FULL-LENGTH SEQUENCING OF HLA CLASS I GENES OF MORE THAN 1000 SAMPLES PROVIDES DEEP INSIGHTS INTO SEQUENCE VARIABILITY

Kathrin Lang[1], Ines Wagner[1], Bianca Schöne[1], Gerhard Schöfl[1], Carolin Zweiniger[1], Sylvia Clausing[2], Yannick Duport[2,3], Nicola Gscheidel[3], Andreas Dahl[2], Jürgen Sauter[4], Vinzenz Lange[1], Irina Böhme[1], Alexander Schmidt[1,4]

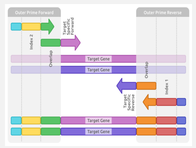
[1]**DKMS Life Science Lab**, Dresden, Germany; [2]CRTD - Center for Regenerative Therapies Dresden, Deep Sequencing Group, Dresden, Germany; [3]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany; [4]DKMS German Bone Marrow Donor Center, Tübingen, Germany

Aim

The vast majority of donor typing relies on sequencing exons 2 and 3 of HLA class I genes (HLA-A, -B, -C). For certain allele combinations this approach fails to report the anticipated “high resolution” (G-code) typing, due to the lack of exon-phasing information. To resolve ambiguous typing results for a haplotype frequency project, we established a whole gene sequencing approach for HLA class I, facilitating also an estimation of the degree of sequence variability outside the commonly sequenced exons.

Methods

Primer Design & PCR

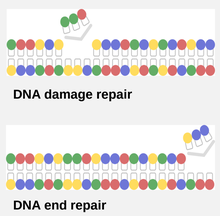


4-primer PCR approach

Pooled barcoded class I long-range PCR amplicons

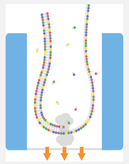
Primers were developed flanking the UTR regions of class I alleles resulting in similar amplicon lengths of 4.2-4.4 kb. Using a 4-primer approach, secondary primers containing barcodes were combined with the gene specific primers to obtain barcoded full-gene amplicons in a single amplification step. Up to 72 amplicons were pooled and purified before NGS library preparation.

NGS Library Preparation for SMRT Sequencing



DNA damage repair

SMRT-bell adaptor ligation

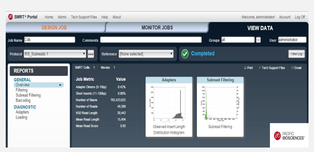


Continuous sequencing of full-length class I amplicons

DNA end repair

For SMRT sequencing library preparation we used the SMRTbell Template Prep KIT 1.0 from Pacific Biosciences following standard protocols. Library preparation included DNA damage repair as well as end repair. SMRT-bell adaptors were ligated to blunt-ended long-range amplicons to bind sequencing primers and facilitate continuous sequencing.

Data Analysis & HLA-Typing



fastQs per Barcode

Pooled amplicons were sequenced full length and phased in single runs on a PacBio RSII instrument. Demultiplexing was achieved using the SMRT Portal from Pacific Biosciences. Sequence analysis and HLA-Typing were performed using NGSengine (GenDx).

Results

Table 1: Resolved ambiguity combinations. The initial ambiguity combinations of the 1003 samples that had successfully been resolved by whole gene sequencing.

Locus	Ambiguities	G-code Typing	# Samples
A	AAANN/AAAPK	02:01:01G/66:01:01G	23
	XCUT/XMWE	29:02:01G/31:01:02	17
	AAAPH/XCUT	29:02:01G/33:03:01G	3
	ADJPV/MW	25:01:01G/26:08	1
	DPS/WCH	01:01:01G/11:01:01G	1
	GPE/XR	68:01:01G/68:02:01G	1
B	TFGP/XSTR	15:01:01G/35:01:01G	137
	AAARX/UTEN (WHPK/XXTP)	15:01:01G/40:01:01G	101
	AAARW/BAXH (WWWK/XYDZ)	40:01:01G/49:01:01G	15
	WHSE/XXWR	35:01:01G/52:01:01G	13
	BRVK/XYDC	07:02:01G/41:02:01	13
	UVTJ/XYDV	35:01:01G/41:02:01	8
	SHRK/UXCM	07:05:01G/08:01:01G	5
	NDAS/TFGT	15:01:01G/15:03:01G	3
	BCG/XZUZ	07:02:01G/48:01:01G	2
	TWRN/YEJT	35:08:01/58:01:01	2
	TWRR/XYYR	35:03:01G/58:01:01	2
	WGEY/WHSE	52:01:01G/58:01:01	1
	AAASR/FJDZ	53:01:01G/56:01:01	1
	ACENJ/ADNFG	35:01:01G/58:01:01	1
C	ABHPT/ZCXH (JGGY/SKMJ)	07:01:01G/07:02:01G	609
	AAATT/ZMSC	12:03:01G/16:01:01G	28
	CSSY/ZMSC (WREH/YVCV)	12:02:01G/16:01:01G	7
	AAATS/XV (WREF/WREG)	12:03:01G/16:02:01G	5
	AAASU/AAASW (ADJTM/AAASW)	03:02:01G/03:03:01G	3
	AAATD/AAATN	05:01:01G/08:01:01G	1

