PHASED FULL-LENGTH SMRT SEQUENCING OF HLA-DPB1

Kathrin Lang[1], **Gerhard Schöfl**[1], Carolin Zweiniger[1], Maarten Penning[2], Erik Rozemuller[2], Sylvia Clausing[3], Yannick Duport[3,4], Nicola Gscheidel[4], Sylke Winkler[4], Vinzenz Lange[1], Irina Böhme[1], Alexander Schmidt[1,5]

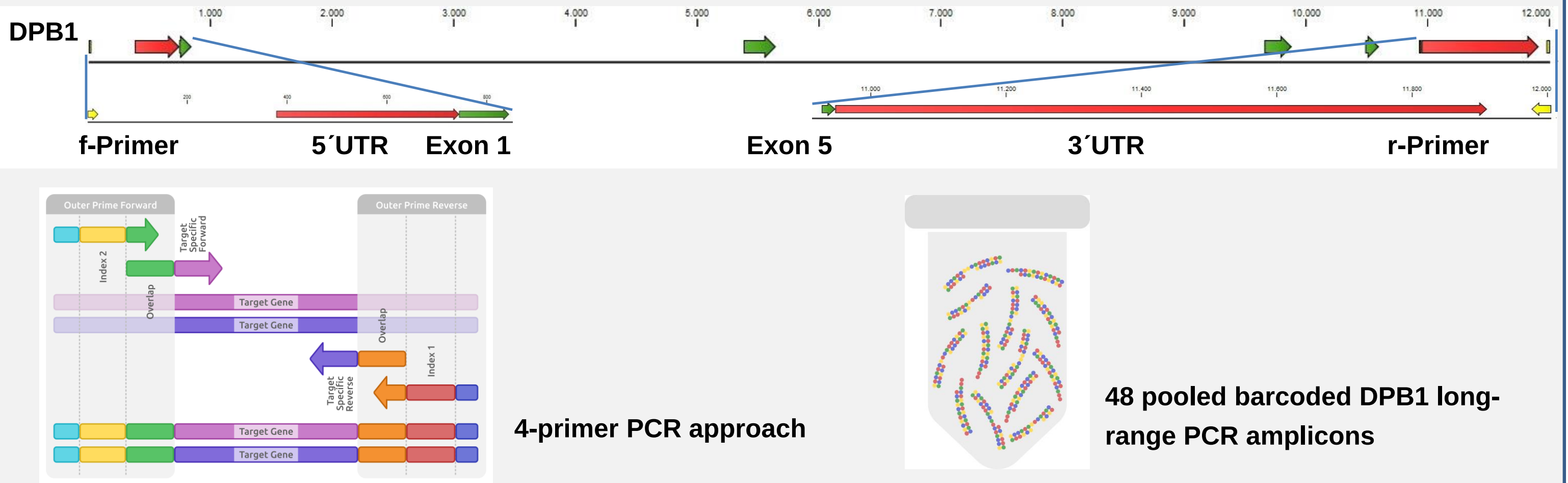
[1]**DKMS Life Science Lab**, Dresden, Germany; [2]GenDx, Utrecht, The Netherlands; [3]CRTD - Center for Regenerative Therapies Dresden, Deep Sequencing Group, Dresden, Germany; [4]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany; [5]DKMS German Bone Marrow Donor Center, Tübingen, Germany

Aim

In contrast to exon-based HLA-typing approaches, whole gene genotyping crucially depends on full-length sequences submitted to the IMGT/HLA database. Currently, full-length sequences are known for only 12 out of 550 HLA-DPB1 alleles (as of July 2015). Here, we present a whole-gene sequencing approach for DPB1 that allows full phase resolution to facilitate further exploration of the allelic structure at this locus.

