GENDX

personalizing diagnostics

Solution for Unresolvable New Alleles: Combining the Advantages of NGS Long and Short Reads for HLA and KIR Typing

Currently, the common practice for NGS HLA typing is to apply short read sequencing. The Illumina and Ion Torrent systems result in relatively short reads (up to ~500 bp) of high quality. This enables accurate base calling and in the vast majority of samples, this results in allele level typing of HLA genes. However, phasing of heterozygous positions over long distances is not always possible, sometimes causing genotype ambiguities. Long read sequencing, such as Oxford Nanopore or PacBio technology, could potentially solve the phasing between these heterozygous positions. However, up to now, the quality of long reads is not high enough for reliable HLA typing.

Materials & Methods

Whole gene HLA-F and KIR2DL1 amplifications were performed using NGSgo-AmpX (GenDx). Part of the amplicon was processed in the NGSgo® library preparation, with the Illumina compatible workflow (GenDx), and sequenced on an Illumina MiSeq platform. The library was sequenced using MiSeq V2 reagents, with application of paired-end sequencing (2x151 bp). Next to this, the amplicon was subjected to sequencing on the MinION, using the ligation sequencing Kit 1D in combination with the native barcoding kit 1D. The final library was applied to an R9.4 flow cell and run for 48 hours. The MinION was controlled with MinKNOW software v1.4.2 and base calling was performed with Albacore v2.0.1. Subsequently, the data from each workflow was analyzed in NGSengine separately as well as combined in one analysis.

Results

Analysis of the HLA-F Illumina data resulted in a high-quality result. The best matching typing result was found to be HLA-F*01:01:01:01, 01:01:01:08 with one mismatch. The two alleles differ at only two positions: genomic 534 (intron 2) and genomic 2698 (intron 6). In addition, the sample showed heterozygosity at position cDNA 417 (codon 139, TAC▸TAA, Tyr▸ Stop). The largest distance between two of the heterozygous positions was 1906 nucleotides. With a median insert size of 625, it was not possible to determine the phasing (Figure 1a), therefore no conclusion can be drawn about the phasing of the stop codon with the other heterozygous positions.

When the MinION data was analyzed, the same best matching typing result was obtained, showing the same three heterozygous positions. However, due to the higher noise of the data, other positions were called as heterozygous and several InDels were found (Figure 1b). Even though the reads covered the whole amplicon, the additional heterozygous positions resulted in broken phasing.

As NGSengine offers the option to analyze combinations of datasets, we combined reads of the 2 datasets to combine the advantages of both technologies. The short reads were used for high-quality base calling at each position and phasing over short distances (<800 bp), whereas the long reads facilitated phasing over long distances. Balancing the ratio between long and short reads enabled high-quality base determination at each position with long-distance phasing. As shown in Figure 1c, a combination of 20,000 paired-end short reads with 120 long reads resulted in high quality, fully phased data. This made clear that the sample contained an HLA-F*01:01:01:01 and a new allele which differs by one nucleotide from HLA-F*01:01:01:08.

The example in Figure 2 shows a whole-gene KIR2DL1 typing result. Sequencing with Illumina (Figure 2a) resulted in high-quality base calling (with exception of a region in intron 1), but ten phasing regions remained, potentially resulting in cis-trans ambiguities. The triangles point to positions where the sequence differs from the best matching alleles, indicating that the sample contains a new allele.

Sequencing the same amplicon with MinION resulted in a fully phased consensus sequence (Figure 2b). However, additional mismatches were displayed, which were not real mismatches but an artefact caused by the noise. In this example, the two alleles were nicely balanced, but a risk of such high noise levels is that in case of unbalanced amplifications, the second allele might disappear in the noise. Combination of the MiSeq and MinION data (100,000 and 100 reads, respectively), resulted in fully phased, high-quality data (Figure 2c) and therefore a reliable typing result.

Conclusion

- If heterozygous positions are too far apart to be phased with short read NGS technologies, combined analyses with long read sequence data is beneficial.
- Combining Oxford Nanopore and Illumina sequencing data results in high-quality, fully phased data

Figure 1. Sequencing of new HLA-F gene by short and long reads.

