Comprehensive haplotyping of the HLA gene family using nanopore sequencing

Sebastian Johansson1, Szilveszter Juhos1, David Redin2, Afshin Ahmadian2, Max Käller2,*

1 Stockholm University, Department of Biochemistry and Biophysics, Science for Life Laboratory, SE-171 65 Solna, Sweden

2 Royal Institute of Technology (KTH), School of Engineering Sciences in Chemistry, Biotechnology and Health, Department of Gene Technology, Science for Life Laboratory, SE-171 65 Solna, Sweden

* Corresponding author: Tel: +46 8 5248 1426; E-mail: max.kaller@scilifelab.se

ABSTRACT

The HLA gene family is the most polymorphic loci in the human genome; it encodes for the major histocompatibility complexes (MHC) which mediates the immune response in terms of cellular interactions with antigens. Compatibility between HLA alleles is thus of great medical interest for recipients of allogeneic transplantations. Traditional serological techniques to evaluate compatibility are now being replaced by more accurate DNA sequencing-based methods. However, short read sequencing data typically result in collapsed sequences representing a mixture of variants from native haplotypes. In addition, most previous studies have been limited to a few highly polymorphic exons of various HLA genes. Here we present haplotype-resolved full-length sequencing of the six most clinically relevant MHC Class I and Class II genes, to characterize the haplotypes of eight reference individuals, using a single MinION flow cell. The results show that full-length sequencing of single molecules enables haplotypes to be resolved to the highest degree of accuracy (four-field resolution). In this study, a majority of the alleles were classified with four-field resolution and could be verified through previously published genotyping studies. These results support the notion that nanopore sequencing could be a viable solution for highly accurate clinical evaluation of histocompatibility.
INTRODUCTION

The main role of proteins encoded by the human leukocyte antigen (HLA) gene complex is to present antigens and is thus of great clinical importance in organ transplant matching as well as many other research questions where autoimmunity is relevant. The nucleotide sequences of HLA alleles discovered to date are referenced in the IMGT/HLA database (1), containing more than 18 000 alleles. The vast majority of the alleles reported in the IMGT database are referring to highly polymorphic parts of the HLA genes: exons 2 and 3 for class-I genes (HLA-A, -B, and -C), and exon 2 of class-II genes (HLA-DRB1, -DQB1 and -DPB1). These variants often differ by single nucleotides located in coding or non-coding positions. The HLA genotyping nomenclature (e.g. HLA-A*03:01:01:02N) corresponds to the level of variation for given genes within the HLA family; specified by serological type (first field), protein-coding variations (second field), synonymous DNA substitutions (third field), non-coding variants (fourth field), and regulatory variants (alphabetical suffix).

Although Sanger DNA sequencing could accurately read up to a thousand bases, it cannot resolve individual haplotypes from a certain individual and thereby distinguish (i.e. phase) which genetic variations belongs to each allele. Despite technological developments in the era of massively parallel sequencing, the sheer number of HLA alleles previously found makes this genotyping challenging. The widely used Illumina sequencing platform features reads that are too short to directly phase regions of heterozygous SNPs that are too far apart (2), resulting in a need for high sequencing coverage and potentially error prone in silico solutions to correctly assign alleles (3). De novo sequence assembly methods are less prepared to produce multiple alleles (as they usually collapse haplotypes to a single allele), and it is difficult to resolve common errors in sequencing data, like allele dropout or chimeric PCR products (4). Also, non-coding regions are commonly overlooked but can be important since they may have a role in splicing and expression profile. Methods for linking short reads are starting to be used (5), in particular haplotype-resolved sequencing of HLA-A gene (6). Long read solutions provided by Pacific Biosciences (7) and Oxford Nanopore (8) technology can produce thousands of continuous bases that are comparable in length to the HLA genes. However, to overcome the relatively low quality of bases produced by nanopore sequencing a significantly high coverage has to be obtained to reliably interrogate single nucleotide variations. An attractive side of this approach is the short sequencing time and relatively low data footprint. Sequencing and genotyping of HLA-A and HLA-B using Oxford Nanopore technology has previously been demonstrated with promising results (9). Also, the Pacific Biosciences instrument has showed its feasibility in genotyping HLA Class 1 loci (10) as well as the HLA-DPB1 gene (11).
In this study we implemented a library preparation and nanopore sequencing routine as well as a bioinformatics pipeline that generates long reads spanning through the whole genomic regions of HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1 and HLA-DPB1 genes. Our data shows that we can resolve the alleles of these genes from eight individuals with high resolution (predominantly with four-field characterization) and near-perfect concordance for four of the genes in relation to external references (12, 13).

