Bioinformatics Strategies, Challenges, and Opportunities for Next Generation Sequencing-Based HLA Genotyping

Steffen Klasberg Vineeth Surendranath Vinzenz Lange Gerhard Schöfl

DKMS Life Science Lab, Dresden, Germany

Keywords

Next generation sequencing · Human leukocyte antigen · Genotyping

Abstract

The advent of next generation sequencing (NGS) has altered the face of genotyping the human leukocyte antigen (HLA) system in clinical, stem cell donor registry, and research contexts. NGS has led to a dramatically increased sequencing throughput at high accuracy, while being more time and cost efficient than precursor technologies. This has led to a broader and deeper profiling of the key genes in the human immunogenetic make-up. The rapid evolution of sequencing technologies is evidenced by the development of varied short-read sequencing platforms with differing read lengths and sequencing capacities to long-read sequencing platforms capable of profiling full genes without fragmentation. Concomitantly, there has been development of a diverse set of computational analyses and software tools developed to deal with the various strengths and limitations of the sequencing data generated by the different sequencing platforms. This review surveys the different modalities involved in generating NGS HLA profiling sequence data. It systematically describes various computational approaches that have been developed to achieve HLA genotyping to different degrees of resolution. At each stage, this review enumerates the drawbacks and advantages of each of the platforms and analysis approaches, thus providing a comprehensive picture of the current state of HLA genotyping technologies.

© 2019 The Author(s)
Published by S. Karger AG, Basel

Introduction

The exceptionally polymorphic human leukocyte antigen (HLA) system contains numerous genes that encode key components of the human adaptive immune system [1]. Allelic concordance of these genes between patients and donors is essential for transplantation outcome in allogeneic haematopoietic stem cell transplantation (HSCT) and to a lesser extent in solid organ transplantation [2–4]. Thus, genotyping the class I genes HLA-A, -B and -C, and the class II genes HLA-DRB1, -DQB1, and -DPB1 is routinely performed in clinics and by stem cell donor registries.

Currently, more than 21,000 named alleles are reported in the central public reference sequence repository for HLA allele variation for these 6 genes alone (IPD-IMGT/HLA release 3.36.0 [5]). As such, identifying the HLA genotype of a sample is a non-trivial task. Historically, serology-based typing has been replaced by several DNA-based typing methods of varying resolution and accuracy, such as sequence-specific oligonucleotide probes or sequence-specific primer PCR, and finally, sequence-based typing (SBT) based on Sanger sequencing. Whilst these methods remain in use to different degrees, they are being increasingly replaced by next generation sequencing (NGS) methods.

The hallmark of the various NGS technologies is massively parallel sequencing of single DNA molecules (clonal sequencing) [6]. This allows heavy multiplexing of genes and samples, and thereby a dramatic reduction in cost per sample. In addition, due to the clonal nature of NGS, the inherent phasing ambiguities that ail Sanger se-

quencing are avoided (a single signal trace for both alleles delivered by Sanger sequencing does not allow putting heterozygous positions in phase) [7, 8]. NGS technologies have brought many advancements to the field of HLA genotyping within relatively few years [9–12]. These advances have led to a demonstrated 3.5-fold reduction in genotyping error rates compared to Sanger SBT [12]. The scalability of NGS has allowed moving beyond HLA by developing reliable high-throughput genotyping algorithms for other immunogenetically important loci such as MICA and MICB, blood groups [13], or the KIR genes [14]. By now, NGS technologies are firmly entrenched as the preferred method for high-resolution HLA typing in high-throughput immunogenetic laboratories for, e.g., stem cell donor registries [9, 15–17], but also in many basic research contexts, and they are making inroads into the clinical and diagnostic labs [18–21].

Concomitantly, NGS has injected a heavy dose of bioinformatics into the field of HLA genotyping. The amounts of data generated by NGS platforms are typically magnitudes larger compared to Sanger sequencing. The high volume and complex nature of NGS datasets necessitate non-trivial IT infrastructure and expert bioinformatics skills [6]. Although some commercially supported “plug-n-play” software solutions exist for NGS data, they typically cover standard use cases, utilise data from well-established NGS platforms, and generally cater to routine clinical diagnostic usage needs. Dealing with data from emerging NGS platforms like Pacific Biosciences’ (PacBio) or Oxford Nanopore’s (ONT) long-read sequencers in the context of HLA genotyping remains an area of active exploratory research with no accredited support beyond academic proof-of-concept analysis solutions.

This review seeks to discuss the various bioinformatics strategies and challenges associated with NGS-based HLA genotyping in general, as well as the particular benefits and issues connected to the different NGS platforms used in conjunction with HLA typing. Finally, we discuss the likely impact of emerging sequencing technologies on the field of HLA bioinformatics.

Results

Sample Preparation

At the base of any NGS HLA typing experiment stands the preparation and sequencing of the template DNA. The nature of the subsequent bioinformatic approaches and the achievable level of typing resolution are dictated by the sample preparation methods and sequencing platforms used. Among sample preparation methods used for HLA typing we can distinguish between HLA-targeted approaches (e.g., PCR-based target amplification, hy-

bridisation-based capture techniques) and HLA-untargeted approaches (e.g., whole-genome sequencing [WGS], whole-exome sequencing [WES], transcriptome sequencing).

The most commonly used technique for target enrichment for HLA genotyping is PCR (Fig. 1A) [22]. Many studies and genotyping efforts have used short-range PCR on exons, which are then sequenced in their entirety on a short-read sequencing platform such as the Illumina platforms or the Ion Torrent [7, 23]. Long-range PCR has been used to generate full gene constructs which subsequently can be fragmented for sequencing on a short-read sequencing platform [24] or sequenced in whole on a long-read sequencing platform such as provided by ONT or PacBio [25–28]. Samples may be multiplexed by adding barcode sequences to the PCR products, thereby enabling the cost-efficient sequencing of hundreds and even thousands of samples in a single sequencing run [7, 23].

A second approach to enrich specifically for HLA sequences are hybridisation-based target sequence capture techniques with HLA-specific probes. Successful efforts have also been made in enriching for the entire MHC region and deriving HLA genotypes from these data. Probes can either be in solid phase, as with SCE (array-based sequence capture and enrichment) [29], or in solution as with RSE (region specific extraction), which uses pull-down of hybridisation oligomers attached to magnetic beads [30]. There have been recent efforts to use an SSO-like technique in combination with NGS, saturated tiling capture sequencing (STC-Seq), for genotyping of genes using dense chip-based probes that capture the coding regions of HLA [31].

Capture-based enrichment strategies can also be applied to exomes (WES) [32–34]. These methods yield sequences of all exons in the human genome only limited by the availability of reference annotations. It has been demonstrated that WES data can yield reliable HLA genotyping [10, 35, 36]. The increased feasibility of sequencing genomes (WGS) [37, 38] using short-read and long-read platforms, or a combination of both, has led to availability of whole-genome data from which HLA genotypes can be derived [39–41]. However, due to the usually only moderate level of sequence coverage, HLA genotyping from WGS does not achieve the same accuracy as HLA-targeted approaches. Whole-transcriptome sequencing (RNA-Seq) yields data corresponding to the entirety of the human transcriptome from which HLA genotypes can be inferred [42, 43]. To be useful for HLA genotyping, data derived from these HLA-untargeted sequencing strategies require additional algorithmic filtering steps to extract the HLA-specific reads (Fig. 1B).