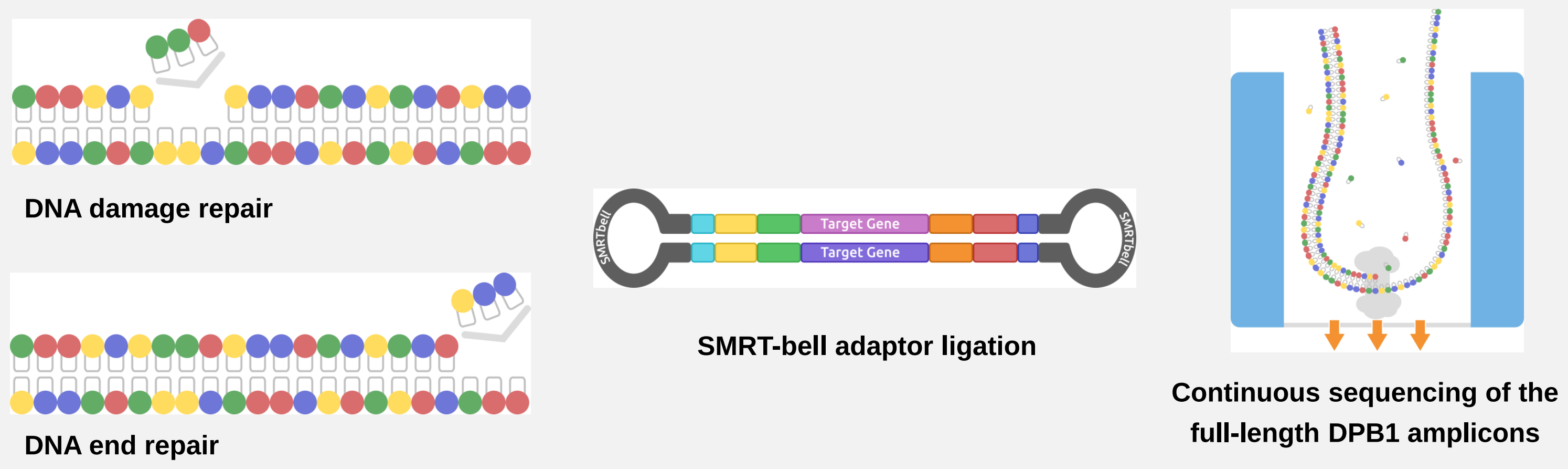
Methods

Primer Design & PCR



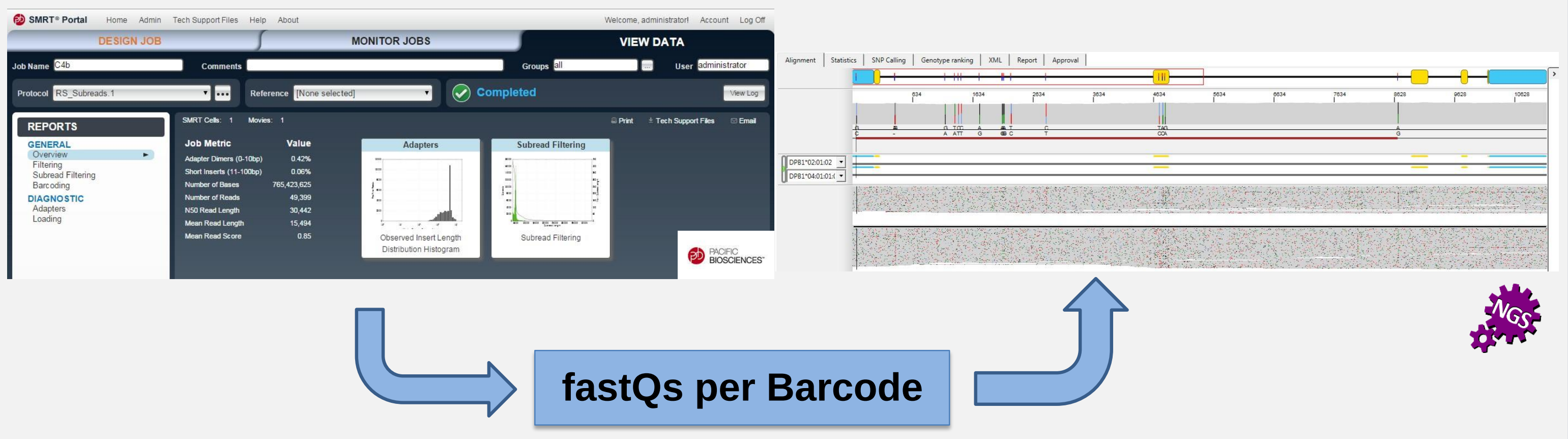
Primers were developed flanking the UTR regions of DPB1 resulting in a 12 kb amplicon. Using a 4-primer approach, secondary primers containing barcodes were combined with the gene-specific primers to obtain barcoded full-gene amplicons in a single amplification step. 48 amplicons were pooled and purified before NGS library preparation.

NGS Library Preparation for SMRT Sequencing


DNA damage repair
DNA end repair
Target Gene
Target Gene
SMRT-bell adaptor ligation
Continuous sequencing of the full-length DPB1 amplicons

For SMRT sequencing library preparation we used the SMRTbell Template Prep KIT 1.0 from Pacific Biosciences following standard protocols. Library preparation includes DNA damage repair as well as DNA 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



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

Results

We analyzed DPB1 for a set of 48 randomly picked donor samples. With 3 exceptions due to PCR failure, all genotype assignments conformed to previous typing results based on exon 2 and 3 short read sequencing. Allelic proportions for SMRT-sequencing-derived heterozygous positions were evenly distributed for all samples (range 0.4 - 0.6), suggesting unbiased long-range amplifications.