**MATERIALS & METHODS**

**Samples**

Extracted human DNA from eight individuals (CEPH Family 1463 members NA12891, NA12892, NA12877, NA12878, NA12882, and CEPH Family 1362 members NA11992, NA11993, NA10860, see Supplementary Figure 1 for pedigree) were obtained from the Coriell Institute (NJ, USA).

**Long range PCR of HLA target genes**

Primers spanning the entire region of HLA Class I genes (HLA-A, HLA-B, HLA-C) and Class II genes (HLA-DQB1, HLA-DRB1 and HLA-DPB1) were synthesized according to Hosomichi et al (2). These genes are all known to be highly polymorphic, hence the primers were designed to anneal to conserved regions and produce amplicons spanning both exonic and intronic regions. The estimated sizes of the amplicons were 3398 bp (HLA-A), 4296 bp (HLA-B), 4440 bp (HLA-C), 7118 bp (HLA-DQB1), 11899 bp (HLA-DRB1) and 13605 bp (HLA-DPB1). Long range PCR was carried out for each gene in separate 50 µl reactions using the PrimeSTAR® GXL kit (Takara Bio, CA, USA), and 25 ng of genomic DNA for HLA-A, -B, -C, -DRB1 and -DPB1 and 50 ng for HLA-DQB1, respectively. Small alterations to the manufacturer’s reaction composition were applied: 1X PrimeSTAR GXL Buffer, 200 µM and 400 µM of dNTPs (for HLA-A, -B, -C, and HLA-DQB1, -DRB1, -DPB1 respectively), 1.25 U PrimeStar GXL DNA Polymerase, 200 nM forward primer and reverse primer per reaction. PCR conditions for HLA-A, -B and -C were as follows: 94°C for 2 min, [98°C for 15 s, 60°C for 30 s, 68°C for 4 min] × 30 cycles, 68°C for 10 min, 4°C hold. PCR conditions for HLA-DQB1, -DRB1 and -DPB1 were as follows: 94°C for 2 min, [98°C for 15 s, 60°C for 30 s, 68°C for 10 min] × 30 cycles, 68°C for 10 min, 4°C hold. The PCR products were purified using an in-house automated protocol for polyethylene glycol precipitation and magnetic bead capture (14). The Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific, MA, USA) was used for concentration measurements and the Fragment Analyzer system (Advanced Analytical Technologies, IA, USA) for evaluating fragment lengths.
Amplicon pooling and adaptor ligation

Six HLA amplicons were pooled for every individual according to a molar ratio of 1:1:1:1:2:4:4 for HLA A:B:C:DQB1:DRB1:DPB1. The volume of each pool was brought to 45 µL using SpeedVac (Thermo Fisher Scientific). Ends were repaired and dA-tailed using the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, MA, USA) and in accordance to protocols from Oxford Nanopore (for R9 and SQK-NSK007). End-repaired and dA-tailed pools were purified with Agencourt AMPure XP beads (Beckman Coulter, CA, USA) with a 1:2 bead to sample ratio, followed by ligation of an adaptor sequence (for barcoding) using the Blunt/TA Ligase Master Mix (New England Biolabs) and a subsequent AMPure XP purification with a 1:2.5 bead to sample ratio.