Given that a majority of HLA genotyping is done using targeted sequencing by way of PCR, some of the issues ailing PCR-based genotyping approaches are briefly con-

Fig. 1. A generalised workflow for NGS-based HLA genotyping. **A** Sequence data are generated by HLA-specific target enrichment, mostly short- or long-range PCR. Long PCR products are fragmented for sequencing by short-read technologies. **B** Reads specific to the HLA genes are retrieved by mapping. **C** Three examples of HLA genotyping algorithms: read mapping, as implemented by PHLAT or OptiType; consensus-sequence-based typing, as implemented by HLA-VBSeq; and graph-based typing, as implemented by HLA*PRG-LA or Kourami.

sidered here. PCR has been known since decades to be affected by amplification bias as a consequence of differences in primer binding or differences in length or GC content between a pair of alleles [44]. This amplification bias, if not corrected, can result in erroneous genotyping [7, 45–47]. An extreme variant of amplification bias is the complete failure of one allele to amplify resulting in allelic dropout (ADO). Low DNA quality and poor sample integrity have been shown to increase ADO rates [17, 48]. Certain DNA structures known as G-quadruplexes (G4 tracts) have also been shown to result in ADO [49], though there has been no investigation of G4 structures in the HLA system. Another pitfall of amplicon-based sequencing is the PCR-mediated formation of cross-over products. Here, reads incorporate portions of both alleles present in the sample into a single sequence. Cross-over formation tends to depend on PCR conditions (e.g., number of cycles) and initial template concentration [50]. Cross-over products can occur in relatively high fractions and may confound the ability to call the correct alleles unless filtered out. A final class of PCR artefacts potentially leading to false genotyping results is DNA replication slippage. Replication slippage mutations occur frequently within low-complexity regions such as short tandem repeats (STR) and longer homopolymeric stretches during the elongation phase of the PCR and result in the random addition or deletion of one or more repeat units [51]. PCR-induced STR length mutation has been shown to

lead to a loss of heterozygotic genotyping [52] and to confound the discrimination of allelic differences in the HLA region in particular cases [53]. In a well optimised PCR setup, however, those artefacts will only affect particular, mostly intronic, regions or a low fraction of samples [9]. Therefore, proof of principle studies may well ignore these effects. However, when the highest accuracy is required as for clinical genotyping or reference sequence submission, software needs to be aware of these PCR limitations and compensate for them or at least point the analyst to the troublesome region.

Sequencing Platforms

Sequencing platforms shape the requirements on downstream analysis via achievable sequencing depths, read lengths and platform-dependent error profiles. Fundamentally, NGS platforms may be classified based on the read lengths they are capable of generating. Short-read sequencers generate reads with lengths in the order of a few hundred bases. Long-read sequencers generate reads with lengths of more than ten thousand bases.

In the short-read segment, the available platforms differ in the directionality of sequencing possible for a given target region and the achievable read lengths. The various Illumina platforms sequence by optical readout of the incorporation of fluorescent nucleotides coupled to a reversible terminator by a DNA polymerase [54]. They are capable of paired-end sequencing, making possible a bi-

directional read through for a target region and yield maximum read lengths of 2×300 bases (MiSeq), 2×250 bases (HiSeq, NovaSeq), or 2×150 bases (iSeq, MiniSeq, NextSeq). The Illumina MiSeq and HiSeq instruments have been demonstrated to be capable of HLA genotyping in a high throughput fashion by multiplexing thousands of samples [7, 9, 17, 21, 23, 55, 56]. The Ion Torrent platform sequences by measuring the hydrogen ion release subsequent to nucleotide incorporations by the DNA polymerase [57] and yields reads of 400–600 bases. Read through is, however, only possible in the 5' to 3' direction. The Ion Torrent has been shown to be capable of HLA genotyping using reads with an average length of ~275 bases across all six classical HLA loci [58–60]. In contrast to Illumina, Ion Torrent suffers from an increased homopolymer-associated indel error rate [61].

The long-read segment has seen rapid developments over the last few years, making now feasible the direct sequencing of target regions of 10–15 kb allowing for the coverage of complete HLA genes. For HLA typing purposes, two major platforms are beginning to be utilised. At the time of writing, however, both platforms are still disadvantaged by a high per read error rate in the order of 10–14%. PacBio's RSII and Sequel platforms, which work by real time monitoring of the incorporation of fluorescent nucleotides [62, 63], have demonstrably been used for HLA genotyping [25, 26, 28, 60]. ONT devices, set apart from all other platforms in that they do not need a DNA polymerase for sequencing, work by sensing electrical signals as DNA molecules pass through a protein pore [64, 65]. The ONT MinION has been successfully used to sequence and characterise HLA class I alleles with rapid turnaround times [27, 66, 67].

Read Attribution

Once raw reads have been obtained from a sequencing experiment, HLA genotyping pipelines will follow either a two-step or a single-step read attribution approach.

If the data derive from HLA-untargeted sequencing libraries or if amplicons/captured sequences from different genomic locations have been pooled for sequencing, reads first need to be correctly attributed to a specific genomic location. Only in a second step will reads be matched to a collection of known reference alleles for that genomic location. For HLA-targeted sequences, typically only the second step will be necessary. Both of these steps involve well known bioinformatics algorithms, the success of which crucially depends on factors such as read coverage, read quality, and reference quality.

Three basic approaches for read attribution are (i) sequence alignment, (ii) de novo assembly, and (iii) read mapping. With sequence alignment, reads that originate from HLA-targeted sequencing of short genomic stretches may be matched directly to reference sequences that

are similar in length and expected to be highly homologous, using common local (e.g., Smith-Waterman) or global (e.g., Needleman-Wunsch) alignment algorithms. Such alignment algorithms may be appropriate, for instance, for matching amplified exon sequences against a known set of reference exons [7]. De novo assembly refers to the algorithmic reconstruction of the source sequences from read fragments without recourse to a reference. The resulting consensus sequences, however, again need to be matched to a reference by alignment or mapping. In contrast to alignment, read mapping is appropriate whenever the individual sequence reads are (much) shorter than the references. Read mapping typically involves two steps: a fast initial heuristic is used for identifying one or more top mapping locations on the reference for each read, followed by local alignment optimised for accuracy.

The majority of HLA genotyping algorithms crucially rely, at least at some point, on mapping reads to one or more reference sequences. In the context of the HLA system, there are specific challenges associated with read mapping:

First, if sequence data need to be assigned to their region of origin, read cross-mapping from or to homologous sequences may confound correct read attribution. Different HLA genes sometimes share conserved, highly similar regions, and several class I and class II pseudogenes exist. Mappers use heuristics to estimate “mapping quality” and “uniqueness of mapping.” Arbitrarily excluding or including non-uniquely mapping reads risks biased read coverage at conserved sequence stretches. Some typing algorithms that rely on extracting HLA reads from untargeted WGS, WES, or RNA-Seq data such as seq2HLA [68], HLAforest [69], or PHLAT [70] start with reference-based read mapping followed by various quality filtering steps to discard spurious reads with low alignment scores. Some algorithms relegate the problem to downstream analysis, e.g. HLA-VBSeq [71] extracts all reads mapped to any HLA locus in a human reference genome as well as all unmapped reads and realigns the extracted reads to a set of reference sequences in a second step. Others attempt to safeguard against read misattribution by scoring each read against each candidate HLA gene individually, attributing a read to the highest-scoring HLA gene [35], or by restricting analyses to reads uniquely aligned to the target gene [72].