To verify PacBio read data, we also conducted standard 2x250 paired-end shotgun sequencing (Illumina MiSeq). Despite the high per-read raw error rates typical for SMRT sequencing (~15%), this comparison indicates an overall high level of agreement between the two sequencing technologies (**Figure 1**). Nevertheless, discrepancies arise at known problematic genomic positions and within specific sequence motifs (e.g. microsatellites and homopolymer stretches).

We describe novel intronic sequence variation for 5 previously described whole-length DPB1 alleles (**Table 1**). Additionally, we gathered whole-length sequences for 9 DPB1 alleles with so far unknown introns (**Table 2**). One of these alleles (HLA-DPB1*131:01) is classified as rare (**Figure 2**).

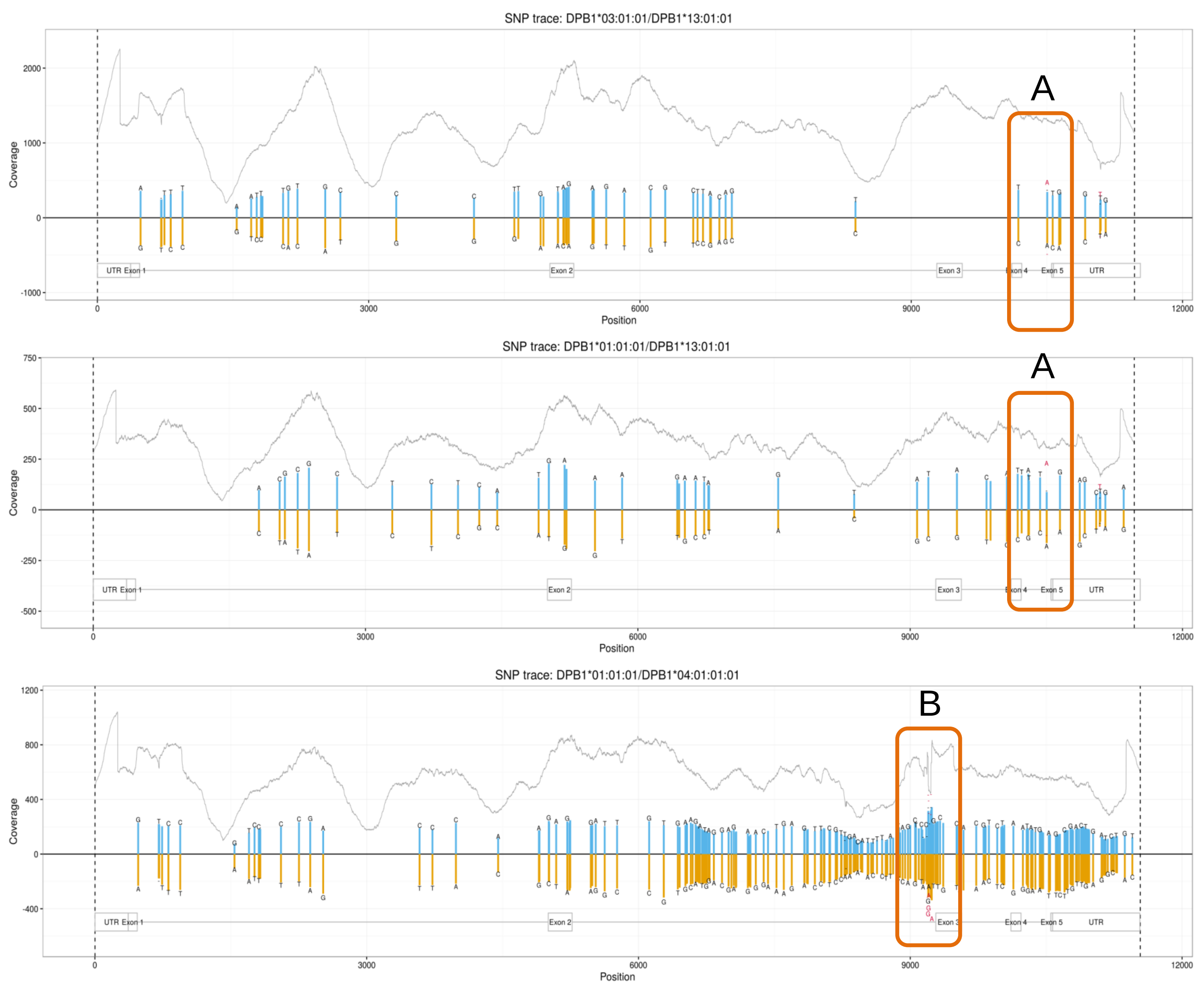


Figure 1: Overlay of PacBio-derived whole-gene phased DPB1 sequences and Illumina-based short reads. Most allelic differences are confirmed by short read sequences. Discrepancies at microsatellite-containing positions (B) and homopolymer stretches (A).