Long-range barcoding PCR

Long-range barcoding PCR was conducted for each pool using the PrimeSTAR® GXL kit, 25 ng input material and 2 µL ONT barcode (BC01-BC08), using a unique barcode for each sample pool, with the following cycling conditions: 94° C for 2 min, [98° C for 15 s, 60° C for 30 s, 68° C for 4 min] × 30 cycles, 68° C for 10 min, 4° C hold. Following PCR, the samples were purified with AMPure XP beads (bead to sample ratio 0.4X), and concentrations and amplicon lengths were checked using Qubit HS and Fragment Analyzer, respectively. 100 ng of each sample pool was combined into an ultra-pool comprising all eight individuals and their six HLA-amplicons. The ultra-pool was brought down to 100 µL using a vacuum concentrator.

Library preparation and sequencing

The SQK-LSK208 kit was used to prepare the 2D sequencing library (Oxford Nanopore). First, the ends were repaired and dA-tailed using the same kit as the previous adaptor ligation step; 14 µL Ultra II Endprep reaction buffer and 6 µL Ultra II Endprep enzyme mix, incubated for 15 min at 20° C followed by 15 min at 65° C. Following a 1:2.5 bead to sample ratio AMPure XP clean-up, hair-pin adaptors were ligated according to the 2D sequencing kit. The subsequent streptavidin bead enrichment was also performed according to the protocol, except for the buffer used which was exchanged for the binding buffer included in the Dynabeads® kilobaseBINDER™ kit (Thermo Fisher Scientific, MA, USA). Approximately 200 ng of the sample pool was loaded on a SpotON flow cell version R9.4. For sequencing the MinKNOW v1.3.25 software was used, along with sequencing script NC_48Hr_Sequencing_Run_FLO-MIN106_SQK-LSK208.py. The sequencing was run for 20 hours and the online basecaller Metrichor was used for basecalling.
Bioinformatic workflow

Poretools (15) was used to generate FASTQ files from base-called FAST5 files. The resulting FASTQ files were separated at this base call step by barcodes, and only 2D reads were used for HLA typing. The genotyping workflow was implemented in a groovy-based domain specific language called Nextflow (16). The sub-processes of the workflow are calling other software like BWA (17), SAMtools (18) or custom python scripts. Nextflow not only connects workflow compartments, but records execution time, amount of input for output data and other process statistics in a timeline and trace file. The former is for visualizing the flow in HTML format, the latter is a tab-delimited text file recording resource usage. The reported benchmark timings were calculated from this text file. The following steps were included in the automated bioinformatic workflow called Icing. The workflow itself is open software under MIT license, and is available at https://github.com/NationalGenomicsInfrastructure/icing.

Read alignment with alternative contigs

All 2D reads corresponding to each barcode were mapped to the genomic sequences of each locus separately using BWA-MEM in an alternative contig-aware manner and saved to a BAM file. In this way, reads that are mappable to multiple alleles will be assigned to each of the corresponding alignments. The alignment parameters for BWA were selected to make sure the reads are aligned with edit distance, but are still processed in few minutes (Supplementary Table 1).

Read filtering

To compensate for sequencing errors, the raw alignment BAM file was filtered so that only reads (reference-based alignments) with a number of mismatches (edit distance) less than 5% of the total read length were kept. In addition, as sequence contigs are expected to be of similar length of the amplicons, a minimum length required for reads were set to 2,000 bp.

Allele selection for consensus generation

For each position in the alignments, the mean and median coverage is calculated using NumPy (19) to be used for consensus generation. In this process, alignments with uneven coverage are discarded, storing only the 16 most promising candidates in a new BAM file.

Generating consensus candidates

From this BAM file a pileup file is generated by SAMtools, and a consensus is generated by selecting the most frequent base for the position (ignoring bases with a quality less than 10). A consensus candidate is not reported for cases where >10% of positions feature lower coverage than predefined minimums.
Consensus sequences shorter than predefined minimum lengths, based on expected amplicon size, are also discarded. To compensate for stochastic errors in sequencing (leading to noise in the data and erroneous base calls), the minimum coverage required for each consensus sequence adjusted according to the sequencing depth (Supplementary Table 2).