One of the most popular tools for the common scenario of mapping genomic short-read data is the Burrows-Wheeler aligner (BWA) with the MEM algorithm [73]. BWA-MEM can be used for high-coverage amplicon data from targeted sequencing as well as for WGS data and is especially suited for NGS reads longer than ~70 bp. Other popular DNA short-read mappers are Bowtie2 [74] and the commercial aligner NovoAlign (<http://www.novocraft.com/>). The latter is reported to be

the most accurate but also a slow option. For RNA-seq data, the most popular splice-aware mappers are STAR [75] and HISAT2 [76]. Long-read data, as produced by ONTs' nanopore sequencing or by *PacBio's* SMRT sequencing require other mapping solutions to ensure resource- and time-efficient mapping and to cope with their high single-read error rates. Minimap2 [77] is a new specialised tool for very fast processing of long reads up to several kilobases. An alternative mapper optimised for high mapping sensitivity in the face of error-prone reads that can be used for both *PacBio* and *ONT* data, is Graph-Map [78].

The parameterisation of these tools is, like the choice of the tool itself, often a compromise between speed and accuracy. The first step of mapping, the determination of the correct genomic location, is crucial for all subsequent analyses and needs to be parameterised with care. The majority of mapping tools are developed for mapping short reads onto long, often genome-sized references. Mapping data from targeted sequencing or the remapping of reads to specific HLA genes should therefore be carried out using more stringent criteria to obtain the best matching position of each read. More reliable mapping positions result in more accurate local alignments of reads, especially in repetitive regions or in the large HLA sequence space. Another consideration is the handling of duplicated reads. Because duplications are a frequent sequencing artefact of WGS experiments, it is important and common practice to remove or mark them prior to downstream analyses. However, this approach leads to highly biased results in targeted sequencing experiments in favour of spurious reads.

Allele Assignment

The final step of HLA genotyping is the inference of the HLA allele combination in the reference allele repository that best explains a set of sequence reads.

A peculiarity of the HLA typing field is, that, what can be validly considered an adequate allelic assignment is often a matter of the stated goal of an HLA genotyping approach. Here, as with the nomenclature of HLA alleles (Box 1), the hierarchy of functional importance of HLA polymorphisms has shaped many HLA genotyping algorithms. For applications aiming to be used in clinical contexts, resolving the antigen recognition domains (ARD) is often considered sufficient. Consequently, many proposed algorithms report at G-group or at two-field resolution even where sequence data covering the whole gene as well as full-length reference data are available (online suppl. Table 1; for all online suppl. material, see www.karger.com/doi/10.1159/000502487). For instance, although OptiType [79] or HLA*PRG [80] are set up to work with WGS data, they will only consider reads that map to the core exons for genotyping. Even for algo-

Box 1. HLA nomenclature

History and rationale of the HLA nomenclature

HLA genotyping has evolved in the context of transplantation medicine from serology-based techniques to high resolution NGS-based technologies. This context shaped the current nomenclature for describing HLA alleles, the objectives of modern sequence-based typing methods, the current state of the reference database for HLA allele sequences, and, consequently, contributes to many of the challenges for HLA genotyping using next-generation sequencing.

The HLA nomenclature is standardised and regulated by the World Health Organisation (WHO) Nomenclature Committee for Factors of the HLA System. Each allele designation starts with a gene name followed by an asterisk and at least two colon-separated numeric fields (e.g., HLA-A*02:02:01:02). The first numeric field differentiates allele families that often correspond to serologically defined antigens. The second field distinguishes alleles within an allele family that differ in one or more non-synonymous base substitutions in any of the exons. The third field differentiates alleles that share the same protein sequence but differ in synonymous base substitutions. The fourth field denotes differences in the intronic regions and the 5' and 3' untranslated regions (UTRs). Non-expressed or null-alleles are denoted by the suffix "N," e.g., HLA-C*04:09N. Additional suffixes exist to describe alleles with low expression levels (L), alleles which express soluble proteins (S), or alleles with questionable expression status (Q).

rithms that are designed to allow for full-resolution typing accuracy, benchmarking results are often only reported at two-field resolution, leaving open the question how well they perform in practice at the higher levels of resolution (compare, e.g. xHLA [41]). Still, while many extant typing algorithms by design fall short of this ambition, there is an increasing number of attempts to arrive at fully phased, whole-gene consensus sequences with unambiguous four field typings.

Irrespective of the level of genotyping resolution that is aimed for, presently, we can identify three fundamental approaches to HLA allele assignment: (i) the "traditional" read matching to the reference allele repository, often followed by varied scoring and refinement schemes to hone typing accuracy and precision (Fig. 1C, upper panel). (ii) de novo consensus sequence generation from sequence reads followed by reference allele matching (Fig. 1C, middle panel). (iii) Read mapping to graph-based structures that succinctly encode nucleotide and structural variation in a single rich reference structure (e.g., the genome graphs [81]) (Fig. 1C, lower panel). Each of these approaches has distinct advantages and challenges.

Currently, a large majority of HLA genotyping tools employ some form of matching individual sequence reads to the reference allele sequences followed by some scoring strategy to account for mapping uncertainty and signal noise. Here an important challenge lies in the exception-

ally high level of sequence diversity in the HLA region. This generates a large search space composed of many highly similar potential candidate sequences. Such a search space is similar in nature to genomic low-complexity regions, which mappers are generally poorly equipped to deal with. Thus, for any given short read, there are potentially many equally good matches and most off-the-shelf mappers are optimised for speed and not for exhaustively reporting all possible matches. To work around such problems, several existing HLA-typing approaches rather arbitrarily reduce the search space, e.g., OptiType by ignoring putatively rare alleles (i.e., those with no reported allele frequency on AlleleFrequency.net) [79, 82] or HLA-VBSeq by ignoring all partially known alleles in IPD-IMGT/HLA [71]. More sophisticated approaches that guarantee that all equally good matches in the database are considered are implemented in xHLA [41] and in HLA*PRG [80, 83]. The former relies on a precomputed protein-level multiple sequence alignment of all known HLA alleles to expand an initial read alignment to all equivalent sequence segments that share the same protein sequence. HLA*PRG represents all known allele-level variation of the six classical HLA genes in population reference graphs (PRGs) [84] and allows a precise quantification of the mapping ambiguities for each read (discussed below).

A second challenge, which starts to matter if a genotyping algorithm aims for higher than two-field resolution, is the fragmentary nature of the reference sequence data contained in the IPD-IMGT/HLA database with complete coverage for the core exons 2 and 3 (or exon 2 alone in case of class II molecules) but variable coverage of the other exons and non-coding parts of the alleles (Box 2; online suppl. Fig. 1).

Here, the challenge becomes to define a best matching allele. Consider, for instance, cases where the correct target allele is represented only by its exon sequences. In a whole-gene sequencing approach, mappers may nonetheless assign reads derived from the introns to other fully sequenced alleles in the database if the unknown introns of the target alleles are sufficiently similar to the known introns. Depending on the proportions of known to unknown sequence and the number of mismatches, genotyping algorithms that do not take corrective actions may favour completely known reference alleles over incomplete reference alleles. Or consider a situation where two candidate alleles are defined in different segments of the gene. With regard to one allele, we find a mismatch in exon 4, with regard to the other allele there is a mismatch in exon 5, but neither allele has been defined for both exons. In such a case, there is simply no possibility for an unequivocal attribution of the reads to one or the other allele.