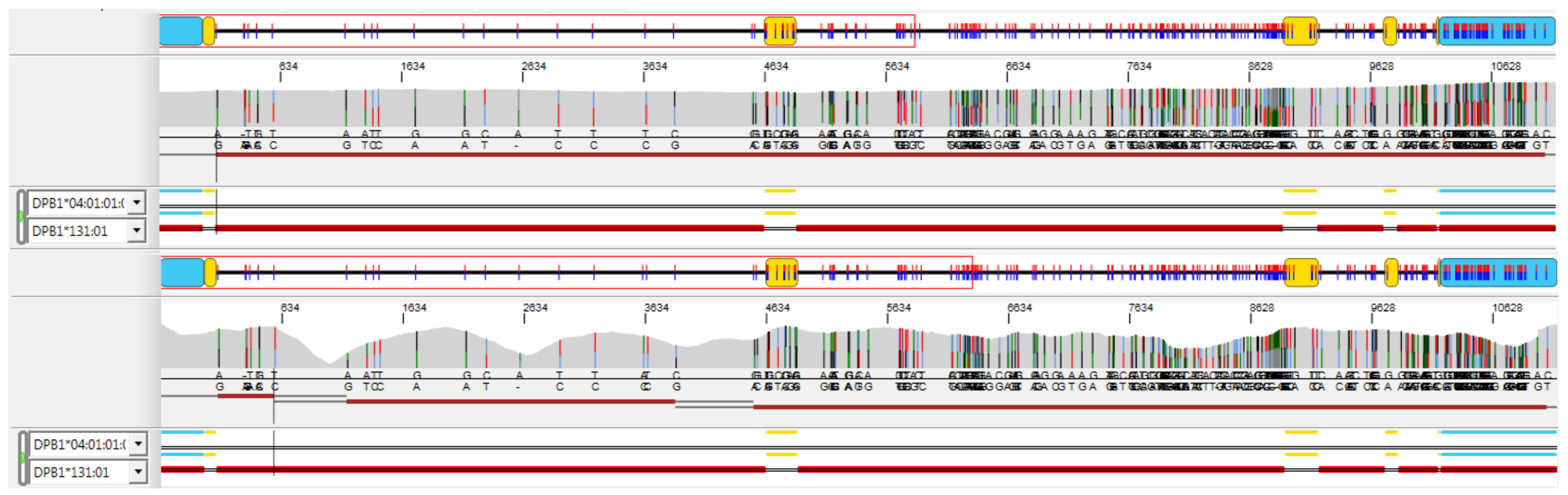


Figure 2: Coverage plots of DPB1*131:01 for PacBio (upper panel) and shotgun reads (lower panel) in NGSengine (GenDx). For such long genes, phase resolution becomes difficult with shotgun sequences. In contrast, whole-length PacBio reads can be reliably phased, even if most reference sequences (IMGT/HLA) miss intronic information.

Table 1: Novel intronic variation for known full-length DPB1 alleles.

DPB1-Allele	Intron	Reference position	Old Base	New Base	# Samples	Shotgun
02:01:02	2	6984	G	A	1	Confirmed
02:01:02	2	5156	A	G	1	Confirmed
04:01:01:01	1	983	C	T	3	Confirmed
	1	4307	C	T		
04:01:01:01	1	1031	C	T	1	Confirmed
04:01:01:01	2	6069	T	G	1	Pending
	3	9632	G	A		

Table 2: Newly full-length sequenced DPB1 alleles.

DPB1-Allel	# Samples
01:01:01	3
09:01:01	1
13:01:01	4
14:01:01	1
17:01	2
79:01	1
104:01	1
105:01	1
131:01	1

Conclusion

Here we present a whole gene amplification and sequencing workflow for DPB1 alleles utilizing single molecule real-time (SMRT) sequencing from Pacific Biosciences. Validation of consensus sequences against known exonic sequences highlights the reliability of this technology. This workflow will facilitate amending the IMGT/HLA Database for DPB1.

DKMS Life Science Lab

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