**Final genotyping**

Characterization of genotypes was done in three steps. First, it was required that the consensus sequence mapped to primary exons (exon 2 and 3 for Class-I and exon 2 only for Class-II loci) with a minimum number of mismatches. Second, the candidate alleles generated in this step were investigated further by mapping against other exons, also with a minimum number of mismatches. Finally, intronic and UTR regions were considered for matching against sequences available in the IMGT/HLA database. In cases where only matches were found towards alleles where only exonic regions are available in the database, reported candidates were manually discarded after the bioinformatic pipeline, as such candidates by definition do not result in haplotypes of four-field resolution.
RESULTS & DISCUSSION

To test the accuracy of nanopore sequencing on the six most important HLA genes (HLA-A, -B, -C, HLA-DRB1, -DQB1, and HLA-DPB1) we examined eight well-characterized Coriell samples. Long range PCR was used to amplify the whole genes (ranging from 3,398 bp to 13,605 bp) before library preparation was performed, applying different barcodes for each individual. The amplicons were pooled and sequenced using a single MinION flow cell. The basecaller produced 171,711 high-quality 2D consensus reads with an average sequence length of 3.63 kb, the longest read was 15.55 kb and the total 2D consensus yield was 668.03 Mb. The basecaller was able to discriminate barcodes for 166,331 reads (96.8% of total).

Table 1. Sequencing statistics and haplotype coverage. The number of full-length reads supporting the called haplotypes (in Table 2) are specified for each allele (with single values for homozygous individuals).

<table>
<thead>
<tr>
<th></th>
<th>NA10860 (BC 1)</th>
<th>NA12878 (BC 2)</th>
<th>NA12891 (BC 3)</th>
<th>NA12882 (BC 4)</th>
<th>NA11992 (BC 5)</th>
<th>NA11993 (BC 6)</th>
<th>NA12877 (BC 7)</th>
<th>NA12892 (BC 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequencing Reads</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% 2D Reads</td>
<td>34.88%</td>
<td>34.85%</td>
<td>34.96%</td>
<td>34.92%</td>
<td>34.88%</td>
<td>35.02%</td>
<td>34.88%</td>
<td>34.84%</td>
</tr>
<tr>
<td>2D Reads (Mb)</td>
<td>61.0</td>
<td>85.6</td>
<td>48.8</td>
<td>89.5</td>
<td>87.9</td>
<td>99.1</td>
<td>72.2</td>
<td>67.9</td>
</tr>
<tr>
<td>Avg. 2D Read Length</td>
<td>3,563 bp</td>
<td>3,493 bp</td>
<td>3,448 bp</td>
<td>3,560 bp</td>
<td>3,396 bp</td>
<td>3,496 bp</td>
<td>3,476 bp</td>
<td>3,883 bp</td>
</tr>
<tr>
<td><strong>Haplotype Coverage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DQA1</td>
<td>33</td>
<td>12 : 61</td>
<td>11 : 12</td>
<td>0 : 129</td>
<td>27 : 35</td>
<td>17 : 54</td>
<td>18 : 25</td>
<td>337</td>
</tr>
<tr>
<td>HLA-DQA2</td>
<td>74</td>
<td>13 : 55</td>
<td>0 : 10</td>
<td>0 : 134</td>
<td>63 : 03</td>
<td>18 : 88</td>
<td>2</td>
<td>0 : 4</td>
</tr>
<tr>
<td>HLA-DPA1</td>
<td>0 : 6</td>
<td>6 : 14</td>
<td>8 : 8</td>
<td>28 : 28</td>
<td>7 : 10</td>
<td>8 : 10</td>
<td>16</td>
<td>8</td>
</tr>
</tbody>
</table>

As detailed in Table 1, the sequencing coverage and yield was fairly even between samples but particularly high for Class I genes (HLA-A, -B, -C). Differences in haplotype coverage points to an imbalance that varies greatly between loci. The observed deviation is strongly correlated with the length of the genes, resulting from reduced amplification and sequencing efficiency of the longest fragments. In particular, the ONT platform struggled with generating enough sequencing reads of significant length to
genotype and phase the alleles for HLA-DRB1 and HLA-DPB1 genes. For Class I genes and HLA-DQB1, the coverage across the lengths of the amplicons were even across the full length of the genes (Supplementary Fig. 2). However, for the two longest amplicons the sequencing coverage was not even with only a fraction of the reads covering the middle of the fragments (Supplementary Fig. 3).