Box 2. The IPD-IMGT/HLA reference sequence repository

The Reference Sequence Repository

The centrally curated repository for the sequence data of the hyperpolymorphic genes of the HLA system is provided by the IPD-IMGT/HLA database [5]. Every 3 months, it releases a fresh snapshot of all publicly available sequences of the HLA system, officially named by the WHO Nomenclature Committee for Factors of the HLA System. As such it provides the canonical reference sequences against which HLA genotyping is performed. For historical reasons (see Box 1) the IPD-IMGT/HLA database is populated to a large extent by alleles where only certain gene features have been characterised. For many entries in the database, only the antigen recognition domains (ARD) have been characterised, which are encoded by exons 2 and 3 for the class I genes and exon 2 for class II genes.

To deal with this situation, different algorithmic approaches with varying degrees of sophistication have been explored: Many tools dodge the problem completely by focusing on the ARD exons alone. Some tools opt for simple elimination or imputation approaches, for instance, by ignoring all partially known alleles [71] or by replacing missing non-coding sequences with the most similar genomic reference sequences [35]. These approaches clearly lead to a reduction in accuracy and resolution that might not be acceptable or desirable, especially if reads covering more of the gene than the ARD exons are available. Hence, a more sophisticated way to deal with some of the ambiguity introduced by the incomplete reference space is to employ some hierarchical scoring strategy. By penalising mismatches in ARD exons higher than in other exons and by penalising mismatches in other exons higher than in non-coding sequences, the best possible full resolution alleles may be recursively identified. Such strategies are employed, for instance, by xHLA [41] and the commercial software NGSengine (GenDx).

The “consensus approach” to first assemble reads into contigs, which subsequently are matched against the reference alleles, is used by only few HLA genotyping tools such as ATHLATES or HLAreporter [72, 85]. It may be a useful approach for WGS or WES data, and it has the advantage of not initially relying on reference data, hence forging reference allele bias, i.e. the tendency of mappers to underreport data whose underlying DNA does not match a reference allele. On the other hand, the extremely high similarity between many if not most HLA alleles from the same locus, makes it difficult for assemblers to construct contigs that sufficiently accurately represent the underlying alleles. Moreover, by assembling we remove spurious and true variation in the underlying reads by applying a consensus calling algorithm. This will mask ambiguities, thus potentially insinuating higher than

warranted certainty about allelic matches or mismatches, unless “quality scores” for each consensus base are generated and considered at allele assignment.

One of the most exciting developments is currently taking place in the area of graph-based genotyping. The pervasive use of NGS over the last decade has generated a treasure trove of genetic diversity data that has served to highlight the limitations of organising genomic information in a linear reference model. The linear reference model, which encodes genotype information as a sequence of consecutive nucleotides, is prone to reference allele bias, where efforts to characterise a given sequencing read are limited by the data available. This model also is an inefficient way to capture diversity as each new allele is represented as an independent entity with additional metadata required to represent relations across alleles.

An alternative to the linear model of representing sequence data emerged by imagining a set of genotypes as a graph, a network with nodes denoting nucleotides and edges connecting each nucleotide to all its possible predecessor and successor nucleotides [81]. Each genotype variant may then be represented as a walk or path through the graph, with relationships between the variants captured by deviations across paths. Different graph representation concepts have been explored, such as directed graphs, bidirected graphs, and biedged graphs depending on the directionality of connections between the nodes in a graph. These representations allow for an efficient encoding of genomic nucleotide diversity as well as a full description of structural genomic events such as insertions, deletions, inversions, and repeat structures. Graph representations also allow the efficient and elegant summarisation of pangenomes, where a set of related reference sequences are analysed together as a single reference without loss of information. Efforts to explore graph representations for HLA genotyping are still evolving, with two successful implementation attempts to date. HLA*PRG [80, 83] uses the idea of the PRG, which is a probabilistic extension of the idea of a pangenome [80]. Kourami takes a different approach by building a graph of 3-mers, instead of single nucleotides, and constructing an assembly from the graph [86]. Whilst extremely promising and necessary as volumes of genotyping data continue to be generated, graph-based genotyping is still in its infancy. Active areas of research are mitigating the computational demands imposed by very large graph data structures. Efforts are also being put into solutions for deriving the positional information needed to describe genotype variants as this information gets lost in standard graph representations [81].

Irrespective of the allele assignment approach used, there exist a number of sources for ambiguity and potential for mistyping. Here, one of the major promises of using NGS-based technologies is that the phasing ambigu-

ity inherent to Sanger sequencing is eliminated. Sanger capillary sequencing can produce long contiguous reads but mixes the signals from the two chromosomal strands. NGS reads are clonal in nature and adequate algorithms can be used, in principle, to sort even short reads of less than 100 bp into fully phased haplotypes. Haplotypic linkage can, however, only be reliably reconstructed if two heterozygous variants that need to be phased are covered by reads or read pairs. In our experience, once polymorphisms are separated by a long homozygous stretch of more than about 1 kb, resolution of phase between the two chromosomes tends to fail using short-read technologies as read pairs no longer reliably span this distance. Unfortunately, especially in the long HLA Class II genes, genotypes with long homomorphic stretches are far from uncommon, leading to inherently ambiguous genotyping results. Another challenge for typing is the determination of homozygosity versus heterozygosity at a locus. For instance, various sequencing artefacts may introduce spurious signatures of a second allele in homozygous samples. Since truly homozygous HLA loci are relatively infrequent and read artefacts often do not match well against the reference sequences, this is less of a problem in practice. More insidious are ADOs, which may remain completely “silent” by generating technically valid but erroneous homozygous genotype. Finally, there are also differences in the level of “difficulty” that different HLA genes present for HLA typing. Thus, HLA class II genes are generally perceived to be more difficult to type than HLA class I genes. Tools such as OptiType [79] do not implement class II typing in the first place, and benchmarking results from HLA typing tools in general exhibit lower levels of typing accuracy for class II genes than for class I genes.

Indel Issues

Besides nucleotide substitutions, the HLA allele space harbours insertion and deletion (indel) variants, in particular in intronic regions. Indels may present major difficulties that only more recent tools may be able to cope with, starting from indel-aware mappers to genotyping algorithms. For instance, insertions in a read cause issues for graph-based genotyping approaches if they are not already part of the reference graph. In order to be scored correctly the reference graph structure requires a computationally costly modification by addition of new nodes, as it is possible with Kourami. This becomes especially important if indels in a sample identify a true variant not yet described in the HLA reference database used for graph construction.

As discussed in the section Sequencing Platforms, indels can be common artefacts introduced during sequencing. The fraction of indel artefacts on the various popular Illumina short-read platforms is considered to be

relatively low, but it is substantially higher on the Ion Torrent and even more so on present day long-read platforms. All technologies, however, are somewhat prone to artefacts in homopolymeric regions, long stretches of single nucleotide repeats. Depending on the technology, once a homopolymer stretch reaches a certain length, the exact number of nucleotides at such positions may become impossible to ascertain. Unfortunately, there are HLA alleles which differ only in homopolymer length. Examples are the frequent A*03:01:01:01 allele and the rare A*03:21N allele, which differ only by a 7C versus 8C stretch at the beginning of exon 4, or B*51:01:01:01 and B*51:11N, which differ by a 6C versus 7C stretch. These cases are not only difficult to resolve with long-read sequences, but the former has additionally been shown to suffer from read cross-mapping (section 2.3) from a highly homologous HLA-H pseudogene sequence stretch harbouring 8 Cs, potentially leading to systematic mistyping of A*03:01:01:01 even on Illumina platforms [87]. Also, on a long-read sequencing platform such as ONT, currently sometimes even short homopolymers in low complexity motifs as for instance GGGGCCGG (A*68:01:02:02) versus GGGCCCCG (A*68:142N) may not be reliably differentiated. The length of homopolymers can reach up to 30 nucleotides in classical HLA genes (HLA-DRB1*15:02:01:02) and up to 45 in other clinically relevant genes in the MHC (e.g., MICB).

