17/20	17/19/22
DRB1 * 03:02:01	ATCTGA$(GT)_m(GA)_{n1}$CAGAGAGACA$(GA)_{n2}$GCGCCAT	19	19
DRB1 * 04:01:01	ATCTGA$(GT)_m(GA)_{n1}$GCGCCAT	22	21/22
DRB1 * 04:06:01	ATCTGA$(GT)_m(GA)_{n1}$GCGCCAT	22/23	22
DRB1 * 07:01:01	ATCTGAGTGGTGTT$(GT)_m(GA)_{n1}$GCGAGACCGCCAT	7	7
DRB1 * 08:02:01	ATCTGA$(GT)_m(GA)_{n1}$CA$(GA)_{n2}$GCGCGCCAT	17	17
DRB1 * 08:03:02	ATCTGA$(GT)_m(GA)_{n1}$CA$(GA)_{n2}$GCGCGCCAT	17	17
DRB1 * 09:01:02	ATCTGA$(GT)_m(GA)_{n1}$GGAA$(GA)_{n2}$CAGAAAGAGGGAGCGCGCCAT	10	10
DRB1 * 10:01:01	ATCTGACTCT$(GT)_m(GA)_{n1}$GCGCGCCAT	16/17	15/16/17
DRB1 * 11:01:01	ATCTGA$(GT)_m(GA)_{n1}$CAGAGAGACA$(GA)_{n2}$GCGCCAT	23/24/25	25/27/28
DRB1 * 12:02:01	ATCTGA$(GT)_m(GA)_{n1}$CA$(GA)_{n2}$GCGCCAT	28	28/30
DRB1 * 13:01:01	ATCTGA$(GT)_m(GA)_{n1}$CAGAGAGACA$(GA)_{n2}$GCGCCAT	23	22/23
DRB1 * 13:02:01	ATCTGA$(GT)_m(GA)_{n1}$CAGAGAGACA$(GA)_{n2}$GCGCCAT	20/21	20/24/26
DRB1 * 14:05:01	ATCTGA$(GT)_m(GA)_{n1}$CA$(GA)_{n2}$GCGCCAT	25	26/28
DRB1 * 14:54:01	ATCTGA$(GT)_m(GA)_{n1}$CA$(GA)_{n2}$GCGCCAT	24	21/25/34
DRB1 * 15:01:01	ATCTGA$(GT)_m(GA)_{n1}$CA$(GA)_{n2}$CAGAGAGAGGAA$(GA)_{n3}$GCGCCAT	18	17/18/19/20
DRB1 * 15:02:01	ATCTGA$(GT)_m$GAGACA$(GA)_{n1}$CAGAGAGAGGAA$(GA)_{n2}$GCGCCAT	27	28/29
DRB1 * 16:01:01	ATCTGA$(GT)_m(GA)_{n1}$GGAA$(GA)_{n2}$GCGCCAT	18	18

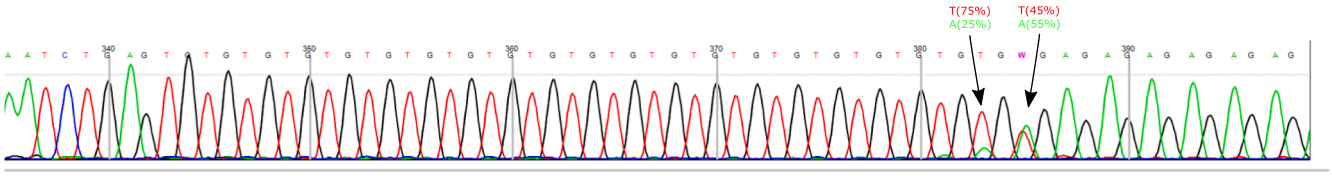

Fig. 4. The forward DNA trace of a DRB5 * 01:02 sample with 22 GT repeats. The double peaks indicated by the arrows are caused by the fragments lacking one or two GTs due to DNA slippage.

heterozygous, resulting in 41 estimated GT repeat lengths. A correlation is found between the estimated GT repeat length and the percentage of reads with the estimated repeat length (Fig. 3, $p < 0.001$, $R^2 = 0.84$), which is similar to the results for DRB5.

One could expect that the differences in repeat length depend on the PCR cycli, as has been shown for PCR crossover [6]. Therefore the effect of different PCR cycli on the repeat length has been studied. The analysis to determine the estimated repeat length has been applied to two samples with different PCR cycli, as described in the Materials and Methodsd. The estimated repeat length was the same for each cycle, hence the number of cycli does not have an influence on the estimated repeat length (data not shown).