A bioinformatic workflow was developed to keep track of statistics and map higher quality 2D reads to all alleles present in the IMGT/HLA database. Generally, using both 1D and 2D reads made genotyping more ambiguous (data not shown), therefore the 1D reads were omitted from the analysis. The high error rate of the platform was taken into account by adjusting the allowed edit distance (i.e. number of mismatches) according to the length of different amplicons (Supplementary Table 2). With solely a sample barcode to go by, filtering of reads was required to identify which gene each read mapped against (See Methods, Supplementary Fig. 4). This process entailed removal of highly erroneous and short reads, after which the most frequent base for each position (and for each allele) could be selected to build a consensus sequence to be matched against the database. The length bias of the ONT platform was evaluated by comparing the amplicon pool length against the sequenced read lengths, and results show that the relative quantity of sequencing reads are skewed towards shorter fragments (Supplementary Fig. 5). The distribution of read lengths (within and between samples) for the filtered 2D reads appeared to be similar to the total read population (Supplementary Fig. 6).

For all samples, all alleles for HLA-A and HLA-B genes were characterized with 100% concordance to external references (Table 2). In addition, 16 (100%) and 15 (93.8%) alleles for HLA-A and HLA-B respectively, were characterized with the highest degree of accuracy (four-field resolution). For HLA-C and HLA-DBQ1 we observed 93.8% concordance to the reference, with one non-concordant call for HLA-C of NA12892 (BC 8) and a missing call for HLA-DBQ1 of NA12882 (BC 4). Given the sequencing depth obtained for the non-concordant call (Table 1), it is likely that errors have been introduced by the polymerase in early stages of the long-range PCR. In the case of HLA-DBQ1 for BC 4, only the DQB1*05:01 allele was identified with 129 reads supporting it while no reads supported the reference DQB1*03:02 allele, suggesting that allele dropout occurred during long range PCR. Findings that are concordant with external references but which seem to defy the typical rules of allele inheritance can also be observed. In the case of family CEPH 1463, neither of the two HLA-C alleles (HLA-C*03:04 / HLA-C*05:01) identified in the father (BC 7) has been passed down to his son (BC 8), who seems to have obtained the HLA-C*04:01:01:06 allele instead (Table 2, Supplementary Fig. 4). Much like the mistyping of HLA-C*04:01 as HLA-C*03:04 for BC 8, it is probable that those alleles are closely related sequentially.
### Table 2. HLA Genotyping results.

Reference genotypes from external sources (12, 13) are underlined and our results are detailed below for each gene and sample. For each gene, the overall concordance relative to the reference, and the proportion of identified alleles with four-field resolution, has been calculated. The results for individuals where no reference is available (gene HLA-DPB1) have not been considered in the calculation for reference concordance. The minimum coverage and edit distance used for each gene is listed in Supplementary Table 2. Cases where an expected allele is missing has been marked (---), and discordant calls have been flagged (■).