Limitations of the various sequencing platforms are not the only source of spurious length variation in low-complexity regions. Both homopolymer tracts and simple repetitive sequences such as microsatellites, or STR, are prone to insertions or deletions of one or more repeat units through a process known as DNA replication slippage during PCR amplification [88]. To an extent, the original template length can be estimated from simple noise models such as highest peak analysis, i.e. the true length of the STR should correspond to the mode of the read length distribution at an STR locus. However, as slippage rates increase with the length of the STR, compounding the effects of sequencer-induced indel errors, such simple models tend to break down [53, 89].

On top of these difficulties, some STR tracts may even be longer than the read length achievable with short-read sequencers. If no read spans the entire STR, there is no way for mappers or aligners to precisely determine the size of the STR during the mapping step. STRs of considerable length and diversity exist, for instance, in intron 2 of HLA-DPB1 as (AAGG)_(4–17) tetranucleotide repeats [90] or in intron 2 of various DRB genes as (GT)_(7–27)(GA)_(5–30) dinucleotide patterns [53]. Fortunately, such long homopolymer tracts and STRs are located mostly in the non-coding parts of a gene, and knowledge of the exact tract length may not be crucial to clinical applications. Consequently, genotyping tools of-

ten exclude or mask such error-prone regions (in non-coding areas) from HLA typing [53].

Novel Alleles

Most of the previously introduced HLA genotyping tools do not report a nucleotide sequence but yield the best matching alleles in the IPD-IMGT/HLA database. As such, the genotyping accuracy and benefit for clinical applications relies highly on the correctness and completeness of the database. Ideally, a genotyping algorithm would also be able to report differences to known alleles and provide a sequence which can be submitted to the IPD to be included in the database. Due to the heterozygosity of the HLA locus, it is not possible to just draw a consensus sequence from a mapping of the reads at a specific locus. However, short reads either need to be assigned to specific target alleles or put in phase by other approaches.

One of the few algorithms that is able to output novel alleles, alongside the best matching known alleles, from WGS data is Kourami [86]. Kourami uses a graph-guided allele assignment of the precomputed MSA of all known alleles, similar to HLA*PRG [80], but allows the graph to be extended by new positions in an MSA derived from the short-read mapping. Another approach for the creation of highly accurate allele sequences out of targeted sequencing is by combining short-read and long-read data. The long reads are used for providing a phased guide alongside which the short reads can be mapped to receive an accurate sequence signal. This approach is, for example, utilised by the DR2S software [91].

Another obstacle for adding novel alleles to the IPD-IMGT/HLA database is the submission process itself. The novel sequence needs to be annotated and submitted to a nucleotide archive database like the European Nucleotide Archive (Embl-ENA). After acceptance at an archive, it can be submitted to IPD together with meta information. The metadata comprise information about the complete HLA genotype of the individual as well as annotations and differences to the closest known allele. This tedious and error-prone process can be automated to a great extent by the submission tool TypeLoader [92].

Reporting NGS-Based HLA Genotyping Results

In many cases, HLA genotyping does not result in one or two true alleles, but rather one or two allele lists as some typing ambiguities cannot be resolved. This becomes an issue when not the full nucleotide sequence of a gene is examined, when the quality of data does not allow a clear-cut interpretation, or when the allele search space has been artificially limited prior to the analysis. A common form of representing such ambiguous genotypes is by using *Multiple Allele Codes* (MAC), also known as *NMPD codes* (<https://bioinformatics.bethematchclini->

cal.org/hla-resources/allele-codes/). With MAC, any currently known combination of allele numbers is encoded as 2–4 letters which are added to an allele name following the first field. As an example, an allele reported as DQB1*02:GKDU can be any allele from the DQB1*02:02 group of alleles or DQB1*02:97. Whilst MAC codes are widely used in particular for communicating donor registry genotyping data, they are not part of the official HLA nomenclature system. Two other possibilities are defined by the HLA nomenclature system. Alleles that share the protein sequence of the ARD can be referred to as P-code by adding a “P” suffix to the second field of the lowest-named HLA allele containing this sequence. Alleles with a common ARD nucleotide sequence can be summarised by adding a “G” suffix to the third field of the lowest-named allele. A drawback of G-codes is, however, that they might also contain null alleles. MAC, P-codes, and G-codes can lead to a loss of accuracy as not all alleles encoded in one group need be valid results from the HLA genotyping. A more flexible, less opaque, albeit somewhat verbose alternative to record genotyping ambiguity are Genotype List (GL) Strings [93]. The GL string format uses a set of operators to denote lists of possible genes, alleles, genotypes, or haplotypes without loss of accuracy. Phasing ambiguities can only be systematically reported using GL strings.

The transmission of HLA genotyping results to collaborators, genotyping registries, or during publication is done in ad hoc data formats, often unstructured, and often idiosyncratic to a genotyping tool or a vendor. The speed at which HLA genotyping technologies have evolved, the varied approaches to data generation, and the sheer wealth of information collected from NGS data, have all made it abundantly clear that simple reports of allele assignments are not nearly future-proof enough. Results may only be replicable or intelligible in the light of additional information on methodology. As a consequence, there have been efforts to develop and establish a data standard that includes not only the allele assignments but also the meta-data pertaining to data collection, data processing, and interpretation. At the foundation of this effort lies the definition of the Minimum Information for Reporting Immunogenomic NGS Genotyping (MIRING), a collection of principles and guidelines for reporting genotyping data [94]. A technical specification that is compliant with the MIRING principles has been developed with the *Histoimmunogenetics Markup Language* [95], a platform-independent electronic data interchange format based on XML. It has explicitly been designed with NGS-based genotyping in mind, allows complete reporting of allele and genotype ambiguity through the use of GL strings, and may incorporate pointers to external reference data, previously registered metadata, additional genotypes of other genes in a sample, and raw sequence reads.

Discussion

The genotyping of HLA genes is of considerable interest mainly in a clinical context but also in purely academic settings. Since the first attempts to use NGS data for HLA typing [96, 97], a plethora of algorithms and tools have been developed, sometimes applying fundamentally different approaches to the task of accurately matching NGS fragments to different HLA allele variants. Some algorithmic approaches are used by a majority of tools, e.g. read mapping as a crucial step to attribute reads to alleles and a deterministic allele scoring or calling directly based on the nucleotide variants. However, a single standard approach to achieve HLA typing by NGS data has not yet evolved. This may be in part due to the somewhat divergent aims of HLA typing for clinical versus academic applications. For the purpose of finding suitable HLA-matched donors for transplantation, typing accuracy is crucial lest a falsely inferred genotype leads to severe health issues for patients. On the other hand, information about the amino acid variation in the antigen binding groove alone is currently considered the gold standard for HSCT as long as null variants can be excluded. Clinical HLA typing demands standardised protocols and is often carried out in targeted experiments based on amplicon sequencing. Standard procedures are commercialised, and special kits are sold together with genotyping software solutions, which seek to give reliable, accurate and reproducible results. In an academic research environment, though, a wrong genotype is not life-threatening. Full allele level resolution, however, might become desirable. HLA typing in academic research often utilise existing non-targeted data from WGS, WES, or transcriptomic sequencing experiments. This setting allows wider active research on novel genotyping algorithms and procedures leading to a greater variety of available tools. However, many of these tools are proofs-of-concept only, with few systematic studies on accuracy and use-cases being carried out apart from benchmarking the new tools.