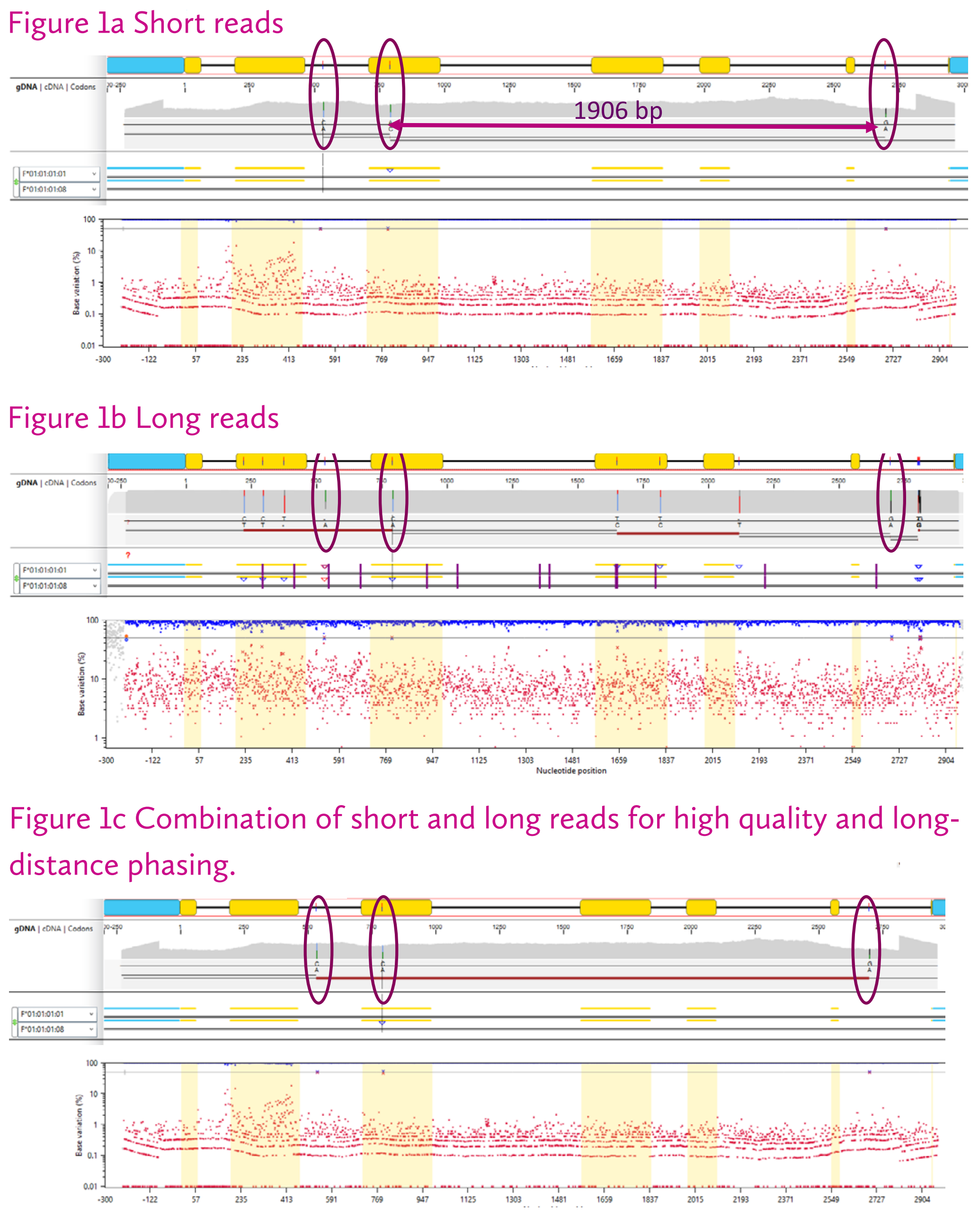
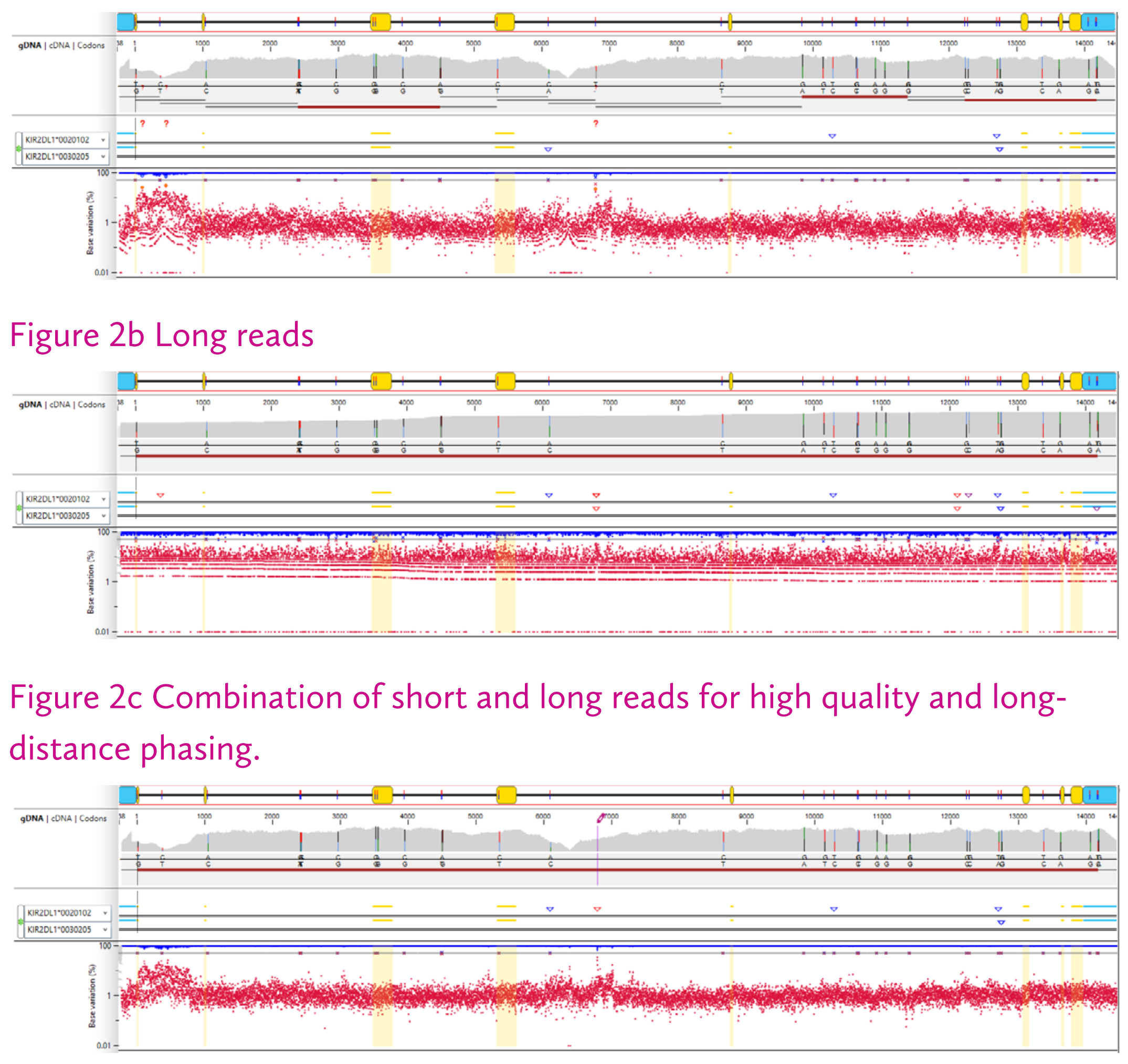


Figure 1a Short reads

Figure 1b Long reads

Figure 1c Combination of short and long reads for high quality and long-distance phasing.

Figure 2. Sequencing of a new KIR2DL1 gene by short and long reads.



2a Short reads

Figure 2b Long reads

Figure 2c Combination of short and long reads for high quality and long-distance phasing.

Rozemuller E.H., van Deutekom H., Bouwmans E.E., Van de Pasch L.A.L., Penning M.
GenDx, Utrecht, The Netherlands

GenDx.com

P182. M19-023 V1 ©2019 GenDx