<table>
<thead>
<tr>
<th>Sample (Barcode)</th>
<th>Family (Role)</th>
<th>HLA-A</th>
<th>HLA-B</th>
<th>HLA-C</th>
<th>HLA-DQB1</th>
<th>HLA-DRB1</th>
<th>HLA-DPB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA11992(BC5)</td>
<td>CEPH 1362 (father)</td>
<td>A<em>01:01 /A</em>02:01</td>
<td>B<em>08:01 /B</em>35:01</td>
<td>C<em>04:01 / C</em>07:01</td>
<td>DQB1*02:01</td>
<td>DRB1*03:01</td>
<td>DPB1*04:02</td>
</tr>
<tr>
<td>NA11993(BC6)</td>
<td>CEPH 1362 (mother)</td>
<td>A<em>26:01 / A</em>29:02</td>
<td>B<em>44:02 /B</em>44:03</td>
<td>C<em>05:01 / C</em>16:01</td>
<td>DQB1*05:01</td>
<td>DRB1*01:01</td>
<td>DPB1*04:02</td>
</tr>
<tr>
<td>NA10860(BC7)</td>
<td>CEPH 1362 (son)</td>
<td>A<em>02:01 /A</em>29:02</td>
<td>B<em>35:01 /B</em>44:03</td>
<td>C<em>04:01 /C</em>16:01</td>
<td>DQB1*05:01</td>
<td>(Homozygous)</td>
<td>DPB1*04:02</td>
</tr>
<tr>
<td>NA12891(BC5)</td>
<td>CEPH 1463 (grandfather)</td>
<td>A<em>01:01 /A</em>24:02</td>
<td>B<em>07:02 /B</em>08:01</td>
<td>C<em>07:01 / C</em>07:02</td>
<td>DQB1*02:01</td>
<td>DRB1*03:01</td>
<td>DPB1*04:01</td>
</tr>
<tr>
<td>NA12892(BC8)</td>
<td>CEPH 1463 (grandmother)</td>
<td>A<em>02:01 /A</em>11:01</td>
<td>B<em>15:01 /B</em>56:01</td>
<td>C<em>01:02 / C</em>04:01</td>
<td>DQB1*05:01</td>
<td>(Homozygous)</td>
<td>DRB1*04:01</td>
</tr>
<tr>
<td>NA12878(BC2)</td>
<td>CEPH 1463 (mother)</td>
<td>A<em>01:01 /A</em>11:01</td>
<td>B<em>08:01 /B</em>56:01</td>
<td>C<em>01:02 / C</em>07:01</td>
<td>DQB1*02:01</td>
<td>DRB1*01:01</td>
<td>DPB1*04:01</td>
</tr>
<tr>
<td>NA12877(BC7)</td>
<td>CEPH 1463 (father)</td>
<td>A<em>02:01 /A</em>03:01</td>
<td>B<em>15:01 /B</em>54:02</td>
<td>C<em>03:04 / C</em>05:03</td>
<td>DQB1*03:01</td>
<td>(Homozygous)</td>
<td>DRB1*04:03</td>
</tr>
<tr>
<td>NA12882(BC4)</td>
<td>CEPH 1463 (son)</td>
<td>A<em>02:01 /A</em>11:01</td>
<td>B<em>15:01 /B</em>56:01</td>
<td>C<em>01:02 / C</em>04:01</td>
<td>DQB1*03:02</td>
<td>(Homozygous)</td>
<td>DRB1*11:193:02</td>
</tr>
</tbody>
</table>

**Four-field Resolution**
- 100% 93.75% 75.00% 56.25% 14.29% 75.00%

**Reference Concordance**
- 100% 100% 93.75% 93.75% 66.67% 75.00%

---

9
The results for the two longest genes are less conclusive, with five and two calls missing or not matching the reference for HLA-DRB1 and HLA-DQB1 respectively (Table 2). Furthermore, the majority of concordant calls for HLA-DRB1 were resolved with three fields of resolution. For HLA-DPB1 the data shows that most of the alleles could be called, although not validated due to missing reference data for four of the individuals. Given the support from 28 full-length reads per haplotype (Table 1), it is unlikely that haplotype calls for HLA-DPB1 (BC 4) would be incorrect. The limited support of reads spanning the entirety of the longest genes, for HLA-DRB1 gene (11.9 kb) in particular, was not sufficient for accurate calling of variants. The sequencing data shows that most of the reads for amplicons over 10 kb did not yield full-length alignments, resulting in a lack of coverage towards the center of the genes (Supplementary Fig. 3), and thus reduced support for consensus generation and subsequent genotyping.

The error profile of the Oxford Nanopore platform is a combination of random errors and a systematic issue when it comes to calling homopolymers. These errors influence the generation of reference alignments and consensus sequences, and parameters of minimum coverage and edit distance needed to be fine-tuned on a case to case basis (Table 1). Looking closer into full-length alignments (Supplementary Fig. 2), the issue of both stochastic and homopolymer errors is evident. However, through significant sequencing depth and alignment coverage across whole genes an error free consensus sequence be generated. As a result, no mistyping of called alleles was attributed to sequencing errors resulting from indels in homopolymer sequences.