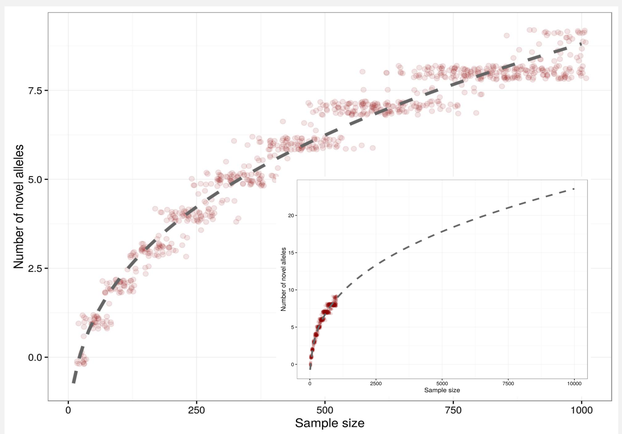


Figure 2: Extrapolation of the number of novel C*07:01:01G and C*07:02:01G allelic variants. The extrapolation was performed by fitting a rarefaction curve to bootstrapped samples based on 609 full-length C*07 sequences.

We successfully performed full-length gene sequencing of 1003 class I samples harboring ambiguous typings (**Table 1**). All consensus sequences for exons 2 and 3 were in full accordance with their MiSeq-derived sequences. Unambiguous allelic resolution was achieved for all samples.

62 samples (2 x HLA-A, 42 x HLA-B, 18 x HLA-C) contained 26 distinct novel alleles. These include one allele with exonic variation, 20 alleles with intronic variation, and 5 alleles with UTR variation (**Table 2**). For verification of these alleles standard 2x250 paired-end shotgun sequencing was conducted on an Illumina MiSeq. 17 alleles have been confirmed (for 7 analysis is pending). 2 sequence variants could not be verified, mainly because discrepancies arise within specific sequence motifs (e.g. homopolymer stretches).

C*07:01:01G and C*07:02:01G were most abundant in our data set (609 sequences) and comprise 24% of the alleles occurring in the German population (DKMS LSL data). Extrapolating our data (**Figure 2**), we expect approximately 500 novel C*07 variants among the 80 million Germans alone.

Figure 1: NGSengine comparison between Shotgun- and SMRT-(PacBio)-sequencing based HLA typing for one HLA-B sample. The novel B*15:01:01:01 shows one base exchange from G to T on Position 2324 (blue triangle). The magnification view displays examples of SMRT sequencing reads harboring the novel „T“ at position 2324 of B*15, when compared with the partner B*35 („G“ at position 2324).

Table 2: Novel sequence variations for class I alleles.

Locus/Allele	NA-Position	Reference Position	Old Base	New Base	# Samples	Shotgun
A*02:01:01:01	Intron 3	1406:9	T	C	1	Confirmed
A*31:01:02:01	Intron 5	2486	C	T	1	Confirmed
B*07:02:01	Intron 5	2380	G	C	1	Confirmed
B*15:01:01:01	Intron 1	159	C	A	11	Confirmed
B*15:01:01:01	Intron 2	691	G	T	1	Confirmed
B*15:01:01:01	Intron 2	691	G	T	1	Confirmed
		709	T	G		
B*15:01:01:01	Intron 3	1310	G	T	2	Confirmed
B*15:01:01:01	Intron 5	2324	G	T	19	Confirmed
B*15:03:01	Intron 3	1465	C	T	2	Confirmed
B*35:01:01:02	Intron 2	563	T	C	2	Not Confirmed
B*40:01:02	5´UTR	-216	A	G	1	Confirmed
B*40:01:02	Intron 1	196	C	A	1	Confirmed
B*49:01:01	Intron 5	2377	T	C	1	Not Confirmed
	Intron 6	2596	A	C		
C*07:01:01:01	Intron 1	203	G	C	1	Pending
C*07:01:01:01	Intron 3	1064	G	A	1	Pending
C*07:01:01:01	Intron 3	1133	T	G	1	Pending
C*07:01:01:01	Intron 3	1473	G	C	1	Pending
		1527	C	T		
C*07:01:01:01	Intron 5	2275	A	G	3	Confirmed
C*07:01:01:01	3´UTR	2987	G	A	1	Confirmed
C*07:02:01:03	Intron 2	710	G	C	1	Confirmed
C*07:02:01:03	Intron 5	2425	G	A	1	Pending
C*07:02:01:03	3´UTR	2995	G	A	1	Pending
C*07:02:01:03	3´UTR	3005	T	C	4	Confirmed
C*12:02:02	3´UTR	3005	T	C	1	Confirmed
C*12:03:01:01	Intron 2	654	C	A	1	Pending
C*12:03:01:01	Exon 4	1691	G	C	1	Confirmed

Conclusion

Here we present a whole gene amplification and sequencing approach for HLA class I genes. The maturity of this approach was demonstrated by sequencing more than 1000 samples, achieving fully phased allelic sequences. Extensive sequencing of one common allele combination hints at the yet to discover diversity of the HLA system outside the commonly analyzed exons.

DKMS Life Science Lab

Blasewitzer Str. 43 * 01307 Dresden * www.dkms.de