Often, HLA typing algorithms that are geared towards using HLA-untargeted data restrict themselves to achieving two-field resolution only, even if the sequence data used in principle allows for three-field or four-field resolution. This may be explained by the fact that many of these tools have been designed in the context of cancer immunotherapy with the express goal to predict HLA class I and class II binding neoepitopes. For this purpose, ARD resolution is sufficient. Other tools additionally restrict the allele search space by, for instance, excluding alleles perceived as too low in frequency to matter.

Both restrictions are also encouraged by the current requirements on resolution of international immunogenetics organisations. For instance, the American Society for Histocompatibility and Immunogenetics (ASHI) con-

siders it sufficient to report genotypes at (i) two-field resolution, (ii) exclude certain null alleles, and (iii) restrict the allelic search space to alleles recorded in the Common and Well-Documented (CWD) alleles catalogue [98]. Especially the restriction to CWD alleles can easily cause problems for patients from populations which are not well represented in the cohorts that formed the basis for the CWD definitions. The restraint to two-field resolution by employing only the ARD exons for genotyping, is to some extent also encouraged by the nature of the HLA sequences. While the exonic sequences are highly diverse, including nucleotide substitutions and indels, the non-coding sequences of HLA do additionally harbour features which are hard to sequence and vexing to analyse, especially long homopolymer stretches, STRs, and structural variation.

Another source for erroneous or ambiguous HLA typing results are difficulties to correctly infer the phase relationship between variants along two HLA alleles. Consider, for instance, a genotype with two adjacent variants that cannot be linked by a single read, either because the distance between the variants is greater than the achievable read or read-pair length, or because they are located on different exons deriving from separate PCR reactions. To some extent, this phasing ambiguity can be mitigated by considering only exact matches to the allele reference repository and restrict results to those alleles. However, this approach will lead to spurious results if the sample harbours a so far undescribed allele that has evolved by recombination. While this is highly unlikely for the ARD regions, it may be encountered more frequently when other exons are considered in high-volume genotyping. A promising solution for this issue is offered by long-read sequencing technologies from, e.g. *Pacific Bioscience* or *Oxford Nanopore Technologies* that routinely achieve read lengths of multiple kilobases. Both, commercial HLA typing software such as *GenDx*' NGSengine® and *Omixon*'s HLA Explore™ and academic software such as HLA*PRG:LA [83] offer support for long-read-based typing, although this is to some extent still considered experimental.

The choice of sequencing platform and experimental setup is ultimately affected by the desired comprehensiveness of the genetic characterisation of a sample. For the characterisation of potential donors during high-throughput registry genotyping, the focus on the ARD region fulfils the purpose of facilitating the identification of well-matched donors. A more comprehensive characterisation seems warranted for clinical samples and confirmatory genotyping for HSCT. This enables ruling out the presence of known or novel null alleles caused by variations outside the ARD regions, even though these are very rare events. The commercial solutions based on short-read technologies fulfil that need, even though class

II alleles may not be fully resolved due to missing exons or phasing issues. For studies aiming for full four-field resolution, the use of long reads may be beneficial. Finally, for the characterisation of novel alleles harbouring long intronic regions, the combination of long and short reads has the potential to provide utmost resolution and accuracy given appropriate algorithms.

In general, current research on the bioinformatics of HLA genotyping explores two major directions. First, more sophisticated methods and algorithms are developed to improve matching to the reference allele repository and resolve phasing and other ambiguities. Recent tools for HLA genotyping from WGS, WES, or RNA-Seq data attempt to do this, for instance, by exploring graph-based representations of the reference data or Bayesian probabilistic methods that may leverage information on allele or haplotype frequencies [35]. Second, advances in sequencing technologies are continuing to change the landscape of HLA typing. The currently available technologies are, however, still error-prone, and analyses need statistical methods to obtain reliable results. As such, we expect that the bioinformatics of HLA typing remains a challenge for the foreseeable future.

Statement of Ethics

The authors have no ethical conflicts to disclose.

Disclosure Statement

The authors have no conflicts of interest to declare.

Funding Sources

No funding was received in the preparation of the manuscript.

Author Contributions

S.K., V.S., V.L., and G.S. researched and wrote the manuscript.

References

- 1 Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet.* 2013;14(1): 301–23.
- 2 Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood.* 2007 Dec;110(13): 4576–83.