In comparison to short read sequencing platforms, the ONT platform enables a significant reduction in computational resources. This stems from having read lengths that are orders of magnitude longer, despite the higher per base sequencing depth that is needed to compensate for errors. The workflow utilized in this study can be performed using a commodity computer, with average run times ranging from 1 to 12 minutes per sample and locus (Supplementary Table 1). In addition, the peak memory usage is about one gigabyte per locus and the disk usage is less than 200 MB for the analysis, while the raw FASTQ data is ~600 MB. For processing of the four loci with high coverage, only about 5 GB of disk space was required for all samples.

Exploring use of the nanopore technology for amplicon sequencing lead to the conclusion that optimizing pooling of differently sized fragments is of key importance. This is done to minimize the bias towards shorter amplicons to be translocated through the nanopores with greater ease and frequency. It is probable that attaching the barcodes by a ligation-based protocol instead of by long-range PCR would be beneficial in keeping the molar ratio more intact. Despite measures to even out the distribution of reads for
amplicons of different lengths, the sequencing run struggled to generate reads spanning the whole length of amplicons over 10 kb in size. A reason why several alleles were not called for the HLA-DPB1 could also because of allelic dropout due to variations which influence the efficiency of PCR primer binding. The use of multiple primer pairs (i.e. amplicons) is likely to decrease the risk of unsuccessful amplification reactions. Regarding the extensive length of amplicons, a potential solution to the issues for HLA-DQB1 and HLA-DPB1 genes would be to redesign the primers and focus on heterogeneous exonic regions rather than the whole genes. Although this could solve the sequencing issues, the resulting genotypes would not be classified with four-field resolution.

This paper showcases the use of the Oxford Nanopore MinION device for phased HLA genotyping of eight individually barcoded reference samples, resulting in characterization of alleles to the highest degree of accuracy (four field resolution). Focusing on the genes which the ONT platform seems to be well suited to cover (i.e. those below 10 kb in length), the rate of concordance relative to external references was 96.9% (62 out of 64). The haplotypes were phased to the highest degree of accuracy in 83.9% (52 out of 62) of these cases, with remaining haplotypes being characterized with three-field resolution. These results go well beyond the accuracy of frequently used clinical tests based on serological testing (microlymphocytotoxicity assays), and they support the idea that nanopore sequencing could be a viable solution in clinical settings where speed and accuracy of HLA typing is of utmost importance.

ACKNOWLEDGEMENTS

The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure (NGI) in Stockholm and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure. This work was supported by the Olle Engkvist Foundation [2015/347] and Stockholms Läns Landsting (LS2016-0764).
REFERENCES