4. Discussion

HLA typing results based on NGS data are in general highly reliable. However, errors introduced during the primary PCR amplification can be identified in individual reads. The random incorporation of a wrong nucleotide during PCR has a low frequency and can be identified in the NGS data as a low level of noise. Systematic errors may influence reliability of the typing result. DNA slippage during PCR is a known phenomenon that generates artefacts observed at repetitive DNA structures, especially when this involves di-nucleotide repeats [1,2]. In the HLA-DRB genes two adjacent dinucleotide repeats can be identified flanking the 3′ end of exon 2. Studying repeats in HLA regions is hampered by heterozygosity of samples, leading to complex data. However, HLA-DRB5 can be used to study the repeats, since many samples have only a single copy of the DRB5 gene. The 18 samples included in this study all are proven to have a single copy of DRB5.

One could reason that different lengths of STRs are due to contamination of other DRB (pseudo) genes. Co-amplification of DRB1, DRB3, DRB4 or any of the DRB pseudogenes would result in additional patterns visible in between the noise (base calling errors) and the real data. One of the standard troubleshooting procedures is to blast sequences with these patterns. This would reveal the gene that was co-amplified. In the DRB5 data for this study no extra patterns were observed, therefore there is no indication of co-amplification, and hence, the different length of STRs cannot arise from co-amplification.

The typical distribution of the read length in the reads reflects the DNA slippage during PCR amplification at a dinucleotide region. Similar patterns in STR fragment length analysis in e.g. chimerism STR methods can be observed [2], indicating that the aberrant lengths are generated during the PCR amplification. In Sanger dideoxy chain termination sequencing these artefacts can be identified by double peaks (Fig. 4). Samples analyzed with both Sanger sequencing and NGS sequencing showed similar results: the read distribution over the repeat lengths reflect the peak patterns in Sanger results at the repeat location (data not shown). This supports the hypothesis that the repeat variation in the reads is caused by the primary PCR, and not introduced by the NGS or Sanger sequencing procedure.

Our results show a clear correlation between the estimated repeat length and the percentage reads with the estimated repeat length.

Different configurations of repeats for specific DRB5 alleles have been shown, which represent new alleles. Unfortunately, the sequences of these alleles cannot be submitted to the IPD-IMGT/HLA database for two reasons. The sequences available do not cover the complete allele, and it is not fully proven that the lengths of the dinucleotide repeats are correct, especially for the longer repeats.

The same method is applied to the DRB1 NGS data of 23 samples, in which we identified 20 different best matching alleles. These alleles cover all DRB1 allele groups. The estimated repeat length of GT in DRB1 shows a similar correlation between percentages estimated read length and repeat length as compared to the DRB5 result (see Fig. 3). For 16/20 best matching alleles, the samples showed estimated repeat lengths which are also reported by the IPD-IMGT/HLA database 3.34. The estimated repeat lengths of 4/20 alleles differ from the length reported in IPD-IMGT/HLA database. In 5/20 alleles different estimated repeat lengths for the same best matching alleles are identified in different samples. This is in agreement with the IPD-IMGT/HLA database, where alleles from the same 3-field DRB1 group have different repeat lengths (Table 2).

If HLA typing analysis methods do not take the described results into account, the repeat region will in many cases be identified as heterozygous, i.e. either a repeat or an indel, which may lead to confusing typing results.

Until we have overcome the PCR generating artefacts, DNA regions of long dinucleotide repeats – in non-coding areas – should be excluded from routine HLA typing to avoid possible misinterpretation.

Acknowledgements

The authors would like to thank Hilde van der Schaar and Roel van Aard for technical assistance and Maarten Penning for carefully reading the manuscript.

References

[1] A. Hosseinzadeh-Colagar, M.J. Haghighatnia, Z. Amiri, M. Mohadjerani, M. Tafrihi, Microsatellite (SSR) amplification by PCR usually led to polymorphic bands: evidence which shows replication slippage occurs in extend or nascent DNA strands, Mol. Biol. Res. Commun. 5 (3) (2016) 167–174.

[2] D. Shinde, Y. Lai, F. Sun, N. Arnheim, *Taq* DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$microsatellites, Nucleic Acids Res. 31 (3) (2003) 974–980.

[3] T. Lyons-Darden, M.D. Topal, Effects of temperature, Mg2+ concentration and mismatches on triplet-repeat expansion during DNA replication *in vitro*, Nucleic Acids Res. 27 (11) (1999) 2235–2240.

[4] G. Andersson, Evolution of the human HLA-DR region, Front Biosci. 3 (1998) d739–45.

[5] J. Robinson, J.A. Halliwell, J.H. Hayhurst, P. Flicek, P. Parham, S.G.E. Marsh, The IPD and IMGT/HLA database: allele variant databases, Nucleic Acids Res. 43 (2015) D423–431.

[6] C.L. Holcomb, M. Rastrou, T.C. Williams, D. Goodridge, A.M. Lazaro, M. Tilanus, H.A. Erlich, Next-generation sequencing can reveal in vitro-generated PCR crossover products: some artifactual sequences correspond to HLA alleles in the IMGT/HLA database, Tissue Antigens 83 (1) (2014) 32–40.