- 3 Loiseau P, Busson M, Balere ML, Dormoy A, Bignon JD, Gagne K, et al. HLA Association with hematopoietic stem cell transplantation outcome: the number of mismatches at HLA-A, -B, -C, -DRB1, or -DQB1 is strongly associated with overall survival. *Biol Blood Marrow Transplant*. 2007 Aug;13(8):965–74.
- 4 Morishima Y, Kashiwase K, Matsuo K, Azuma F, Morishima S, Onizuka M, et al.; Japan Marrow Donor Program. Biological significance of HLA locus matching in unrelated donor bone marrow transplantation. *Blood*. 2015 Feb;125(7):1189–97.
- 5 Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D423–31.
- 6 Kulski JK. Next-Generation Sequencing — An Overview of the History, Tools, and “Omic” Applications. In: Kulski JK, editor. *Next Generation Sequencing - Advances, Applications and Challenges*. London: InTech; 2016. pp. 3–60.
- 7 Lange V, Böhme I, Hofmann J, Lang K, Sauter J, Schöne B, et al. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics*. 2014 Jan;15(1):63.
- 8 Juhos S, Rigó K, Horváth G. On Genotyping Polymorphic HLA Genes — Ambiguities and Quality Measures Using NGS. In: Kulski JK, editor. *Next Generation Sequencing - Advances, Applications and Challenges*. London: InTech; 2016. pp. 369–86.
- 9 Schöfl G, Lang K, Quenzel P, Böhme I, Sauter J, Hofmann JA, et al. 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. *BMC Genomics*. 2017 Feb;18(1):161.
- 10 Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. *J Hum Genet*. 2015 Nov;60(11):665–73.
- 11 Bravo-Egana V, Monos D. The impact of next-generation sequencing in immunogenetics: current status and future directions. *Curr Opin Organ Transplant*. 2017 Aug;22(4):400–6.
- 12 Baier DM, Hofmann JA, Fischer H, Rall G, Stolze J, Ruhner K, et al. Very low error rates of NGS-based HLA typing at stem cell donor recruitment question the need for a standard confirmatory typing step before donor work-up. *Bone Marrow Transplant*. 2019 Jun;54(6):928–930.
- 13 Lang K, Wagner I, Schöne B, Schöfl G, Birkner K, Hofmann JA, et al. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. *BMC Genomics*. 2016 May;17(1):374.
- 14 Wagner I, Schefzyk D, Pruschke J, Schöfl G, Schöne B, Gruber N, et al. Allele-Level KIR Genotyping of More Than a Million Samples: Workflow, Algorithm, and Observations. *Front Immunol*. 2018 Dec;9:2843.
- 15 Cereb N, Kim HR, Ryu J, Yang SY. Advances in DNA sequencing technologies for high resolution HLA typing. *Hum Immunol*. 2015 Dec;76(12):923–7.
- 16 Zhou M, Gao D, Chai X, Liu J, Lan Z, Liu Q, et al. Application of high-throughput, high-resolution and cost-effective next generation sequencing-based large-scale HLA typing in donor registry. *Tissue Antigens*. 2015 Jan;85(1):20–8.
- 17 Yin Y, Lan JH, Nguyen D, Valenzuela N, Takemura P, Bolon YT, et al. Application of High-Throughput Next-Generation Sequencing for HLA Typing on Buccal Extracted DNA: Results from over 10,000 Donor Recruitment Samples. *PLoS One*. 2016 Oct;11(10):e0165810.
- 18 Erlich H. HLA DNA typing: past, present, and future. *Tissue Antigens*. 2012 Jul;80(1):1–11.
- 19 Profaizer T, Lázár-Molnár E, Close DW, Delgado JC, Kumánovics A. HLA genotyping in the clinical laboratory: comparison of next-generation sequencing methods. *HLA*. 2016 Jul;88(1-2):14–24.
- 20 Ingram KJ, Merckens H, O’Shields EF, Kiger D, Gautreaux MD. New HLA alleles discovered by next generation sequencing in routine histocompatibility lab work in a medium-volume laboratory. *Hum Immunol*. 2019 Jul;80(7):465–7.
- 21 Gandhi MJ, Ferriola D, Huang Y, Duke JL, Monos D. Targeted next-generation sequencing for human leukocyte antigen typing in a clinical laboratory: metrics of relevance and considerations for its successful implementation. *Arch Pathol Lab Med*. 2017 Jun;141(6):806–12.
- 22 Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol*. 1987;155:335–50.
- 23 Ehrenberg PK, Geretz A, Baldwin KM, Apps R, Polonis VR, Robb ML, et al. High-throughput multiplex HLA genotyping by next-generation sequencing using multi-locus individual tagging. *BMC Genomics*. 2014 Oct;15(1):864.
- 24 Albrecht V, Zweiniger C, Surendranath V, Lang K, Schöfl G, Dahl A, et al. Dual redundant sequencing strategy: full-length gene characterisation of 1056 novel and confirmatory HLA alleles. *HLA*. 2017 Aug;90(2):79–87.
- 25 Mayor NP, Brosnan N, Midwinter W, McWhinnie AJ, Turner TR, Robinson J, et al. A multiplexed typing strategy for the HLA class II genes HLA-DRB1, -DQB1 and -DPB1 using DNA barcodes and SMRT® DNA sequencing. *HLA*. 2016;87(4):276.
- 26 Mayor NP, Hayhurst JD, Turner TR, Szydlo RM. Better HLA matching as revealed only by next generation sequencing technology results in superior overall survival post-allogeneic haematopoietic cell. *Biol Blood Marrow Transplant*. 2018;24(3):S63–S64.
- 27 Lang K, Surendranath V, Quenzel P, Schöfl G, Schmidt AH, Lange V. Full-Length HLA Class I Genotyping with the MinION Nanopore Sequencer. *Methods Mol Biol*. 2018;1802:155–62.
- 28 Ambardar S, Gowda M. High-Resolution Full-Length HLA Typing Method Using Third Generation (Pac-Bio SMRT) Sequencing Technology. *Methods Mol Biol*. 2018;1802:135–53.
- 29 Pröll J, Danzer M, Stabentheiner S, Niklas N, Hackl C, Hofer K, et al. Sequence capture and next generation resequencing of the MHC region highlights potential transplantation determinants in HLA identical haematopoietic stem cell transplantation. *DNA Res*. 2011 Aug;18(4):201–10.
- 30 Dapprich J, Ferriola D, Mackiewicz K, Clark PM, Rappaport E, D’Arcy M, et al. The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC Genomics*. 2016 Jul;17(1):486.
- 31 Jiao Y, Li R, Wu C, Ding Y, Liu Y, Jia D, et al. High-sensitivity HLA typing by Saturated Tiling Capture Sequencing (STC-Seq). *BMC Genomics*. 2018 Jan;19(1):50.
- 32 Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA*. 2009 Nov;106(45):19096–101.
- 33 Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009 Sep;461(7261):272–6.
- 34 Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. *Nat Methods*. 2010 Feb;7(2):111–8.
- 35 Hayashi S, Yamaguchi R, Mizuno S, Komura M, Miyano S, Nakagawa H, et al. ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data. *BMC Genomics*. 2018 Nov;19(1):790.
- 36 Ka S, Lee S, Hong J, Cho Y, Sung J, Kim HN, et al. HLAScan: genotyping of the HLA region using next-generation sequencing data. *BMC Bioinformatics*. 2017 May;18(1):258.
- 37 Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature*. 2014 Jul;511(7509):344–7.
- 38 van El CG, Cornel MC, Borry P, Hastings RJ, Fellmann F, Hodgson SV, et al.; ESHG Public and Professional Policy Committee. Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics. *Eur J Hum Genet*. 2013 Jun;21(6):580–4.
- 39 Mimori T, Yasuda J, Kuroki Y, Shibata TF, Katsuoaka F, Saito S, et al. Construction of full-length Japanese reference panel of class I HLA genes with single-molecule, real-time sequencing. *Pharmacogenomics J*. 2019 Apr;19(2):136–46.
- 40 Juhos S, Vágó T, Ferriola D, Duke J, Vörös S, Brown BO, et al. Deriving HLA genotyping from whole genome sequencing data using Omixon HLA twin (tm) in G3’s global clinical study. *Hum Immunol*. 2015;76:131.
- 41 Xie C, Yeo ZX, Wong M, Piper J, Long T, Kirkness EF, et al. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc Natl Acad Sci USA*. 2017 Jul;114(30):8059–64.