Supplementary Figure 1. Family pedigree trees and reference HLA genotypes with 2 fields resolution. The upper family tree is part of the 1463 CEPH for UTAH pedigree for samples NA12877, NA12878, NA12891, NA12892 and NA12882, the lower one is part of the 1362 CEPH for UTAH pedigree for NA11992, NA11993 and NA10860. Square nodes are male and circular nodes are female. The reference HLA genotypes are also depicted; left sides are the parental, right sides are the maternal types derived from the parental genotypes. An anomaly to the expected transfer of alleles from parents to offspring can be observed for NA12882 the inherited type for HLA-C should be HLA-C*03:04 or HLA-C*05:01 instead of the reference HLA-C*04:01 (confirmed by our results).
Supplementary Figure 2. Coverage over HLA-A (Class-I) and HLA-DQB1 (Class-II) alleles. (A) For HLA-A, the reads used for consensus generation are spanning through the whole allele, the homopolymer errors appear as dark vertical bars. Some of the untrimmed barcodes are colored mismatches at the beginning of the alignment. The coverage is over 320 reads for most of the positions, overall average coverage for this allele (HLA*02:01:01:01) is 295 reads. (B) Less deep is the coverage for the HLA-DQB1 locus (average 33), but still there are many reads spanning through the whole 7 kb long allele.
Supplementary Figure 3. HLA-DRB1 alignments for allele HLA-DRB1*03:01:01:01 showing a successful (top) and an unsuccessful (bottom) genotyping attempt as visualized with IGV (http://software.broadinstitute.org/software/igv/). Reads colored in shades of red and blue correspond to forward and reverse orientation respectively, and the lighter shades are reads which feature alternative mapping positions. The majority of reads can be found at the ends of the amplicons, with only a few crucial reads spanning through the whole gene. Even in the top case the coverage is barely enough to phase through the locus, and for the bottom case the incorrect genotype was reported.
Supplementary Figure 4. Alignment of reads for HLA-DRB1 of NA12878 (BC 1) after (top) and before (bottom) filtering of reads based on mapping quality and length.
Supplementary Figure 5. Fragment length distribution from Fragment Analyzer (blue) and read length (red) for barcode 1 illustrating the length bias both for the nanopore sequencing and the electrophoresis method. Other samples in the study showed a similar pattern (data not shown).
Supplementary Figure 6. Read-length distribution for all the reads (blue) and for 2D reads only (yellow). The peaks are approximately at the same lengths for both all the reads and for the filtered 2D reads only.
**Supplementary Table 1.** Resource utilization. The workflow can be run on a commodity computer; these benchmarks were run on a Dell XPS 15 9550 laptop with 16 GB memory, 512 GB hard disk (no SSD) and Intel Core i7-6700HQ CPU @ 2.60 GH using 8 threads. The alignment (bwa-mem) step is the most CPU demanding but that is parallelized to use all the available cores. The IO demand and the memory consumption is moderate. Values are calculated for single samples, and averaged over all the eight barcodes. Run parameters used: -a -k 70 -W100 -r10 -A1 -B1 -O1 -E1 -L0.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Wall clock (sec)</th>
<th>Peak memory (MB)</th>
<th>Data read (MB)</th>
<th>Data written (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-A</td>
<td>392 ± 99</td>
<td>981 ± 144</td>
<td>1446 ± 129</td>
<td>172 ± 90</td>
</tr>
<tr>
<td>HLA-B</td>
<td>279 ± 87</td>
<td>962 ± 150</td>
<td>1471 ± 121</td>
<td>152 ± 94</td>
</tr>
<tr>
<td>HLA-C</td>
<td>751 ± 244</td>
<td>980 ± 143</td>
<td>2098 ± 146</td>
<td>182 ± 104</td>
</tr>
<tr>
<td>HLA-DQB1</td>
<td>44 ± 30</td>
<td>986 ± 189</td>
<td>257 ± 116</td>
<td>112 ± 84</td>
</tr>
</tbody>
</table>
**Supplementary Table 2.** Genotype analysis parameters. The minimum coverage (MC) and edit distance was adjusted for each gene. This was done systematically to reduce the ambiguity of each call since the default setting for most cases generated long lists of putative alleles. The parameters were adjusted until the pipeline only generated the two alleles with the most support. For suspected homozygous samples, the minimum coverage parameter was decreased repeatedly to verify the call. For the final genotyping results outlined in Table 1, several of the individuals required custom parameter settings for an unambiguous result to be obtained: HLA-A for BC 3 (MC = 30); HLA-C for BC 3 (MC = 15); HLA-C for BC 8 (MC = 20) and HLA-DPB1 BC 2 (MC = 2).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>HLA-A</th>
<th>HLA-B</th>
<th>HLA-C</th>
<th>HLA-DQB1</th>
<th>HLA-DRB1</th>
<th>HLA-DPB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplicon Length</td>
<td>3,398 bp</td>
<td>4,296 bp</td>
<td>4,440 bp</td>
<td>7,118 bp</td>
<td>11,899 bp</td>
<td>13,605 bp</td>
</tr>
<tr>
<td>Minimum Coverage</td>
<td>40</td>
<td>10</td>
<td>30</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Edit Distance</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>300</td>
<td>600</td>
<td>600</td>
</tr>
</tbody>
</table>