- 42 Orenbuch R, Filip I, Comito D, Shaman J, Pe'er I, Rabadan R. arcasHLA: high resolution HLA typing from RNA seq. *Bioinformatics*. 2019;btz474. doi: 10.1093/bioinformatics/btz474.
- 43 Boegel S, Bukur T, Castle JC, Sahin U. In Silico Typing of Classical and Non-classical HLA Alleles from Standard RNA-Seq Reads. *Methods Mol Biol*. 2018;1802:177–91.
- 44 Walsh PS, Erlich HA, Higuchi R. Preferential PCR amplification of alleles: mechanisms and solutions. *PCR Methods Appl*. 1992 May; 1(4):241–50.
- 45 Veal CD, Freeman PJ, Jacobs K, Lancaster O, Jamain S, Leboyer M, et al. A mechanistic basis for amplification differences between samples and between genome regions. *BMC Genomics*. 2012 Sep;13(1):455.
- 46 Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N. Efficient study design for next generation sequencing. *Genet Epidemiol*. 2011 May;35(4):269–77.
- 47 Curran MD, Williams F, Earle JA, Rima BK, van Dam MG, Bunce M, et al. Long-range PCR amplification as an alternative strategy for characterizing novel HLA-B alleles. *Eur J Immunogenet*. 1996 Aug;23(4):297–309.
- 48 Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. *Brief Bioinform*. 2010 Sep;11(5):484–98.
- 49 Stevens AJ, Taylor MG, Pearce FG, Kennedy MA. Allelic Dropout During Polymerase Chain Reaction due to G-Quadruplex Structures and DNA Methylation Is Widespread at Imprinted Human Loci. *G3 (Bethesda)*. 2017 Mar;7(3):1019–25.
- 50 Lahr DJ, Katz LA. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques*. 2009 Oct;47(4):857–66.
- 51 Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res*. 1992 Jan;20(2):211–5.
- 52 Pereira S, Vayntrub T, Hiraki DD, Cherry AM, Arai S, Dvorak CC, et al. Short tandem repeat and human leukocyte antigen mutations or losses confound engraftment and typing analysis in hematopoietic stem cell transplants. *Hum Immunol*. 2011 Jun;72(6):503–9.
- 53 van Deutekom HW, Mulder W, Rozemuller EH. Accuracy of NGS HLA typing data influenced by STR. *Hum Immunol*. 2019 Jul;80(7):461–4.
- 54 Turcatti G, Romieu A, Fedurco M, Tairi AP. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res*. 2008 Mar;36(4):e25.
- 55 Cao H, Wang Y, Zhang W, Chai X, Zhang X, Chen S, et al. A short-read multiplex sequencing method for reliable, cost-effective and high-throughput genotyping in large-scale studies. *Hum Mutat*. 2013 Dec;34(12):1715–20.
- 56 Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. *Am J Hum Genet*. 2016 Aug;99(2):375–91.
- 57 Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011 Jul; 475(7356):348–52.
- 58 Barone JC, Saito K, Beutner K, Campo M, Dong W, Goswami CP, et al. HLA-genotyping of clinical specimens using Ion Torrent-based NGS. *Hum Immunol*. 2015 Dec;76(12):903–9.
- 59 Segawa H, Kukita Y, Kato K. HLA genotyping by next-generation sequencing of complementary DNA. *BMC Genomics*. 2017 Nov; 18(1):914.
- 60 Suzuki S, Ranade S, Osaki K, Ito S, Shigenari A, Ohnuki Y, et al. Reference grade characterization of polymorphisms in full-length HLA class I and II genes with short-read sequencing on the Ion PGM system and long-reads generated by Single Molecule, Real-time Sequencing on the PacBio platform. *Front Immunol*. 2018 Oct;9:2294.
- 61 Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012 May;30(5):434–9.
- 62 Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, et al. Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol*. 2010;472:431–55.
- 63 Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009 Jan; 323(5910):133–8.
- 64 Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat Biotechnol*. 2012 Feb; 30(4):344–8.
- 65 Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*. 2016 Nov;17(1):239.
- 66 Ammar R, Paton TA, Torti D, Shlien A, Bader GD. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000 Res*. 2015 Jan;4:17.
- 67 Ton KN, Cree SL, Gronert-Sum SJ, Merriman TR, Stamp LK, Kennedy MA. Multiplexed Nanopore Sequencing of HLA-B Locus in Māori and Pacific Island Samples. *Front Genet*. 2018 Apr;9:152.
- 68 Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boissguérin V, et al. HLA typing from RNA-Seq sequence reads. *Genome Med*. 2012 Dec;4(12):102.
- 69 Kim HJ, Pourmand N. HLA typing from RNA-seq data using hierarchical read weighting [corrected]. *PLoS One*. 2013 Jun;8(6):e67885.
- 70 Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics*. 2014 May;15(1):325.
- 71 Nariai N, Kojima K, Saito S, Mimori T, Sato Y, Kawai Y, et al. HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. *BMC Genomics*. 2015;16(S2 Suppl 2):S7.
- 72 Liu C, Yang X, Duffy B, Mohanakumar T, Mitra RD, Zody MC, et al. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res*. 2013 Aug;41(14):e142–142.
- 73 Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 2013;1303.3997.
- 74 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012 Mar;9(4):357–9.
- 75 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013 Jan;29(1):15–21.
- 76 Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015 Apr;12(4):357–60.
- 77 Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018 Sep; 34(18):3094–100.
- 78 Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun*. 2016 Apr;7(1):11307.
- 79 Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*. 2014 Dec;30(23):3310–6.
- 80 Dilthey AT, Gourraud PA, Mentzer AJ, Cereb N, Iqbal Z, McVean G. High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. *PLOS Comput Biol*. 2016 Oct; 12(10):e1005151.
- 81 Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. *Genome Res*. 2017 May;27(5):665–76.
- 82 Szolek A. HLA Typing from Short-Read Sequencing Data with OptiType. New York: Humana Press; 2018. pp. 215–23.
- 83 Dilthey AT, Mentzer AJ, Carapito R, Cutland C, Cereb N, Madhi SA, et al. HLA*PRG: LA – HLA typing from linearly projected graph alignments. *bioRxiv*. 2018 Oct;453555.
- 84 Dilthey A, Cox C, Iqbal Z, Nelson MR, McVean G. Improved genome inference in the MHC using a population reference graph. *Nat Genet*. 2015 Jun;47(6):682–8.
- 85 Huang Y, Yang J, Ying D, Zhang Y, Shotelersuk V, Hirankarn N, et al. HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med*. 2015 Mar;7(1):25.
- 86 Lee H, Kingsford C. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biol*. 2018 Feb; 19(1):16.
- 87 Major E, Rigó K, Hague T, Bérces A, Juhos S. HLA typing from 1000 genomes whole genome and whole exome illumina data. *PLoS One*. 2013 Nov;8(11):e78410.

- 88 Shinde D, Lai Y, Sun F, Arnheim N. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)_n and (A/T)_n microsatellites. *Nucleic Acids Res.* 2003 Feb;31(3):974–80.
- 89 Raz O, Biezuner T, Spiro A, Amir S, Milo L, Titelman A, et al. Short Tandem Repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic Acids Res.* 2019;47(5):2436–2445.
- 90 Klasberg S, Lang K, Günther M, Schober G, Massalski C, Schmidt AH, et al. Patterns of non-ARD variation in more than 300 full-length HLA-DPB1 alleles. *Hum Immunol.* 2019 Jan;80(1):44–52.
- 91 Schöfl G, Lang K, Schmidt AH, Lange V. OR30 Dual redundant reference sequencing (DR2S): a computational approach for phase-defined full-length HLA-gene characterization. *Hum Immunol.* 2016;77:25.
- 92 Schöne B, Fuhrmann M, Surendranath V, Schmidt AH, Lange V, Schöfl G. TypeLoader2: Automated submission of novel HLA and KIR alleles in full length. *HLA.* 2019;93(4):195–202.
- 93 Milius RP, Mack SJ, Hollenbach JA, Pollack J, Heuer ML, Gragert L, et al. Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string. *Tissue Antigens.* 2013 Aug;82(2):106–12.
- 94 Mack SJ, Milius RP, Gifford BD, Sauter J, Hofmann J, Osoegawa K, et al. Minimum information for reporting next generation sequence genotyping (MIRING): guidelines for reporting HLA and KIR genotyping via next generation sequencing. *Hum Immunol.* 2015 Dec;76(12):954–62.
- 95 Milius RP, Heuer M, Valiga D, Doroschak KJ, Kennedy CJ, Bolon YT, et al. Histoimmunogenetics Markup Language 1.0: reporting next generation sequencing-based HLA and KIR genotyping. *Hum Immunol.* 2015 Dec;76(12):963–74.
- 96 Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, Trachtenberg EA, et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens.* 2009 Nov;74(5):393–403.
- 97 Gabriel C, Danzer M, Hackl C, Kopal G, Hufnagl P, Hofer K, et al. Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. *Hum Immunol.* 2009 Nov;70(11):960–4.
- 98 Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, et al. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens.* 2013;81(4):194–203.